Titre: Title:	Advancing Acceleration and Deceleration Detection in Fetal Cardiotocography through Various Deep Learning Architectures
Auteur: Author:	Pingao Wang
Date:	2025
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Wang, P. (2025). Advancing Acceleration and Deceleration Detection in Fetal Cardiotocography through Various Deep Learning Architectures [Thèse de doctorat, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/72085/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/72085/
Directeurs de recherche: Advisors:	Michel C. Desmarais
Programme: Program:	Génie informatique

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Advancing Acceleration and Deceleration Detection in Fetal Cardiotocography through Various Deep Learning Architectures

PINGAO WANG
Département de génie informatique et génie logiciel

Thèse présentée en vue de l'obtention du diplôme de *Philosophiæ Doctor*
Génie informatique

Décembre 2025

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Cette thèse intitulée :

Advancing Acceleration and Deceleration Detection in Fetal Cardiotocography through Various Deep Learning Architectures

présentée par **Pingao WANG**
en vue de l'obtention du diplôme de *Philosophiæ Doctor*
a été dûment acceptée par le jury d'examen constitué de :

Foutse KHOMH, président
Michel DESMARAIS, membre et directeur de recherche
Maxime LAMOTHE, membre
Roger NKAMBOU, membre externe

DEDICATION

To my parents and my supervisor,
for showing me the value of curiosity, persistence, and kindness.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my supervisor, Prof. Michel Desmarais, for his continuous guidance, encouragement, and patience throughout my Ph.D. journey. His scientific insight, rigor, and thoughtful mentorship have been invaluable in shaping my research direction and academic growth.

I would also like to thank Dr. Tian Li and her clinical collaborators for providing the fetal cardiotocography datasets and sharing their medical expertise, which made this research both meaningful and clinically relevant. My sincere thanks go to all the members of our research group at Polytechnique Montréal for their technical support, inspiring discussions, and friendship.

I would like to express my appreciation to Polytechnique Montréal for providing an inspiring academic environment and research culture that greatly supported the completion of this doctoral work.

Finally, I would like to thank my family for their unconditional love, patience, and understanding. Their support has been my source of strength and perseverance throughout this doctoral journey.

RÉSUMÉ

Le monitorage continu du bien-être fœtal pendant le travail demeure l'une des tâches les plus critiques et les plus exigeantes en ressources en obstétrique. La cardiotocographie (CTG), qui enregistre simultanément la fréquence cardiaque fœtale (FHR) et les contractions utérines (UC), constitue la technique standard utilisée à l'échelle mondiale pour évaluer en temps réel l'état du fœtus. Malgré son utilisation généralisée, l'interprétation de la CTG demeure fortement subjective et dépend de l'expérience des cliniciens, qui doivent identifier les épisodes d'accélération et de décélération susceptibles d'indiquer une détresse ou une récupération fœtale. La variabilité interhumaine, l'incohérence des pratiques d'annotation et la disponibilité limitée de spécialistes peuvent ainsi conduire à des décisions tardives ou sous-optimales, en particulier dans les milieux à ressources limitées. Ces défis motivent le développement de systèmes computationnels robustes, interprétables et fondés sur les données, capables d'aider les cliniciens à identifier avec précision et constance les événements critiques.

Cette thèse aborde la problématique de la détection automatique des événements d'accélération / décélération dans la CTG à travers trois études complémentaires formant un cadre méthodologique cohérent. Chaque étude explore une direction d'innovation distincte mais reliée, allant du raffinement post-traitement à la fusion multimodale et à l'apprentissage spectral-temporel, dans le but d'améliorer la précision, la généralisation et l'interprétabilité clinique des systèmes d'analyse automatique.

L'étude 1 présente un réseau de post-traitement fondé sur l'apprentissage résiduel, nommé Event Refinement via Neural Processing (ERNP). Contrairement aux systèmes conventionnels reposant sur des règles heuristiques fixes, ERNP apprend à affiner les masques d'événements générés par les modèles de détection existants au moyen d'un chemin résiduel différentiable. Il intègre des connaissances cliniques dans une architecture neuronale capable de correction contextuelle et de lissage temporel. Ce module différentiable, conçu de manière modulaire, peut être intégré à diverses architectures sans nécessiter le réentraînement de l'ensemble du pipeline. Les expériences menées sur des ensembles de données publics et privés montrent qu'ERNP améliore significativement les scores F1 au niveau événementiel, produisant des frontières plus cohérentes et un alignement plus proche des annotations expertes. Cette approche comble ainsi l'écart entre les règles cliniques non différentiables et l'apprentissage profond de bout en bout en introduisant un mécanisme d'affinement apprenable favorisant la précision temporelle.

L'étude 2 étend l'analyse du signal FHR seul vers un cadre multimodal en intégrant les

signaux obstétriques hétérogènes de la CTG, incluant FHR, UC et les mouvements fœtaux (FM). Quatre stratégies de fusion sont examinées, à savoir la fusion à l'entrée, la fusion précoce, la fusion intermédiaire et la fusion tardive, chacune visant à capturer les dépendances inter-signaux et les interactions temporelles. Un schéma d'entrée multiscalaire est proposé afin de modéliser conjointement les fluctuations à court terme et les tendances physiologiques à long terme, améliorant ainsi la sensibilité aux événements et la robustesse au bruit. Les résultats démontrent que la fusion multimodale améliore systématiquement la précision de classification des segments d'accélération, de décélération et de fond par rapport aux approches monocanal. Cette étude révèle que la combinaison de signaux physiologiques hétérogènes aide le modèle à lever les ambiguïtés des motifs FHR grâce aux informations contextuelles issues de l'activité utérine et des mouvements fœtaux.

L'étude 3 propose une architecture bi-flux nommée FNO-AugUNet, qui combine un UNet unidimensionnel avec un opérateur neuronal de Fourier (FNO). La branche UNet capture la morphologie temporelle locale, tandis que la branche FNO réalise un raisonnement spectral permettant de modéliser les dépendances à longue portée directement à partir des séquences FHR brutes. Un mécanisme de fusion léger aligne et agrège les caractéristiques des deux flux, permettant un apprentissage conjoint des détails temporels et du contexte fréquentiel global. Contrairement aux approches existantes de fusion temps-fréquence reposant sur des représentations spectrogrammes explicites, le réseau proposé effectue un raisonnement fréquentiel implicite au moyen de transformations de Fourier apprenables. Les résultats expérimentaux montrent que cette architecture bi-flux améliore les scores F1 au niveau événementiel d'environ dix points de pourcentage par rapport aux modèles UNet monocanal et permet une réduction d'erreur de huit pour cent sur l'ensemble public CULF-DB. En particulier, le modèle obtient une réduction d'erreur de vingt-deux pour cent pour la détection des décélérations, ce qui confirme l'importance du modèle fréquentiel global pour la dynamique temporelle complexe des signaux FHR.

