UNIVERSITÀ
DI PAVIA

FACOLTÀ DI INGEGNERIA

DIPARTIMENTO DI INGEGNERIA INDUSTRIALE E DELL'INFORMAZIONE

CORSO DI LAUREA MAGISTRALE IN BIOINGEGNERIA

TESI DI LAUREA

EXPLORING FOUNDATION MODELS FOR NEONATAL SEPSIS PREDICTION: A COMPARISON OF DISCRIMINATIVE AND GENERATIVE AI APPROACHES

Candidato: Marzio Cossali

Relatore: Dr. Enea Parimbelli
Correlatore: Prof. Jean Louis Raisaro

Academic Year 2023/24

Abstract (Italiano)

Questo studio si inserisce nel contesto del progetto ADONIS, finalizzato a migliorare la diagnosi precoce della sepsi nei neonati in Svizzera. L'obiettivo principale è stato esplorare l'utilizzo di foundation models per la previsione della sepsi, confrontando tali approcci con metodi di machine learning convenzionali, in particolare con un modello Support Vector Machine (SVM). Per valutare la capacità di adattamento e il potenziale impatto clinico dei modelli proposti, diversi esperimenti sono stati condotti utilizzando diversi setup sperimentali, volti a simulare scenari limite tipici del contesto ospedaliero. Le analisi hanno incluso uno studio approfondito della robustezza e dell'affidabilità delle predizioni al variare del numero di features considerate e della quantità di dati disponibili per l'addestramento, nonché un'analisi delle capacita' di generalizzazione su diversi task attraverso la predizione della Necrotizing Enterocolotis (NEC) mediante un approccio zero-shot.

I dati impiegati nello studio sono stati raccolti da oltre 15.000 pazienti neonati, provenienti da tre dei principali centri ospedalieri svizzeri. La raccolta e l'armonizzazione dei dati sono state effettuate mediante l'infrastruttura BioMedIT, che ha garantito la sicurezza, la condivisione e l'interoperabilità dei dati in conformità con lo schema fornito da SPHN. Tale infrastruttura ha permesso di superare le tradizionali limitazioni legate alla scarsità di dati e alla loro eterogeneità nei contesti multicentrici.

Questo studio si propone di valutare l'efficacia di due approcci distinti per la predizione della sepsi nei neonati, facendo leva sul potenziale impatto di questo tipo di modelli in clinica. Il primo approccio prevede l'elaborazione diretta delle serie temporali grezze utilizzando un foundation model per la classificazione di serie temporali (FORMED), eliminando la necessità di un'estrazione manuale delle features, mentre il secondo utilizza un altro foundation model progettato per dati tabulari (TabPFN).

I risultati indicano che, sebbene sia SVM che TabPFN raggiungano performance comparabili nella previsione della sepsi, TabPFN dimostra una maggiore capacità di generalizzazione, in particolare in scenari con quantità limitate di dati di training. L'analisi zero-shot sul task di NEC conferma ulteriormente la robustezza dei foundation models nel trasferire la conoscenza appresa da un task all'altro. Tuttavia, nel caso di FORMED, un calo delle performance é stato osservato rispetto ai modelli basati su features ingegnerizzate, suggerendo che, nonostante il potenziale vantaggio di eliminare l'estrazione manuale delle features, tale approccio

possa ancora essere limitato se comparato con un approccio classico.

In conclusione, questo studio evidenzia il potenziale dei foundation models nella risoluzione di problemi complessi in ambito clinico, concentrandosi sulla predizione della sepsi nei neonati, offrendo una promettente alternativa ai metodi di machine learning tradizionali. I risultati ottenuti rappresentano un passo importante verso lo sviluppo di sistemi di supporto decisionale più robusti e adattabili, favorendo una migliore cura dei pazienti in ambienti clinici caratterizzati da risorse limitate.

Parole chiave: sepsi, machine learning, time series classification, foundation models, zero-shot learning, few-shot learning

Contents

Acronyms

AI Artificial Intelligence. 13–16, 18

AUROC Area Under the Receiver Operating Characteristic curve. 10–12

BPD bronchopulmonary dysplasia. 23

CA-LOS Community Acquired LOS. 6, 9

CHUV Lausanne University Hospital. 3, 8, 19, 21, 25, 26, 28, 32, 43, 44, 53, 56, 59, 60

CLABSI Central Line-Associated Bloodstream Infections. 6

CLaMs Clinical Language Models. 14, 15

CNN Convolutional Neural Networks. 11

CoNS Coagulase-Negative Staphylococci. 6

CRASH Cultures, Resuscitation, and Antibiotics Started Here. 21–23

DBP diastolic blood pressure. 23, 27

DWH Data Warehouse. 25

ECG electrocardiogram. 8, 9, 46, 47

ED Emergency Department. 17

EHRs Electronic Health Records. 13, 15

EOS Early-Onset Sepsis. 5, 6, 21

FEMRs Foundation models for Electronic Medical Records. 14, 15

FM Foundation Models. 2–4, 12, 13, 16–18, 32, 33, 38–40, 42, 45, 50, 52, 55–62, 65, 66

GBS Group B Streptococcus. 5, 6

HA-LOS Hospital Acquired LOS. 6, 9, 22

HeRO Heart Rate Observation. 8, 9, 21, 32, 62

HR heart rate. 23, 25, 26, 28, 46

HRV Heart Rate Variability. 10–12

ICL In-Context Learning. 42

IVH intraventricular hemorrhage. 23

KiSpi University Children's Hospital Zurich. 3, 19–21, 26, 53, 61, 62

LLMs Large Language Models. 14, 15, 17

LOS Late-Onset Sepsis. 5, 9–12, 20, 32, 42

ML Machine Learning. 1–4, 10, 12, 13, 17, 29, 32, 33, 40, 55, 56, 58–60, 65, 66

MLP Multi-Layer Perceptron. 36

MV MetaVision. 25–27

NEC Necrotizing Enterocolitis. 7, 9, 10, 22, 31, 52, 59

NICU Neonatal Intensive Care Unit. 6, 8, 9, 16, 18, 21, 25, 61, 65

NLP Natural Language Processing. 35, 37

PVL periventricular leukomalacia. 23

ROP retinopathy of prematurity. 23

RR respiratory rate. 23, 26

SBP systolic blood pressure. 23, 27, 46

SPHN Swiss Personalized Health Network. 2, 3, 25

SpO_2 oxygen saturation. 23, 26, 46

SVM Support Vector Machine. 32, 42–44, 47, 56, 58–61, 65

USZ University Hospital of Zurich. 3, 19, 21, 26, 53, 59, 60

Chapter 1

Introduction

1.1 Aims and scope

1.1.1 Background and motivation

Sepsis remains a leading cause of neonatal morbidity and mortality, contributing to over half a million deaths annually worldwide [1]. Despite advances in neonatal care, its early identification remains a significant challenge. Detection and prompt initiation of antimicrobial therapy, along with supportive interventions, are crucial for improving clinical outcomes [2]. However, current diagnostic approaches, which rely on clinicians' evaluation of risk factors, clinical signs, and biomarkers, suffer from limited precision [3]. As a result, the start of treatment is often delayed, while other patients are unnecessarily exposed to antibiotics, contributing to antimicrobial resistance and potential adverse effects [4].

Given the complex and non-specific presentation of neonatal sepsis, clinicians face substantial diagnostic uncertainty in practice. Sepsis is a highly heterogeneous syndrome with no single defining clinical pattern, and its manifestation in neonates can overlap with other common neonatal conditions [5]. This clinical ambiguity, combined with the lack of universally accepted diagnostic guidelines or treatment procedures [4], poses serious challenges for the development of rule-based decision support tools. In this context, the integration of automated, data-driven solutions becomes increasingly important to assist clinicians in recognizing early signs of sepsis and guiding timely intervention.

Machine Learning (ML)-based predictive models have been proposed to enhance early sepsis detection, yet their performance remains constrained by several factors [6–10] Sepsis is a highly heterogeneous syndrome with no single defining biomarker, making it difficult to establish a consistent predictive signature [4]. This challenge is further exacerbated in neonates, whose physiology is dynamic and rapidly evolving. The absence of standardized clinical definitions, the high inter-individual variability in physiological parameters, and the impact of gestational age on biomarker expression introduce additional complexities. Furthermore, the

low temporal resolution of the available clinical data limits the ability of predictive models to capture subtle patterns indicative of sepsis onset. Another critical limitation is the lack of large, standardized datasets collected across multiple clinical centers. Existing studies primarily focus on single-center or limited-cohort datasets, raising concerns about model generalizability and external validity [6].

Addressing these challenges requires a paradigm shift towards more advanced modeling approaches capable of capturing the intricate, non-linear relationships within high-dimensional, heterogeneous clinical data. Traditional ML models, characterized by relatively simple and lightweight architectures, often exhibit limited capacity to capture the complex and dynamic nature of physiological processes. This limitation becomes particularly pronounced in high-variability clinical scenarios such as neonatal sepsis, where vital signs evolve rapidly and clinical presentations are highly heterogeneous. In such contexts, more advanced approaches are required to effectively integrate multimodal clinical inputs, model temporal dependencies, and generalize across diverse patient populations. The adoption of these sophisticated methodologies holds the potential to markedly enhance predictive performance, thereby enabling earlier and more reliable identification of neonatal sepsis and improving clinical outcomes.

Foundation Models (FM)s emerge as a promising solution to these limitations. FMs are large-scale, pre-trained architectures designed to generalize across a wide range of downstream tasks by leveraging broad representations learned from extensive datasets [11]. FMs support task generalization, allowing a single model to be applied across multiple diagnostic or prognostic tasks. This intrinsic flexibility not only reduces development time but also eliminates the need for extensive retraining [12]. As a result, FMs are particularly advantageous in healthcare settings, where data scarcity is a pervasive challenge, especially in the context of rare diseases such as neonatal sepsis [13]. Their ability to generalize from prior knowledge enables effective adaptation to new tasks even in the absence of large labeled datasets, addressing one of the most critical limitations of traditional ML approaches. Their robustness to distribution shifts, whether temporal, institutional, or demographic, enhances their applicability in real-world settings where clinical practices and patient populations vary [14]. Moreover, FMs provide a framework for multimodal data integration, allowing the combination of structured clinical variables, physiological time series, imaging, and textual notes into a unified representation [11]. This capability may foster a more holistic view of the patient's condition, aligning closely with the decision-making processes of clinicians. Collectively, these advantages position FMs as a transformative tool in the development of clinical decision support systems for neonatal sepsis and beyond.

1.1.2 The ADONIS project

The Accelerating Detection Of Neonatal sepsIS (ADONIS) project, funded by the Swiss Personalized Health Network (SPHN), aims to advance the early detection of neonatal sepsis through data-driven approaches. This interdisciplinary initiative brings together experts

from Lausanne University Hospital (CHUV), the University Hospital of Zurich (USZ), and the University Children's Hospital Zurich (KiSpi). The primary objective of the project is to establish the largest neonatal sepsis cohort in Switzerland, enabling a comprehensive investigation of sepsis phenotypes and improving neonatal sepsis care through the development of a highly adaptive ML-based tool for early and accurate detection.

ADONIS leverages the SPHN framework and infrastructure to harmonize and analyze high-resolution, multi-dimensional clinical data. With an unprecedented level of granularity, encompassing over 800 clinical and laboratory variables recorded at intervals as frequent as every minute, the project seeks to characterize the physiological alterations preceding the clinical manifestation of sepsis. By systematically capturing these early physiological changes, ADONIS aims to refine risk stratification, enhance early recognition, and ultimately support timely clinical decision-making.

Within this context, the present research is conducted as part of the ADONIS initiative. The findings generated through this research are expected to contribute to the broader objectives of the project by refining methodologies for early sepsis detection and advancing the integration of ML-driven tools into neonatal intensive care settings.

1.1.3 Objectives and contributions of this thesis

The primary objective of this thesis is to evaluate the potential of state-of-the-art models, in particular FMs, to improve early onset sepsis prediction. Their ability to encode rich, and transferable features positions them as promising tools for complex clinical applications. This study specifically investigates whether such models can outperform traditional ML approaches by more effectively navigating the multidimensional structure of clinical data, accounting for the substantial intra- and inter-patient variability of neonatal populations, and maintaining robust predictive performance in the presence of population and temporal distribution shifts.

To achieve this, two distinct FM-based approaches are explored. The first approach involves processing raw clinical time series using FORMED, a foundation model specifically designed for time series classification [15]. By operating directly on raw signals, this method aims to eliminate the reliance on manual feature engineering, a process that is often labor-intensive and time-consuming. The second approach focuses on tabular data and employs TabPFN, a foundation model tailored for tabular prediction tasks [16]. TabPFN leverages in-context learning and prompt-based conditioning to achieve high adaptability across diverse clinical scenarios.

A further objective of this thesis is to externally validate the sepsis prediction models developed at CHUV within a multicentric setting. This validation will involve clinical data from two additional hospitals in Switzerland, ensuring a comprehensive assessment of models performance in different institutional contexts.

To systematically assess the clinical applicability of FMs, this study proposes a comprehen-

sive evaluation framework designed to simulate realistic clinical scenarios. The framework includes experimental setups that reflect critical constraints commonly encountered in clinical practice, such as limited data availability and distribution shifts across centers. This design enables an in-depth analysis of the strengths and limitations of the proposed FM-based approaches, providing valuable insights into their potential for real-world implementation. Unlike many recent studies that focus primarily on performance benchmarks using standardized datasets [17–19], this work emphasizes the importance of evaluating models in conditions that closely resemble actual clinical environments. In fact, while benchmark datasets can offer insights into average performance, they often overlook key operational challenges, such as model usability and integration feasibility, that are essential for clinical adoption.

The proposed evaluation not only highlights the advantages of FMs in terms of adaptability, robustness, and scalability, but also surfaces important limitations that must be addressed for their successful deployment. In doing so, this study lays the groundwork for a more informed and clinically meaningful application of FMs in neonatal care and contributes to fostering clinician trust in emerging AI-driven tools bridging the gap between ML research and clinical practice.

1.2 Literature review

1.2.1 Neonatal sepsis: clinical aspects and challenges

Sepsis is a complex and heterogeneous syndrome, characterized in both adults and children by a dysregulated host response to infection [2]. According to the first Global Burden of Disease report published in 2020, sepsis affects approximately 50 million individuals annually, leading to more than 11 million deaths worldwide [1].

The neonatal period, defined as the first 28 days of life, represents the highest lifetime risk for sepsis, accounting for over 400,000 annual deaths worldwide. This life-threatening condition manifests in newborns as a systemic inflammatory response to bacterial, viral, or fungal infections, with bloodstream infections constituting the most frequent presentation. In Switzerland, population-based studies estimate an incidence of 1.43 cases per 1,000 live births, demonstrating the substantial burden of this condition even in high-resource settings [20].

The pathophysiology of neonatal sepsis evolves through sequential phases: initial microbial invasion, systemic inflammatory response, and potential progression to septic shock and multi-organ failure. Preterm infants face heightened vulnerability due to immature immune systems, compromised skin/mucosal barriers, and frequent requirement for invasive medical devices. Term infants demonstrate greater immunological competence but remain susceptible to specific pathogens like Group B Streptococcus (GBS) through vertical transmission [20].

The clinical landscape of neonatal sepsis is shaped by two distinct entities: Early-Onset Sepsis (EOS) and Late-Onset Sepsis (LOS), each with unique epidemiological profiles, risk factors, and outcomes.

Early-Onset Sepsis (EOS)

Early-onset sepsis (EOS) is defined by its occurrence within the first 72 hours after birth. In a Swiss cohort study, the estimated national incidence of EOS was 0.28 per 1000 live births [20]. The primary route of infection is vertical transmission from mother to infant during labor and delivery. GBS and *Escherichia coli* have been idenfied as the most frequently implicated pathogens, accounting for 38% and 23% of episodes, respectively [20]. However, the specific causative agent often correlates with the gestational age of the infant. Term infants are more likely to contract EOS from GBS, typically presenting as primary bloodstream infections without maternal chorioamnionitis. Conversely, preterm infants are more susceptible to *E. coli* infections, which often have a strong association with maternal chorioamnionitis. This study indicates that clinically defined maternal chorioamnionitis was observed in 37% of EOS cases, including 70% and 24% of cases caused by *E. coli* and GBS, respectively, and affected almost exclusively preterm infants (94%) [20]. EOS carries a significant mortality rate, reaching 18%, with risk stratification heavily influenced by birth weight categories. Extremely low birth weight infants (<1000g) face 4.6-fold higher sepsis incidence compared to term counterparts, compounded by frequent respiratory failure (63% requiring mechanical

ventilation) and septic shock (22%)[20].

Late-Onset Sepsis (LOS): hospital vs community acquisition

Late-onset sepsis (LOS) is characterized by infections manifesting beyond 72 hours after birth. This condition affects up to 20% of extremely preterm newborns. In the most premature infants, case-fatality rates can reach 50%, depending on the causative pathogen [21–23]. It is further subdivided into Hospital-Acquired LOS (HA-LOS) and Community-Acquired LOS (CA-LOS), based on whether the infection is contracted within a healthcare setting or in the community. In Switzerland, the estimated national incidence of Hospital Acquired LOS (HA-LOS) and Community Acquired LOS (CA-LOS) was 0.86 and 0.28 per 1000 live births, respectively [20].

Hospital-Acquired LOS (HA-LOS) Hospital-Acquired LOS (HA-LOS) primarily affects preterm infants, with a median gestational age of 29 weeks. It is often associated with significant comorbidities and prolonged stays in the Neonatal Intensive Care Unit (NICU). The predominant pathogens in HA-LOS differ from those in EOS, with Coagulase-Negative Staphylococci (CoNS) being the leading cause, accounting for 40% of episodes, followed by *S. aureus* (16%) and *E. coli* (16%) [20]. Risk factors for HA-LOS include central venous catheterization (OR 4.2) and prolonged NICU stay (>14 days). The mortality rate is 12%, with extreme prematurity and septic shock being key prognostic determinants [20].

Central Line-Associated Bloodstream Infections (CLABSI) represent a critical iatrogenic complication, accounting for 47% of HA-LOS cases. These infections develop when pathogens colonize intravascular catheters, with CoNS and *Staphylococcus aureus* being predominant culprits. CLABSI patients typically exhibit extreme prematurity (gestational age <28 weeks), prolonged hospitalization, and multiple comorbidities requiring intensive care support [20].

Community-Acquired LOS (CA-LOS) Community-Acquired LOS (CA-LOS) primarily affects term male infants, with a male:female ratio of 3:1. The pathogen profile also differs, with GBS (30%), *E. coli* (24%), and *S. pneumoniae* (11%) being the most common. The clinical presentation of CA-LOS often involves meningitis (23%) and urinary tract infections (18%) [20]. In the Swiss cohort, mortality was 0%, reflecting later onset and more robust host defenses.

Advancements in obstetric care have contributed to a decline in perinatal infections, reducing the incidence of EOS in high-income countries. However, this positive trend has been accompanied by a marked increase in healthcare-associated infections, particularly among preterm and critically ill term neonates, who now represent a growing proportion of NICUs population. As a result, HA-LOS has emerged as the most prevalent form of neonatal sepsis in these settings. In light of this epidemiological transition, the present study specifically targets HA-LOS cases, with the objective of developing a clinical decision support tool optimized for the early detection of sepsis in this high-risk subgroup. By focusing on HA-LOS,

the study aims to address the most pressing clinical need in current neonatal care, where timely diagnosis remains critical for improving outcomes. Furthermore, survivors of neonatal sepsis are at risk of long-term complications, including cerebral palsy, as well as visual, auditory, and cognitive impairments [24]. As a result, neonatal sepsis imposes a significant medical, societal, and economic burden.

Limitation in the diagnosis of neonatal sepsis

Neonatal sepsis appears with non-specific clinical signs, making it particularly difficult to differentiate from non-infectious conditions and other pathologies with overlapping symptoms. One such condition is Necrotizing Enterocolitis (NEC), which shares several clinical features with sepsis, including apnea, lethargy, temperature instability, and feeding intolerance [20]. This symptomatic overlap further complicates early recognition and can delay appropriate treatment. Sepsis is further characterized by rapid progression to multi-organ dysfunction, high mortality, and significant morbidity [25]. At symptom onset, commonly used laboratory tests, such as white blood cell indices and acute phase reactants, exhibit low positive predictive value for neonatal sepsis [26]. Additionally, blood cultures, the gold standard for diagnosing bloodstream infections, require 12–36 hours to yield results and have limited sensitivity [27]. These diagnostic limitations, combined with the subtle and often ambiguous early clinical signs, can delay recognition and lead clinicians to adopt a low threshold for initiating empirical antibiotic therapy. Consequently, neonatal infection or "rule out sepsis" is one of the most frequent diagnoses in neonatal intensive care units, with antibiotics among the most commonly prescribed medications [28]. However, in the majority of cases, sepsis is ultimately ruled out, resulting in significant antibiotic overuse. This excessive exposure increases the risk of colonization with antibiotic-resistant bacteria and disrupts early-life microbiota, potentially leading to long-term adverse health effects [29].

The prediction and definition of organ dysfunction in newborns remain challenging due to the complex and dynamic physiological changes occurring during early life [30]. Additionally, factors such as gestational age, congenital conditions unique to this population, and medical interventions required to sustain vital functions in preterm neonates significantly influence organ function. These complexities contribute to the lack of a universally accepted definition of neonatal sepsis [4]. Most studies define sepsis based on bloodstream infections confirmed by a positive blood culture, as bacteremia and the associated inflammatory response can harm the developing brain and lead to lifelong disabilities, even in the absence of detectable organ dysfunction [24]. Furthermore, dysfunction of the respiratory and cardiovascular systems significantly increases mortality and morbidity in neonates with bloodstream infections [31], underscoring the importance of early diagnosis and treatment before the onset of clinically apparent organ dysfunction.

In this context, the development and clinical integration of advanced diagnostic systems capable of managing multiple pathologies with overlapping clinical presentations, including both sepsis and NEC, could prove highly beneficial. These systems may offer more accurate, early differentiation between conditions, reduce diagnostic uncertainty, and support timely,

targeted therapeutic decisions, thereby improving outcomes and reducing unnecessary interventions.

HeRO Score

Among the most advanced predictive monitoring tools currently employed, the Heart Rate Observation (HeRO) score has emerged as a state-of-the-art approach for identifying neonates at risk of sepsis. Developed through the analysis of electrocardiogram (ECG) data acquired from bedside monitors, the HeRO score quantifies subtle alterations in heart rate characteristics (HRC), specifically decreased variability and transient decelerations, which are often indicative of early-stage sepsis. These computed features are transformed into a single numerical index, reflecting the fold increase in a patient's probability of experiencing clinical deterioration due to sepsis within the subsequent 24 hours [32].

The efficacy of the HeRO score has been demonstrated in a multicenter randomized trial involving 3,003 very low birth weight infants, in which the continuous display of the score to clinicians resulted in a more than 20% reduction in mortality. By serving as an early warning system, the HeRO score enables timely clinical decision-making, allowing for earlier therapeutic interventions that improve patient prognosis. This outcome underscores the transformative potential of predictive monitoring in neonatal care, positioning the HeRO score as a pivotal advancement in sepsis detection and management [32].

Limitations of the HeRO Score and challenges in clinical implementation While the integration of the HeRO score into NICU protocols represents a significant advancement in predictive monitoring, its clinical adoption remains restricted to a limited number of hospitals, including CHUV. Several factors hinder its widespread implementation in routine neonatal care.

One of the primary limitations of the HeRO score is the reliability of its predictions. While studies have demonstrated its utility in predicting nosocomial infections in neonates, prospective research has also highlighted its false positives rate [3]. Although an elevated HeRO score often precedes the clinical manifestation of sepsis, enabling timely medical intervention, the high false positive rate raises concerns [32]. An excessive number of false alarms not only increases the workload for healthcare staff but may also lead to a well-documented phenomenon known as alarm fatigue. Alarm fatigue arises when clinicians are repeatedly exposed to spurious alerts, leading to desensitization and a diminished sense of urgency in responding to subsequent alarms, which can result in delayed or missed responses to critical events. This phenomenon can have critical implications for patient safety, particularly in the context of sepsis detection, where early intervention is essential. Thus, when developing a predictive model for clinical use, it is essential to establish an acceptable false alarm rate in consultation with clinicians. Nonetheless, in the event of a false positive, physicians can still assess the patient's condition and act accordingly. For instance, if a neonate is not intubated, does not have a catheter, or is nearing discharge, the clinician may reasonably choose to disregard the alert.

A second major barrier to the broader adoption of the HeRO score is its cost. The system is marketed as an additional device that must be integrated with existing ECG monitoring equipment. This financial burden limits its accessibility in many healthcare settings. Consequently, the development of a predictive model capable of detecting sepsis using routinely collected hospital data could offer a more cost-effective alternative. Such an approach has the potential to improve neonatal care by reducing false alarms while simultaneously lowering implementation costs.

1.2.2 Necrotizing Enterocolitis (NEC)

Together with sepsis, NEC represents one of the most critical conditions that significantly contributes to morbidity and mortality in NICUs, with mortality rates exceeding 50% in extremely low birth weight infants requiring surgical intervention [33]. Like sepsis, NEC is a life-threatening syndrome characterized by a highly heterogeneous clinical presentation and a multifactorial pathogenesis. While sepsis results from systemic infection and dysregulated immune response, NEC involves severe intestinal inflammation and necrosis, often co-occurring or overlapping with septic states, complicating both diagnosis and management. Similar to sepsis, the pathogenesis of NEC is multifactorial, involving immature intestinal defenses, dysbiosis of the gut microbiota, and dysregulated immune responses, which collectively create a "perfect storm" for mucosal injury [34]. Despite advances in neonatal care, NEC continues to challenge clinicians due to its unpredictable onset, diagnostic ambiguities, and long-term problems such as neurodevelopmental delays and short-gut syndrome [35].

NEC typically manifests within the first two weeks of life, with symptoms ranging from feeding intolerance and abdominal distension to fulminant septic shock [33, 35]. The modified Bell staging system remains the cornerstone for diagnosis and severity classification [33]. Stage I (suspected NEC) presents with nonspecific signs such as temperature instability and gastric residuals, whereas Stage II (definite NEC) includes radiographic evidence of pneumatosis intestinalis or portal venous gas. Stage III (advanced NEC) is characterized by intestinal perforation, pneumoperitoneum, and hemodynamic instability necessitating surgical intervention [33].

Imaging plays a pivotal role in diagnosis. Abdominal radiographs revealing pneumatosis intestinalis have 44–72% sensitivity, while ultrasonography improves detection of portal venous gas and intramural fluid collections [35]. However, diagnostic ambiguity persists, as NEC-like symptoms can overlap with sepsis or spontaneous intestinal perforation. A study conducted on Swiss cohort highlights the significant burden of NEC in neonates with LOS [20]. Among the infections analyzed, NEC accounted for approximately 6% of all LOS cases, with the majority of these cases being HA-LOS rather than CA-LOS. Notably, NEC was associated with an increased risk of mortality, with a hazard ratio of 1.91 (95% CI: 0.51-7.10) compared to other infection sites.

1.2.3 Machine learning for neonatal care

Novel approaches to improve recognition of neonatal sepsis are leveraging advances in technology, particularly in the fields of biomarker discovery, non-invasive monitoring, and machine learning. Recent advancements in ML technology offer an unprecedented capacity to analyze the vast amounts of multidimensional data collected daily as a by-product of care and to model complex problems. This provides unique opportunities to identify patterns and insights that clinicians might miss, improving the detection of conditions where diagnostic accuracy is still a challenge. Additionally, ML technology supports the development of advanced clinical decision support systems, enhancing the efficiency and effectiveness of healthcare delivery.