Ensemble, ces trois études mettent en évidence deux axes complémentaires pour une détection robuste des événements A/D, à savoir le raffinement post-traitement et l'innovation architecturale du front-end. Le cadre proposé offre une feuille de route systématique pour combiner des mécanismes d'affinement différentiables, une intégration multimodale et une modélisation spectral-temporelle au sein d'un paradigme d'apprentissage unifié. Il démontre que la modélisation précise, interprétable et généralisable de la dynamique FHR nécessite à la fois une sensibilité morphologique locale et une conscience contextuelle globale.

Au-delà du monitorage fœtal, les contributions de cette thèse offrent des pistes généralisables pour la modélisation des séries temporelles médicales. Le paradigme de post-traitement

différentiable illustré par ERNP peut combler l'écart entre les règles cliniques heuristiques et l'inférence neuronale pour d'autres biosignaux tels que l'ECG ou l'EEG. De même, les stratégies multicanales et multiescales proposées constituent une base solide pour l'intégration de modalités physiologiques diverses dans des applications de santé comme la stadification du sommeil, la détection de crises épileptiques ou le monitorage en soins intensifs. Enfin, le cadre spectral-temporel bi-flux fondé sur les opérateurs neuronaux de Fourier offre une approche générale pour capturer efficacement les dépendances globales dans les données biomédicales séquentielles.

Dans une perspective plus large, cette thèse s'inscrit dans la philosophie de l'intelligence artificielle au service du bien commun, telle qu'endossée par les Nations Unies. En permettant une analyse automatisée, équitable et interprétable des données obstétricales, les méthodes proposées contribuent à améliorer les résultats maternels et néonataux, en particulier dans les régions disposant de ressources médicales limitées. En définitive, ce travail montre comment les innovations en apprentissage profond peuvent constituer non seulement des avancées académiques, mais également des outils significatifs permettant de relier le développement algorithmique, la pratique clinique et l'équité en santé à l'échelle mondiale.

ABSTRACT

Continuous monitoring of fetal well-being during labor remains one of the most critical and resource-intensive tasks in obstetrics. Cardiotocography (CTG), the simultaneous recording of fetal heart rate (FHR) and uterine contraction (UC) signals, is the standard technique used worldwide to assess fetal condition in real time. Despite its ubiquity, CTG interpretation is highly subjective and depends on the experience of clinicians, who must identify acceleration and deceleration (A/D) patterns that may indicate fetal distress or recovery. Human variability, inconsistent annotation practices, and limited availability of specialists often lead to delayed or suboptimal decisions, especially in resource-limited settings. These challenges motivate the development of robust, interpretable, and data-driven computational systems that can assist clinicians in identifying critical fetal events accurately and consistently.

This thesis addresses the challenge of automatic A/D event detection in fetal CTG through three complementary studies that together form a coherent methodological and conceptual framework. Each study explores a distinct yet related direction of innovation, ranging from post-processing refinement to multimodal fusion and spectral-temporal representation learning, with the goal of improving accuracy, generalization, and clinical interpretability of automated analysis systems.

Study 1 introduces a residual-learning-based post-processing network named Event Refinement via Neural Processing (ERNP). Unlike conventional systems that rely on fixed heuristic criteria, ERNP learns to refine coarse event masks generated by existing detection models through a differentiable residual pathway. It integrates clinical priors into a neural architecture capable of context-aware correction and temporal smoothing. The network operates as a modular differentiable component and can be integrated into existing architecture without the need to retrain the entire pipeline. Experiments on both public and private datasets show that ERNP substantially improves event-level F1-scores, yielding smoother boundaries and higher consistency with expert annotations. This approach bridges the gap between non-differentiable clinical post-processing and end-to-end deep learning by introducing a learnable refinement mechanism that enhances temporal precision.

Study 2 extends the analysis from single-signal FHR modeling to a multimodal setting by integrating heterogeneous obstetric signals of CTG, including FHR, UC, and fetal movement (FM). Four fusion strategies are investigated, namely input-fusion, early-fusion, intermediate-fusion and late-fusion, each designed to capture cross-signal dependencies and temporal interactions. A unified multiscale input scheme is proposed to jointly model short-term waveform

fluctuations and long-term physiological trends, thereby enhancing both event sensitivity and robustness to noise. Results demonstrate that multimodal fusion consistently improves the classification accuracy of acceleration, deceleration, and background segments compared with single-channel baselines. This study reveals that combining heterogeneous physiological signals allows the model to disambiguate ambiguous FHR patterns by leveraging contextual uterine and fetal information.

Study 3 proposes a dual-stream architecture named FNO-AugUNet, which couples a one-dimensional UNet with a Fourier Neural Operator (FNO) stream. The UNet branch captures local temporal morphology, while the FNO branch performs spectral-domain reasoning to learn long-range dependencies directly from raw FHR sequences. A lightweight fusion mechanism aligns and aggregates features from both streams, enabling joint learning of temporal detail and global frequency context. Unlike existing time–frequency fusion methods that rely on explicit spectrogram representations, the proposed network performs implicit frequency modeling through learnable Fourier transformations. Empirical results show that this dual-stream design improves event-level F1 by approximately ten percentage points compared with single-stream UNet models and achieves an eight percent error reduction rate over existing baselines on the public CULF-DB dataset. In particular, the model achieves a twenty-two percent error reduction rate in deceleration detection, confirming the value of global spectral modeling for complex temporal dynamics in FHR signals.

Together, these three studies highlight two complementary directions for robust A/D detection: post-processing refinement and front-end architectural innovation. The proposed framework provides a systematic roadmap for combining differentiable refinement mechanisms, multimodal feature integration, and spectral-temporal reasoning within a unified learning paradigm. It demonstrates that accurate, interpretable, and generalizable modeling of FHR dynamics requires both local morphological sensitivity and global contextual awareness.

Beyond fetal monitoring, the contributions of this thesis provide generalizable insights for medical time-series modeling. The differentiable post-processing paradigm exemplified by ERNP can bridge heuristic clinical rules and neural inference for other biosignals such as ECG or EEG, where rule-based interpretation remains dominant. Similarly, the multimodal and multiscale fusion strategies offer a foundation for integrating diverse physiological modalities in healthcare applications such as sleep staging, seizure detection, and intensive care monitoring. Finally, the dual-stream spectral-temporal modeling framework based on Fourier Neural Operators establishes a general approach for efficiently capturing global dependencies in sequential biomedical data.

From a broader perspective, this thesis aligns with the philosophy of Artificial Intelligence for Good as endorsed by the United Nations. By enabling automated, equitable, and interpretable analysis of obstetric data, the proposed methods contribute to improving maternal and neonatal health outcomes, particularly in regions with limited medical expertise. Ultimately, this work demonstrates how innovations in deep learning can serve not only as academic advances but also as meaningful tools for social good, bridging algorithmic development, clinical practice, and global health equity.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ACRONYMS

A/D	Acceleration and Deceleration events in Cardiotocography
AUC (AUROC)	Area Under the Receiver Operating Characteristic curve
CNN	Convolutional Neural Network
CTG	Cardiotocography
CULF-DB	Public CTG dataset
CTU-UHB	Public CTG dataset, from CTU-UHB Intrapartum Cardiotocography Database
ERR	Error Reduction Rate (relative error decrease vs. a baseline)
FHR	Fetal Heart Rate
FFT	Fast Fourier Transform
FNO	Fourier Neural Operator
FNO-AugUNet	UNet augmented with a Fourier Neural Operator stream (dual-stream)
FN	False Negative
FP	False Positive
LSMP	Long- and Short-term Median-based Post-processing
PPV	Positive Predictive Value (Precision)
ROC	Receiver Operating Characteristic
SEN	Sensitivity (Recall)
SOTA	State of the Art
TN	True Negative
TP	True Positive
UC	Uterine Contraction
UNet	U-shaped Convolutional Neural Network architecture
7AH-SYSU	Private institutional dataset (rename if your official name differs)

CHAPTER 1 INTRODUCTION

1.1 Background and Motivation

Childbirth is one of the most critical and resource-intensive aspects of healthcare [1]. In many countries, especially in developed regions, the costs associated with maternal delivery are significantly high due to the need for continuous monitoring, risk management, and specialized medical personnel [2,3]. For patients, this often translates into high insurance premiums and out-of-pocket expenses. For healthcare systems, it presents a constant challenge in balancing safety, cost-efficiency, and availability of expert care.

Obstetricians and gynecologists (OB/GYNs) operate under immense clinical pressure. They are required to interpret complex physiological signals such as fetal heart rate (FHR) and uterine contractions (UC)—in real time, make rapid decisions, and ensure both maternal and fetal well-being [4]. The shortage of specialists in many regions, including remote or underserved areas, further amplifies this burden, sometimes leading to delayed or suboptimal care [5].

In this context, Artificial Intelligence (AI) and deep learning (DL) technologies offer a promising solution [6–8]. By automating parts of the interpretation process and assisting clinicians in identifying high-risk patterns, AI technologies can enhance the accuracy, efficiency, and accessibility of obstetric care [9]. In particular, AI-driven analysis of cardiotocography (CTG) signals has the potential to support early detection of abnormal fetal conditions, reduce human variability in interpretation, and relieve part of the cognitive workload from OB/GYN practitioners [10]. Ultimately, such innovations may contribute to improved clinical outcomes, more equitable maternal care, and reduced systemic costs.

1.1.1 Fetal-CTG analysis & interpretation

Fetal continuous cardiotocography (Fetal-CTG) is a widely used technique for monitoring fetal heart rate, uterine contractions and fetal movement (FM) during pregnancy and labor [11].

However, the interpretation of Fetal-CTG can be challenging due to variability in equipment vendors and the experience of midwives and obstetricians. To standardize Fetal-CTG interpretation, national scientific organizations, such as the National Institute of Child Health and Human Development (NICHD) and the American College of Obstetricians and Gynecologists (ACOG), have proposed standardized guidelines [12, 13]. Despite this, high inter-

and intra-observer variability in Fetal-CTG interpretation remains a challenge [14].

1.1.2 Acceleration / Deceleration detection in Fetal-CTG

Of the five Fetal-CTG medical features (Baseline, Variability, Acceleration, Deceleration, and Sinusoidal Pattern), deceleration is the most predictive of acidemia and significant morbidity risk when combined with tachycardia [15]. Failure to recognize acceleration and deceleration (A/D) can lead to serious consequences, including fetal death, while over-interpretation of A/D may result in unnecessary caesarean sections.

Despite being a key reference factor, A/D detection is currently under-studied. Several reports [16–19] have shown the unsatisfactory performance of existing automated diagnosis algorithms for A/D detection.

In recent years, deep learning techniques have been increasingly applied due to their remarkable performance in various pattern recognition tasks [20]. Among them, some related works [21–23] have proposed deep learning solutions proposed to recognize the patterns in A/D. However, most of them focus on predicting fetal postpartum diagnosis directly from Fetal-CTG signals without recognizing the Fetal-CTG medical features. From a clinical perspective, these rough end-to-end approaches lack maturity, explainability, and interpretability.

1.2 Research Objectives

This study aims to improve the automatic detection accuracy and clinical applicability of acceleration/deceleration events in fetal heart rate monitoring and to address the problems of weak generalization and sensitivity to label bias in current deep learning models by introducing a guidable post-processing mechanism and a multimodal fusion strategy. Could also mention providing better replicability and benchmarks by providing open source code.

1.3 Overview of Proposed Studies

This paper focuses on acceleration/deceleration event detection and relies on three main research directions to systematically improve the performance and practicality of the automatic analysis system.

There are three main studies proposed in this paper as follows.

- **Study 1:** Develop a residual-learning-based post-processing method named Event Re-

finement via Neural Processing (ERNP) to improve and refine the traditional rule-based heuristic used in most of the existing related works;

- **Study 2:** Explore a multimodal network architecture with four types of fusion methods that integrates FHR, UC, and FM signals with multiscale input strategies to improve the recognition capability of A/D events;
- **Study 3:** Propose a dual-stream architecture named *FNO-AugUNet*, in which a 1D-UNet captures local temporal morphology while a Fourier Neural Operator (FNO) stream models global dependencies directly from raw FHR signals.

Study 1 focuses on the post-processing module and is designed with strong adaptability and modular plug-and-play functionality. Rather than replacing any module from existing deep learning solutions, it refines their outputs through a residual-learning-based strategy. This modular design ensures that the proposed refinement method can be seamlessly integrated with a wide range of existing architectures without retraining the entire pipeline.

In contrast, Study 2 targets the structural innovation of the front-end backbone network. By explicitly modeling the effective integration of multiple input signals, including FHR, UC, and FM, it addresses the limitations of single-signal approaches and leverages the complementary physiological information provided by multimodal monitoring.

In addition to spatial and multimodal integration, Study 3 explores the complementary relationship between local temporal modeling and global contextual modeling in fetal heart-rate (FHR) analysis. We propose a dual-stream architecture, termed *FNO-AugUNet*, in which a 1D-UNet stream captures fine-grained temporal morphology while an FNO stream performs spectral-domain transformations on the same FHR signal to model long-range dependencies. A lightweight fusion mechanism aligns and aggregates the logits from both streams, enabling joint learning of local waveform transitions and global temporal context. Unlike our previous time–frequency fusion design that relied on explicit spectrogram inputs, the proposed method performs implicit frequency-domain reasoning through learnable Fourier transformations within the network. This design substantially improves event-level detection accuracy and robustness compared with single-stream and spectrogram-based baselines, highlighting the effectiveness of implicit spectral modeling for physiological signal analysis.

Publication Status. The work corresponding to Study 1 has been submitted to the *IEEE Journal of Biomedical and Health Informatics (JBHI)* and is currently under peer review. The work corresponding to Study 2 has been submitted to *Biomedical Signal Processing and Control (BSPC)* and is under review. Study 3 is being prepared for journal submission.