Recent advances in ML have demonstrated significant potential for early sepsis detection in preterm infants by analyzing high-resolution physiological signals. Traditional approaches such as logistic regression, Naïve Bayes, and k-Nearest Neighbors (kNN), have been explored, with moderate success in classifying sepsis versus control cases based on physiological monitoring signals [9, 10, 36]. These models achieved a mean accuracy of 79% and precision of 82% when identifying sepsis 3 hours before clinical deterioration, highlighting their capacity to detect subtle physiological shifts preceding overt symptoms [6]. More recently, advanced algorithms, including Extreme Gradient Boosting (XGBoost), have demonstrated improved predictive performance by leveraging high-resolution time-series data. In [7], Peng et al. achieved an Area Under the Receiver Operating Characteristic curve (AUROC) of 0.88, demonstrating its effectiveness in combining multiple physiological features. In another study, an explainable ML model has been developed for predicting both LOS and NEC, delivering hourly risk assessments with a median time gain of up to 10 hours before clinical diagnosis [10]. Despite these encouraging results, comparing predictive performance across studies remains inherently challenging due to significant variability in study design, model evaluation protocols, and patient cohorts. Many of the existing studies rely on relatively small, single-center datasets, which limits generalizability and hinders cross-study comparisons [6, 7, 10]. To the best of our knowledge, a large, standardized cohort of neonatal sepsis cases suitable for rigorous model benchmarking has yet to be established, underscoring the need for multicenter studies and harmonized data collection efforts.

A critical aspect of sepsis prediction models is the selection of relevant features that reflect early physiological alterations. Studies have consistently highlighted the importance of Heart Rate Variability (HRV), respiratory parameters, and motion features, which are derived from continuous physiological monitoring. Changes in HRV have been widely studied as early indicators of sepsis. Reduced variability and transient decelerations often precede clinical deterioration, reflecting autonomic nervous system dysregulation [9]. In [8], León et al. introduced a novel approach for early detection of LOS in preterm infants, leveraging visibility graph (VG) analysis of HRV data. Beyond traditional HRV measures, they incorporated VG indices (mean degree, clustering coefficient, transitivity, assortativity) to capture the complexity of HRV dynamics. The results demonstrated that logistic regression, using a feature set including VG features, achieved an AUROC of 87.7% in the six hours

preceding the start of antibiotics, with predictive potential (AUROC above 70%) as early as 42 hours before the start of antibiotics. Respiratory instability, characterized by an increased tendency towards apnea, irregular breathing patterns, and altered oxygen saturation levels, is another key marker of sepsis. Several studies have incorporated features such as respiratory rate variability, inspiration-to-expiration ratio, and entropy measures of respiratory waveforms to enhance prediction models [6]. Lethargy, or the absence of normal infant movement, is a well-documented clinical sign of sepsis. Studies have developed methods to derive motion-related features from electrocardiogram (ECG) and chest impedance (CI) signals, quantifying movement patterns and their disruptions. Features such as cumulative motion duration, frequency of motion events, and signal entropy have demonstrated predictive value [7]. Although integrating motion features with HRV and respiratory parameters has been shown to significantly improve the predictive capabilities of the model, the performance of these motion features was poorer. The cause could be attributed to the nature of the ECG signal since during the acquisition process the signal is processed to filter out any motion-related artifacts [9].

However, these models highly rely on the extracted features based on domain knowledge, with a good model interpretability but at the cost of loosing opportunities to uncover unknown but potentially essential information. Another notable approach is the use of deep learning-based models, which can automatically extract relevant patterns from raw physiological data, reducing the reliance on handcrafted features. While these models hold promise, their complexity and the need for large, high-quality datasets remain key challenges.

Convolutional Neural Networks (CNN) have emerged as a powerful class of models for learning spatial and temporal patterns from physiological time series. In the context of neonatal sepsis prediction, Peng et al. proposed DeepLOS [37], a CNN-based model specifically designed for sepsis prediction using only heart-related signals. The study included 128 preterm infants, of whom 60 had blood-culture-proven LOS, and was evaluated using a 5-fold patient-independent cross-validation strategy. DeepLOS achieved an AUROC of 0.77 on the full dataset and 0.67 on an age-matched subset, outperforming a baseline ResNet model without attention mechanisms. These results highlight the effectiveness of CNN-based models in leveraging HRV patterns for early sepsis detection.

In contrast, Recurrent Neural Networks (RNNs), and particularly their advanced variants such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks, have proven highly effective in modeling temporal dependencies within sequential data. These architectures have been shown to be particularly suitable for the analysis of physiological signals, where long-term dependencies may contain critical diagnostic information. Notably, Leon et al. investigated the use of RNNs for early detection of LOS, employing a feature-based approach instead of directly processing raw signals [38]. Their model incorporated a combination of both LSTM and GRU layers. The model was trained on a dataset of 259 infants, utilizing HRV-derived features as input instead of raw heart rate waveforms. By incorporating handcrafted features extracted from HRV signals, the model aimed to capture temporal trends associated with sepsis onset. Their model demonstrated strong performance

in sepsis prediction, achieving an AUROC of 90.4% when identifying LOS up to six hours prior to clinical diagnosis. This result highlights the potential of RNN-based architectures to effectively capture temporal dependencies in structured physiological data for early sepsis detection.

Limitations of using machine learning for sepsis detection

Numerous studies in both adults and children have sought to identify predictive signatures of sepsis based on specific patterns of clinical and biological alterations [39, 40]. However, only a small subset of these studies have demonstrated sufficient predictive performance to support clinical implementation. Among the ML-based sepsis alert systems that have been integrated into bedside care, randomized controlled trials and post-implementation studies have shown limited patient benefit [40]. In some cases, these tools have even resulted in unintended negative consequences, such as alarm fatigue, unnecessary sepsis evaluations, and antibiotic overuse.

Several factors contribute to the modest performance and lack of clinical translation of these models. First, many algorithms are developed under the assumption of a universal sepsis signature, despite sepsis being a highly heterogeneous syndrome with distinct clinical and biological subtypes [5, 41, 42]. The lack of well-defined sepsis phenotypes, combined with the inclusion of both suspected and confirmed infection cases, reduces the signal-to-noise ratio, further complicating model development. Second, distinguishing sepsis-related alterations from normal physiology and inflammatory responses is particularly challenging in neonates due to the profound physiological changes occurring during the transition to extrauterine life and postnatal development, as well as the influence of preexisting conditions and comorbidities. The failure to account for this evolving physiology represents a significant limitation in the design of neonatal sepsis prediction models. Third, many predictive models have been trained on limited patient cohorts [6, 43], or a restricted set of clinical variables [8, 9, 36], limiting their robustness and applicability. Despite their predictive potential, these features which were typically used present challenges. HRV can be influenced by gestational age and other comorbidities, respiratory instability may arise from non-infectious causes, and motion-related signals can be affected by external factors such as caregiver interventions. These limitations highlight the need for robust models capable of distinguishing sepsis-specific patterns from normal physiological variability. Furthermore, training models using data obtained only after sepsis becomes clinically apparent offers minimal added value for clinicians. Fourth, due to the inherent difficulty of assembling harmonized, high-quality, multicentric neonatal datasets, many existing studies rely on single-center databases [6, 9], which raises concerns regarding model generalizability. Finally, overrepresentation of sepsis cases in training datasets and the use of inappropriate performance metrics for imbalanced data, such as accuracy, often lead to an overestimation of model effectiveness [8, 38]. Addressing these challenges is critical for advancing the clinical applicability of ML-based neonatal sepsis prediction models.

Therefore, adopting a FM approach may offer a potential solution to several of the limita-

tions associated with traditional machine learning models. Rather than relying on a fixed, universal signature of sepsis, FMs could support the identification of distinct patient subpopulations and disease phenotypes, thereby facilitating the recognition of sepsis endotypes. This capacity to cluster similar patterns of physiological perturbation may enhance early detection while reducing false positives and avoiding unnecessary interventions. Furthermore, the adaptability of FMs to different hospitals, populations, and clinical settings could improve robustness to distribution shifts [14, 44], one of the key challenges faced by conventional ML models trained on single-center datasets. While further validation is necessary, these characteristics suggest that FMs might enable the development of scalable and flexible patient representations, with the potential to improve the accuracy, generalizability, and clinical relevance of decision-support systems for neonatal sepsis.

1.2.4 Foundation models in healthcare

The rapid evolution of Artificial Intelligence (AI) in healthcare has ushered in a new paradigm centered on FMs. Models of this kind, epitomized by breakthroughs like ChatGPT in natural language processing [45] and AlphaFold in protein structure prediction [46], are now being actively explored for their potential to revolutionize clinical decision-making, operational efficiency, and patient outcomes. In the context of Electronic Health Records (EHRs), foundation models aim to unify structured data (e.g. lab values, diagnoses) and unstructured clinical text (e.g., progress notes, radiology reports) into cohesive representations that enable robust predictions and insights.

Foundation models

The advent of FMs represents a paradigm shift in AI, offering unprecedented versatility in adapting ML systems to solve multiple tasks and opening new possibilities for various applications, including healthcare. Defined as "models trained on broad data at scale that can be adapted to a wide range of downstream tasks" [11], FMs are characterized by their ability to learn from vast amounts of diverse data and adapt to a large range of downstream tasks with minimal task-specific training. These systems serve as a versatile base for numerous applications, offering a promising approach for tackling complex problems like sepsis prediction in neonatal care.

FMs are characterized by an initial training on massive unlabeled datasets using objectives that force the system to infer missing information. For text, this includes masked language modeling [47] where the model predicts obscured words from context. Vision-language models like CLIP [48] extend this paradigm by learning joint embeddings from image-text pairs, demonstrating that cross-modal pretraining enhances generalization. Large models like GPT-3 [45] with 175 billion parameters exhibit emergent few-shot learning abilities absent in smaller variants. These unprecedented adaptive capabilities can be attributed to the scalability of the new architectures used to define such models, transformers above all [49], the diversity and sheer volume of datasets used during the model train, and the computational

resources available [11].

One of the most remarkable aspects of foundation models is their capacity for few-shot and zero-shot learning. Few-shot learning allows the model to perform well on new tasks with only a small number of examples, called "shots", while zero-shot learning enables the model to tackle entirely new tasks without any specific training examples. The ability to perform tasks without explicit training examples (zero-shot) or with minimal demonstrations (few-shot) emerges from the model's pretraining on diverse data distributions. For instance, GPT-3 achieves competitive performance on clinical question answering by conditioning on relevant context passages [45]. In [11], Bommasani et al. hypothesize that this capability stems from the model's internal representation space capturing latent relationships between medical concepts during pretraining.

These capabilities could prove particularly valuable in medical applications where labeled data may be scarce or difficult to obtain [50]. In sepsis prediction, for example, few-shot learning could enable rapid adaptation to neonatal populations with distinct physiological baselines. At least in theory, such models, pretrained on heterogeneous patient data, might infer risk patterns from few examples of confirmed cases, significantly reducing data collection burdens.

In the realm of healthcare, foundation models are can be categorized into Clinical Language Models (CLaMs) and Foundation models for Electronic Medical Records (FEMRs) [13].

CLaMS

Clinical Language Models (CLaMs) are a specialized class of Large Language Models (LLMs) tailored to process and generate biomedical text. Unlike general-purpose LLMs like GPT-4 [51], which are trained on internet-scale corpora, CLaMs undergo domain-specific pretraining on clinical notes, scientific literature, and medical ontologies [13]. This specialization enhances their ability to parse jargon-laden narratives, extract entities (e.g., medications, diagnoses), and generate contextually accurate text. These capabilities are critical for applications such as automated documentation, patient communication, and decision support.

CLaMs typically employ transformer-based architectures, which use self-attention mechanisms to model long-range dependencies in text. For example, models like BioBERT [19] and ClinicalBERT [52] are initialized with weights from general-domain BERT [47] and fine-tuned on clinical text from MIMIC-III or proprietary hospital datasets. In recent years, research in medical AI has shifted from specialized models toward large-scale CLaMs, leveraging extensive pretraining to enhance adaptability and performance across diverse medical tasks. Two notable models leading this shift are MEDITRON-70B [17] and Med-PaLM [53], each contributing uniquely to the field. MEDITRON-70B, an open-source model, builds on Llama-2-70B, further pre-trained on a specialized medical corpus including PubMed articles, clinical guidelines, and general-domain texts. Med-PaLM, developed by Google Research, is a closed-source model based on PaLM (540B parameters) and fine-tuned for medical tasks via instruction prompt tuning. However, as noted in [13], human evaluations indicated that

while impressive, most of these models are trained on narrowly scoped datasets, limiting their utility for tasks requiring cross-institutional generalization.

CLaMs for time series analysis Based on the astonishing performance achieved by CLaMs, recent advancements have explored the adaptation of LLMs for time series analysis in healthcare. A notable example is MedTsLLM, introduced by Chan et al., which leverages the power of pretrained LLMs to analyze multimodal medical time series data [54]. This innovative approach treats both language and time series as sequences with similar underlying patterns, enabling the model to harness the extensive knowledge and reasoning capabilities of LLMs for complex physiological signal analysis. MedTsLLM employs a reprogramming layer to align embeddings of time series patches with the pretrained LLM's embedding space, effectively integrating raw time series data with rich contextual information in the form of text [54]. This framework not only performs traditional classification tasks but also extends to more nuanced analyses such as semantic segmentation, boundary detection, and anomaly detection in physiological signals. By incorporating patient-specific information into the text prompts and developing novel methods to handle multiple covariates, MedTsLLM demonstrates superior performance across various medical domains, including electrocardiograms and respiratory waveforms [54]. This approach represents a significant step forward in harnessing the potential of LLMs for medical time series analysis, offering a more holistic and context-aware method for interpreting complex physiological data.

FEMRs

Foundation Models for Electronic Medical Records (FEMRs) are large-scale AI pretrained models designed to process structured EHRs data like sequences of diagnoses, medications, lab results, and procedures, into unified patient representations. These embeddings condense a patient's medical history into fixed-dimensional vectors, enabling predictions across hundreds of endpoints (e.g., readmission risk, treatment response) using simple downstream models [13].

FEMRs typically use transformer architectures to model temporal sequences. In [18], Li et al. introduced BEHRT, a model that processes International Classification of Diseases (ICD) codes, globally used to represent diagnoses and medical conditions in EHRs, as token sequences, analogous to words in a sentence, enabling the prediction of future diagnoses. Other approaches, tried to integrate structured data with clinical text using multimodal transformers [55]. A key challenge, however, is the heterogeneity of EHRs data: lab values, vital signs, and free-text notes vary in frequency, scale, and semantic meaning, complicating the creation of cohesive input representations [13]. Recent advancements have pushed the boundaries of healthcare AI with models capable of processing complex, multimodal electronic health data. A groundbreaking example is ETHOS, introduced by Renc et al., which represents a significant leap forward in the application of transformer architectures to healthcare analytics [12]. ETHOS innovatively adapts the transformer model, originally designed for natural language processing, to analyze high-dimensional, heterogeneous, and

episodic health data [12]. The model's core strength lies in its ability to process Patient Health Timelines (PHTs), represented as detailed, tokenized records of health events, to predict future health trajectories. ETHOS employs a zero-shot learning approach, eliminating the need for labeled data or task-specific fine-tuning, which has been a major bottleneck in healthcare AI deployment. This capability allows ETHOS to perform a wide range of predictive tasks, from mortality risk assessment to length of stay estimation, without additional training. The model's architecture enables it to handle noisy, inconsistent data typical of real-world electronic health records, demonstrating robustness that is crucial for large-scale AI applications in healthcare [12]. However, a critical limitation lies in its handling of out-of-vocabulary (OOV) tokens. The model's reliance on a fixed vocabulary derived from the training data means that it struggles to effectively process new or unseen medical concepts, potentially leading to inaccurate predictions or requiring resource-intensive retraining. This constraint poses a significant challenge for real-world deployment, particularly in resource-limited healthcare settings where retraining is often infeasible.

FEMRs for time series analysis Recent advancements in time series analysis have led to the development of foundation models that leverage large-scale pretraining to enhance forecasting capabilities. Notable contributions include decoder-only architectures trained on extensive time-series data, achieving zero-shot forecasting performance comparable to state-of-the-art supervised models. Chronos introduces a novel approach by tokenizing time series data similarly to natural language, leveraging transformer-based models to significantly outperform traditional statistical and deep learning methods [56]. MOMENT extends this by providing an open-source suite of pretrained models, capturing complex temporal dependencies to improve generalizability across domains [57]. The "One Fits All" approach explores the transferability of pretrained language models to time series tasks, demonstrating competitive performance in forecasting, classification, and anomaly detection, with self-attention mechanisms functioning similarly to principal component analysis [58]. Additionally, TimesFM, a large-scale time series foundation model, has been introduced to generalize across multiple forecasting benchmarks while leveraging a self-supervised pretraining strategy to enhance adaptability and robustness [59]. Despite these innovations, most of these models remain primarily focused on forecasting, leaving time series classification relatively underexplored. Addressing this gap could extend the impact of foundation models to broader applications, including critical domains such as healthcare.

Limitations of foundation models

Foundation models have garnered significant attention in AI research due to their ability to generalize across diverse tasks. However, their integration into healthcare, particularly in high risk settings like NICUs, presents several challenges. NICUs are specialized hospital departments dedicated to the care of critically ill or premature newborns, where patients are continuously monitored and clinical conditions can deteriorate rapidly. In such settings, the application of FMs must be approached with caution, ensuring that their deployment aligns with strict clinical standards.

A primary concern is the substantial computational resources required by FMs. Training and deploying these models necessitate advanced hardware, such as high-performance GPUs, leading to elevated energy consumption and operational costs. This poses a barrier for healthcare institutions with limited budgets and infrastructure. Models with billions of parameters, such LLMs, present significant challenges for hospital environments with limited computational resources. Training or fine-tuning these expansive models is often unfeasible due to the substantial infrastructure required. Consequently, alternative strategies like prompt tuning have emerged, allowing adaptation of LLMs to specific tasks without extensive retraining [60]. Similarly, Retrieval-Augmented Generation (RAG) systems enhance model responses by integrating external knowledge bases, thereby reducing the need for exhaustive internal training [61]. However, it's important to note that even inference operations with these large-scale models can be resource-intensive, further complicating their practical deployment in clinical settings.

A significant challenge lies in the lack of comprehensive benchmarks tailored to the complexities of clinical diagnostics. Traditional evaluation metrics often rely on medical question-answering datasets, which may not accurately reflect the nuanced decision-making processes inherent in clinical settings. For instance, a study assessed the performance of LLMs in providing clinical recommendations within the Emergency Department (ED). The researchers evaluated GPT-4's ability to predict admission status, radiological investigation requests, and antibiotic prescriptions based on clinical notes. While the model demonstrated promising results, achieving accuracies of 86.5%, 41.5%, and 84.0% for the respective tasks. These results, in addition to being in line with the performance achieved by clinicians, also highlighted the challenges LLMs face in complex diagnostic reasoning [62]. Similarly, another study explored the use of LLMs to assess clinical acuity in the ED by classifying patient histories based on severity. The model accurately identified higher acuity cases, suggesting potential in triage applications. However, the study underscored that LLMs should complement, not replace, clinical judgment, emphasizing the necessity for robust evaluation frameworks to ensure safe integration into clinical workflows [63]

In the realm of time-series analysis, particularly within critical care, the scarcity of large, well-annotated datasets poses a significant obstacle to model development and assessment. Addressing this gap, Burger et al. introduced a harmonized dataset combining multiple critical care datasets to enhance patient diversity and model robustness [44]. This initiative aims to establish a foundation for training and evaluating large-scale time-series models, facilitating research into sequence modeling, transfer learning, and generalization across diverse clinical settings.

One key limitation of FMs in clinical applications is their reduced interpretability. While their complex, deep architectures contribute to improved predictive performance, they also obscure the underlying decision-making process compared to traditional ML models. In clinical settings, where transparency and accountability are critical for trust and adoption, this lack of interpretability poses a significant barrier. Clinicians must be able to understand, verify, and justify the model's predictions, especially when used in high-stakes environments

like NICUs. To address this issue, explainability tools—such as attention visualization, must be integrated into the model evaluation pipeline. These tools can help elucidate which features contribute most to a given prediction, providing essential insights for clinicians and ensuring that the model's outputs align with established medical reasoning.

Accessibility also remains a pressing issue. Many FMs are not fully open-sourced; while some may offer open weights, they often lack complete transparency. In the medical domain, data privacy regulations further restrict the sharing of models trained on sensitive information, limiting collaborative advancements and external validation efforts [13].

Certification and regulatory approval of FMs for clinical use present additional challenges. The European Union's Artificial Intelligence Act aims to establish a legal framework for AI applications, including those in healthcare. However, the act's implications for the certification of FMs remain under discussion, highlighting the need for clear guidelines to ensure their safe and effective deployment in medical settings [64].

In summary, while FMs hold promise for advancing healthcare AI applications, addressing these limitations is crucial to facilitate their integration into clinical practice.

Chapter 2

Materials and Methods

2.1 Data and preprocessing

2.1.1 Patients and data acquisition

This study is a retrospective, multicenter investigation conducted in Switzerland and approved by the Ethics Committee of the Canton de Vaud (CER 2022-00528). The study enrolled over 15,000 newborns, defined as infants younger than 28 days for those born at a gestational age of 37 weeks or greater, and as infants with a postmenstrual age, defined as the sum between gestational age and plus postnatal age, below 44 weeks for those born before 37 weeks. Patients were drawn from three hospital centers: Lausanne University Hospital (CHUV), University Children's Hospital Zurich (KiSpi), and University Hospital Zurich (USZ), with enrollment periods spanning from January 1, 2007 to June 30, 2023, January 1, 2015 to June 30, 2023, and March 25, 2019 to June 30, 2023, respectively. Patients with any form of consent refusal were excluded from the study.

Table 2.1 summarizes the characteristics of the patient population included in this study. A notable variation in gestational age distribution was observed across the three participating centers. Among the two hospitals in Zurich, USZ primarily admits more premature neonates, with a median gestational age of 35.3 weeks, compared to 36.9 weeks at CHUV. Conversely, KiSpi treats relatively older neonates, with a median gestational age of 38.4 weeks. This trend is also reflected in birth weight distributions, as shown in Figure 2.1, which illustrates the expected correlation between higher gestational age and greater birth weight.

The presence of two hospitals within the same city (Zurich) introduces a potential confounding factor that must be considered when interpreting these data. Unlike Lausanne, where all neonates are admitted to CHUV, in Zurich, premature infants are initially admitted to USZ and are later transferred to KiSpi only in cases of severe complications. KiSpi, in contrast to USZ, is a highly specialized center dedicated to the treatment of neonates and pediatric patients, handling only the most critical cases or those requiring high-risk surgical interven-

Table 2.1: Demographic characteristics of neonatal patients across the three centers

	Hospital		
	CHUV (N=11297)	KiSpi (N=1540)	USZ (N=2524)
Sex (%)			
Female	4851 (42.9%)	668 (43.4%)	1102 (43.7%)
Male	6446 (57.1%)	872 (56.6%)	1422 (56.3%)
Gestational age (weeks)			
Mean	36.1	37.7	34.86
Median	36.9	38.4	35.3
90% CI	[28 - 41.1]	[31.3 - 41.14]	[26.1 - 40.9]
Length of stay (days)			
Mean	16.8	31.1	16.6
Median	7	17	7
90% CI	[1 - 66]	[2 - 115.6]	[2 - 72.9]
Death (%)			
Yes	372 (3.3%)	144 (9.4%)	55 (2.2%)
In-hospital	213 (0.9%)	66 (4.2%)	30 (1.2%)
Birth weight (kg)			
Mean	2.185	2.904	1.800
Median	2.090	3.000	1.700
90% CI	[0.729 - 3.820]	[1.311 - 4.000]	[0.635 - 3.410]
Sepsis (%)			
Yes	295 (2.5%)	25 (1.6%)	98 (3.9%)
EOS	72 (24.4%)	0 (0%)	40 (40.8%)
LOS	223 (75.6%)	25 (100%)	58 (59.2%)
NEC (%)			
Yes	67 (0.6%)	10 (0.6%)	18 (0.7%)
Antibiotics (5 days) (%)			
Yes	1298 (11.2%)	401 (25.8%)	216 (8.6%)

tions. As a result, it is not surprising that the gestational age distribution at KiSpi is shifted toward older neonates. This patient selection bias is further reflected in clinical outcomes, including higher mortality rates and increased antibiotic usage at KiSpi. As confirmed by KiSpi's clinical staff, the complexity of cases managed at their institution, coupled with the high volume of daily surgical procedures, justifies the extensive use of antibiotics and the elevated mortality rate. Regarding sepsis prevalence, the overall incidence across all centers was approximately 3%, which is consistent with previous studies conducted in Switzerland [20]. The prevalence of LOS was slightly lower, at 2%. A notably lower prevalence of sepsis

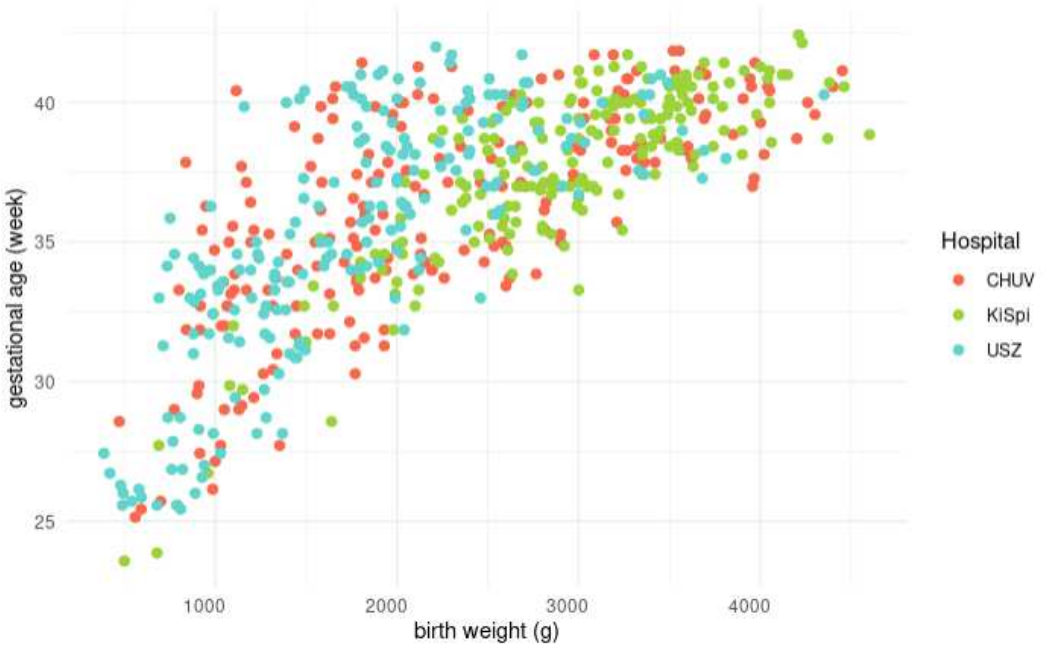


Figure 2.1: Scatter plot of birth weight versus gestational age across the three centers. This plot highlights differences in patient distributions between centers. Notably, the CHUV distribution spans both USZ and KiSpi, where patients are more distinctly separated between the two hospitals.

cases was observed at KiSpi where no cases of EOS were recorded. This pattern is likely attributable to the previously described patient transfer dynamics, where extremely premature neonates, typically associated with a higher risk of EOS, are primarily treated at USZ before transfer to KiSpi. An apparent contradiction arises in the data, where higher mortality at KiSpi coincides with a lower incidence of sepsis. However, as emphasized by the clinical staff, sepsis is not among the leading causes of mortality at KiSpi, further supporting the hypothesis that the high mortality rate is driven by the complexity of the cases admitted rather than by sepsis-related complications.

The HeRO score was not included in the present study, as it is not used in the neonatal units at either USZ or KiSpi. Currently, CHUV is the only center among the three that routinely employs it in clinical practice. To ensure a meaningful comparison with this established clinical tool and to provide insights into its potential impact within NICU, the analysis was restricted to newborns with a gestational age of less than 32 weeks, the target population for which the HeRO score was originally developed.