Task definitions across studies. Although all three studies in this dissertation share the common goal of identifying acceleration and deceleration patterns from FHR signals, they differ in the granularity of prediction.

Temporal segmentation task. Study 1 and Study 3 are formulated as temporal segmentation tasks, where the model predicts a sample-wise segmentation mask at each time step , thereby localizing the onset and offset of individual A/D events.

Sequence-level classification task. In contrast, Study 2 is designed as a sequence-level classification task, in which an entire FHR segment is assigned to a single class label representing its dominant morphological pattern. Unlike segmentation-based formulations, this task does not involve temporal localization of individual A/D events. Conceptually, segmentation and classification are closely related learning paradigms; however, in this dissertation, the term segmentation is strictly used to denote sample-wise temporal localization, while classification refers to decision-making at the sequence or event level.

The difference in task formulation across studies primarily reflects the distinct research motivations behind each. Study 2, which focuses on multimodal fusion, aims to investigate how multiple physiological signals (FHR, UC, and FM) can be effectively integrated within a CNN-based framework. By formulating this problem as a direct sequence-level classification task, the design isolates the core representational capacity of CNNs for multimodal integration, without the confounding effects of fine-grained temporal localization algorithms. In contrast, Study 1 and Study 3 emphasize precise event localization and temporal segmentation of accelerations and decelerations, targeting improved clinical interpretability and generalization across recording conditions. These differing perspectives, one emphasizing methodological exploration and the other focusing on temporal precision, naturally motivate the adoption of different task types across the three studies.

Evaluation overview. We evaluated the performance of the proposed methods on public and private datasets and systematically compared them with existing state-of-the-art methods. They were conducted on CULF-DB, CTU-UHB, and our private large scale clinical datasets. A unified post-processing process and evaluation metrics were used to ensure consistency and fairness in comparisons.

Reproducibility and transparency. In addition, this thesis emphasizes the importance of reproducibility and transparency, which are often lacking in prior works. All codes, including data pre-processing, baseline implementations, and proposed methods, are released in open-source repositories. This enables fellow researchers to faithfully reproduce our results,

conduct fair comparisons, and build upon a unified benchmarking framework. Beyond the specific task of fetal heart rate acceleration/deceleration detection, we also aim to establish a generalizable paradigm for medical time-series processing. By combining modular post-processing refinement and multi-modal multi-scale modeling, the proposed framework is not tied to a particular dataset or clinical signal, but can be adapted to diverse biomedical monitoring contexts such as ECG or EEG. We believe this dual commitment to reproducibility and generalization not only strengthens the scientific value of our work but also contributes to the broader research community by fostering rigorous evaluation standards and facilitating the translation of AI methods into clinical practice.

1.4 Contributions

- Proposed a new residual-learning-inspired post-processing method, significantly improving accuracy;
- Tested a fusion strategy for three CTG signals (FHR, UC and FM), and incorporated a multi-scale fusion strategy to form a new deep learning network architecture for A/D detection;
- Introduced a dual-stream time–frequency fusion framework based on Fourier Neural Operators (FNO) that achieves state-of-the-art performance in acceleration/deceleration event detection by integrating spectral and temporal features within a unified architecture.
- Reproduced state-of-the-art performance, conducted comparisons, and established a benchmark;
- Open-sourced the codes, which facilitates reproducibility of experimental results for the research community and promotes fair benchmarking and transparent comparisons across future studies;

1.5 Thesis Outlines

The organization of the thesis is as follows:

- Chapter 2 provides a comprehensive literature review. We summarize the evolution of automated FHR analysis from rule-based approaches to deep learning methods, highlight state-of-the-art techniques, and identify open challenges that motivate the research questions of this thesis.

- Chapter 3 introduces the datasets used in this work, including both public benchmarks and a large-scale private clinical dataset. We describe their characteristics, annotation procedures, and pre-processing strategies, which form the foundation for developing and evaluating the proposed methods.
- Chapter 4 presents the first study, Event Refinement via Neural Processing (ERNP). This residual-inspired post-processing framework is designed to enhance segmentation-based A/D detection pipelines. We detail its methodology, experimental evaluation, and the insights gained from comparisons with existing post-processing strategies.
- Chapter 5 presents the second study, a multi-modal multi-scale deep convolutional neural network. This chapter explores the integration of multiple physiological signals and multi-scale input strategies, demonstrating how structural innovations can further improve the accuracy and robustness of A/D detection.
- Chapter 6 presents the third study, a dual-stream time–frequency fusion network that introduces Fourier Neural Operators to model frequency-domain context and combines them with temporal UNet representations. Extensive experiments demonstrate that this architecture significantly enhances event-level accuracy.
- Chapter 7 concludes the thesis by summarizing the key findings, reflecting on their broader implications, and discussing both limitations and avenues for future research. We also consider the philosophical and social perspective of "AI for Good" in the context of maternal healthcare.

CHAPTER 2 LITERATURE REVIEW

2.1 Fetal-CTG automatic analysis methods

Fetal-CTG automatic analysis methods (AAMs) have been studied for decades since the FIGO guideline introduced in 1986. Most early works were aligned with clinical guideline, which means baseline computation and A/D detection were essential for developing any AAMs system [24].

For rule-based algorithms, 11 methods are re-implemented and evaluated in a study [16], but the accuracies of these methods were not satisfactory. For AAMs related systems, studies [18, 25] presented comprehensive comparison results for five existing systems, which showed that these systems cannot improve decision-making results for clinicians. Even worse, the rate of miscarriages had risen due to the high false positive rate [26, 27].

CTU-UHB [24, 28] is the earliest related open source dataset, which uses the pH, pCO2, pO2, base excess and computed BDecf of umbilical artery blood sample as criteria of fetal postpartum diagnosis. In this dataset, the information of occurrences of acceleration and deceleration is annotated. After CTU-UHB was introduced, most research has been based on it and/or similar-structured private datasets. For examples, in a study [22] a rule-based and Deep Gaussian mixture model was employed, and in another study [29], a hybrid model based on One Dimension Convolutional Neural Network (1D-CNN) and bidirectional Gate Recurrent Unit (GRU) achieved 95.15% accuracy to predict whether the fetus had hypoxia. A study [30] stated that no system based on machine learning or deep learning has yet been evaluated in prospective trials.

2.2 Applications of Deep Learning to FHR and UC

With the advent of deep-learning techniques, end-to-end diagnosis related research became the main stream, focusing on predicting postpartum outcomes from raw Fetal-CTG data and some clinical information (include women's age, week of gestation, known fetal diseases, etc).

Several recent studies have explored the direct prediction of fetal hypoxia from FHR signals using machine learning models. As summarized in a comprehensive review by Alharbi et al. [31], both classical machine learning algorithms (e.g., SVM, Random Forest) and deep learning architectures (e.g., CNNs, RNNs) have been applied to classify entire FHR recordings into hypoxic or non-hypoxic outcomes. While these methods show promise in risk stratifica-

tion, especially using features such as baseline variability or frequency-domain entropy, they typically treat the FHR signal as a single undifferentiated input and output only a binary or scalar classification.

However, such approaches lack temporal granularity and clinical interpretability. In practice, clinicians need not just a final hypoxia risk score but actionable insights about when and where within the FHR recording suspicious patterns—such as decelerations or prolonged abnormal segments—occur. Without localizing such events, direct hypoxia classification models fall short in supporting real-time obstetric decision-making. This gap motivates our focus on event-level A/D segmentation, which provides both interpretable output and the potential for early warning of pathological trends.

2.3 Acceleration / Deceleration Detection: From Rule-based to Deep Learning

One of the most representative studies of early explorations in paper [32] presented a hand-crafted method to classify if a segment of Fetal-CTG is deceleration. However, the method required experts to manually calculate four values, including variability, duration, depth and area of the segmented Fetal-CTG, as the input attributes to fed their classifier. Therefore, the method is impossible to be automated.

The best state-of-the-art A/D detection method was reported in the work [33]. The authors reported experimental results demonstrate that the proposed method achieves the accelerations' F1-score of 78.82%, the decelerations' F1-score of 79.10%, the baseline difference of 2.61 bpm, the synthetic inconsistency coefficient of 49.99%. Another A/D detection method was reported in a paper [16], obtained a sensitivity/F1-scores of 87/70% for acceleration classification and of 93/73% for deceleration. Both works were based on hospital's private dataset, which contained expert annotations. The authors also extended the algorithm to calculate baseline in the study [34].

With the advent of deep-learning techniques, end-to-end diagnosis related research became the mainstream. In particular, one-dimensional (1D) signal segmentation models based on the UNet structure have been widely used in A/D detection tasks. At the same time, the fixed process of LSMP was proposed and is still used today [35–37].

Zhong et al. [38] proposed the CTGNet method, which is the first to use the UNet segmentation architecture to solve the A/D detection task. This architecture is commonly used in the medical imaging domain. Most of the subsequent deep learning solutions refer to it. In their work, this UNet architecture directly outputs the A/D results without introducing any post-processing methods.

Then, Mujun Liu et al. [35] proposed the LSMP approach for the first time for the purpose of refining the UNet segmentation outputs. Their proposed method named EMAU-Net contains the efficient channel attention module, which includes a "non-local attention" module at the deepest part of the UNet architecture, i.e., the bridge between the encoder and decoder. All subsequent related works have used LSMP as post-processing after segmentation to improve the accuracy. The same research team then proposed ELCResU-Net [39], which uses residual layers and channel shortcut connections based on the UNet architecture to improve A/D detection capabilities.

Qingjian et al. [36] developed the ETCNN model which contains the Channel-Residual (C-Res) and the Channel Cross fusion with Transformer (CCT) modules at the connection between the encoding and the decoding stack in the UNet architecture. This is the first time that we have seen research that attempts to use the self-attention mechanism in this field. Minghan et al. [37] proposed the MTU-Net3+ model. It uses the UNet3+ architecture, which is an enhanced version of the UNet architecture that allows for complex connection between the encoding and the decoding stacks. They innovatively used the self-attention mechanism and Bidirectional LSTM model to directly model the baseline as the second stream and assist the A/D segmentation stream.

All the above mentioned state-of-the-art methods use exactly the same post-processing approach, which involves median filtering, rule-based heuristics, and baseline estimation. However, these approaches introduce several limitations, which our proposed method seeks to overcome.

2.4 Residual learning in medical imaging and signal processing

The idea of using residual learning for refinement tasks has shown wide applicability in medical image analysis and time-series classification. For example, ResNet-based architectures have been widely adopted in various medical image recognition tasks, as summarized in the survey [40]. Residual convolutional modules have also been utilized for ECG-based arrhythmia screening [41], while residual learning strategies have proven effective for medical image denoising and reconstruction [42].

These studies highlight the benefit of decoupling coarse prediction generation from fine-grained correction, particularly in contexts where precise expert supervision is expensive but imprecise rule-based methods are widely available. Our proposed ERNP method extends this paradigm into fetal monitoring by learning a residual correction function that maps the candidate zones generated by Long- and Short-Term Median Filter Post-processing (LSMP),

which is introduced in Section 4.1.1, to clinically consistent A/D events. This design not only improves detection accuracy, but also enables flexible integration with any segmentation backbone or post-processing heuristic.

2.5 Recent Trends in Deep Learning-Based FHR Analysis

While one-dimensional U-Net–based architectures and LSMP-style post-processing have become well-established in acceleration and deceleration detection [35,36,38], more recent studies have shifted their focus toward improving robustness, interpretability, and deployment readiness rather than proposing entirely new backbone architectures [43–45]. Current efforts emphasize reducing false positives caused by short-term fluctuations, improving temporal coherence of detected events, and increasing consistency across heterogeneous recording conditions [46–48]. These trends reflect a gradual transition from purely architectural innovation toward system-level refinement, a perspective that aligns with the motivation of the refinement strategy proposed in this dissertation. Two representative directions that exemplify these recent trends are briefly discussed below.

2.5.1 Generative approaches

Beyond discriminative segmentation models, recent studies have explored generative approaches to address data scarcity and signal corruption in fetal monitoring. Conditional GANs and diffusion-based models have been investigated for synthetic FHR signal generation or for imputing missing signal segments caused by sensor dropout [49–53]. These methods primarily target data augmentation and signal restoration rather than direct event-level A/D detection. As such, they are complementary to segmentation-based detection pipelines and remain largely orthogonal to the refinement and post-processing challenges addressed in this dissertation.

2.5.2 Federated learning and privacy-preserving FHR modeling

An emerging research direction in medical time-series analysis is federated learning, which enables collaborative model training across institutions without sharing raw patient data [54, 55]. This paradigm is particularly relevant for FHR analysis, where privacy constraints, heterogeneous acquisition protocols, and limited data availability often hinder centralized learning [56]. A recent study has demonstrated the feasibility of federated learning frameworks for CTG-based fetal health assessment [57]. While federated optimization is not explicitly investigated in this dissertation, the proposed segmentation and refinement models are compatible

with federated training settings, making this an important direction for future extension.

2.6 Summary and Research Gaps

In summary, the literature on automated FHR A/D detection has advanced considerably, progressing from handcrafted features and rule-based systems to deep learning–based segmentation and sequence-level classification models. While these developments have demonstrated significant improvements, several research gaps remain. These gaps highlight both methodological challenges and practical considerations that need to be addressed in order to move toward clinically deployable systems.

2.6.1 Lack of event-level context modeling

Most existing approaches, particularly those based on segmentation networks such as UNet or its variants, treat A/D detection as a per-timepoint segmentation task. As a result, event boundaries may be fragmented, short fluctuations may be misclassified as true events, and clinically meaningful temporal patterns may be lost. Event-level context modeling, such as explicitly representing onset, duration, and offset relationships, remains underexplored. Without this capability, models risk producing outputs that are technically accurate at the time step level but clinically unreliable in practice.