For sepsis, the time of diagnosis, typically called "Cultures, Resuscitation, and Antibiotics Started Here (CRASH) moment" [7], was defined as the precise time at which a positive blood culture was obtained. Given the critical importance of data quality, it was essential to rigorously exclude cases of sample contamination, often arising from inadequate disinfection

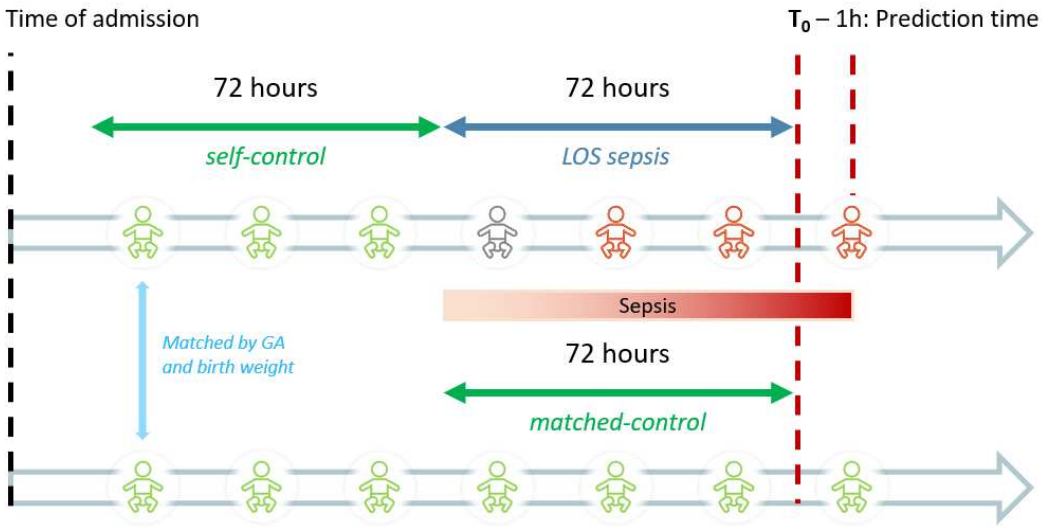


Figure 2.2: Patients stratification and data extraction strategy. Patients were divided into three groups: (i) Sepsis, including 72-hour time series ending 1 hour before the confirmed sepsis diagnosis (CRASH moment); (ii) Self-control, consisting of the 72-hour window immediately preceding the sepsis segment for the same patient, serving as an intra-patient control; and (iii) Matched-control, comprising healthy patients matched to sepsis cases based on gestational age (GA) and birth weight. For matched controls, time series were aligned using the corresponding sepsis case's diagnosis time to extract the same 72-hour window.

during sample collection, by having all positive blood cultures reviewed by an experienced physician, with any suspicious cases removed from the dataset. Newborns diagnosed with proven HA-LOS, defined as a positive blood culture occurring more than 72 hours after admission, were assigned to the case group. In contrast, the control group comprised newborns with no recorded death, no proven sepsis, no positive diagnosis of NEC, defined as Bell Stage $\geq$ 2a, and less than 120 hours of antimicrobial treatment.

To construct the dataset for sepsis prediction, physiological time series data were extracted from these HA-LOS cases using a 72-hour window preceding the CRASH event. Data collection was truncated one hour before the diagnosis time to simulate a clinically relevant prediction scenario and prevent leakage of post-diagnostic information. The 72-hour interval was selected based on clinical consensus, reflecting a conservative approach designed to fully capture the period in which infection could plausibly have developed. Given the absence of a precise timestamp indicating sepsis onset, this strategy aims to ensure that the infection is very likely to have emerged within the defined window, although symptoms may have appeared more recently.

To enhance the robustness and validity of the study design, two types of controls were incorporated: "matched-controls" and "self-controls". This dual-control approach facilitates a comprehensive assessment of physiological differences associated with sepsis onset by cap-

turing both inter-individual and intra-individual variability. This methodology is consistent with previously established frameworks in the literature [7], and has proven effective for sepsis detection studies.

Each sepsis case was paired with a control patient based on gestational age, postnatal age, and birthweight, with the underlying assumption that patients with similar characteristics would exhibit comparable temporal dynamics in the absence of sepsis. The matching process was performed in two steps. Initially, the search was restricted to a subset of patients, called "supercontrols", who met all control inclusion criteria and had no prior diagnosis of confounding secondary pathologies such as retinopathy of prematurity (ROP), bronchopulmonary dysplasia (BPD), periventricular leukomalacia (PVL), intraventricular hemorrhage (IVH). When a suitable match could not be found within this subgroup, the search was extended to include "imperfect controls." For every matched control, an "equivalent CRASH moment" was computed to ensure that the analysis period corresponded to a similar postmenstrual age as that of the sepsis case.

In addition, a "self-control" period was defined for each case by selecting the 72-hour interval preceding the 72-hour time window leading up to sepsis diagnosis. This approach allows the same patient to serve as their own control, providing a baseline representation of their physiological state before the onset of infection. By comparing a patient's data before sepsis with the period immediately preceding the clinical diagnosis, this method helps isolate the temporal dynamics associated with disease progression, reducing the influence of inter-individual variability. Collectively, these matching strategies are designed to minimize bias and ensure that the model is trained on a dataset that accurately reflects the complex temporal and individual variability inherent in neonatal sepsis.

A schematic summary of the cohort stratification and data extraction process is presented in Figure 2.2. The final dataset comprised a total of 471 patients, equally distributed among sepsis, self-controls, and matched-control patients.

Multivariate physiological monitoring in neonatal care

In the realm of medical diagnostics, clinicians routinely synthesize information from multiple sources to form comprehensive assessments of patient health. This integrative approach is crucial, as it mirrors the complexity of human physiology and the multifaceted nature of diseases. With time series data, this practice translates into the utilization of multivariate time series, where multiple physiological variables are monitored over time. The importance of leveraging multivariate time series in medical diagnostics is underscored by the fact that diseases often manifest through intricate interactions among various physiological parameters.

During clinical care, multiple physiological signals were continuously monitored for each patient. For this study, the selected signals included heart rate (HR), oxygen saturation (SpO_2), systolic blood pressure (SBP), diastolic blood pressure (DBP), and respiratory rate (RR). These signals provide a comprehensive, multi-dimensional representation of the pa-

tient's physiological state, aligning with the approach used by clinicians in neonatal health assessment. To further illustrate the variability of these signals across different centers, detailed visualizations of their trajectories are provided in Appendix B. Each plot presents the temporal evolution of a specific physiological signal, stratified by gestational age, across all participating centers, offering additional insight into inter- and intra-institutional differences.

Heart rate serves as a critical indicator of autonomic nervous system function, reflecting compensatory mechanisms in response to stress, infection, or hemodynamic instability. Oxygen saturation provides insight into pulmonary and cardiovascular efficiency, while blood pressure measurements reflect circulatory function and perfusion status. Additionally, respiratory rate and variability in breathing patterns are strong markers of respiratory distress and metabolic adaptation, both of which are crucial in neonatal critical care. The integration of multiple physiological parameters is essential for capturing the complex interactions that underlie systemic responses to infection. By incorporating signals from different physiological systems, this approach enhances the model's ability to detect early signs of clinical deterioration and capture subtle, nonlinear patterns that may indicate sepsis onset. This multimodal strategy is particularly relevant in neonatology, where patient variability is high and isolated features may fail to provide sufficient predictive power. Through the fusion of cardiovascular and respiratory parameters, the model can better approximate clinical decision-making processes, improving early detection and risk stratification in neonatal sepsis prediction.

BioMedIT

One of the primary challenges in multicentric studies is the secure sharing of patient data, as privacy regulations and institutional policies impose strict limitations on data accessibility. Ensuring data confidentiality while enabling collaborative research requires robust infrastructures that facilitate controlled data exchange without compromising patient privacy. In this study, all data were securely stored and processed within BioMedIT [65], a federated and secure computing infrastructure specifically designed to support privacy-preserving biomedical research in compliance with Swiss legal and ethical standards. BioMedIT enables end-to-end encrypted data transfer, with data from the participating centers shared in RDF format to ensure standardized interoperability. The platform enforces controlled access through two-factor authentication and provides isolated computational environments, ensuring the confidentiality and integrity of health data across institutions.

All analyses were conducted exclusively within the BioMedIT environment, ensuring adherence to data protection policies. The software tools used for data processing and analysis included Visual Studio Code (version 1.98) and RStudio (version 2023.12.1), while the programming languages employed were Python (version 3.10) and R (version 4.4.2).

2.1.2 Data harmonization

Despite the implementation of robust security measures, sharing medical data across multiple hospitals remains a complex challenge due to heterogeneous data storage formats and

institution-specific nomenclatures. Each hospital follows its own data management protocols, necessitating a standardization process before integration within BioMedIT. To address these challenges, the Swiss Personalized Health Network (SPHN) has developed a national framework aimed at standardizing the semantic representation of health data, ensuring interoperability and compliance with FAIR (Findable, Accessible, Interoperable, Reusable) principles.

In this study, a knowledge graph approach was adopted to harmonize data from the three participating hospitals. The graph structure was defined by the SPHN RDF schema [66], which encodes semantic relationships to facilitate consistent data exchange and integration. This schema relies on widely recognized external medical terminologies, including ICD-10-GM [67], SNOMED-CT (Swiss extension 2021) [68], ATC (2016-2024) [69], UCUM (2024) [70], LOINC (version 2.76) [71], and CHOP (2013-2024) [72], the Swiss surgical classification catalogue. To store, access, and query the standardized dataset, GraphDB (version 10.8.0) [73] was employed as the primary database management system. SPARQL, a query language specifically designed for RDF (Resource Description Framework) data, was used to extract and analyze relevant information while preserving the semantic relationships embedded within the knowledge graph. The flexibility of GraphDB's SPARQL interface enabled efficient data retrieval, supporting complex queries essential for the integration of heterogeneous clinical datasets.

A comprehensive data harmonization process was undertaken by all three participating hospitals to ensure semantic interoperability across institutions. This process required considerable effort from all stakeholders. On one hand, each hospital performed a detailed mapping of institution-specific datasets to align with the SPHN schema and integrate them into the defined graph structure. On the other hand, rigorous quality control procedures were implemented to detect errors or inconsistencies, ensuring a final dataset that remained consistent across all centers. Several factors, including heterogeneous hospital infrastructures and ambiguities within the adopted standard, contributed to slow down this process.

A typical hospital infrastructure consists of multiple interoperable systems designed to handle various types of clinical data. While interoperability is often ensured, integrating heterogeneous data sources remains a non-trivial challenge, requiring substantial effort from data managers. At CHUV, for instance, data were retrospectively collected from two main systems: the Data Warehouse (DWH) and MetaVision (MV), a clinical information system specifically designed to support critical care workflows, particularly within NICUs. As MV is the primary source for neonatology data, some of its entries are later remapped into the DWH.

One notable challenge is the archiving of physiological signals, where the same parameter can be recorded multiple times using different methods. This inconsistency introduces challenges related to data reliability, as different measurement devices might have varying degrees of precision and accuracy, thereby adding further complexity to data analysis. For instance, HR can be recorded using three distinct methods, each varying in precision. The most accurate

measurement is obtained when the patient is connected to an electrocardiogram (ECG), from which HR is directly extracted. If the patient's condition is not critical and continuous ECG monitoring is unnecessary, HR is manually recorded by the nursing staff and documented in MV. In this case, HR sampling is discontinuous and irregular, as nurses decide when to measure and record the data. Additionally, when a patient presents symptoms suggestive of sepsis, HR may be sampled more frequently, whereas for stable patients, sampling occurs at longer intervals. The third method estimates HR from the SpO_2 signal. Although this approach is less precise, it is always available, providing an approximate HR value when direct measurement is not feasible. A similar sampling variability is observed for RR. In critically ill patients requiring intubation, RR is continuously measured by the mechanical ventilator. In contrast, for non-intubated patients, RR is manually assessed and documented by healthcare professionals in MV.

To address these discrepancies and integrate multiple data sources into a single, unified signal, an algorithm was developed to prioritize measurements from the most accurate source. For example, if an ECG-derived HR value was available, all less accurate measurements were discarded. The ranking of measurement methods for each physiological signal was determined in collaboration with a clinical team, ensuring that the final dataset maintained optimal accuracy and clinical relevance.

Beyond the technical challenges of integrating heterogeneous hospital systems, clear and consistent standardization is essential for effective data harmonization. In the case of the SPHN schema, its flexible design introduced ambiguity in the mapping of clinical concepts, delaying dataset definition. A key example involves the encoding of administered medications: although all centers adhered to the schema, they employed different terminologies. CHUV used SNOMED-CT to capture detailed substance-level information, KiSpi adopted ATC for its broader classification, and USZ inconsistently applied both. According to USZ data managers, this inconsistency arose from the presence of three independent systems used for drug classification, one of which enables automatic mapping from ATC to SNOMED-CT. This example highlights the inherent complexity of constructing a well-curated and consistent multicentric dataset. Even when adhering to a common standard, differences in local infrastructure, terminology preferences, and system configurations can lead to significant discrepancies. This underscores the need for close coordination between institutions and rigorous standardization protocols when working with large-scale clinical data across multiple centers.

2.1.3 Data preprocessing and cleaning

In this study, we opted to retain the raw physiological signals without applying any filtering or smoothing techniques. This decision was motivated by the nature of sepsis-related patterns, which often manifest as transient and short-lived variations in physiological signals, occurring on a timescale shorter than one minute. Given that our dataset consists of minute-by-minute sampled data, applying smoothing or averaging techniques could lead to the loss of crucial

information indicative of sepsis onset. Filtering the signals might inadvertently attenuate or eliminate these subtle but clinically relevant patterns, thereby reducing the model's ability to detect early signs of sepsis.

Missing values

Through the MV portal, nurses could document specific measurements obtained for individual patients. The presence of multiple measurement modalities, sometimes dependent on the patient's clinical condition, resulted in irregular signal sampling. To obtain a uniformly sampled time series with one-minute resolution, missing data were imputed using a last observation carried forward approach, also known as "forward fill". For signals sampled at a low frequency, such as SBP and DBP, the resulting time series exhibited the typical pattern of a piecewise constant function.

Data normalization

Following the imputation of missing values, all signals were normalized using a statistical approach to ensure appropriate feature scaling. Normalizing input variables prior to model training is essential to maintain a balanced contribution of different features, preventing biases related to differences in measurement scales. A global normalization strategy was applied, wherein all available data across patients were aggregated. Mean and standard deviation were computed across the entire cohort, and each signal was standardized accordingly.

Feature engineering

Each physiological signal was initially segmented into non-overlapping 1-hour intervals, with each segment serving as a potential input for the model. Subsequently, a set of statistical measures, including the mean, median, and standard deviation, was computed for each 1-hour segment.

To assess the distributional characteristics of the extracted features, skewness (*skew*) was computed, providing insights into the asymmetry of the data. For example, symptoms such as lethargy, which are often associated with sepsis, may result in a respiratory rate distribution that is skewed toward lower values, manifesting as a negative skewness index [7]. In addition, kurtosis (*kurt*) was calculated to further characterize the distribution, while an entropy-based analysis was conducted to quantify the degree of disorder within the signal. Given the nonspecific and heterogeneous nature of sepsis, which is often characterized by the absence of well-defined repeated patterns, three different entropy measures were employed to capture varying aspects of signal complexity: Approximate Entropy (*ApEn*) [74], Sample Entropy (*SampEn*) [75], and Fuzzy Entropy (*FuzzEn*) [76]. By leveraging these three complementary entropy measures, we aimed to comprehensively assess the complexity of physiological signals and their potential relationship with sepsis-related dysregulation.

A major challenge in working with preterm neonates is managing the substantial intra- and

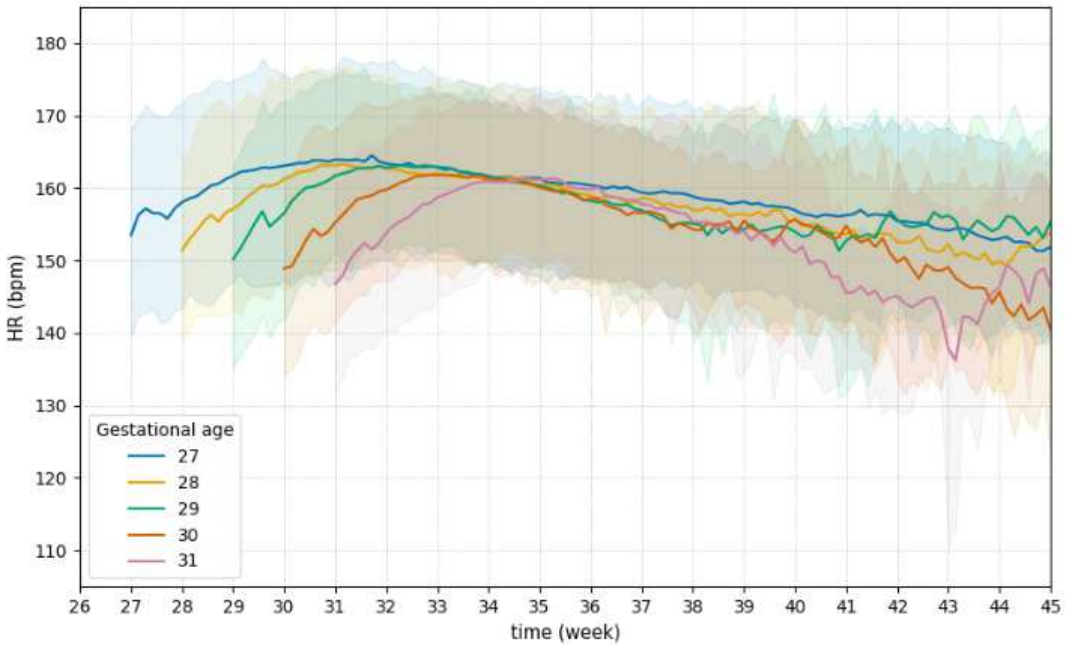


Figure 2.3: Heart rate (HR) trajectories stratified by gestational age in the CHUV cohort. Each line represents the mean HR trend for a specific gestational age, with shaded areas indicating the 90% confidence interval. Notably, the wide confidence intervals highlight the substantial intra-group variability observed across all gestational age groups.

inter-individual variability inherent in their physiological data. As highlighted by Giannoni et al., the evolution of key vital signs during the early weeks of development is strongly influenced by the neonate's gestational age [20]. This trend is further confirmed in Figure 2.3, which illustrates the distribution of HR across different gestational age groups in the CHUV dataset. The figure clearly demonstrates distinct HR patterns depending on gestational age, aligning with the observations reported by Giannoni et al. . To account for this variability, the Mahalanobis distance (*dist*) was employed as a metric to assess an individual's similarity within a group, with the distance measure weighted according to the estimated variability among patients of the same gestational age.

Apnea is one of the most common symptoms observed in neonates developing sepsis, often manifesting in the hours leading up to diagnosis [6, 7]. However, detecting apnea events from minute-by-minute physiological signals presents a significant challenge, as these episodes typically last for less than one minute, making them difficult to capture when analyzing signals at a fixed temporal resolution. Traditional feature extraction approaches may fail to retain this critical information, potentially reducing the model's ability to recognize early signs of deterioration. To address this limitation, two additional variables were computed for each signal to quantify the average number of spikes (*spikes*) and drops (*drops*) within each 1-hour segment, effectively capturing fluctuation patterns that may correspond to clinically relevant events, such as apnea episodes. . A third variable was also added to quantify the

total number of events considering both spikes and drops (*outliers*). These features allows the model to retain information about transient, high-frequency events that might otherwise be lost in the aggregation process, thereby improving its ability to detect subtle, short-term physiological disruptions indicative of sepsis onset.

Furthermore, traditional ML models based on tabular data lack an internal mechanism to capture the temporal dynamics of patients' physiological state. To mitigate this limitation, each computed feature was aggregated over a variable time window extending up to 72 hours. This approach provides the model with both the recent patient history, represented by features from the last hour, and the longer-term history, thereby enabling the model to distinguish between isolated measurements and emerging trends or anomalies.

All the features and their descriptions are summarized in Table 2.2.

Table 2.2: List of features included in the study
*Upstream features are calculated over last 6h, 12h, 24h, 48h, 72h, and overlapping windows.

Type	1-hour features	Upstream features*
Raw features		
Demographics	GA, BW	
Raw data	HR_{mean}, HR_{median}	$HR_{median\_avg\_last\{t\}h}$, $HR_{median\_sd\_last\{t\}h}$
	SpO_{2mean}, $SpO_{2median}$	...
	SBP_{mean}, SBP_{median}	...
	DBP_{mean}, DBP_{median}	
	RR_{mean}, RR_{median}	...
Statistical features		
SD	HR_{std}	
	SpO_{2std}	
	SBP_{std}	
	RR_{std}	
Skewness	HR_{skew}	$HR_{skew\_avg\_last\{t\}h}$, $HR_{skew\_sd\_last\{t\}h}$
	SpO_{2skew}	...
	SBP_{skew}	...
	RR_{skew}	...
Kurtosis	HR_{kurt}	$HR_{kurt\_avg\_last\{t\}h}$, $HR_{kurt\_sd\_last\{t\}h}$
	SpO_{2kurt}	...
	SBP_{kurt}	...
	RR_{kurt}	...
Entropy	HR_{ApEn}, HR_{SampEn}, HR_{FuzzEn}	$HR_{ApEn\_avg\_last\{t\}h}$, $HR_{ApEn\_sd\_last\{t\}h}$, $HR_{SampEn\_avg\_last\{t\}h}$, $HR_{SampEn\_sd\_last\{t\}h}$, $HR_{FuzzEn\_avg\_last\{t\}h}$, $HR_{FuzzEn\_sd\_last\{t\}h}$
	SpO_{2ApEn}, $SpO_{2SampEn}$, $SpO_{2FuzzEn}$	...
	SBP_{ApEn}, SBP_{SampEn}, SBP_{FuzzEn}	...
	RR_{ApEn}, RR_{SampEn}, RR_{FuzzEn}	...
Domain knowledge features		
Mahalanobis	HR_{dist}	$HR_{dist\_avg\_last\{t\}h}$, $HR_{dist\_sd\_last\{t\}h}$
	SpO_{2dist}	...
	SBP_{dist}	...
	RR_{dist}	...
Spikes and drops	HR_{drops}, HR_{spikes}, $HR_{outliers}$	$HR_{drops\_avg\_last\{t\}h}$, $HR_{drops\_sd\_last\{t\}h}$, $HR_{spikes\_avg\_last\{t\}h}$, $HR_{spikes\_sd\_last\{t\}h}$, $HR_{outliers\_avg\_last\{t\}h}$, $HR_{outlier\_sd\_last\{t\}h}$
	SpO_{2drops}, $SpO_{2spikes}$, $SpO_{2outliers}$	...
	SBP_{drops}, SBP_{spikes}, $SBP_{outliers}$	...
	RR_{drops}, RR_{spikes}, $RR_{outliers}$	...

Trend	HR_{slope_4}, HR_{slope_6}, HR_{slope_12}, $\overline{HR}_{slope\_24}$, $\overline{HR}_{slope\_48}$ SpO_{2slope_4}, $\overline{SpO}_{2slope\_6}$, SpO_{2slope_12}, $\overline{SpO}_{2slope\_24}$, $\overline{SpO}_{2slope\_48}$ SBP_{slope_4}, SBP_{slope_6}, SBP_{slope_12}, $\overline{SBP}_{slope\_24}$, $\overline{SBP}_{slope\_48}$ RR_{slope_4}, RR_{slope_6}, RR_{slope_12}, $\overline{RR}_{slope\_24}$, $\overline{RR}_{slope\_48}$

2.1.4 Definition of the NEC dataset for secondary task evaluation

To further investigate the generalization capabilities of the tested models beyond the primary task of sepsis prediction, we extended our evaluation to an additional clinically relevant classification task: the prediction of NEC. The goal of this analysis was to assess whether the models could adapt to a distinct but related clinical condition, thereby providing insight into their applicability across multiple diagnostic tasks in neonatal care. The decision to focus on NEC was made in close collaboration with a team of neonatologists, prioritizing tasks with direct implications for clinical decision-making. While secondary tasks such as mortality prediction are commonly used in benchmarking studies, they often fall short in reflecting the real-world impact of predictive models. In contrast, NEC represents a critical neonatal emergency with significant morbidity and mortality, thus offering a more clinically meaningful evaluation scenario.

A definitive diagnosis of NEC requires radiographic imaging, which is typically performed in an emergency setting. According to clinical experts, radiographic confirmation is generally obtained within 30 minutes of clinical suspicion. Given this rapid diagnostic turnaround, we adopted the same data extraction pipeline used for sepsis prediction to construct the NEC dataset. To ensure diagnostic specificity, patients classified as Stage I NEC, which classify suspected cases, were excluded from the analysis. Furthermore, NEC may occasionally result from an intestinal bacterial infection, leading to bloodstream involvement and positive blood cultures. As noted by Giannoni et al. [20], this overlap can result in dual diagnoses of NEC and sepsis. To avoid data leakage and ensure that the NEC prediction task reflected distinct pathological signatures, all patients diagnosed with both conditions were excluded from the cohort. This conservative selection strategy was employed to eliminate potential bias arising from shared features in the sepsis training set.

Following these inclusion and exclusion criteria, the final NEC dataset comprised 127 patients.

2.2 Baseline model

2.2.1 Machine Learning Model

To assess the impact of complex models, such as FMs, on predictive performance, we first studied a simpler traditional ML model, in order to establish a reference point for average performance on this specific task.

A previous internal study conducted at CHUV compared several commonly used classifiers with different classification mechanisms, including extreme gradient boosting (XGBoost), k-Nearest Neighbors (KNN), Logistic Regression (LR), and Support Vector Machine (SVM). These classifiers have been extensively utilized in prior studies for LOS prediction and are generally accepted as reliable benchmarks [6–8]. XGBoost is an optimized implementation of gradient-boosted decision trees, designed for parallel computation and incorporating regularization techniques to improve generalization. SVM, on the other hand, is a kernel-based classifier capable of performing effective non-linear classification by mapping input features into a high-dimensional space, making it particularly well-suited for tasks with complex feature interactions. To enhance model performance, feature selection was conducted using a Lasso (Least Absolute Shrinkage and Selection Operator) regression. This process enabled the identification of the most informative features while reducing noise and dimensionality. Model evaluation was carried out using a 10-fold cross-validation strategy. The average performance across folds showed that XGBoost and SVM outperformed all other classifiers, achieving an accuracy of 0.77 ($\pm$0.04) and 0.78 ($\pm$0.06), respectively, compared to 0.74 ($\pm$0.06) for LR and 0.64 ($\pm$0.05) for KNN.

A detailed performance analysis, conducted in collaboration with a team of medical experts, was performed to select the most appropriate model based on clinically relevant criteria. A primary requirement was that the false positive (FP) rate should be lower than that observed in current clinical practice. According to clinicians, the integration of a predictive model into neonatal care should primarily aim to reduce the high number of false alarms generated by existing scoring systems, such as the HeRO score. Lowering the false positive rate is beneficial for several reasons. First, it would alleviate the burden on nursing staff by minimizing unnecessary alerts, thereby allowing them to focus on truly critical cases. Second, it would help reduce the overuse of antibiotic therapy, which is the primary treatment for sepsis. In fact, when administered to healthy neonates due to false alarms, unnecessary antibiotic exposure can contribute to antibiotic resistance and other health complications later in life. Beyond false positives, additional factors were considered to ensure the model's reliability and clinical utility. Model interpretability played a crucial role, as transparent decision-making fosters trust among healthcare professionals. Computational efficiency and ease of implementation were also evaluated to facilitate seamless integration into clinical workflows. Based on these criteria, SVM was identified as the most suitable baseline model.

For consistency, the same hyperparameters used in the previous study were adopted for this analysis. The model was implemented using the Scikit-learn library (version 1.6.1) [77] in

Python, with hyperparameters set to C=0.5 and a sigmoid kernel function.