2.6.2 Multimodal fusion design remains rudimentary

Although some recent studies have begun to incorporate the UC signal alongside the FHR, the design of multimodal architectures remains relatively basic. Most prior work adopts early or late concatenation strategies, which fail to capture the complex physiological relationships between modalities. In reality, the temporal interaction between FHR, UC, and FM is critical to interpreting fetal well-being, for example, to distinguish between decelerations associated with contractions and those occurring independently. Advanced multimodal fusion mechanisms, such as adaptive attention, cross-modal transformers, or dynamic weighting schemes, are still largely absent in this domain. This limits the ability of current models to take advantage of the full richness of available monitoring data.

2.6.3 Post-processing remains separate and non-differentiable

A limitation of current pipelines is the reliance on rule-based, non-differentiable post-processing modules. Median filtering, baseline estimation, and heuristic thresholds are typically applied

after model inference to enforce clinical plausibility. However, because these operations are external to the training process, the model cannot adapt its predictions in a way that optimizes for final event-level accuracy. This misalignment introduces a gap between training objectives and evaluation outcomes, often leading to suboptimal performance. Moreover, non-differentiable post-processing limits the possibility of end-to-end optimization, which has been a key factor in the success of using deep learning in other medical domains. Bridging this gap through differentiable or learnable refinement strategies represents an important direction for future research.

2.6.4 Deployment challenges: signal loss and data quality

Finally, most existing studies evaluate methods on datasets under controlled conditions, but they rarely account for real-world deployment challenges. In clinical practice, monitoring signals often suffer from missing signals (e.g., absent UC or FM traces), motion artifacts, sensor displacements, or intermittent signal dropout. These issues can severely degrade performance if the model is not designed to handle incomplete or noisy input. Moreover, data quality varies significantly across hospitals, vendors, and acquisition protocols, but systematic investigations into robustness under such variability remain scarce. Developing models that are resilient to data degradation and adaptable to heterogeneous clinical environments is therefore a critical, but largely overlooked, research gap.

2.6.5 Conclusion

Taken together, these four research gaps underscore the need for a new generation of methods that move beyond per-timepoint segmentation, simplistic multimodal fusion, and rigid post-processing pipelines. Addressing these challenges will not only improve algorithmic performance but also enhance clinical interpretability, robustness, and trustworthiness—key prerequisites for real-world adoption of AI in fetal monitoring.

CHAPTER 3 DATA SETS

3.1 Public dataset 1: CULF-DB

We adopt a widely used open-access FHR dataset [58,59] from the Catholic University of Lille France (CULF-DB) for model training and evaluation. 66 FHR recordings were collected, organized, and annotated. Data is collected from pregnant women between 36 and 41 weeks of gestation. They are selected from 12,000 deliveries between 2011 and 2016. The Fetal-CTG monitoring machine is the Avalon FM40 and FM50 (Philips Healthcare, Amsterdam, The Netherlands). The provider of this dataset is Saint Vincent de Paul Maternity Hospital of Lille Catholic University. The average recording duration of this dataset is 90 minutes, with a minimum of 30 minutes and a maximum of 7 hours. The signal was sampled at 4 Hz. Based on experience, we downsampled all signals to 1 Hz in our experiments. Each recording was reviewed and annotated by four experts through consensus.

There are 544 acceleration events and 827 deceleration events of CULF-DB. The histograms of these two events' duration is shown in Figure 3.1. From the figures, we know that the duration of acceleration events is generally shorter than that of deceleration events. Overall, their distribution patterns are similar. However, we can also observe that a small number of events exhibit durations significantly longer than the average, introducing variability that increases the difficulty of accurate A/D detection.

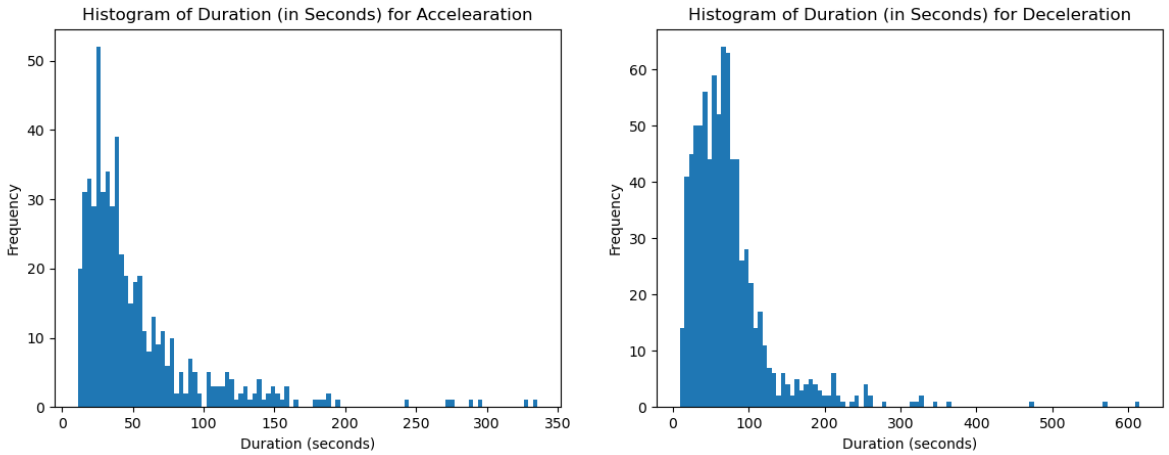

Figure 3.1 The histograms of acceleration and deceleration events' duration in seconds of the public dataset CULF-DB. The left and right figures correspond to the histogram of A and D respectively.

3.2 Public dataset 2: CTU-UHB

We also use the publicly available CTU-UHB intrapartum CTG database for external validation and performance comparison. This dataset has been widely adopted in CTG and FHR research, providing a useful benchmark for cross-study comparisons.

The CTU-UHB database comprises 552 CTG recordings selected from an original pool of 9,164 intrapartum monitoring records collected between 2010 and 2012. There are 824 acceleration events and 2259 deceleration events in it. Each recording contains two parallel signals: fetal heart rate (FHR) and uterine contraction (UC), both sampled at 4 Hz. The recordings were truncated to at most 90 minutes before delivery. To ensure signal quality, only those traces satisfying a minimum completeness threshold (e.g., $\geq 50\%$ availability in rolling 30-min windows) and relevant clinical criteria (e.g. singleton pregnancy, gestational age > 36 weeks) were retained. The raw CTU-UHB recordings have been augmented with manual event annotations in subsequent studies.

In our experiments, we preprocess CTU-UHB signals to match our internal dataset. Firstly, we downsample from 4 Hz to 1 Hz. Secondly, we apply the same segmentation, normalization, and denoising pipeline as for CULF-DB. We reserve CTU-UHB solely for external validation (i.e. no fine-tuning allowed), to assess generalization. When reporting results, we compute both event-level and time-step-wise metrics under the same post-processing rules as for other datasets.

CTU-UHB's relatively small size and older acquisition conditions may lead to domain mismatch with our internal data. There may be differences in sensor model, sampling protocols, noise characteristics, or annotation bias. Since CTU-UHB includes publicly studied annotations and has been reused in multiple works, it also carries inter-study annotation variability and potential bias. We interpret CTU-UHB validation results cautiously, focusing on trend consistency rather than absolute numbers.

3.3 Our Large Scale Private Dataset: 7AH-SYSU

To develop a deep learning model for accurate and precise A/D detection, we curated a large dataset of Fetal-CTG recordings with gold standard A/D annotations. The recordings were collected over a two-year period by obstetricians from the Obstetrics and Gynecology Department of the Seventh Affiliated Hospital Sun Yat-sen University. And the annotations was cross-reviewed by senior obstetricians.

The dataset comprises 667 Fetal-CTG graphs from intrapartum clinical sessions, with each

graph consisting of at least one 20-minute recording page. The signal was sampled at 1 Hz. In total, the dataset includes 2,944 pages with annotations for three classes: background, acceleration, and deceleration. A total of 4,018 acceleration and 2,805 deceleration events were recorded.

An example of an annotated Fetal-CTG with three signal channels, which are FHR, UC and FM, is shown in Figure 3.2. In this figure, the blue bounding box represent a Deceleration, while the red bounding box represent an Acceleration.

It should be noted that our focus was on collecting and annotating deceleration recordings as they are crucial for training machine learning models. In the first half of the data collection, we included only complete recordings with either acceleration or deceleration events. In the second half, we included only complete recordings with at least one deceleration within a 20-minute duration. All Fetal-CTG records with A/D features were labeled.

The annotation and cross-checking process was performed independently by clinicians and the recordings were presented in random order to the doctors. It is worth mentioning that all doctors involved in the annotation and cross-checking process had over 10 years of experience. The resulting dataset comprises high-quality signal data with gold standard A/D annotations, which we believe will be invaluable for developing and evaluating deep learning models for Fetal-CTG analysis.

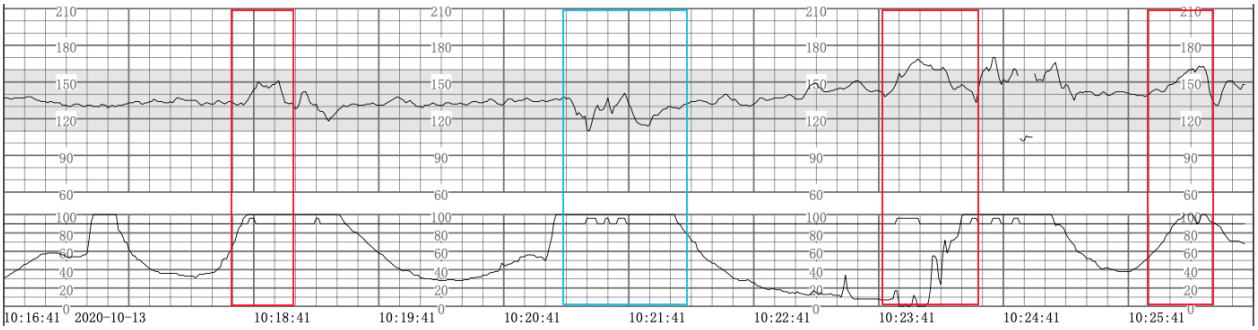

Figure 3.2 An example of a annotated Fetal-CTG. The blue bounding boxes are Deceleration. The red bounding box is an Acceleration.

3.4 Data Set Summary

Table 3.1 Summary of the datasets used in this study.

Dataset	Signals	#Records	Labels	Used in Chapter
CULF-DB	(FHR)	66	544 A / 827 D	Ch. 4, Ch. 6
CTU-UHB	(FHR, UC)	552	824 A / 2259 D	Ch. 5
7AH-SYSU	(FHR, UC, FM)	667	4,018 A / 2,805 D	Ch. 5, Ch. 6

All three datasets collectively cover a broad spectrum of clinical and signal characteristics. CULF-DB provides a compact but well-annotated open-access dataset for model development and benchmark comparison. CTU-UHB offers a publicly available external validation resource with independent acquisition conditions and multi-center variability, which is crucial for testing model generalization. Our large-scale private 7AH-SYSU dataset constitutes the main training corpus, characterized by high-quality clinical annotations from senior obstetricians and richer multimodal information (FHR, UC, and FM). By integrating these datasets, we ensure that the proposed methods are evaluated across diverse domains, thereby enhancing both robustness and clinical applicability.

3.5 Ethics and Anonymization

All datasets used in this thesis comply with relevant ethical standards and privacy protection regulations. The publicly available datasets (CULF-DB and CTU-UHB) were obtained from open-access repositories and used strictly for research and non-commercial purposes under their respective licenses. Both databases were fully anonymized by their providers prior to public release, containing no personal identifiers or metadata that could link the recordings to individual subjects.

For our private clinical dataset (7AH-SYSU), ethical approval was obtained from the Ethics Committee of the Seventh Affiliated Hospital of Sun Yat-sen University (Shenzhen, China). All recordings were collected in accordance with the Declaration of Helsinki, with informed consent obtained from all participants prior to data acquisition. Patient-identifying information (such as name, hospital ID, or admission number) was permanently removed during data export from the hospital system. Each record was assigned a random alphanumeric identifier to ensure full anonymity. The dataset is stored on secure, access-controlled institutional servers, and all analyses were performed on de-identified data only.

No patient or clinical staff can be identified, directly or indirectly, from any of the figures or data presented in this thesis. The use of anonymized clinical data was approved solely for scientific and educational research purposes. All experiments were conducted following the ethical standards of Polytechnique Montréal and the partner institution's data governance policies.

CHAPTER 4 A RESIDUAL-INSPIRED EVENT REFINEMENT NETWORK FOR FETAL HEART RATE ACCELERATION/DECELERATION DETECTION

4.1 Introduction

According to clinical standards (e.g., NICHD guidelines), five key diagnostic features—Baseline, Variability, Acceleration, Deceleration, and Sinusoidal Patterns—form the core criteria by which clinicians assess Fetal-CTG signals. Among these five key diagnostic features, acceleration (A) and deceleration (D) events in FHR signals are key indicators of fetal health and potential distress. And, deceleration is the most predictive of fetal hypoxia, particularly when accompanied by tachycardia [15]. Failure to recognize acceleration and deceleration can lead to serious consequences, including fetal death, while over-interpretation of them may result in unnecessary cesarean sections.

Clinicians rely on well-defined protocols to manually identify acceleration and deceleration events in FHR signals, which involves detecting deviations from a baseline and classifying events according to predefined criteria. Despite this, high inter- and intra-observer variability in Fetal-CTG interpretation remains a challenge [14].

For overcoming the above problems, automatic acceleration and deceleration (A/D) detection is expected to improve efficiency, reduce inter-observer variability, and provide real-time clinical decision support. However, several reports have shown that the performance of existing automated diagnosis algorithms for A/D detection is unsatisfactory [16–19]. In recent years, deep learning techniques have been increasingly applied due to their remarkable performance in various pattern recognition tasks [20]. Various deep learning solutions [35–37] have been proposed for this A/D detection task, showing promising results.