2.3 Foundation model for time series analysis

2.3.1 FORMED

Feature extraction and selection are fundamental steps in training an ML model; however, these steps often represent the primary bottleneck in the model development process. It is estimated that more than half of the total time required to create a new model is typically spent on data preprocessing [78]. Handling missing values, detecting outliers, and filtering signals are just a few of the essential preprocessing steps that must be performed before the actual training phase. In the medical domain, this process becomes even more complex due to the inherent challenges associated with clinical data. Extensive consultations with medical professionals are required to determine which features hold biological relevance, further extending the time required for model deployment in clinical settings.

Recent advancements in ML models have enabled the direct processing of raw data, significantly accelerating the overall development pipeline [79]. New models, such as FMs, not only reduce the reliance on extensive feature engineering but also improve generalization across a broader range of tasks while often enhancing overall predictive performance. For this reason, the first model selected for evaluation is FORMED (Foundation mOdel Repurposed for Medical timE series Diagnosis), an FM for time series classification specifically designed to handle both inter- and intra-dataset heterogeneity in medical time-series data while requiring minimal task-specific adaptation [15]. FORMED introduces a three-stage framework that systematically transforms pre-trained, general-purpose time-series forecasting models into robust medical time-series classifiers. This approach effectively balances the need to preserve general temporal pattern recognition capabilities while simultaneously acquiring domain-specific knowledge for medical classification tasks.

Repurposing framework

Recent research in time series modeling has predominantly focused on time series forecasting, with models such as GPT4TS [58] and Chronos [56] advancing the field. Later on, many existing models, originally designed for predicting future trajectories of time series, have been adapted to solve secondary tasks, such as anomaly detection or time series segmentation, as exemplified by the MOMENT framework [57]. Therefore, the task of time series classification has remained comparatively underexplored.

FORMED introduces a paradigm shift in this area by establishing a general framework for repurposing forecasting models into robust classifiers for medical time-series data. This framework is structured into three key stages: pre-training, repurposing, and adapting, ensuring that models retain their temporal pattern recognition capabilities while being optimized for medical classification tasks.

Pre-training The pretraining stage is designed to develop general temporal representation capabilities by leveraging non-medical time series data. This phase aims to establish a temporal feature extractor f, also called backbone (or encoder) model, capable of recognizing fundamental patterns such as periodicity, trend components, and anomaly signatures, that are presumed to be transferable across domains, including medical applications. The underlying hypothesis assumes that all time series share a common semantic structure, irrespective of their domain or specific characteristics. It is postulated that time series data follows an inherent "language", governed by universal patterns that can be learned and transferred from one domain to another. During this stage, the model learns to project input $x \in \mathbb{R}^L$ from the last L steps of a univariate time series to its latent representation $u \in \mathbb{R}^D$, preserving temporal dependencies through attention mechanisms. Subsequently, a forecasting head g is used to predict the next N steps $\hat{x} \in \mathbb{R}^N$ in the forecasting horizon:

$$g \circ f : \mathbb{R}^L \to \mathbb{R}^N \tag{2.1}$$

This training phase is typically the most computationally and resource-intensive, requiring access to well-curated, diverse general time series datasets spanning multiple domains. This aspect is particularly critical in healthcare, where hospitals often lack the computational infrastructure, data availability, and specialized personnel required to pretrain such models from scratch. Given these constraints, pretraining is often performed using open-source foundation models that have already been extensively validated to leverage their pre-established feature extraction capabilities. In this sense, one of the distinguishing features of FORMED is its modular architecture, making it well-suited for clinical deployment. The framework is entirely independent of the encoder used for feature extraction, allowing users to select the most appropriate backbone model based on their specific domain requirements. This flexibility enhances the model's adaptability, ensuring its applicability across diverse medical time series classification tasks. In the case of FORMED, the chosen backbone model is TimesFM [59], which serves as the foundation for its representation learning.

Repurposing During the repurposing stage, the forecasting model is transformed into a medical time-series classifier through targeted modifications to its architecture. To adapt the backbone model for classification, the forecasting head g is removed and replaced with a classification head h. This structural modification allows the model to transition from predicting future time-series values to performing classification, while preserving the integrity of the feature extraction network f. The new model takes multivariate time series data $X \in \mathbb{R}^{C \times T}$ with C channels and T time steps, and returns the predicted label $\hat{y} \in \Delta^K$ where $\Delta^K = \left\{ d \in [0,1]^K : \sum_{i=1}^K d_i = 1 \right\}$ are the predicted class probabilities for K classes. Since TimesFM was designed for univariate time series, the channels of multivariate time series data are treated as independent and processed separately. Two extra parameters $E \in \mathbb{R}^{C \times D}$ and $Q \in \mathbb{R}^{K \times D}$ are included for indicating the task-specific channels and classes, respectively, with D as the dimension of the latent space. This is also a dynamic mapping as C, T and K may vary across datasets:

$$h \circ f : \mathbb{R}^{C \times T} \times \mathbb{R}^{C \times D} \times \mathbb{R}^{K \times D} \to \Delta^K \tag{2.2}$$

Once the model architecture has been adjusted, the framework mandates training the newly introduced classification head on a diverse set of medical time-series datasets $\mathcal{D}^{med}$. This step ensures that the model is adapted to the medical domain while retaining the general temporal representation capabilities learned during pretraining. Specifically, the backbone model is kept frozen, allowing only the task-specific layers to be fine-tuned. By preserving the pretrained backbone, the model maintains its ability to extract generic temporal features from time series data, while the newly trained components adapt the output representations to the specific requirements of the clinical classification tasks. Specifically, to specialize FORMED, five medical time-series datasets were selected to construct a specialized medical cohort [15]. These datasets encompass a wide range of physiological signals, particularly focusing on cardiac and neurological activity, such as ECG and EEG, which are among the most frequently analyzed modalities in clinical practice. Further details regarding the selected datasets and the applied preprocessing strategies are provided in Appendix C.

Adapting The final stage of the framework enables efficient customization for new medical datasets $\mathcal{D}^{new}$ through parameter-efficient tuning. During this phase, the model is adapted to new datasets or tasks, by constructing constructing dataset- and task-specific parameters such as $E' \in \mathbb{R}^{C' \times D}$ and $Q' \in \mathbb{R}^{K' \times D}$. This allows the model to specialize in previously unseen tasks that were not encountered during pretraining or repurposing. The limited number of trainable parameters ensures adaptability to complex medical classification problems, even in cases when training data is scarce, or availability is low, such as in the case of rare pathologies. Additionally, this characteristic is particularly advantageous in hospital settings, where computational resources are often constrained. By minimizing the number of parameters requiring adaptation, the model can be fine-tuned efficiently, significantly reducing both training time and resource consumption. However, due to its architecture, the model cannot be applied in a zero-shot learning setting, as certain parameters remain task-dependent. Instead, the framework aligns more closely with a few-shot learning paradigm, where limited labeled data is sufficient to effectively specialize the model for a new classification task.

Model architecture

The FORMED model is built upon a transformer-based architecture, which is inherently designed around the attention mechanism [49]. Widely employed for analyzing complex sequential data, such as natural language or time series, this architecture enables the flexible processing of input sequences while ensuring scalability across varying data modalities. The structural design is inspired by the latest advancements in Natural Language Processing (NLP) models, leveraging the adaptability and efficiency of transformer architectures to enhance time-series representation learning and classification.

Feature extractor FORMED leverages a pretrained foundation model backbone, initially trained on general time-series forecasting tasks, to enhance its adaptability for medical time-series classification. The selected model, TimesFM, is a decoder-only foundation model specifically designed for time-series forecasting [59]. By integrating transformer-based archi-

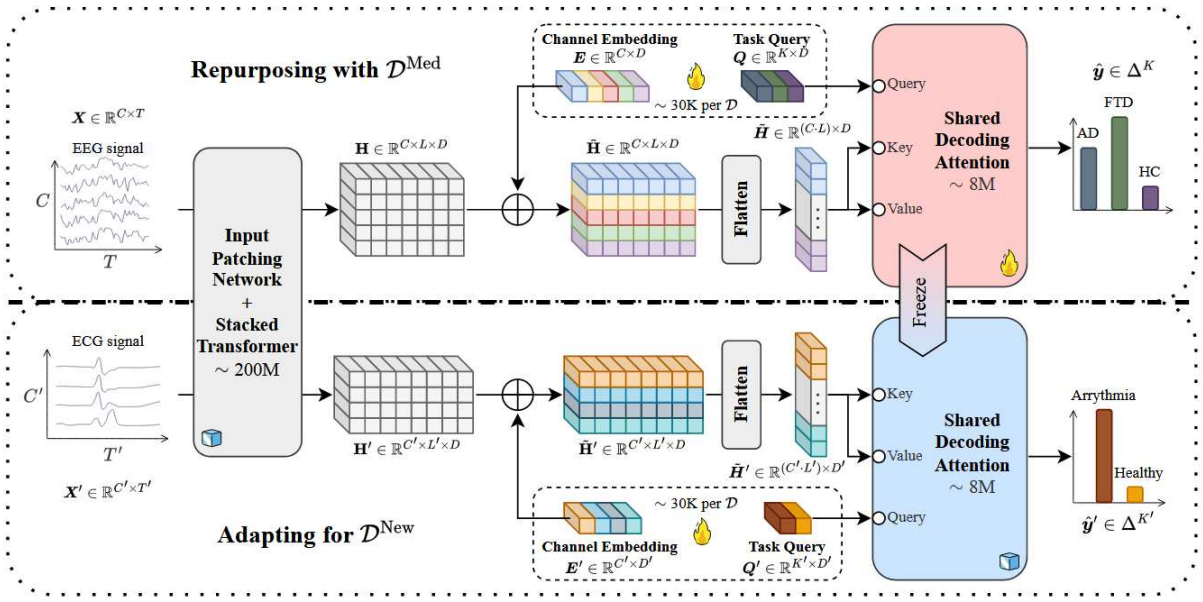


Figure 2.4: The architecture of FORMED in repurposing (top) and adapting mode (bottom). The backbone foundation model acts as a feature extractor and remains frozen all the time. This figure is adapted from the original FORMED paper [15]

tectures, TimesFM effectively captures the complex dynamics of time-series data, providing a robust and generalizable representation that facilitates its subsequent adaptation to medical applications. TimesFM is designed to address the challenges of time-series forecasting across multiple domains, including retail, finance, manufacturing, healthcare, and the natural sciences. The model is pretrained on an extensive corpus comprising 100 billion real-world time points, allowing it to learn diverse temporal patterns, including trends, seasonality, and varying time granularities, which are critical for effective forecasting. A key challenge in training foundation models for time-series forecasting lies in acquiring sufficiently large and diverse datasets to ensure broad generalizability. To overcome this limitation, the pretraining corpus for TimesFM was constructed using data sourced from three major repositories: Google Trends, Wikipedia Pageview statistics, and synthetically generated time-series data. This approach ensures that the model is exposed to a wide range of temporal structures, enabling it to learn representations that are both domain-agnostic and transferable across various forecasting tasks.

Input patching network To effectively process time-series data, TimesFM employs an input patching mechanism that segments continuous time-series input tokens that can be processed by the transformer layers. First, the input $x_{1:L}$ is broken into contiguous non-overlapping patches of length p. The j-th patch can be denoted as $\tilde{y}_j = y_{p(j-1)+1:pj}$. Then, each patch is processed by a residual block which map it into a vector of size D. The residual block is essentially a Multi-Layer Perceptron (MLP) block with a residual skip connection. In addition to the input, a binary padding mask $m_{1:L}$ is supplied to specify which timestamps

must be ignored. The j-th input token $t_j \in \mathbb{R}^D$ can be denoted as:

$$t_j = \text{InputResidualBlock}(\tilde{y}_j \odot (1 - \tilde{m}_j)) + \text{PE}_j \tag{2.3}$$

where PE_j denotes the j-th positional encoding vector.

This approach, inspired by methodologies used in NLP, facilitates the model's ability to learn complex temporal dependencies and patterns within the data.

Stacked transformer At the core of TimesFM lies a stacked transformer architecture, specifically a decoder-only model. Each of these layers contains one standard multi-head self-attention layer followed by a feed-forward network. This design choice allows the model to focus on autoregressive forecasting tasks, predicting future values based on past observations. The self-attention mechanism within the transformer enables the model to weigh the importance of different time steps, capturing both short-term fluctuations and long-term trends in the data. As is standard for time series forecasting with transformers, a causal self-attention mask is used to avoid leaking information from the future. Each output token can only attend to input tokens that come before it in the sequence. This can be described as follows:

$$u_j = \text{StackedTransformer}((t_1, m_1), ..., (t_j, m_j)) \tag{2.4}$$

for all $j = 1, ..., N$.

By stacking multiple transformer layers, TimesFM enhances its capacity to model intricate temporal relationships, leading to more accurate forecasting outcomes.

Ouput prediction network The final component of TimesFM is the output projection layer g, which transforms the internal representations learned by the transformer into forecasts $\hat{x}$. This layer transforms the latent representations, also called embeddings, back into the time-series domain, generating predictions that reflect the temporal characteristics of the original input signal. One key difference in TimesFM is that input patch length doesn't need to be equal to the output patch length. To do so, another residual block is used to map the output tokens to the original space: The design ensures that the model's outputs are interpretable and directly applicable to real-world forecasting scenarios.

$$\hat{x}_{pk+1:pj+h} = \text{OutputResidualBlock}(u_j) \tag{2.5}$$

Attention-based classifier The FORMED model introduces an attention-based classifier designed to handle the complexities of medical time series classification, particularly addressing issues of inter-dataset heterogeneity, variable-length time series, and diverse diagnostic targets. The classification mechanism is composed of three key components: channel embeddings E, task queries Q, and shared decoding attention, which together allow FORMED to overcome the rigid structural constraints of traditional classifiers, allowing for efficient adaptation to new medical datasets with minimal fine-tuning.

Channel embeddings The channel embedding mechanism plays a crucial role in distinguishing between different time series channels. Given that medical time series often comprise multivariate signals, embedding the channel information ensures that the model can effectively differentiate between physiological signals originating from different sources. The channel embeddings are learnable parameters $E \in \mathbb{R}^{C \times D}$ stored in a lookup table, mapping dataset names to embedding vectors. These vectors are then added to the feature tokens $H \in \mathbb{R}^{C \times L \times D}$ extracted from the pre-trained backbone, enabling the model to retain information about the channel structure while allowing for flexibility in processing datasets with varying numbers of channels:

$$\tilde{H}_{:,i,:} = H_{:,i,:} \oplus E$$

Task query The label query mechanism introduces a task-specific component to the classifier. Each medical classification task requires a different set of labels, which may not be predefined during pretraining. FORMED addresses this by learning label-specific embeddings $Q \in \mathbb{R}^{K \times D}$, known as label queries, which guide the classification process. These task-specific embeddings help the model focus on the relevant features associated with each diagnostic category, ensuring that the classification head can adapt dynamically to different tasks. By employing learnable label queries, FORMED maintains its generalizability, effectively handling new classification tasks with only minimal retraining.

Shared decoding attention The shared decoding attention mechanism forms the core of the FORMED classification process. It utilizes a transformer decoder layer, similar to those employed in object detection and image classification, but adapted specifically for medical time series. This layer enables each label query to attend to the feature representations of the time series data $\tilde{H}$ through multi-head attention. By leveraging this shared attention mechanism, the model establishes a global context for classification, capturing both temporal dependencies and cross-channel relationships. The output of this mechanism is further processed through a residual block, producing classification logits that can be transformed into probability distributions using softmax or sigmoid functions, depending on the classification task:

$$\hat{y} = \text{ResidualBlock}(\text{MultiHeadAttention}(Q = Q, K = \tilde{H}, V = \tilde{H})) \tag{2.6}$$

Customized FORMED architectures for neonatal sepsis prediction

Sepsis is a complex and multifactorial condition that affects multiple physiological systems simultaneously, resulting in coordinated changes across several vital signs. For instance, sepsis can cause simultaneous alterations in heart rate, blood pressure, and respiratory rate [6–8]. This multivariate nature necessitates the joint analysis of different physiological channels to capture the intricate interdependencies that signal the onset or progression of the disease. Analyzing these variables collectively, rather than independently, enhances the model's ability to detect subtle patterns indicative of the onset or progression of such conditions. However, many existing FMs for time series forecasting are designed to process

different channels independently, potentially overlooking the interdependencies among variables. To address this limitation, recent advancements have focused on adapting univariate time series FMs for multivariate data. For example, Liu et al. proposes a method to enable FMs to effectively handle multivariate time series data by incorporating generalized prompt tuning techniques [80].

Considering this specific limitation inherent to FMs for time series forecasting, including TimesFM, the model architecture was modified accordingly. Due to constraints in computational resources and the unavailability of the original pretraining data, we chose not to alter the underlying structure of TimesFM directly. Instead, modifications were introduced within the FORMED model prior to the classification head. In order to effectively integrate information from the various channels, which are processed independently by TimesFM, an additional self-attention layer was incorporated, following a strategy similar to that described in [80]. In this configuration, the transformer module is applied to the embedding matrix $H \in \mathbb{R}^{C\times L\times D}$ by treating the L different patches as distinct data points in $\mathbb{R}^{C\times D}$, while interpreting the channel dimension C analogously to the sequence of tokens in an ordinary transformed layer. The transformer then outputs a patch in $\mathbb{R}^{C\times D}$, and these outputs are subsequently concatenated into a single array $H' \in \mathbb{R}^{C\times L\times D}$, thereby enabling the combined representation of multivariate information for improved classification performance.

Beyond physiological time series, incorporating demographic information is vital for personalized and accurate medical assessments. Gestational age, in particular, plays a significant role in neonatal care, influencing the interpretation of vital signs and the assessment of developmental progress (see Appendix B). Integrating such demographic factors into predictive models aligns with clinical practices, where patient-specific characteristics are considered alongside physiological measurements to inform diagnosis and treatment decisions. This multimodal approach ensures that models are sensitive to individual variability, leading to more accurate and tailored healthcare interventions. To achieve this, gestational age information was encoded through a dedicated gestational age embedding layer. Following the same approach used for channel embeddings, these embeddings $G \in \mathbb{R}^{V\times D}$, where V represents the number of different values for gestational age, were added to the feature token $H \in \mathbb{R}^{C\times L\times D}$ to generate the augmented feature token $\tilde{H}$. Given the gestational age embedding vector $g = G_{i,:}$, where i represents the gestational age value for patient j, we can compute $\tilde{H}$ as follows:

$$\tilde{H} = H + g \tag{2.7}$$

for all patients $j = 1, ..., N$.

As previously introduced, two distinct strategies were explored to enhance the capacity of the model to process multivariate physiological signals in the context of neonatal sepsis prediction. The first involved the integration of a self-attention layer to capture cross-channel dependencies among the different physiological signals. The second introduced gestational age as an additional conditioning variable via a dedicated embedding layer, thereby enabling the model to contextualize signal patterns with respect to developmental maturity. To further enhance the model's ability to integrate multimodal information, both approaches

were combined by incorporating gestational age directly into the self-attention mechanism in the form of a prompt. Specifically, the gestational age embedding vector g was introduced as an additional token within the self-attention layer, modifying the way multivariate time-series signals are processed.

Given an input sequence of L patches, each with C channels, the feature tokens before self-attention is represented as $H \in \mathbb{R}^{C \times L \times D}$, where D denotes the latent space dimension. The gestational age prompt was incorporated by expanding this representation to include an additional prompt token, yielding an augmented input:

$$\tilde{H} = [H; g] \in \mathbb{R}^{(P+1) \times C \times D} \tag{2.8}$$

where the concatenation operation ensures that the gestational age information is treated as an integral part of the attention computation. The self-attention mechanism is then applied as follows:

$$H' = \text{MultiHeadAttention}(Q = \tilde{H}, K = \tilde{H}, V = \tilde{H}) \tag{2.9}$$

Then, the last token embedding of H', which corresponds to the token used to prompt the model, is removed to obtain the output $H \in \mathbb{R}^{C \times L \times D}$. By including the gestational age embedding in this formulation, the model modulates how individual signals are processed, allowing it to adjust feature extraction based on gestational age.

2.3.2 TabPFN

The application of FMs to tabular data presents unique challenges that have historically limited their success in this domain. Unlike text or image data, tabular datasets exhibit a high degree of heterogeneity, not only in feature types and distributions but also in dataset size, structure, and semantics. This variability significantly complicates the development of generalizable models capable of transferring knowledge across tasks and domains. Traditional ML methods, while effective in narrow, well-defined scenarios, often suffer from poor out-of-distribution generalization and limited transferability. Moreover, conditioning FMs on the specific context of a tabular prediction task is inherently difficult due to the vast number of possible data configurations and problem formulations.

TabPFN (Tabular Prior-data Fitted Network) represents a significant advancement in this context, introducing a FM specifically designed to handle the structural and statistical complexity of tabular data [16]. By learning from millions of synthetically generated datasets that simulate real-world variability, TabPFN achieves improved generalization and predictive performance, particularly in medium-sized datasets with up to 10,000 samples. This approach marks a paradigm shift in tabular data modeling by enabling fast, accurate predictions without task-specific retraining, thereby addressing key limitations of prior methods in real-world applications such as clinical decision support.

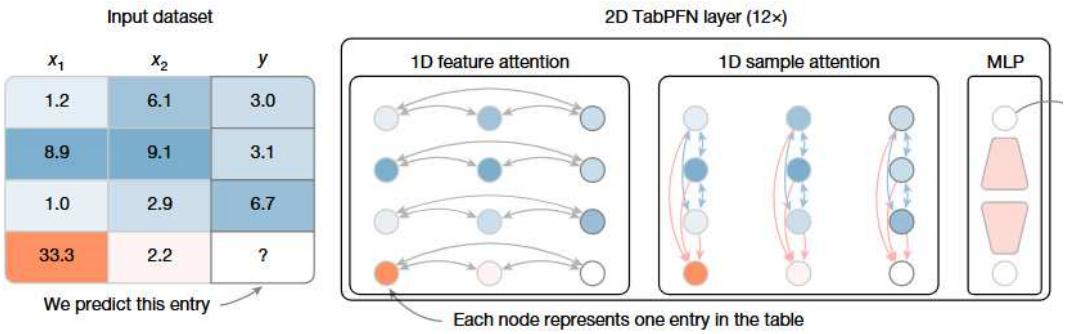


Figure 2.5: The TabPFN architecture. This figure illustrates the two-way attention mechanism, highlighting feature-wise (left) and sample-wise attention (right). This figure is adapted from the original TabPFN paper [16]

Synthetic data based on causal models for model pretraining

A critical challenge in developing robust models for tabular data is the scarcity of large, diverse datasets necessary for comprehensive training. TabPFN addresses this by generating synthetic datasets through structural causal models, which simulate complex data-generating processes. This methodology allows the model to learn a wide array of functional relationships and patterns, thereby enhancing its generalization capabilities across various real-world scenarios and circumventing privacy concerns associated with real-world data collection. Training on millions of these synthetic datasets ensures that the synthetic prior captures essential aspects of real tabular data distributions, including non-linear relationships, feature correlations, and diverse noise patterns.

This pretraining approach also confers practical advantages during downstream application. By learning from diverse data distributions, the model becomes inherently capable of handling common data irregularities, such as missing values and varying feature scales. TabPFN streamlines the data preprocessing phase by automatically managing missing values, encoding categorical variables, and normalizing features when provided with raw tabular data. This automation contrasts with traditional models that typically require manual intervention for such preprocessing tasks prior to training. By eliminating the need for labor-intensive data preparation, TabPFN not only reduces the potential for human error but also accelerates the overall model development process. This capability is particularly advantageous in clinical settings, where datasets often contain incomplete information and time is of the essence.

Model architecture

TabPFN leverages a transformer-based neural network architecture specifically designed for tabular data. Unlike conventional transformer models, which are optimized for sequential data, TabPFN is tailored to the two-dimensional structure of tabular datasets by employing a novel two-way attention mechanism. This mechanism enables each cell within a table to

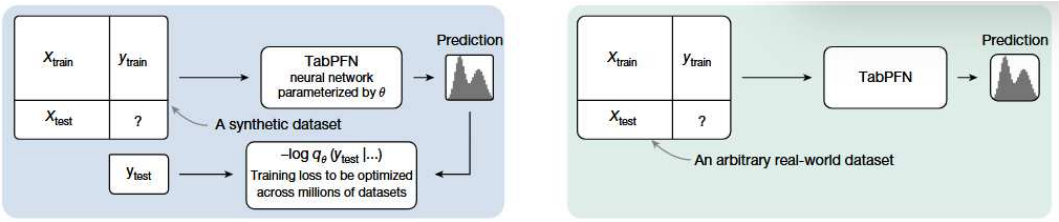


Figure 2.6: An high-level overview of TabPFN pre-training and usage pipelines. This figure is adapted from the original TabPFN paper [16]

attend to other features within the same row (sample-wise attention) as well as to the same feature across different rows (feature-wise attention). The core of the TabPFN model consists of 12 transformer layers, each optimized for tabular data processing. The architecture is designed to efficiently handle small-sample learning scenarios by incorporating several key modifications aimed at improving memory efficiency and computational scalability. Key innovations include half-precision layer normalizations, flash attention mechanisms for reduced memory footprint, and activation checkpointing to enable training on large synthetic datasets.

In-context learning

A central innovation of TabPFN is its integration of In-Context Learning (ICL), a mechanism originally developed for large language models, which enables the simultaneous execution of task learning and prediction within a single forward pass. This design eliminates the need for distinct training and testing phases and provides a scalable alternative to traditional retraining paradigms. ICL addresses a fundamental limitation in the application of FMs to tabular data, namely, the challenge of task contextualization across heterogeneous datasets. Unlike conventional models that require parameter optimization specific to each task, TabPFN conditions its predictions on a given support set, effectively treating it as a prompt that encapsulates the statistical and semantic properties of the task at hand.

Through the use of attention mechanisms, TabPFN dynamically conditions its predictions on the provided training set, effectively treating it as a prompt, thereby allowing the model to generalize to previously unseen tasks without modifying its internal parameters. This capability is particularly advantageous in clinical contexts, where datasets are often small, heterogeneous, and task-specific, and where rapid deployment without retraining is essential. By circumventing the need for explicit task specialization, ICL serves as the cornerstone of TabPFN's adaptability, enabling robust performance in data-scarce environments, such as healthcare, and facilitating broader applicability of FMs to the tabular data domain.

2.3.3 Experimental setup

Both FMs were evaluated LOS sepsis prediction task, and their performance was compared against that of a baseline SVM model. While the FORMED model is described as open

source, direct access to its pre-trained weights was not available at the time of this study. Consequently, a reverse engineering strategy was adopted to approximate the original model configuration and enable comparative analysis. This approach allowed for the evaluation of FORMED's architecture and adaptation capabilities within the constraints of reproducibility and accessibility, providing insight into its practical deployment potential in clinical settings.

For TimesFM, FORMED's backbone foundation model, the pre-trained weights were directly obtained from the publicly available version released by Google on the Hugging Face platform. To ensure the validity of the downloaded weights, a series of sanity checks were conducted by replicating the results reported in the original paper [59]. Specifically, the model was tested on artificially generated signals to assess its ability to predict time-series trends. Appendix D, and specifically Figure D.1, illustrates that TimesFM accurately forecasts periodic signals, particularly in low-noise conditions. Regarding the classification head, direct access to the pre-trained weights was not available, necessitating a complete retraining of the head module. The newly trained model's performance was then compared with the results reported in the original paper [15]. This process was computationally intensive, as it required downloading all datasets $\mathcal{D}^{med}$ used in the pretraining phase and re-training the classification head from scratch. The computational burden associated with this process presents a major bottleneck, particularly in resource-constrained environments such as hospitals, where model usability is a fundamental prerequisite. While it was feasible to re-train the classification head due to its relatively small number of parameters, re-training the entire model, including TimesFM, would have been impractical given the limited numbers of GPUs available for this project.

In contrast, the implementation and use of TabPFN was considerably more straightforward. This is largely attributed to the fact that TabPFN offers a Scikit-Learn API, facilitating seamless adoption within existing machine learning pipelines. This compatibility enhances its usability, allowing for easy integration into pre-existing codebases and enabling rapid experimentation. Additionally, the pre-trained weights for TabPFN were readily available and downloaded directly from Hugging Face, eliminating the need for extensive re-training efforts.

Model evaluation and hyperparameter tuning

All models were evaluated on the CHUV dataset using a 5-fold cross-validation (CV) strategy. To ensure a fair comparison across models, the same folds were maintained consistently throughout all experiments. To prevent data leakage during model evaluation, normalization was applied within the CV loop. Specifically, for each fold, the mean and standard deviation were computed exclusively on the training set. These statistics were then used to normalize both the training and corresponding test sets. This approach ensures that no information from the test data influenced the training process, thereby preserving the integrity of the model evaluation.

Both SVM and TabPFN were trained using the feature-engineered dataset, whereas FORMED

was the only model trained directly on the raw time series data. This distinction underscores a fundamental difference in data processing approaches, potentially impacting model interpretability and performance. Moreover, the SVM model was trained using the same set of hyperparameters identified in a previous study conducted at CHUV on the same patient cohort. These hyperparameters were selected based on a 5-fold CV procedure; however, it is important to note that the data were split differently from the current experimental setup. For TabPFN, no retraining was required due to the model's inherent structure, and consequently, no hyperparameter tuning was necessary. Instead of following the conventional machine learning training paradigm, TabPFN was conditioned on the entire training set, which was provided as a prompt to the model. In contrast, FORMED required hyperparameter optimization, which was performed using a grid search approach. For hyperparameter tuning we employed a nested cross-validation strategy was adopted, consisting of 5 outer splits and 3 inner splits. In each outer loop iteration, an additional inner loop was executed to optimize the model's hyperparameters. The combination of hyperparameters that achieved the highest average performance across the inner folds was selected, and the corresponding model, the one achieving the highest AUROC, was then evaluated on the test set of the respective outer fold. Importantly, no refitting was performed on the entire training set, ensuring that the evaluation remained unbiased and that no data leakage occurred during testing. This approach allowed for systematic hyperparameter tuning without artificially inflating model performance.

All experiments were conducted using an NVIDIA Quadro P6000 GPU. To ensure reproducibility and efficient tracking of experimental results, MLflow (version 2.21.1) was employed to log and monitor all conducted experiments. MLflow's interface enables systematic tracking of model evaluations while providing real-time insights into performance metrics during training. Additionally, the platform facilitated a comparative analysis of model performance across different experiments through graphical visualizations and summary tables. All models were implemented using Scikit-learn (version 1.6.1) and PyTorch (version 2.6) using the Python programming language. For hyperparameter optimization, RayTune (version 2.44.1) was utilized, allowing for an efficient and scalable grid search process. RayTune, a Python-based library compatible with PyTorch, enables large-scale experiment execution and hyperparameter tuning. By leveraging parallelization, multiple search iterations were executed simultaneously on the same GPU, significantly accelerating the hyperparameter optimization process.

Chapter 3

Results

3.1 Neonatal sepsis prediction traditional machine learning vs. foundation models

Table 3.1 presents the performance of compared classifiers on the sepsis prediction task. Notably, SVM and TabPFN demonstrate comparable results, with TabPFN achieving a slightly higher AUROC (83.34 ± 2.36) and precision (75.10 ± 2.58) relative to SVM's AUROC (80.65 ± 2.26) and precision (73.81 ± 3.52). This similarity is particularly striking given that TabPFN is only conditioned on sepsis data. In fact, TabPFN does not undergo additional training for this specific task, highlighting the potential of pretrained foundation models to adapt effectively to new datasets. In contrast, FORMED consistently underperforms the other two models, exhibiting a lower AUROC (72.49 ± 6.77) and precision (69.15 ± 3.99). One possible explanation for this discrepancy is that FORMED processes raw time-series inputs, whereas SVM and TabPFN both rely on feature-engineered data. The direct use of unfiltered signals may introduce noise or artifacts that degrade classification performance.

3.1.1 Further investigation of the FORMED model and its predictions

Assessing FORMED's repurposing srategy for medical time series

To evaluate the applicability of FMs to clinical time series classification, we investigated FORMED, a framework that repurposes a pretrained forecasting model for classification tasks. Unlike traditional models that rely on manually engineered features, FORMED extracts representations directly from raw signals. This design is grounded in the assumption that FORMED's backbone model, TimesFM, can generate clinically meaningful embeddings suitable for classification, despite being originally developed for general-purpose time series

Table 3.1: LOS sepsis prediction results using different classifiers. Numbers are reported as percentage mean (standard deviation) computed across the outer folds of the nested cross-validation. Metrics include accuracy, precision, recall, F1 score, AUROC, and AUPRC, enabling a comprehensive comparison of classifiers performance.

Model	**Accuracy**	**Precision**	**Recall**	**F1 score**	**AUROC**	**AUPRC**
SVM	76.02 (3.41)	73.81 (3.52)	73.58 (4.04)	73.52 (3.83)	80.65 (2.26)	71.79 (6.85)
FORMED	70.49 (1.99)	69.15 (3.99)	63.89 (2.60)	63.94 (3.17)	72.49 (6.77)	61.15 (7.05)
TabPFN	**77.31** (2.39)	**75.10** (2.58)	**75.12** (3.01)	**75.19** (2.73)	**83.34** (2.36)	**75.61** (5.35)

forecasting. Given this assumption, we first assessed TimesFM's forecasting performance by evaluating its forecasting performance on physiological signals from our cohort.

During the repurposing phase, FORMED was fine-tuned for medical classification tasks using five different datasets $\mathcal{D}^{med}$ encompassing typical physiological signals such as ECG and EEG. The model demonstrated strong performance on these classification tasks, suggesting that the extracted feature representations were indeed informative. To further investigate this, we analyzed TimesFM's ability to predict the temporal evolution of these physiological signals. Figure D.2, Appendix D presents the forecasting results for ECG waveforms from the PTB-XL dataset, which was selected due to its superior classification performance among the evaluated datasets. Overall, TimesFM successfully captures the general trend of ECG signals. However, it fails to accurately predict high-amplitude variations, such as the characteristic peaks of ECG waveforms. This limitation does not necessarily indicate model failure; rather, in cases where sharp fluctuations are not well captured, TimesFM converges towards the mean value of the signal, which remains a reasonable approximation.

An additional factor influencing TimesFM's performance is the periodicity of the input signals. As previously observed when testing on synthetic time-series data (Figure D.1, Appendix D), TimesFM performs optimally with quasi-periodic signals. To further validate this, we tested the model on the physiological signals from neonates in this study. The results are shown in Figure 3.1 (see Figure D.3, Appendix D for additional examples). Similar to previous experiments, TimesFM accurately predicts the trend of most signals. However, a notable difference is that these physiological signals are not inherently periodic. Additionally, the sampling frequency of the signal plays a critical role. While TimesFM effectively captures the general trend of HR and SpO_2, it struggles with low-frequency signals, such as SBP, where it again tends to converge towards the mean.

Training FORMED with different modes for sepsis prediction

FORMED was originally designed with the objective of minimizing the number of trainable parameters (see Section 2). This strategy was initially applied to the task of sepsis prediction. In a first experiment evaluating FORMED's generalization capabilities, the model was employed in the *adapting* phase to adapt it for the sepsis classification task. However,

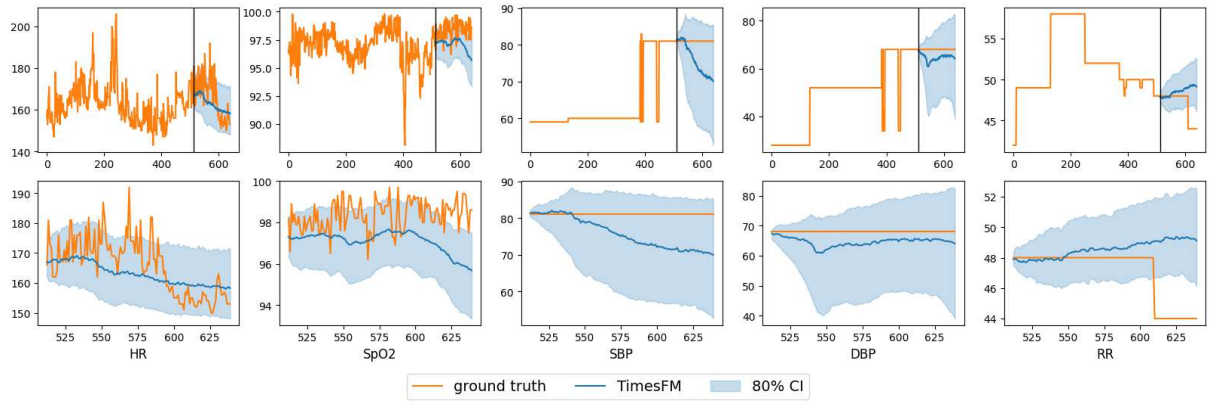


Figure 3.1: Forecasted vital signals for a randomly selected patient. The bottom row plots zoom in on the forecasting horizon for better visibility.

as shown in Table 3.2, FORMED failed to achieve satisfactory performance, reaching an accuracy of only 67.13 (± 6.15).

To further investigate FORMED's performance, a series of additional experiments were conducted, testing different training strategies. First, the repurposing phase was re-executed while incorporating the sepsis dataset to better align the learned feature representations (*repurposing*). Next, a fine-tuning approach was applied, where the model was initialized using the weight configuration obtained after the original repurposing phase and subsequently trained on the sepsis dataset (*fine-tuning*). Unlike the adapting phase, in which only a small subset of parameters was optimized, fine-tuning involved training the entire model while leveraging the knowledge acquired during repurposing. Lastly, the model was trained from scratch, using only the sepsis dataset (*training*). The results of these experiments are summarized in Table 3.2. Notably, training the model from scratch led to the highest performance, achieving an AUROC of 72.49 (± 6.77) and a precision of 69.15 (± 3.99). These findings confirm the hypothesis that FORMED does not benefit from prior knowledge acquired during the classification of ECG and ECG signals. Moreover, incorporating the sepsis dataset into the repurposing phase did not yield any performance improvements.

Architectural limitation and modifications of FORMED for sepsis prediction

As shown in Table 3.1, FORMED consistently underperforms compared to SVM and TabPFN. Several factors may contribute to this reduced predictive capability. One key limitation lies in the fixed-length context window. TimesFM, and consequently FORMED, can only process time series with a maximum length of 512 timestamps. In the context of sepsis prediction, this corresponds to approximately 9 hours of minute-by-minute time series data prior to diagnosis. In contrast, SVM and TabPFN were trained on an engineered dataset containing features extracted from windows extending up to 72 hours before diagnosis. This significant difference in temporal context suggests that FORMED may not have access to sufficient predictive information, potentially limiting its performance.

Table 3.2: LOS sepsis prediction results using different versions of FORMED. Numbers are reported as percentage mean (standard deviation) computed across the outer folds of the nested cross-validation. Metrics include accuracy, precision, recall, F1 score, AUROC, and AUPRC, enabling a comprehensive comparison of classifiers performance.

Experiment	**Accuracy**	**Precision**	**Recall**	**F1 score**	**AUROC**	**AUPRC**
adapting	67.13 (6.15)	54.91 (6.73)	54.13 (5.33)	55.20 (6.83)	54.38 (5.72)	45.17 (7.98)
repurposing	56.52	55.87	56.46	55.17	61.25	44.35
fine-tuning	56.69 (9.87)	53.10 (8.45)	53.11 (8.44)	52.25 (9.26)	57.03 (10.12)	44.85 (8.54)
training	**70.49** (1.99)	**69.15** (3.99)	**63.89** (2.60)	**63.94** (3.17)	**72.49** (6.77)	**61.15** (7.05)

Table 3.3: LOS sepsis prediction results with SVM using different time windows. Numbers are reported as percentage mean (standard deviation) computed across the outer folds of the nested cross-validation. Metrics include accuracy, precision, recall, F1 score, AUROC, and AUPRC, enabling a comprehensive comparison of classifiers performance.

Experiment	**Accuracy**	**Precision**	**Recall**	**F1 score**	**AUROC**	**AUPRC**
72h	76.02 (3.41)	73.81 (3.52)	73.58 (4.04)	73.52 (3.83)	80.65 (2.26)	71.79 (6.85)
12h	75.12 (1.70)	73.14 (1.85)	73.04 (2.13)	72.76 (1.76)	81.70 (1.35)	75.99 (4.26)

To investigate this hypothesis, we restricted the temporal window used for feature extraction in the engineered dataset to 12 hours and subsequently compared the performance of the SVM model trained on this reduced dataset with its original 72-hour configuration. Results are presented in Table 3.3. As expected, a slight decrease in performance was observed; however, the model trained on a limited window still achieved comparable results, with an accuracy of 75.12 ($\pm$ 1.70) versus 76.02 ($\pm$ 3.41) in the original setting. This finding aligns with previous internal studies, which identified features extracted 6 to 12 hours before sepsis diagnosis as among the most predictive, whereas those obtained 24, 48, or 72 hours prior contributed less to classification performance.

Another important consideration in FORMED's design is its assumption of channel independence. Since TimesFM processes only univariate time series, this assumption translates into applying the feature extraction network independently to each signal. While previous studies have supported this assumption in various applications [80], its implications in biomedical time-series classification may limit overall classification performance, as the model might fail to capture clinically relevant interdependencies across input signals. To address this, we explored several modifications to FORMED's architecture. First, an additional self-attention layer was introduced before the final classification step to allow for information exchange between channels (*self-attenton*). Second, instead of modifying the architecture, we incorporated an embedding layer to encode gestational age, a feature explicitly used by SVM and TabPFN but not previously integrated into FORMED (*gestational age*). Finally, both

Table 3.4: LOS sepsis prediction results with FORMED using different experimental settings. Numbers are reported as percentage mean (standard deviation) computed across the outer folds of the nested cross-validation. Metrics include accuracy, precision, recall, F1 score, AUROC, and AUPRC, enabling a comprehensive comparison of classifiers performance.

Experiment	Accuracy	Precision	Recall	F1 score	AUROC	AUPRC
Original	70.49 (1.99)	**69.15** (3.99)	63.89 (2.60)	63.94 (3.17)	**72.49** (6.77)	**61.15** (7.05)
Gestational age	69.43 (3.18)	66.47 (4.99)	63.22 (4.12)	64.34 (4.51)	71.45 (5.43)	57.21 (6.84)
Self attention	**70.92** (2.81)	68.48 (3.22)	**66.19** (3.48)	**66.27** (3.82)	72.41 (5.63)	59.83 (8.24)
Prompt	69.64 (3.33)	66.27 (4.47)	64.08 (3.84)	64.50 (4.11)	72.18 (5.46)	59.00 (6.27)

modifications were combined to maximize the extracted information from the input signals (*prompt*). All experiments were conducted using the sepsis dataset, following the same experimental setup described earlier. Given that training the model from scratch yielded the best performance in previous evaluations (see Table 3.2), the same training strategy was adopted for this analysis.

The results, summarized in Table 3.4, indicate that these modifications did not significantly improve FORMED's predictive performance.

3.1.2 Robustness analysis of the foundation model approach with TabPFN

The results presented in Table 3.1 indicate that both SVM and TabPFN achieve comparable performance in predicting neonatal sepsis, aligning with findings reported in previous studies [6–8]. However, despite similar overall performance, the two models employ fundamentally different approaches. Determining which model is better suited for potential clinical implementation remains a complex task that requires further investigation. The standard performance metrics used in this study, while informative, do not provide a comprehensive assessment of a model's suitability for real-world deployment in hospitals. The adoption of complex models such as TabPFN should be justified by distinct advantages in performance that cannot be attained using classical machine learning models such as SVM. Moreover, additional factors, such as the reliability of predictions, model robustness to distributional and temporal shifts, and adaptability to various clinical settings, must be taken into account when considering model deployment. To address these concerns, an in-depth analysis of both models was conducted to highlight the strengths and weaknesses of each approach.

One of the most time-consuming processes in machine learning model development is feature extraction. This step not only requires substantial resources, both in terms of time and specialized personnel, but is also inherently influenced by the specific model being tested. As a result, there is a risk of designing a dataset that is tailored to a particular model, limiting its generalizability to other tasks. In hospital settings, where resources are often constrained due to public funding, optimizing workflow efficiency is crucial to maintaining a high standard

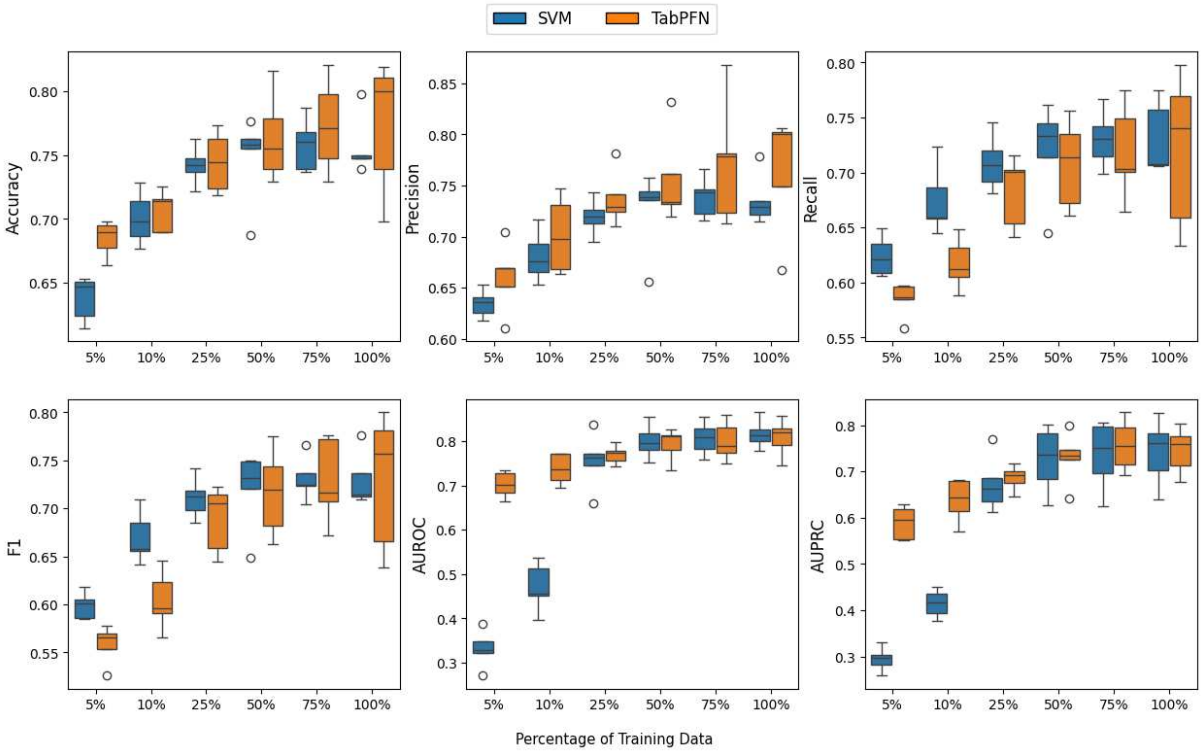


Figure 3.2: Comparison of the performance of SVM (blue) and TabPFN (orange) as a function of training data size. Boxplots represent the distribution of performance metrics across outer splits of a cross-validation procedure, with the median and interquartile range shown to illustrate both central tendency and variability.

of care. Furthermore, in the medical field, feature extraction should ideally be supervised by clinical experts to ensure the inclusion of clinically meaningful variables. However, this requirement imposes an additional burden on physicians. A major advantage of modern FMs is their ability to automatically extract features, thereby reducing the reliance on manual feature engineering. In the case of TabPFN, however, feature extraction is not performed internally, as the model still relies on pre-processed features. Therefore, its adoption must be justified by other benefits. To assess the robustness of TabPFN, two additional experiments were conducted.

The first experiment evaluated model performance under varying amounts of training data. As before, a 5-fold cross-validation approach was employed, where the proportion of training data was systematically adjusted for each fold. To mitigate potential sample bias, this sampling procedure was repeated n times to cover the whole original training set. The final performance metrics of all n repeats were averaged within each training data fraction. The results, presented in Figure 3.2, demonstrate that TabPFN consistently outperforms SVM across all metrics, particularly when trained on less than 25% of the available data. Moreover, TabPFN exhibits higher precision, which is a key requirement for clinical adoption, as it helps

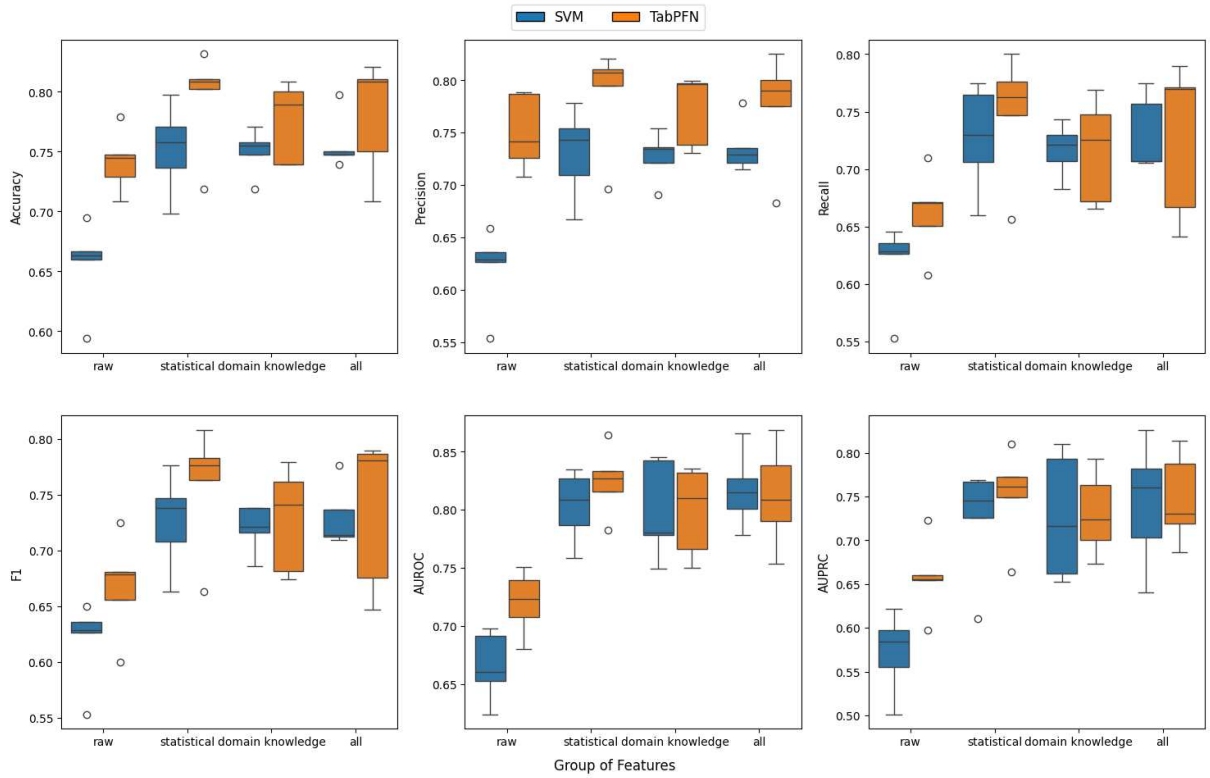


Figure 3.3: Comparison of the performance of SVM (blue) and TabPFN (orange) as a function of feature group. Boxplots represent the distribution of performance metrics across outer splits of a cross-validation procedure, with the median and interquartile range shown to illustrate both central tendency and variability.

reduce the occurrence of false alarms, a major concern for healthcare professionals. While no significant differences are observed when the full training dataset is utilized, these findings suggest that TabPFN possesses a superior ability to generalize from limited data compared to SVM, which is particularly valuable in clinical settings involving small datasets.

In the second experiment, we aimed to evaluate the performance of SVM and TabPFN under varying levels of feature complexity. To achieve this, the extracted features were categorized into three groups, arranged in increasing order of complexity and computational effort required for their definition, as detailed in Table 2.2 (see Section 2).

Figure 3.3 presents a comparative analysis of the two models across six evaluation metrics. The results indicate that TabPFN consistently outperforms SVM across all metrics and feature sets. Notably, TabPFN demonstrates strong predictive performance even when trained exclusively on raw features, surpassing SVM in both median performance and variability. This finding suggests that TabPFN does not require extensive feature engineering to achieve reliable predictions, potentially simplifying the entire data preprocessing pipeline. In contrast, SVM exhibits a stronger dependence on domain-specific features, requiring additional

Table 3.5: NEC prediction results using different classifiers trained on sepsis prediction task (as described in Section 2) in a zero-shot setting. Metrics include accuracy, precision, recall, F1 score, AUROC, and AUPRC, enabling a comprehensive comparison of classifiers performance.

Model	Accuracy	Precision	Recall	F1 score	AUROC	AUPRC
SVM	78.05	75.64	76.46	76.00	84.94	71.22
FORMED	78.66	76.40	**77.78**	76.92	81.21	67.30
TabPFN	**82.32**	**82.52**	77.12	**78.77**	**91.63**	**83.59**

preprocessing steps and extensive collaboration with clinical experts.

3.1.3 Assessing models generalization on an unseen clinical task

FMs have demonstrated remarkable adaptability across multiple tasks by leveraging large-scale pretraining and flexible architectures. Assessing their generalization capabilities is crucial to understand their potential clinical impact, particularly in neonatology, where predictive modeling could enhance decision-making and optimize patient care. To rigorously assess the adaptability of the evaluated models beyond the primary objective of sepsis prediction, an additional classification task involving the prediction of NEC was introduced (see Section 2).

Table 3.5 presents the results of NEC prediction in a zero-shot setting, where all three models were initially trained on sepsis prediction task (as described in Section 2) before being directly applied to the entire NEC dataset. In the case of TabPFN, the model was conditioned on sepsis data, meaning that the sepsis dataset was provided as a prompt. Despite not being explicitly trained on sepsis data, TabPFN outperforms both SVM and FORMED, achieving the highest accuracy, AUROC, and AUPRC. These results highlight the model's ability to generalize to a novel classification task with minimal or no additional training. In contrast, SVM and FORMED exhibit comparable performance, with SVM achieving an accuracy of 78.05 and FORMED reaching 78.66. However, an important observation is that NEC classification appears to be inherently less challenging than sepsis prediction. According to clinical experts, NEC is typically characterized by more pronounced clinical manifestations, making it easier to identify. This observation is further supported by Figure B.6, Appendix B, which compares the average signal trajectories of patients across the two tasks. In particular, the average signals for NEC patients show a greater deviation from the control group, whereas this distinction is less evident in sepsis cases. Furthermore, both datasets exhibit significant inter-individual variability, which adds another layer of complexity to the classification task. Despite NEC being an easier condition to classify, TabPFN demonstrates a clear advantage due to its ability to adapt to new tasks through prompt-based conditioning. This result underscores the potential of pre-trained foundation models to enhance clinical decision-making by generalizing across different medical conditions

with minimal task-specific tuning.

3.2 External validation

To assess the generalizability of the models beyond the development cohort, external validation was performed using data from a distinct clinical center. External validation is a critical step in evaluating the robustness and applicability of predictive models, as it enables the assessment of performance on populations that may differ in terms of demographics, clinical practices, and data distributions. This step is particularly important in neonatal care, where inter-center variability can be substantial.

Once the models were tested and internally validated on the CHUV cohort, they underwent external validation using data from USZ. Data from KiSpi were not included in this study due to the limited number of sepsis cases available after applying the inclusion criteria. As shown in Table 2.1, KiSpi primarily treats neonates with a higher gestational age compared to CHUV and USZ. One of the inclusion criteria for this study required patients to have a gestational age below 32 weeks, which resulted in only two sepsis cases at KiSpi. Given this small sample size, these data were not considered for external validation.

Based on the results obtained locally on CHUV data, only SVM and TabPFN were selected for further validation, as FORMED did not achieve performance comparable to the baseline model (see Table 3.1). Both models were trained on CHUV data and subsequently tested on USZ data.

To assess model performance under conditions that mimic a real-world clinical setting, two experimental strategies were implemented. In the first experiment, SVM was evaluated using the same preprocessing pipeline as CHUV, meaning that the scaling parameters were derived from CHUV data. This approach simulates an offline deployment scenario, where the model, having no prior exposure to the new cohort, relies solely on the training phase without any additional rescaling procedure. In the second experiment, the SVM model was recalibrated using scaling parameters derived from USZ data, allowing for an assessment of the model's adaptability to distribution shifts. While this second approach does not fully represent a real-world deployment scenario, it provides insight into the model's generalization and adaptability. These two experiments were conducted only for SVM, as TabPFN internally manages normalization and missing value imputation. To obtain a benchmark performance for sepsis prediction on the external USZ data, both models were also trained/conditioned directly on USZ data to establish a reference for their expected performance.

Table 3.6 presents the results of this analysis. TabPFN outperforms SVM, demonstrating consistent performance regardless of whether it was conditioned on CHUV or USZ data. This finding highlights the robustness of TabPFN in adapting to new data distributions without explicit recalibration. SVM's performance is highly dependent on the chosen calibration strategy. Without recalibration, a significant drop in performance is observed, indicating that SVM cannot effectively transfer knowledge from one cohort to another without additional

adjustment.

Table 3.6: Performance comparison of SVM and TabPFN on USZ data (N=149) under different validation settings. Both models were initially trained on CHUV data and tested on USZ data, with SVM evaluated under two different rescaling strategies: (1) using CHUV-based scaling parameters and (2) rescaling with USZ-based scaling parameters. Additionally, both models were trained directly on USZ data to establish a benchmark for external validation.

	Model	**Accuracy**	**Precision**	**Recall**	**F1 score**	**AUROC**
zero-shot	SVM[1]	55.72 (11.08)	42.44 (6.77)	**89.33** (9.47)	57.27 (7.50)	64.35 (9.13)
	SVM[2]	72.10 (10.58)	56.92 (14.39)	59.50 (23.07)	56.46 (18.41)	68.69 (12.59)
	TabPFN	**76.25** (5.42)	**64.11** (8.75)	62.22 (13.13)	**62.55** (9.89)	**72.56** (6.85)
retrained	SVM	71.79 (10.29)	56.70 (14.12)	**58.00** (18.09)	56.66 (15.17)	68.05 (11.43)
	TabPFN	**77.91** (6.05)	**70.41** (10.71)	52.22 (14.72)	**59.50** (13.51)	**71.18** (8.18)

Chapter 4

Discussion

4.1 Comparison of traditional machine learning and foundation models

The primary objective of this study was to investigate the potential of introducing advanced modeling paradigms, specifically, FMs, within the domain of clinical research, with a focus on neonatal sepsis prediction. Several strategies were explored to assess whether these novel approaches could address long-standing limitations associated with traditional ML techniques. Among these challenges are suboptimal generalization capabilities, limited robustness to distributional shifts across patient populations and care settings, reduced adaptability to multiple tasks, and a strong reliance on extensive feature engineering, a process that is both time- and resource-intensive, often requiring considerable domain expertise, particularly in healthcare settings.

While the application of FMs in clinical settings is promising and their potential remains substantial, our findings underscore the importance of cautious interpretation. The observed performance gains were not uniform and depended heavily on both the model architecture and the underlying data characteristics. In particular, the benefits of adopting FM-based approaches appear to be contingent upon factors such as pretraining strategies, input data modalities, and the nature of the downstream task. To systematically assess these aspects, two distinct FM paradigms were evaluated in this study: one operating directly on raw medical time series and another utilizing an innovative in-context learning approach tailored for tabular data. The results revealed key insights into the strengths and limitations of each method, highlighting the current capabilities of FMs while also identifying critical areas for further investigation.

4.1.1 From forecasting to classification: repurposing raw-signal foundation models

In this study, we investigated FORMED, a FM-based classification framework that operates directly on raw physiological time series, with the goal of circumventing the need for manual feature engineering. This approach aligns with recent trends in ML, where end-to-end architectures are increasingly employed to extract meaningful representations directly from unprocessed input data. Despite these theoretical advantages, the empirical results obtained using FORMED were consistently inferior to those achieved by a much simpler SVM classifier trained on handcrafted features. This discrepancy raises important questions regarding the limitations of current FM architectures for clinical time series classification, particularly in the context of neonatal sepsis detection.

One contributing factor to this suboptimal performance may be the inefficacy of the transfer learning process. Specifically, FORMED did not appear to leverage the information acquired during the repurposing phase. One possible explanation lies in the nature of the signals used in the repurposing phase, which differ significantly from those used for sepsis prediction. While both are related to cardiovascular activity, the physiological signals collected in this study differ markedly from standard ECG waveforms in both morphology and sampling frequency. For example, ECGs used in FORMED's repurposing phase are typically sampled at 500 Hz, whereas the sepsis-related signals used in this study were recorded at a much lower frequency, specifically minute-by-minute. As a result, the underlying temporal structure and spectral content of the signals differ substantially, potentially limiting the model's ability to generalize learned patterns to this new domain.

Moreover, the nature of sepsis presents additional challenges. An exploratory analysis using FORMED's backbone model, TimesFM, to forecast sepsis-related physiological signals revealed that the model could effectively capture the general trend of the signal while filtering out substantial amounts of noise. This finding suggests that the model is capable of modeling low-frequency components while smoothing signal fluctuations. However, unlike tasks such as arrhythmia detection, where periodic and highly structured patterns in ECG signals can be efficiently leveraged by trend-based forecasting models, sepsis is inherently more heterogeneous and lacks regular, easily identifiable temporal signatures. While capturing signal trends may prove sufficient in certain applications, it may not adequately reflect the subtle physiological changes that precede sepsis onset. Clinical events frequently associated with sepsis, such as apnea or lethargy, may not be detectable through trend estimation alone, especially when these manifestations are brief, irregular, or context dependent. Despite these limitations, previous internal investigations conducted at CHUV have shown that trend-related features rank among the most important variables for neonatal sepsis classification, as revealed through feature importance analyses. Accordingly, our evaluation of FORMED proceeded under the assumption that the latent representations extracted by TimesFM, despite being based primarily on trend information, could still capture clinically relevant patterns sufficient for sepsis classification. This hypothesis underpinned the rationale for further investigating FORMED's applicability to the task, despite its reliance on a

model originally pretrained for general-purpose time series forecasting.

To further evaluate FORMED's adaptability to the neonatal sepsis prediction task, two complementary analyses were conducted. The first set of analyses investigated the impact of different training modalities. Specifically, we evaluated whether including the sepsis dataset alongside the original set of publicly available physiological time series could improve model performance. However, this strategy did not yield substantial gains. A plausible explanation for this outcome lies in the imbalance between the datasets: for example, PTB-XL contains over 10,000 samples, whereas the sepsis dataset comprises only 471 patients. As dataset size was not explicitly considered during weight updates, the model may have disproportionately adapted to the larger datasets, thus failing to sufficiently internalize representations relevant to sepsis. This hypothesis is supported by training dynamics observed via MLFlow, which revealed indications of overfitting on the dominant datasets and limited generalization to the target clinical task.

The second analysis aimed to address architectural limitations of FORMED. Two targeted modifications were explored to improve its capacity to model multivariate clinical signals and patient-specific information. First, an additional self-attention layer was introduced before the final classification head to mitigate the univariate processing constraint imposed by the backbone. Second, a gestational age embedding was incorporated to condition the model on a clinically relevant covariate known to influence sepsis progression. However, none of the altered configurations demonstrated clear improvements in predictive performance. The limited dataset size likely contributed to the lack of benefit from the increased model complexity introduced by the self-attention mechanism. Indeed, learning curves showed early signs of overfitting, suggesting that the model was unable to generalize effectively under these constraints.

Finally, an additional and often overlooked factor pertains to the translational applicability of pretrained models. During the repurposing stage, FORMED was trained predominantly on time series data from adult patients. However, preterm neonates exhibit unique physiological characteristics and developmental trajectories, marked by high inter- and intra-patient variability. This physiological heterogeneity, compounded by gestational age differences and ongoing maturation processes, may limit the effectiveness of transfer learning from adult to neonatal populations. Although the architecture of FORMED is theoretically designed to accommodate such adaptation through lightweight fine-tuning, our findings suggest that this capability was insufficient to achieve satisfactory classification performance in the neonatal sepsis context.

Taken together, these observations highlight several limitations that must be addressed before raw-data-based FMs like FORMED can be effectively deployed in clinical neonatal care. While promising in principle, the current instantiation of this approach requires further refinement to fully realize its potential for early sepsis detection in neonates.

4.1.2 Evaluating tabular foundation models for sepsis prediction

The second modeling strategy explored in this study focused on assessing the performance and robustness of a FM tailored for tabular data. To this end, we systematically compared TabPFN with a SVM, taken as baseline, across a variety of experimental configurations. Given that both models rely on hand-engineered features, the utility of TabPFN must be substantiated either through superior predictive performance or through improved efficiency and adaptability.

Across all experiments, TabPFN consistently outperformed the SVM, particularly under constrained conditions such as reduced training data availability. These results suggest that FMs like TabPFN may offer a compelling alternative to traditional ML approaches in clinical settings, where data scarcity and limited computational resources are common. Notably, TabPFN maintained high performance even when trained on a minimal set of raw features. This observation is particularly relevant in healthcare environments, where the collection of annotated data, especially in the context of rare diseases, requires substantial input from domain experts, thereby increasing both development time and personnel workload.

One of the distinguishing advantages of TabPFN lies in its practical usability. As a plug-and-play solution, TabPFN abstracts away the complexity of model pretraining and large-scale dataset curation. In contrast, conventional ML workflows often place the full burden of model development, feature engineering, and deployment on the same practitioner. TabPFN's accessibility, for example through platforms like Hugging Face, requires only minimal setup and is particularly suited for clinical contexts where ML expertise may be limited.

Additionally, model performance was evaluated across varying levels of feature complexity. TabPFN demonstrated comparable predictive accuracy even when limited to basic statistical features, indicating that its performance does not strongly depend on the inclusion of advanced, domain-specific variables. These results suggest that the feature engineering process can be substantially simplified without compromising model effectiveness. Minimizing the reliance on iterative feature selection and reducing the need for extensive clinical input may lead to more efficient model development and better allocation of resources within hospital settings.

In summary, the performance and usability of TabPFN emphasize the potential advantages of integrating FMs into clinical predictive workflows. Its ability to operate effectively under typical clinical constraints, such as limited data, minimal preprocessing, and resource scarcity, positions it as a viable candidate for broader adoption in real-world neonatal care settings.

4.1.3 Generalizability of the ML and foundation model approaches on additional tasks and datasets

This study highlights the transformative potential of FMs in redefining ML workflows within clinical settings. Traditionally, each clinical prediction task has required the development

and training of task-specific models, often constrained by limited data availability and high development costs. In contrast, FMs offer a paradigm shift by enabling knowledge transfer across tasks without the need for retraining, thus reducing the time and resources required for model development. From an institutional perspective, the adoption of flexible, general-purpose models may represent a more cost-effective and sustainable strategy than investing in multiple narrowly tailored solutions. This shift toward generalizable modeling frameworks could foster greater collaboration across hospitals and research institutions, promoting the development of shared models capable of generalizing across patient populations and clinical contexts. Such an approach not only enhances model robustness but also contributes to improved reproducibility and quality of clinical research.

To further investigate the generalization capacity of FMs, we evaluated model performance on NEC prediction, using a zero-shot learning framework. Results from this evaluation reaffirmed the superior generalization capabilities of FM-based approaches. Notably, the baseline SVM model also achieved competitive performance, likely due to the pathophysiological overlap between sepsis and NEC, as both conditions are associated with bacterial infections and may present with similar temporal patterns in physiological signals. Despite this, TabPFN consistently outperformed the SVM across evaluation metrics, demonstrating its potential for application in complex and data-constrained clinical scenarios, such as the early detection of rare neonatal diseases.

To assess the generalization capabilities of the tested models, an external validation was conducted using USZ data. This evaluation aimed to simulate a realistic deployment scenario, where models trained on data from one clinical center are applied to a different patient population without additional retraining. While TabPFN consistently maintained high performance, the SVM model exhibited considerable sensitivity to the preprocessing pipeline, particularly the normalization strategy employed. This variability highlights a key limitation of traditional ML approaches, which often rely on strong assumptions about data distribution and require careful calibration to new environments. The superior generalization performance of TabPFN can be attributed to its ability to extract more abstract and robust representations of the input data, enabling it to remain effective across varying clinical settings. In contrast, the SVM model showed performance degradation when applied directly to USZ data using scaling parameters derived from the CHUV cohort, indicating a lack of inherent robustness to distribution shifts. This limitation is likely due to underlying differences in patient populations across the two centers, particularly in gestational age and birth weight, as illustrated in Appendix A (Figures A.1 and A.2). These findings underscore the importance of developing and validating models that are not only accurate but also robust to the heterogeneity inherent in real-world clinical data.

4.2 Key findings

This study was conducted within the framework of the ADONIS project, which aims to reduce the incidence of neonatal sepsis in Switzerland. As part of this effort, we compiled a

large multicenter dataset comprising over 15,000 neonatal patients from three of the country's most prominent hospitals. To our knowledge, this represents the largest cohort of neonatal patients ever collected for a study of this kind in Switzerland.

The availability of such a comprehensive dataset enabled the training and evaluation of multiple models on a robust sample, addressing one of the primary limitations in the current literature, namely, the scarcity of training data in neonatal sepsis research [6]. Unlike previous studies, which have often been constrained by limited data, our work provides reliable estimates of model performance across a wide patient population. Furthermore, the dataset includes numerous additional variables, such as laboratory results and medication records, that were not considered in this analysis but may be leveraged in future research and extended to different clinical prediction tasks.

A feature engineering pipeline tailored to neonatal sepsis was developed and validated both internally on data from CHUV and externally using data from USZ. The ability to externally validate a model developed at CHUV using an independent cohort represents a major strength of this study, offering insight into the generalizability of the model across heterogeneous data distributions. Moreover, the collaboration between the three participating centers facilitated data sharing and laid the groundwork for defining a standard of interoperability among institutions.

The application of two distinct FMs for time series analysis in this study provided valuable insights into both the potential and the limitations of these approaches in the clinical domain. Rather than solely emphasizing improved performance, this investigation focused on evaluating the practical impact of FMs under conditions that closely mirror real-world constraints, including limited training data and reduced feature complexity. The findings underscore the importance of rigorous validation and careful adaptation of FMs to specific clinical use cases before deployment. When properly assessed and adapted, FMs may offer a promising strategy to reduce the number of task-specific models required in hospital settings. By enabling generalization across multiple prediction tasks, even those traditionally hindered by data scarcity, FMs could support a more scalable and resource-efficient approach to clinical decision support.

Finally, the close collaboration with CHUV and continuous feedback from the clinical team enabled us to evaluate the adaptability of these models to alternative, clinically relevant use cases. This synergy between ML development and clinical insight was critical in assessing the feasibility and translational potential of FM-based approaches in neonatal intensive care.

4.3 Limitations and challenges

One of the primary limitations associated with the use of FMs is the significant computational cost required to deploy them. Unlike traditional ML models such as SVM, which can be executed efficiently on standard CPU hardware, FMs, especially those of large scale, necessitate advanced technological components, such as GPUs, to perform inference within a

clinically acceptable time frame. This requirement becomes particularly critical in scenarios that demand real-time predictions to support patient care. Such equipment entails substantial costs for hospitals, which are often publicly funded institutions. The procurement of new hardware typically involves public tender processes, which, combined with the rapid obsolescence of GPU technology, can pose logistical and financial challenges for long-term deployment.

This study focused on low-frequency sampled time series collected in the NICU. Due to the high-stakes nature of the NICU environment, patients are under continuous monitoring and require frequent interventions from clinical staff. Given the fragile condition of neonates, this includes a high degree of interaction by both nurses and caregivers. As such, the presence of artifacts in the recorded signals cannot be excluded. Currently, the model does not have access to metadata regarding clinical interventions, such as manual handling of the patient, feeding by the mother, or episodes of crying, which may affect physiological measurements like oxygen saturation. These events can introduce noise that is difficult to disambiguate without contextual information. Due to the complexity involved, no preprocessing or artifact removal was performed on the raw time series in this study.

Another limitation concerns the composition of the study cohort. Due to current inclusion criteria, patients admitted at KiSpi were excluded, resulting in the omission of data from over 1,500 individuals. Future work should aim to remove gestational age-based inclusion criteria to expand the applicability of the model to the entire neonatal population, thereby simulating deployment across the full neonatal ward and not only in high-risk units.

Finally, an important limitation relates to the regulatory landscape. According to the European Medical Device Regulation (Regulation EU 2017/745), all predictive models must be certified as medical devices before deployment in clinical settings. This process is already complex for simpler models such as SVM due to the vagueness of current regulatory guidelines. In the case of large-scale FMs, certification is even more challenging. These models are typically pretrained by third parties on large, often unspecified datasets, raising concerns about data privacy and transparency. Furthermore, issues related to data ownership and intellectual property rights further complicate the certification process. Recent developments such as the European AI Act (Regulation EU 2024/1689) are being introduced to address these regulatory gaps, but significant challenges remain before widespread clinical adoption of FM-based models becomes feasible.

Finally, a critical limitation concerns the current regulatory landscape. According to the European Medical Device Regulation (Regulation EU 2017/745), all predictive models intended for clinical use must undergo formal certification as medical devices prior to deployment. This process is already intricate for relatively simple models, such as SVM, largely due to ambiguities in existing regulatory guidelines. The certification of large-scale FMs, however, presents even greater challenges. These models are typically pretrained by third parties on large and often undisclosed datasets, raising concerns about data provenance, privacy, and transparency. Additionally, unresolved issues regarding data ownership and intellectual

property rights further complicate the pathway to regulatory compliance.

A further obstacle to the clinical adoption of FMs lies in their limited interpretability. Due to their complex and opaque architectures, FMs often function as "black boxes", making it difficult for developers, clinicians, and regulators to fully understand their decision-making processes. This lack of transparency poses a significant barrier to certification, as current medical device regulations emphasize the importance of explainability and risk assessment to ensure patient safety. While ongoing efforts, including the European AI Act (Regulation EU 2024/1689), aim to address some of these regulatory gaps, substantial challenges remain before FM-based models can be reliably certified and deployed in real-world clinical environments.

4.4 Future directions

To allow for a comprehensive evaluation of the models' generalization capabilities, inclusion criteria should be revised to encompass the entire patient cohort, including data from KiSpi. An additional performance analysis should be conducted by comparing the models' predictive capabilities to clinical practice, including benchmarks such as the HeRO score. Prior to initiating a clinical trial, it is essential to conduct a comprehensive evaluation of the model's "live" performance under conditions that closely reflect clinical practice. This includes simulating the true prevalence of sepsis within both training and testing datasets, thereby ensuring that performance metrics accurately represent real-world deployment scenarios. Following this intermediary validation step, a clinical trial could be launched, initially for the SVM model, with the aim of assessing its effectiveness in a real-world setting and ultimately obtaining certification for clinical use.

Further improvements can also be pursued by further customizing the tested model architectures. In the case of FORMED, the feature extraction module could be refined to better specialize in medical time series. The FORMED framework is modular by design, allowing for flexible substitution of the feature encoder. With the increasing number of FMs for time series forecasting published recently [56, 57, 81], it becomes feasible to tune and select architectures that are most suitable for clinical signals. For TabPFN, various architectural adaptations could be explored to reduce the model's reliance on external feature engineering. Research in this direction is already ongoing [82], aiming to make the model more robust and adaptable to raw clinical tabular data.

To enhance the model's awareness of a patient's clinical state and to minimize misinterpretation of signal artifacts caused by clinical interventions, a multimodal modeling approach could be employed. The substantial volume of heterogeneous data collected across the three centers enables the potential integration of multiple modalities, including laboratory results and medication records. Recent studies have introduced architectures capable of modeling a patient's complete clinical history by integrating information from diverse sources [12]. This strategy would allow models to better reflect the complex reasoning process of clini-

cians, potentially providing more actionable insights for patient management. Moreover, the integration of such models into a multi-agent framework could facilitate the development of decision-support tools within hospital information systems. Rather than being limited to sepsis prediction alone, such agents could assist with various clinical tasks in real time, offering a scalable solution for intelligent, data-driven decision-making across the healthcare continuum [83].

Chapter 5

Conclusion

This study demonstrates the potential of using FMs for sepsis prediction in neonatal care, with a particular focus on their applicability to real-world clinical settings. By evaluating two distinct FM-based approaches, one operating directly on raw physiological time series and the other on tabular data, we provide a comprehensive assessment of their strengths and limitations in comparison to traditional ML methods. Our findings show that, while FM-based models can alleviate the need for manual feature engineering and exhibit superior generalization capabilities, particularly under data-constrained scenarios, their performance remains highly dependent on the specific model architecture and the nature of the pretraining strategy.

In particular, our experiments revealed that raw-data-based models like FORMED, despite being pretrained on unrelated datasets, can still reach performance levels comparable to classical models such as SVM. However, they failed to consistently outperform traditional approaches, underscoring the challenges of transferring representations learned from non-clinical data to complex biomedical tasks. On the other hand, TabPFN, a FM tailored for tabular data, demonstrated improved robustness and adaptability, even in zero-shot settings, suggesting that FMs can significantly enhance the model development pipeline and reduce resource consumption, especially in terms of time, cost, and personnel, if appropriately chosen and applied.

Nevertheless, this study also highlights key limitations. Despite their promise, FMs are still in an early stage of adoption in healthcare, and their effectiveness remains highly context-dependent. The variability observed in model performance across tasks and institutions emphasizes the need for careful adaptation and thorough validation before clinical deployment. Our results suggest that FMs should not be viewed as plug-and-play solutions but rather as powerful tools that require rigorous evaluation and tuning to meet the demands of high-risk environments such as NICUs.

Overall, while further research and optimization are necessary to fully realize the advantages

of FMs in this domain, the results of this study contribute to an emerging paradigm in healthcare ML prioritizing generalizability, efficiency, and ease of integration. Future work should focus on refining these models, addressing remaining challenges such as distribution shifts and variability in clinical data, and exploring their applicability to a wider range of diagnostic tasks.

Chapter 6

Acknowledgments

Desidero esprimere la mia sincera gratitudine a tutti i membri del laboratorio di Clinical Data Science e a tutto il personale clinico dell'Ospedale di Losanna che mi ha accompagnato lungo questo percorso. Il loro supporto e contributo sono stati fondamentali per la realizzazione di questa tesi. Un ringraziamento particolare va al Professor Jean Louis Raisaro per avermi offerto l'opportunità di svolgere il mio lavoro di tesi all'interno del suo laboratorio, permettendomi di confrontarmi con un problema reale e di mettere alla prova le competenze acquisite durante il mio percorso accademico.

Desidero inoltre ringraziare il Professor Enea Parimbelli, mio relatore, per il costante supporto fornito durante tutto il mio percorso, dalla ricerca della sede per il mio periodo di studi all'estero fino alla redazione finale della presente tesi.

Bibliography

[1] Kristina E. Rudd, Sarah Charlotte Johnson, Kareha M. Agesa, Katya Anne Shackelford, Derrick Tsoi, Daniel Rhodes Kievlan, Danny V. Colombara, Kevin S. Ikuta, Niranjan Kissoon, Simon Finfer, Carolin Fleischmann-Struzek, Flavia R. Machado, Konrad K. Reinhart, Kathryn Rowan, Christopher W. Seymour, R. Scott Watson, T. Eoin West, Fatima Marinho, Simon I. Hay, Rafael Lozano, Alan D. Lopez, Derek C. Angus, Christopher J. L. Murray, and Mohsen Naghavi. Global, regional, and national sepsis incidence and mortality, 1990-2017: analysis for the Global Burden of Disease Study. *Lancet (London, England)*, 395(10219):200–211, January 2020.

[2] Luregn J. Schlapbach, R. Scott Watson, Lauren R. Sorce, Andrew C. Argent, Kusum Menon, Mark W. Hall, Samuel Akech, David J. Albers, Elizabeth R. Alpern, Fran Balamuth, Melania Bembea, Paolo Biban, Enitan D. Carrol, Kathleen Chiotos, Mohammod Jobayer Chisti, Peter E. DeWitt, Idris Evans, Cláudio Flauzino De Oliveira, Christopher M. Horvat, David Inwald, Paul Ishimine, Juan Camilo Jaramillo-Bustamante, Michael Levin, Rakesh Lodha, Blake Martin, Simon Nadel, Satoshi Nakagawa, Mark J. Peters, Adrienne G. Randolph, Suchitra Ranjit, Margaret N. Rebull, Seth Russell, Halden F. Scott, Daniela Carla De Souza, Pierre Tissieres, Scott L. Weiss, Matthew O. Wiens, James L. Wynn, Niranjan Kissoon, Jerry J. Zimmerman, L. Nelson Sanchez-Pinto, Tellen D. Bennett, Society of Critical Care Medicine Pediatric Sepsis Definition Task Force, and Juliane Bubeck Wardenburg. International Consensus Criteria for Pediatric Sepsis and Septic Shock. *JAMA*, 331(8):665, February 2024.

[3] Fatiha Bennaoui, Abdessamad Lalaoui, Nadia El Idrissi Slitine, Nabila Soraa, and Fadl Mrabih Rabou Maoulainine. The HeRO score: Enhancing prognosis and predicting nosocomial infections in newborns: Insights from the neonatal intensive care unit. *Journal of Neonatal-Perinatal Medicine*, 17(1):57–62, 2024.

[4] Eleanor J. Molloy, James L. Wynn, Joseph Bliss, Joyce M. Koenig, Fleur M. Keij, Matt McGovern, Helmut Kuester, Mark A. Turner, Eric Giannoni, Jan Mazela, Marina Degtyareva, Tobias Strunk, Sinno H. P. Simons, Jan Janota, Franz B. Plotz, Ages van den Hoogen, Willem de Boode, Luregn J. Schlapbach, Irwin K. M. Reiss, and on behalf of the Infection, Inflammation, Immunology and Immunisation (I4) section of the ESPR. Neonatal sepsis: need for consensus definition, collaboration and core outcomes. *Pediatric Research*, 88(1):2–4, July 2020.

[5] Joseph A. Carcillo, Robert A. Berg, David Wessel, Murray Pollack, Kathleen Meert, Mark Hall, Christopher Newth, John C. Lin, Allan Doctor, Tom Shanley, Tim Cornell, Rick E. Harrison, Athena F. Zuppa, Ron W. Reeder, Russell Banks, John A. Kellum, Richard Holubkov, Daniel A. Notterman, and J. Michael Dean. A Multicenter Network Assessment of Three Inflammation Phenotypes in Pediatric Sepsis-Induced Multiple Organ Failure. *Pediatric Critical Care Medicine*, 20(12):1137–1146, December 2019.

[6] Laura Cabrera-Quiros, Deedee Kommers, Maria K. Wolvers, Laurien Oosterwijk, Niek Arents, Jacqueline van der Sluijs-Bens, Eduardus J. E. Cottaar, Peter Andriessen, and Carola van Pul. Prediction of Late-Onset Sepsis in Preterm Infants Using Monitoring Signals and Machine Learning. *Critical Care Explorations*, 3(1):e0302, January 2021.

[7] Zheng Peng, Gabriele Varisco, Xi Long, Rong-Hao Liang, Deedee Kommers, Ward Cottaar, Peter Andriessen, and Carola van Pul. A Continuous Late-Onset Sepsis Prediction Algorithm for Preterm Infants Using Multi-Channel Physiological Signals From a Patient Monitor. *IEEE Journal of Biomedical and Health Informatics*, 27(1):550–561, January 2023. Conference Name: IEEE Journal of Biomedical and Health Informatics.

[8] Cristhyne Leon, Guy Carrault, Patrick Pladys, and Alain Beuchee. Early Detection of Late Onset Sepsis in Premature Infants Using Visibility Graph Analysis of Heart Rate Variability. *IEEE journal of biomedical and health informatics*, 25(4):1006–1017, April 2021.

[9] Rohan Joshi, Deedee Kommers, Laurien Oosterwijk, Loe Feijs, Carola van Pul, and Peter Andriessen. Predicting Neonatal Sepsis Using Features of Heart Rate Variability, Respiratory Characteristics, and ECG-Derived Estimates of Infant Motion. *IEEE journal of biomedical and health informatics*, 24(3):681–692, March 2020.

[10] Marisse Meeus, Charlie Beirnaert, Ludo Mahieu, Kris Laukens, Pieter Meysman, Antonius Mulder, and David Van Laere. Clinical decision support for improved neonatal care: The development of a machine learning model for the prediction of late-onset sepsis and necrotizing enterocolitis. *The Journal of Pediatrics*, 266:113869, 2024.

[11] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir

Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the Opportunities and Risks of Foundation Models, July 2022. arXiv:2108.07258 [cs].

[12] Pawel Renc, Yugang Jia, Anthony E. Samir, Jaroslaw Was, Quanzheng Li, David W. Bates, and Arkadiusz Sitek. Zero shot health trajectory prediction using transformer. *npj Digital Medicine*, 7(1):1–10, September 2024. Publisher: Nature Publishing Group.

[13] Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A. Pfeffer, Jason Fries, and Nigam H. Shah. The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine*, 6(1):1–10, July 2023. Publisher: Nature Publishing Group.

[14] Lin Lawrence Guo, Ethan Steinberg, Scott Lanyon Fleming, Jose Posada, Joshua Lemmon, Stephen R. Pfohl, Nigam Shah, Jason Fries, and Lillian Sung. EHR foundation models improve robustness in the presence of temporal distribution shift. *Scientific Reports*, 13(1):3767, March 2023. Publisher: Nature Publishing Group.

[15] Nan Huang, Haishuai Wang, Zihuai He, Marinka Zitnik, and Xiang Zhang. Repurposing Foundation Model for Generalizable Medical Time Series Classification, October 2024. arXiv:2410.03794.

[16] Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, January 2025. Publisher: Nature Publishing Group.

[17] Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. MEDITRON-70B: Scaling Medical Pretraining for Large Language Models, November 2023. arXiv:2311.16079 [cs].

[18] Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. BEHRT: Transformer for Electronic Health Records. *Scientific Reports*, 10(1):7155, April 2020.

[19] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, February 2020. arXiv:1901.08746 [cs].

[20] Eric Giannoni, Philipp K. A. Agyeman, Martin Stocker, Klara M. Posfay-Barbe, Ulrich Heininger, Ben D. Spycher, Sara Bernhard-Stirnemann, Anita Niederer-Loher, Christian R. Kahlert, Alex Donas, Antonio Leone, Paul Hasters, Christa Relly, Thomas Riedel, Claudia Kuehni, Christoph Aebi, Christoph Berger, Luregn J. Schlapbach, and Swiss Pediatric Sepsis Study. Neonatal Sepsis of Early Onset, and Hospital-Acquired and Community-Acquired Late Onset: A Prospective Population-Based Cohort Study. *The Journal of Pediatrics*, 201:106–114.e4, October 2018.

[21] Daniel K. Benjamin and Barbara J. Stoll. Infection in Late Preterm Infants. *Clinics in Perinatology*, 33(4):871–882, December 2006.

[22] Elsa Kermorvant-Duchemin, Sophie Laborie, Muriel Rabilloud, Alexandre Lapillonne, and Olivier Claris. Outcome and prognostic factors in neonates with septic shock. *Pediatric Critical Care Medicine: A Journal of the Society of Critical Care Medicine and the World Federation of Pediatric Intensive and Critical Care Societies*, 9(2):186–191, March 2008.

[23] Barbara J. Stoll, Nellie Hansen, Avroy A. Fanaroff, Linda L. Wright, Waldemar A. Carlo, Richard A. Ehrenkranz, James A. Lemons, Edward F. Donovan, Ann R. Stark, Jon E. Tyson, William Oh, Charles R. Bauer, Sheldon B. Korones, Seetha Shankaran, Abbot R. Laptook, David K. Stevenson, Lu-Ann Papile, and W. Kenneth Poole. Late-onset sepsis in very low birth weight neonates: the experience of the NICHD Neonatal Research Network. *Pediatrics*, 110(2 Pt 1):285–291, August 2002.

[24] Luregn J. Schlapbach, Maude Aebischer, Mark Adams, Giancarlo Natalucci, Jan Bonhoeffer, Philipp Latzin, Mathias Nelle, Hans Ulrich Bucher, Beatrice Latal, and Swiss Neonatal Network and Follow-Up Group. Impact of sepsis on neurodevelopmental outcome in a Swiss National Cohort of extremely premature infants. *Pediatrics*, 128(2):e348–357, August 2011.

[25] Kirsten Glaser, Christoph Härtel, Claus Klingenberg, Egbert Herting, Mats I. Fortmann, Christian P. Speer, Hans J. Stensvold, Zuzana Huncikova, Arild E. Rønnestad, Martin M. Nentwich, Andreas Stahl, Olaf Dammann, Wolfgang Göpel, and German Neonatal Network, the Norwegian Neonatal Network Investigators, and the Infection, Inflammation, Immunology and Immunisation section of the European Society for Paediatric Research. Neonatal Sepsis Episodes and Retinopathy of Prematurity in Very Preterm Infants. *JAMA network open*, 7(7):e2423933, July 2024.

[26] Jennifer Valeska Elli Brown, Nick Meader, Kath Wright, Jemma Cleminson, and William McGuire. Assessment of C-Reactive Protein Diagnostic Test Accuracy for Late-

Onset Infection in Newborn Infants: A Systematic Review and Meta-analysis. *JAMA pediatrics*, 174(3):260–268, March 2020.

[27] Alexa Dierig, Christoph Berger, Philipp K. A. Agyeman, Sara Bernhard-Stirnemann, Eric Giannoni, Martin Stocker, Klara M. Posfay-Barbe, Anita Niederer-Loher, Christian R. Kahlert, Alex Donas, Paul Hasters, Christa Relly, Thomas Riedel, Christoph Aebi, Luregn J. Schlapbach, Ulrich Heininger, and Swiss Pediatric Sepsis Study. Time-to-Positivity of Blood Cultures in Children With Sepsis. *Frontiers in Pediatrics*, 6:222, 2018.

[28] Reese H. Clark, Barry T. Bloom, Alan R. Spitzer, and Dale R. Gerstmann. Reported medication use in the neonatal intensive care unit: data from a large national data set. *Pediatrics*, 117(6):1979–1987, June 2006.

[29] Steven L. Raymond, Jaimar C. Rincon, James L. Wynn, Lyle L. Moldawer, and Shawn D. Larson. Impact of Early-Life Exposures to Infections, Antibiotics, and Vaccines on Perinatal and Long-term Health and Disease. *Frontiers in Immunology*, 8:729, 2017.

[30] Matthew McGovern, Eric Giannoni, Helmut Kuester, Mark A. Turner, Agnes van den Hoogen, Joseph M. Bliss, Joyce M. Koenig, Fleur M. Keij, Jan Mazela, Rebecca Finnegan, Marina Degtyareva, Sinno H. P. Simons, Willem P. de Boode, Tobias Strunk, Irwin K. M. Reiss, James L. Wynn, Eleanor J. Molloy, and Infection, Inflammation, Immunology and Immunisation (I4) section of the ESPR. Challenges in developing a consensus definition of neonatal sepsis. *Pediatric Research*, 88(1):14–26, July 2020.

[31] James L. Wynn and Richard A. Polin. A neonatal sequential organ failure assessment score predicts mortality to late-onset sepsis in preterm very low birth weight infants. *Pediatric Research*, 88(1):85–90, July 2020.

[32] Karen D. Fairchild. Predictive monitoring for early detection of sepsis in neonatal ICU patients. *Current Opinion in Pediatrics*, 25(2):172–179, April 2013.

[33] Xiaohan Hu, Hansi Liang, Fang Li, Rui Zhang, Yanbo Zhu, Xueping Zhu, and Yunyun Xu. Necrotizing enterocolitis: current understanding of the prevention and management. *Pediatric Surgery International*, 40(1):32, January 2024.

[34] Annette Gawron Roberts, Noelle Younge, and Rachel Gottron Greenberg. Neonatal Necrotizing Enterocolitis: An Update on Pathophysiology, Treatment, and Prevention. *Paediatric Drugs*, 26(3):259–275, May 2024.

[35] Barrie S. Rich and Stephen E. Dolgin. Necrotizing Enterocolitis. *Pediatrics In Review*, 38(12):552–559, December 2017.

[36] Martin Stocker, Imant Daunhawer, Wendy Van Herk, Salhab El Helou, Sourabh Dutta, Frank A. B. A. Schuerman, Rita K. Van Den Tooren-de Groot, Jantien W. Wieringa, Jan Janota, Laura H. Van Der Meer-Kappelle, Rob Moonen, Sintha D. Sie, Esther De Vries,

Albertine E. Donker, Urs Zimmerman, Luregn J. Schlapbach, Amerik C. De Mol, Angelique Hoffmann-Haringsma, Madan Roy, Maren Tomaske, René F. Kornelisse, Juliette Van Gijsel, Frans B. Plötz, Sven Wellmann, Niek B. Achten, Dirk Lehnick, Annemarie M. C. Van Rossum, and Julia E. Vogt. Machine Learning Used to Compare the Diagnostic Accuracy of Risk Factors, Clinical Signs and Biomarkers and to Develop a New Prediction Model for Neonatal Early-onset Sepsis. *Pediatric Infectious Disease Journal*, 41(3):248–254, March 2022.

[37] Zheng Peng, Gabriele Varisco, Rong-Hao Liang, Deedee Kommers, Ward Cottaar, Peter Andriessen, Carola van Pul, and Xi Long. DeepLOS: Deep learning for late-onset sepsis prediction in preterm infants using heart rate variability. *Smart Health*, 26:100335, 2022.

[38] Cristhyne Leon, Patrick Pladys, Alain Beuchee, and Guy Carrault. Recurrent Neural Networks for Early Detection of Late Onset Sepsis in Premature Infants Using Heart Rate Variability. In *2021 Computing in Cardiology (CinC)*, pages 1–4, Brno, Czech Republic, September 2021. IEEE.

[39] Kim Huat Goh, Le Wang, Adrian Yong Kwang Yeow, Hermione Poh, Ke Li, Joannas Jie Lin Yeow, and Gamaliel Yu Heng Tan. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. *Nature Communications*, 12(1):711, January 2021.

[40] Andrew Wong, Erkin Otles, John P. Donnelly, Andrew Krumm, Jeffrey McCullough, Olivia DeTroyer-Cooley, Justin Pestrue, Marie Phillips, Judy Konye, Carleen Penoza, Muhammad Ghous, and Karandeep Singh. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. *JAMA internal medicine*, 181(8):1065–1070, August 2021.

[41] L. Nelson Sanchez-Pinto, Tellen D. Bennett, Emily K. Stroup, Yuan Luo, Mihir Atreya, Juliane Bubeck Wardenburg, Grace Chong, Alon Geva, E. Vincent S. Faustino, Reid W. Farris, Mark W. Hall, Colin Rogerson, Sareen S. Shah, Scott L. Weiss, and Robinder G. Khemani. Derivation, Validation, and Clinical Relevance of a Pediatric Sepsis Phenotype With Persistent Hypoxemia, Encephalopathy, and Shock*. *Pediatric Critical Care Medicine*, 24(10):795–806, October 2023.

[42] Sivasubramanium V. Bhavani, Matthew Semler, Edward T. Qian, Philip A. Verhoef, Chad Robichaux, Matthew M. Churpek, and Craig M. Coopersmith. Development and validation of novel sepsis subphenotypes using trajectories of vital signs. *Intensive Care Medicine*, 48(11):1582–1592, November 2022.

[43] Wongeun Song, Se Young Jung, Hyunyoung Baek, Chang Won Choi, Young Hwa Jung, and Sooyoung Yoo. A Predictive Model Based on Machine Learning for the Early Detection of Late-Onset Neonatal Sepsis: Development and Observational Study. *JMIR Medical Informatics*, 8(7):e15965, July 2020.

[44] Manuel Burger, Fedor Sergeev, Malte Londschien, Daphné Chopard, Hugo Yèche, Eike

Gerdes, Polina Leshetkina, Alexander Morgenroth, Zeynep Babür, Jasmina Bogojeska, Martin Faltys, Rita Kuznetsova, and Gunnar Rätsch. Towards Foundation Models for Critical Care Time Series, November 2024. arXiv:2411.16346.

[45] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners, July 2020. arXiv:2005.14165 [cs].

[46] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021.

[47] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019. arXiv:1810.04805 [cs].

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, February 2021. arXiv:2103.00020 [cs].

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, August 2023. arXiv:1706.03762 [cs].

[50] Stefano Woerner and Christian F. Baumgartner. Navigating Data Scarcity using Foundation Models: A Benchmark of Few-Shot and Zero-Shot Learning Approaches in Medical Imaging, August 2024. arXiv:2408.08058 [cs].

[51] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea

Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu,

Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. GPT-4 Technical Report, March 2024. arXiv:2303.08774 [cs].

[52] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission, November 2020. arXiv:1904.05342 [cs].

[53] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera Y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, August 2023.

[54] Nimeesha Chan, Felix Parker, William Bennett, Tianyi Wu, Mung Yao Jia, James Fackler, and Kimia Ghobadi. MedTsLLM: Leveraging LLMs for Multimodal Medical Time Series Analysis, August 2024. arXiv:2408.07773 [cs].

[55] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Michaela Hardt, Peter J. Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Yi Zhang, Gerardo Flores, Gavin E. Duggan, Jamie Irvine, Quoc Le, Kurt Litsch, Alexander Mossin, Justin Tansuwan, De Wang, James Wexler, Jimbo Wilson, Dana Ludwig, Samuel L. Volchenboum, Katherine Chou, Michael Pearson, Srinivasan Madabushi, Nigam H. Shah, Atul J. Butte, Michael D. Howell, Claire Cui, Greg S. Corrado, and Jeffrey Dean. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1(1):18, May 2018.

[56] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the Language of Time Series, November 2024. arXiv:2403.07815.

[57] Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A Family of Open Time-series Foundation Models, October 2024. arXiv:2402.03885 [cs].

[58] Tian Zhou, PeiSong Niu, Xue Wang, Liang Sun, and Rong Jin. One Fits All:Power General Time Series Analysis by Pretrained LM, October 2023. arXiv:2302.11939.

[59] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting, April 2024. arXiv:2310.10688 [cs].

[60] Brian Lester, Rami Al-Rfou, and Noah Constant. The Power of Scale for Parameter-Efficient Prompt Tuning, September 2021. arXiv:2104.08691 [cs].

[61] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, April 2021. arXiv:2005.11401 [cs].

[62] Christopher Y. K. Williams, Brenda Y. Miao, Aaron E. Kornblith, and Atul J. Butte. Evaluating the use of large language models to provide clinical recommendations in the Emergency Department. *Nature Communications*, 15(1):8236, October 2024.

[63] Ethan Goh, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Joséphine A. Cool, Zahir Kanjee, Andrew S. Parsons, Neera Ahuja, Eric Horvitz, Daniel Yang, Arnold Milstein, Andrew P. J. Olson, Adam Rodman, and Jonathan H. Chen. Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial. *JAMA Network Open*, 7(10):e2440969, October 2024.

[64] Felix Busch, Jakob Nikolas Kather, Christian Johner, Marina Moser, Daniel Truhn, Lisa C. Adams, and Keno K. Bressem. Navigating the European Union Artificial Intelligence Act for Healthcare. *npj Digital Medicine*, 7(1):210, August 2024.

[65] BioMedIT Network. Biomedit network - secure it infrastructure for health-related data. `https://www.biomedit.ch/home.html`, 2025. Accessed: 2025-04-16.

[66] Swiss Personalized Health Network (SPHN). Schemascope - sphn rdf schema explorer. `https://schemascope.dcc.sib.swiss/`, 2025. Accessed: 2025-04-16.

[67] Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM). *Internationale statistische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme, 10. Revision, German Modification (ICD-10-GM)*, 2024.

[68] SNOMED International and Swiss Confederation. *SNOMED CT Swiss Extension, Release 2021*, 2021.

[69] WHO Collaborating Centre for Drug Statistics Methodology. *Guidelines for ATC classification and DDD assignment 2024*. Oslo, Norway, 27th edition, 2023.

[70] Gunther Schadow and Clement J. McDonald. *The Unified Code for Units of Measure (UCUM), Version 2024*, 2024.

[71] Regenstrief Institute. *LOINC Version 2.76*, 2023.

[72] Bundesamt für Statistik (BFS). *Schweizerische Operationsklassifikation (CHOP), Version 2024*, 2024.

[73] Ontotext. *GraphDB: The Semantic Graph Database*, 2024.

[74] Steven M. Pincus. Approximate entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences*, 88(6):2297–2301, 1991.

[75] Joshua S. Richman and J. Randall Moorman. Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology-Heart and Circulatory Physiology*, 278(6):H2039–H2049, 2000.

[76] Hamed Azami and Javier Escudero. Improved fuzzy entropy for biomedical signal analysis: Application to heart rate variability. *IEEE Transactions on Biomedical Engineering*, 63(12):2638–2646, 2016.

[77] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. *scikit-learn: Machine Learning in Python*, 2011.

[78] Erhard Rahm and Hong Hai Do. Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin*, 23(4):3–13, 2000. Estimates that data preparation tasks may take up to 80% of the time in a data mining project.

[79] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015.

[80] Mingzhu Liu, Angela H. Chen, and George H. Chen. Generalized Prompt Tuning: Adapting Frozen Univariate Time Series Foundation Models for Multivariate Healthcare Time Series, November 2024. arXiv:2411.12824.

[81] Azul Garza, Cristian Challu, and Max Mergenthaler-Canseco. TimeGPT-1, May 2024. arXiv:2310.03589 [cs].

[82] Shi Bin Hoo, Samuel Müller, David Salinas, and Frank Hutter. The Tabular Foundation Model TabPFN Outperforms Specialized Time Series Forecasting Models Based on Simple Features.

[83] Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. AgentClinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments, 2024. Version Number: 4.

[84] Yihe Wang, Nan Huang, Taida Li, Yujun Yan, and Xiang Zhang. Medformer: A Multi-Granularity Patching Transformer for Medical Time-Series Classification, October 2024. arXiv:2405.19363 [eess].

Appendix A

Demographics distribution across centers

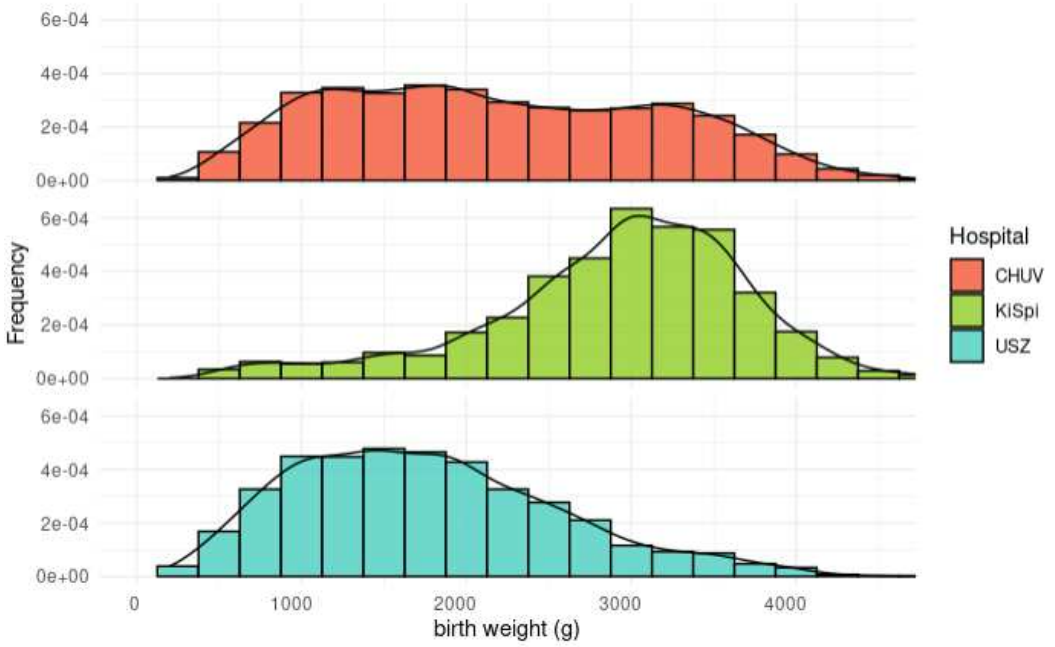


Figure A.1: Distribution of birth weight across the three centers. This figure shows the birth weight distribution for neonatal patients from CHUV (top), KiSpi (middle), and USZ (bottom). Differences in distributions across centers highlight variations in patient populations.

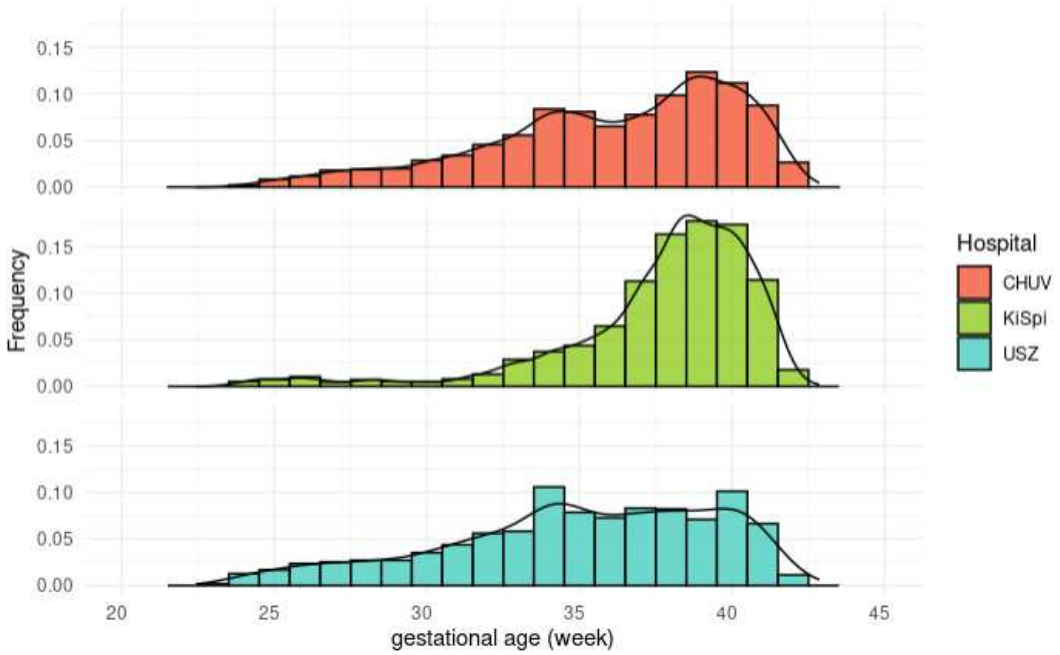
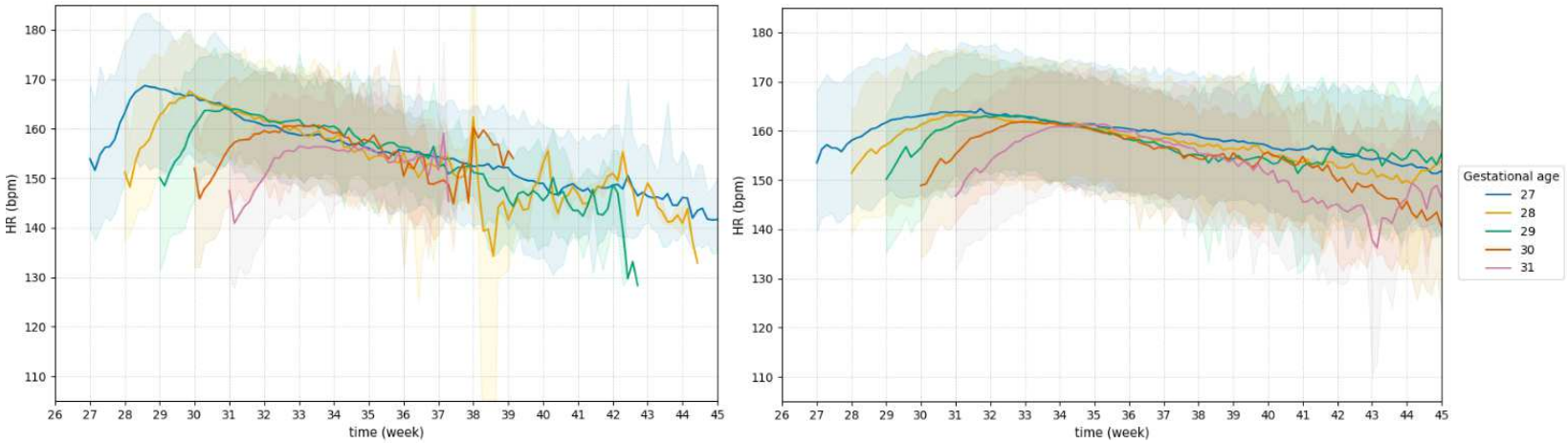


Figure A.2: Distribution of gestational age across the three centers. This figure shows the birth weight distribution for neonatal patients from CHUV (top), KiSpi (middle), and USZ (bottom). Differences in distributions across centers highlight variations in patient populations.

Appendix B

Vitals trajectories across centers


HR (bpm)
time (week)
Gestational age
27
28
29
30
31

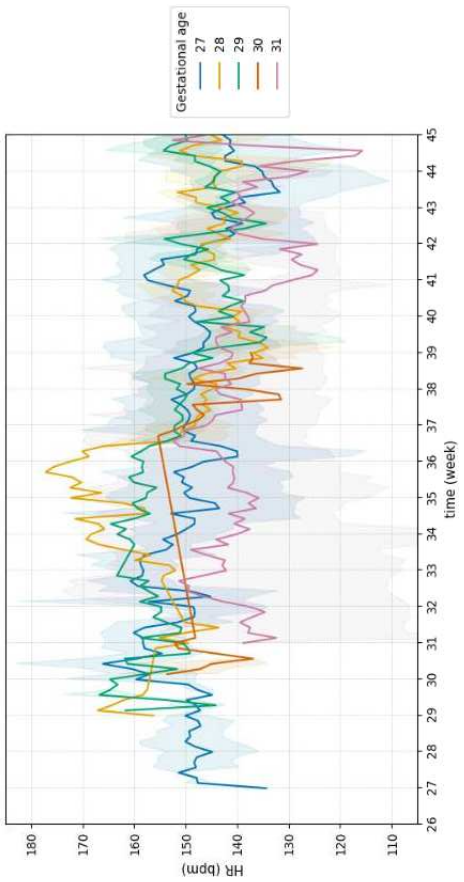
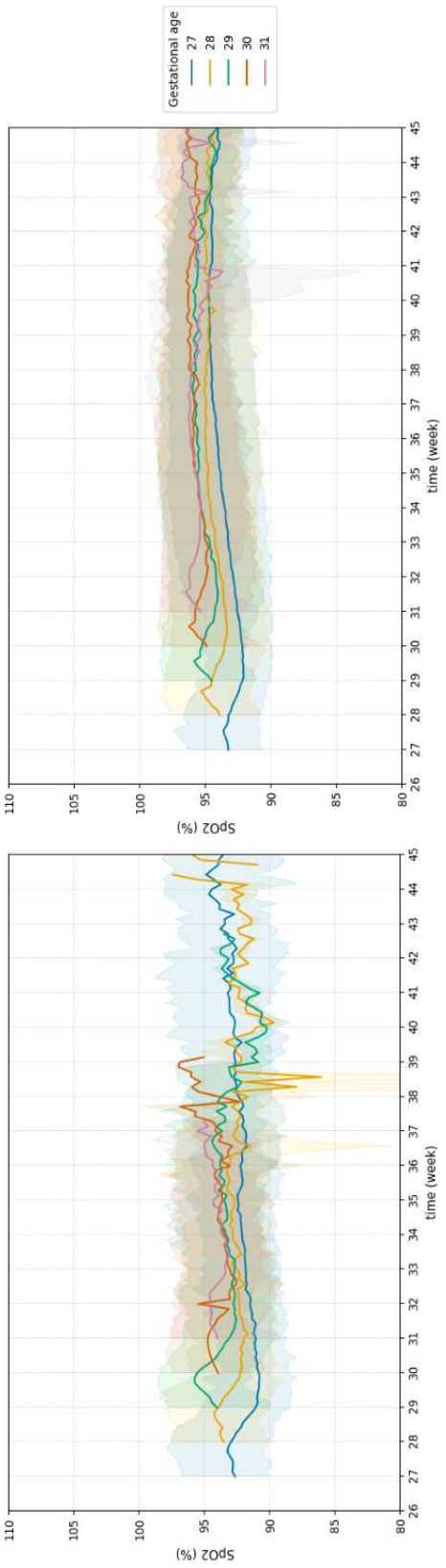


Figure B.1: Heart rate (HR) trajectories stratified by gestational age across three hospitals. From top to bottom, the plots correspond to CHUV, USZ, and KiSpi


Gestational age
27
28
29
30
31
SpO2 (%)
time (week)
110
105
100
95
90
85
80
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45

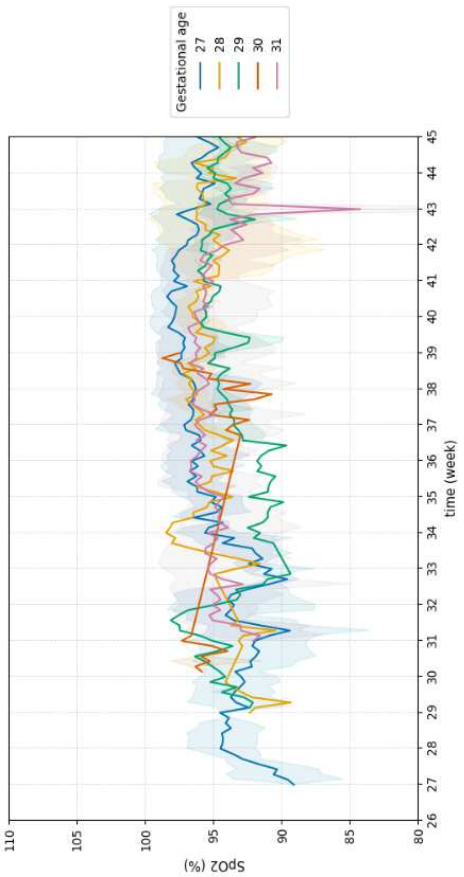
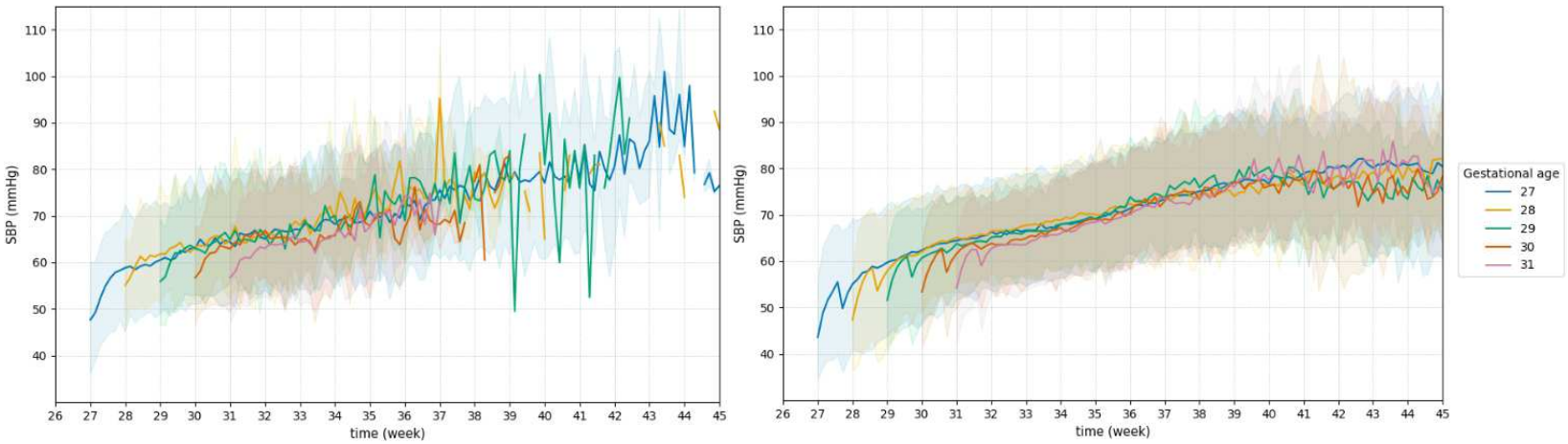


Figure B.2: Oxygen saturation (SpO_2) trajectories stratified by gestational age across three hospitals. From top to bottom, the plots correspond to CHUV, USZ, and KiSpi


SBP (mmHg)
time (week)
Gestational age
27
28
29
30
31

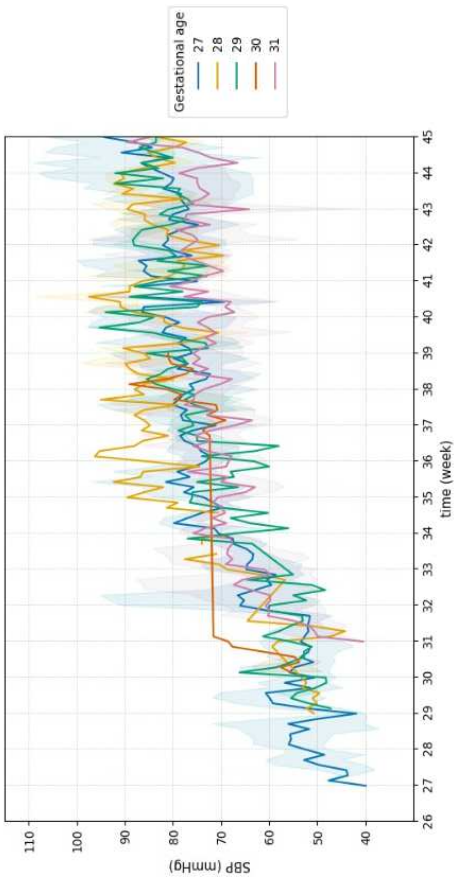


Figure B.3: Systolic blood pressure (SBP) trajectories stratified by gestational age across three hospitals. From top to bottom, the plots correspond to CHUV, USZ, and KiSpi

Gestational age
27
28
29
30
31
DBP (mmHg)
time (week)
DBP (mmHg)
time (week)

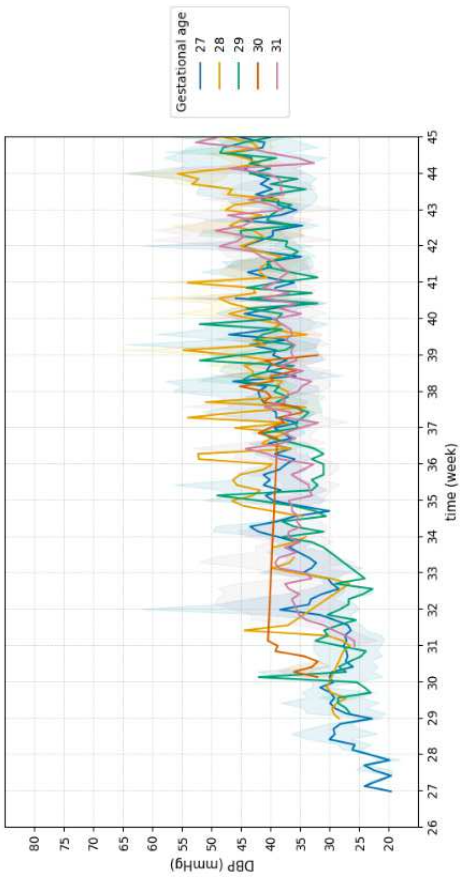
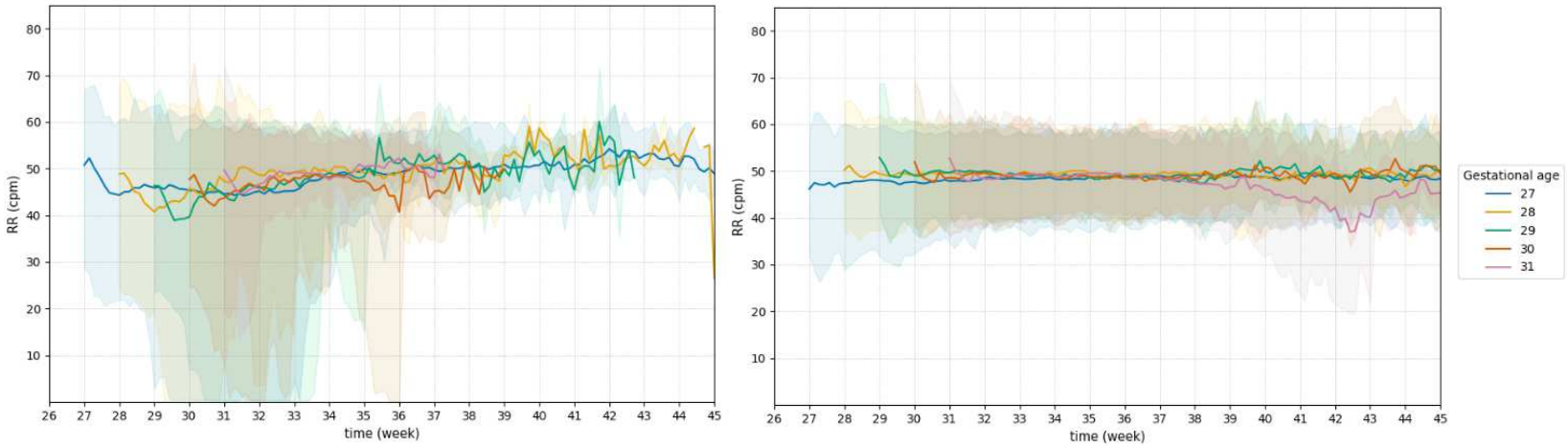


Figure B.4: Diastolic blood pressure (DBP) trajectories stratified by gestational age across three hospitals. From top to bottom, the plots correspond to CHUV, USZ, and KiSpi


RR (cpm)
time (week)
RR (cpm)
time (week)
Gestational age
27
28
29
30
31

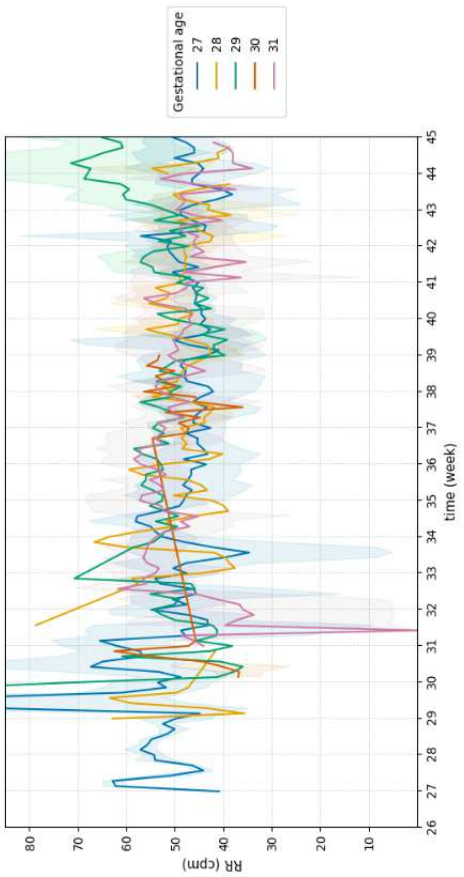


Figure B.5: Respiratory rate (RR) trajectories stratified by gestational age across three hospitals. From top to bottom, the plots correspond to CHUV, USZ, and KiSpi

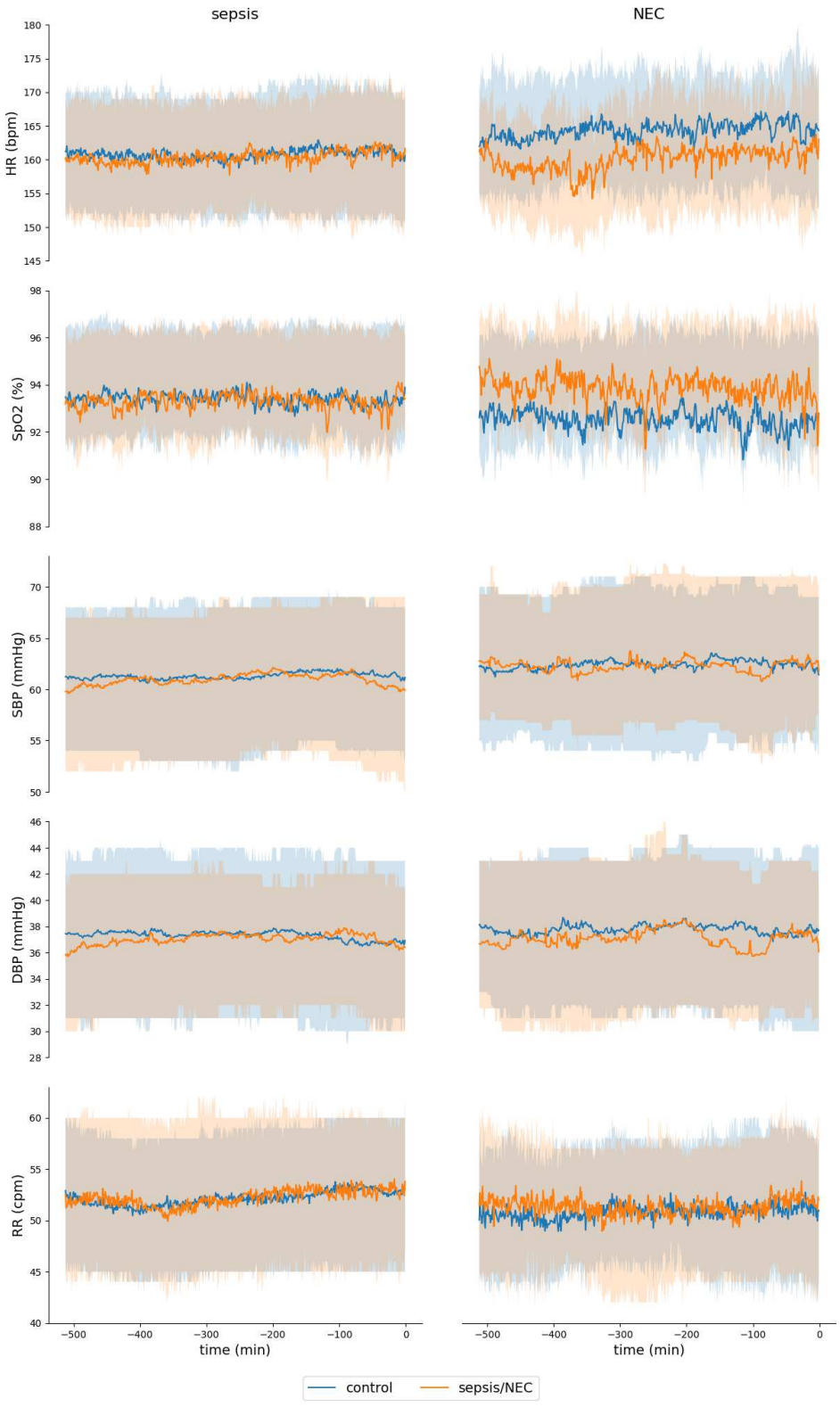


Figure B.6: Comparison of vital signals between sepsis and NEC. The left column shows sepsis data, while the right column shows NEC data, both grouped by class.

Appendix C

Medical datasets used for FORMED pretraining

Here we provide the details of the datasets used as the medical time series cohort for FORMED repurposing. The datasets are publicly available, and we follow the pre-processing and splitting procedures as in [84].

APAVA The Alzheimer's Patients' Relatives Association of Valladolid (APAVA) dataset, is a public EEG time series dataset with 2 classes and 23 subjects, including 12 Alzheimer's disease patients and 11 healthy control subjects. On average, each subject has 30.0 ± 12.5 trials, with each trial being a 5-second time sequence consisting of 1280 timestamps across 16 channels.

TDBrain The TDBrain dataset, is a large permission-accessible EEG time series dataset recording brain activities of 1274 subjects with 33 channels. Each subject has two trials: one under eye open and one under eye closed setup. The dataset includes a total of 60 labels, with each subject potentially having multiple labels indicating multiple diseases simultaneously.

ADFTD The Alzheimer's Disease and FronTotemporal Dementia (ADFTD) dataset, is a public EEG time series dataset with 3 classes, including 36 Alzheimer's disease (AD) patients, 23 Frontotemporal Dementia (FTD) patients, and 29 healthy control (HC) subjects. The dataset has 19 channels, and the raw sampling rate is 500Hz. Each subject has a trial, with trial durations of approximately 13.5 minutes for AD subjects (min=5.1, max=21.3), 12 minutes for FD subjects (min=7.9, max=16.9), and 13.8 minutes for HC subjects (min=12.5, max=16.5).

PTB The PTB dataset, is a public ECG time series recording from 290 subjects, with 15 channels and a total of 8 labels representing 7 heart diseases and 1 health control. The

raw sampling rate is 1000Hz. For this paper, we utilize a subset of 198 subjects, including patients with Myocardial infarction and healthy control subjects.

PTB-XL The PTB-XL dataset, is a large public ECG time series dataset recorded from 18,869 subjects, with 12 channels and 5 labels representing 4 heart diseases and 1 healthy control category. Each subject may have one or more trials. The raw trials consist of 10-second time intervals, with sampling frequencies of 100Hz and 500Hz versions.

Appendix D

Illustrative examples for TimesFM

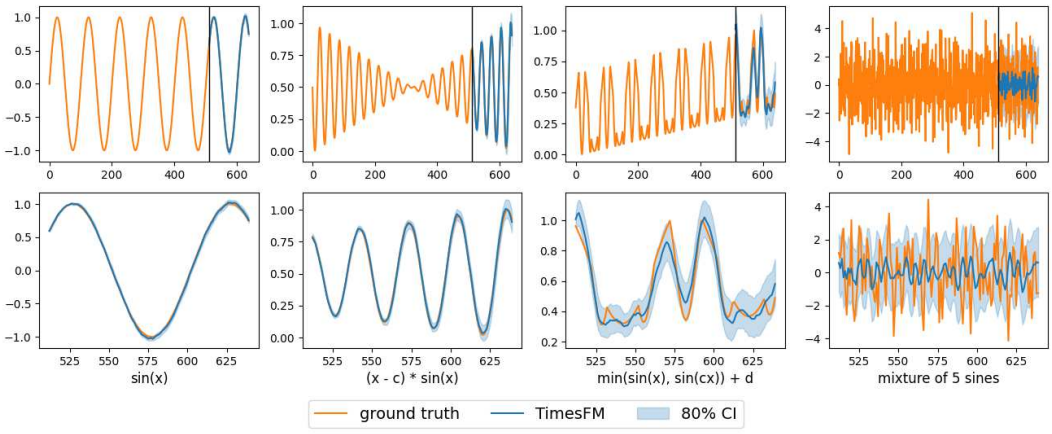


Figure D.1: Forecasts visualized on synthetic curves. The bottom row plots zoom in on the prediction horizon for the sake of clarity.

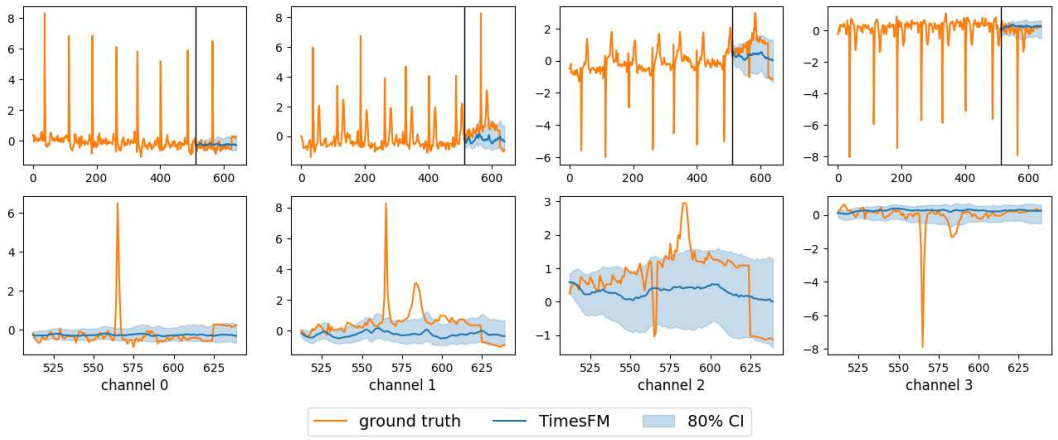
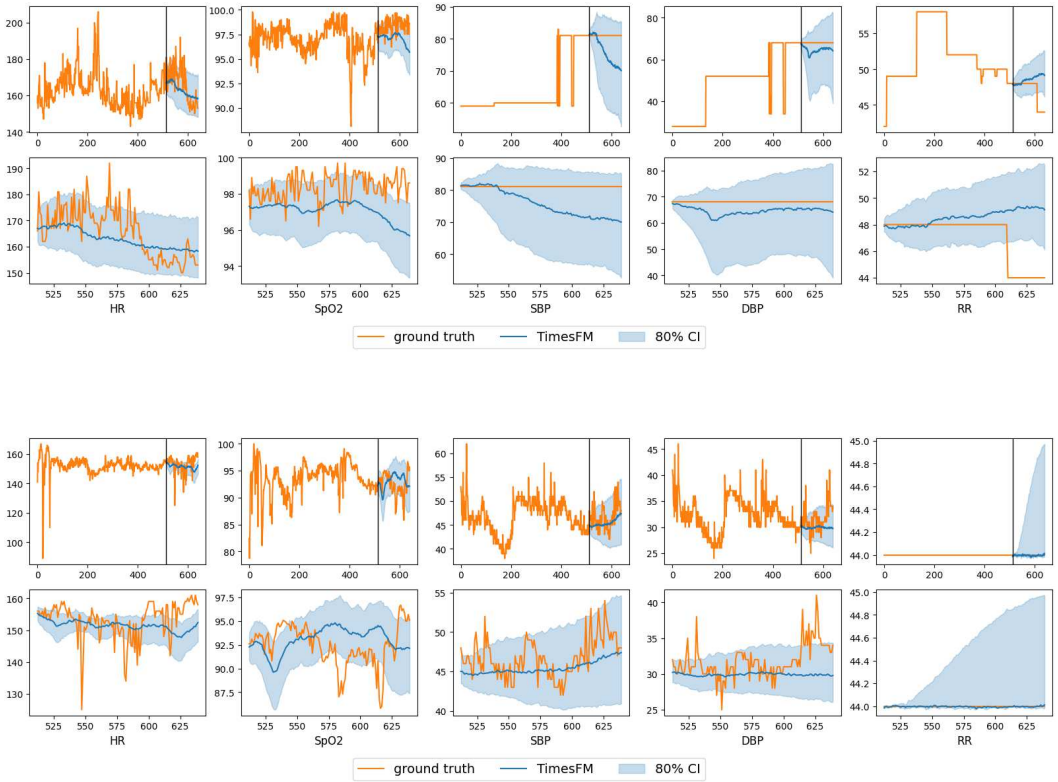


Figure D.2: Forecasts visualized on PTB-XL dataset. The bottom row plots zoom in on the prediction horizon for the sake of clarity.



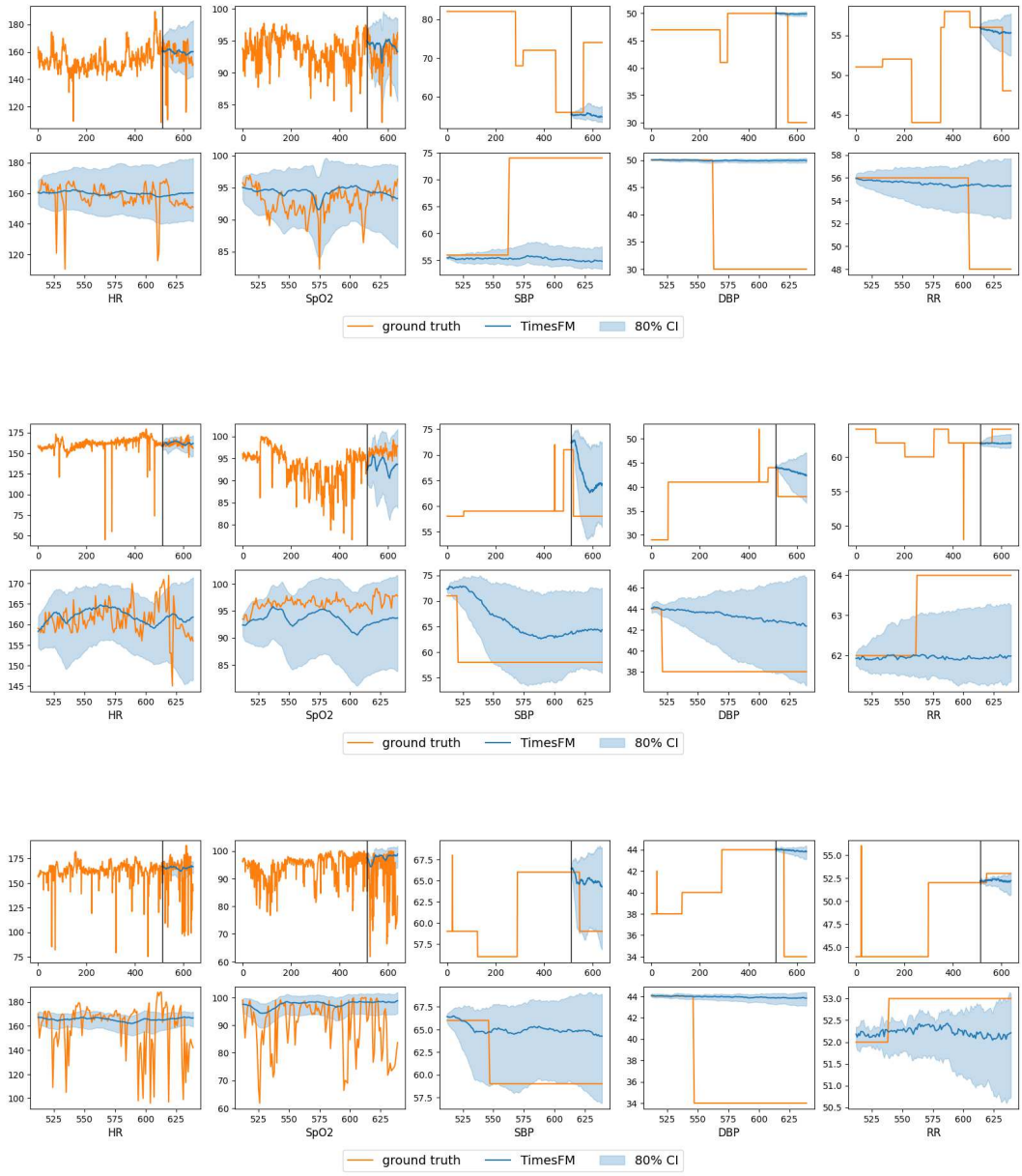


Figure D.3: Forecasts visualized on vital signals from 5 randomly drawn patients. The bottom row plots zoom in on the prediction horizon for the sake of clarity.