4.1.1 Challenges

Most state-of-the-art deep learning solutions for acceleration/deceleration (A/D) detection follow a common three-step segmentation pipeline, here referred to as the LSMP (Long- and Short-Term Median Filter Post-processing) paradigm. This pipeline consists of:

1. Per-timestep classification: A deep segmentation model assigns a class label—acceleration, deceleration, or background—to each timepoint, producing coarse A/D region masks. This is a typical temporal segmentation task.

2. Baseline estimation via median filtering: Two median filters with short and long temporal windows are applied to the raw FHR signal to remove transient fluctuations and estimate the baseline. The lower envelope of the filtered signal is treated as the baseline curve. Detected A/D regions are temporarily removed during this step, and the baseline is interpolated across these gaps.
3. Rule-based event determination: Threshold-based heuristics are applied on the estimated baseline to determine final A/D events. These rules typically enforce constraints on amplitude and duration to filter out clinically irrelevant fluctuations.

The detailed algorithmic procedures and parameter settings of this pipeline can be found in [35].

Despite its widespread adoption in the literature [35–37], the LSMP-based pipeline presents two key limitations that warrant further refinement.

The first issue, which we denote as semantic mismatch, lies in a subtle but important misalignment between the training objective and the actual usage of the model predictions. UNet-based segmentation models are typically trained using per-timestep labels derived from precise expert annotations. However, LSMP—like many heuristic post-processing steps—only requires coarse candidate zones as inputs, and does not preserve or leverage the precise boundary information encoded during training. This mismatch introduces a representational gap that hinders end-to-end optimization and limits detection reliability.

Furthermore, We suspect that the downstream steps in the LSMP-based pipeline heavily rely on a baseline (BL) estimated via median filtering. Final A/D event decisions are made by applying amplitude and duration thresholds relative to this BL. This rule-based logic assumes that a reliable baseline can be extracted from noisy and fluctuating signals using simple filtering, which is often not the case. We empirically show that even when constructing the BL directly from expert-labeled A/D, the resulting A/D event detection yields an average F1-score of only 70.33%. This finding reinforces our claim: focusing on the task of training the segmentation (step 1) based on the A/D task is misaligned with the goal of optimally segmenting the data in the pipeline for step 3. training the segmentation model to reproduce expert labels does not necessarily translate to optimal performance under a baseline-driven post-processing framework.

4.1.2 Lack of reproducibility in existing methods

Another important challenge in this field is the lack of reproducibility across studies. Despite using the same public datasets and claiming similar model architectures, reproduced

results often show large discrepancies, raising concerns about the validity and comparability of reported findings. In particular, most published state-of-the-art (SOTA) methods do not provide open-source code, and their papers omit essential implementation details—especially for pre-processing and post-processing procedures—which are critical for performance replication.

For example, the EMAU-Net [35] model was re-evaluated in both the MTU-Net3+ [37] and ETCNN [36] papers, reporting deceleration F1 scores of 84.62% and 74.83%, and acceleration F1 scores of 50.00% and 70.37%, respectively. These substantial differences, despite using the same dataset and the same model, illustrate how the absence of reproducible pipelines undermines the credibility of comparisons in this field.

To address this issue, we make available the full source code of our entire pipeline, including all SOTA model re-implementations, data pre-processing steps, and post-processing routines. This open approach ensures that all reported results are reproducible and facilitates more rigorous benchmarking for future research.

4.1.3 Our approach

To address the limitations identified in existing A/D detection pipelines, we propose a robust and universally applicable refinement architecture named Event Refinement via Neural Processing (ERNP) inspired by the residual learning paradigm. ERNP serves as a lightweight, segmentation-model-agnostic refinement module that improves the accuracy and clinical reliability of A/D event detection by correcting the output of any LSMP paradigm approach from the original FHR signals.

To support reproducibility, we also provide full open-source implementations of our method, including all pre-processing and segmentation model components.

The overall framework consists of two stages:

1. Rough processing by a deep learning segmentation model followed by LSMP: A deep learning segmentation model predicts per-timestep A/D probabilities based on expert annotations. These predictions are processed by an LSMP approach, which produces binary candidate masks for A/D regions.

2. Residual Refinement via ERNP: ERNP takes both the raw FHR signal and the LSMP output as inputs, treating the LSMP mask as an informative prior. It learns to refine these predictions into expert-aligned A/D events using a combination of temporal convolutions and self-attention.

Our proposed ERNP enhances A/D event detection by leveraging a 1D convolutional layer (Conv-1D), a self-attention module, and a rule-based final processing step. The Conv-1D layer captures local temporal features, while the self-attention module enables modeling of global dependencies and contextual patterns. While recent models such as ETCNN [36] and EMAU-Net [35] have incorporated Conv-1D and self-attention within end-to-end segmentation architectures, ERNP is distinct in both function and placement.

Although ERNP operates as a post-processing module, its architecture is deeply inspired by the concept of residual learning. Rather than replacing the outputs of existing LSMP paradigm approaches, ERNP treats these outputs as informative priors and learns to refine them using the original FHR signal as contextual support. This design mirrors the core idea of residual networks, where a lightweight refinement module learns only the correction needed to bridge the gap between the coarse prediction and the desired output.

ERNP addresses three key error modes commonly observed in LSMP paradigm pipelines:

1. Semantic mismatch: By explicitly modeling the refinement process as a residual learning task—correcting coarse LSMP outputs toward expert-level annotations—ERNP bridges this semantic divide.

2. False positives: Non-event regions may be falsely classified as A/D events due to minor FHR fluctuations. ERNP suppresses these by residual learning refinement.

3. Contextual inconsistency: Classical methods may identify acceleration events in regions dominated by decelerations, which contradicts physiological plausibility. ERNP mitigates this by using self-attention mechanisms that model inter-event dependencies.

Experimental results demonstrate that integrating ERNP into existing segmentation pipelines consistently improves performance across multiple SOTA models and datasets, validating its effectiveness as a general-purpose refinement tool.

Our main contributions are as follows:

- **A novel post-processing method ERNP:** We introduce ERNP, a residual-inspired and self-attention-based refinement module that enhances A/D event predictions by learning contextual dependencies and correcting typical segmentation errors. It operates independently of specific model architectures.

- **Comprehensive benchmarking with and without ERNP:** We conduct a thorough comparative study evaluating multiple state-of-the-art deep learning models un-

der both the existing LSMP-based pipeline and our proposed ERNP framework. This allows us to quantify the generality and benefits of ERNP across architectures.

- **Fully open-source implementation:** To ensure reproducibility and facilitate future research, we release complete code for all model implementations, data pre-processing, and both LSMP and ERNP post-processing pipelines. The codebase is available at: `https://github.com/pingaowang/ernp_fhr_ad_detection`.

4.2 Methods

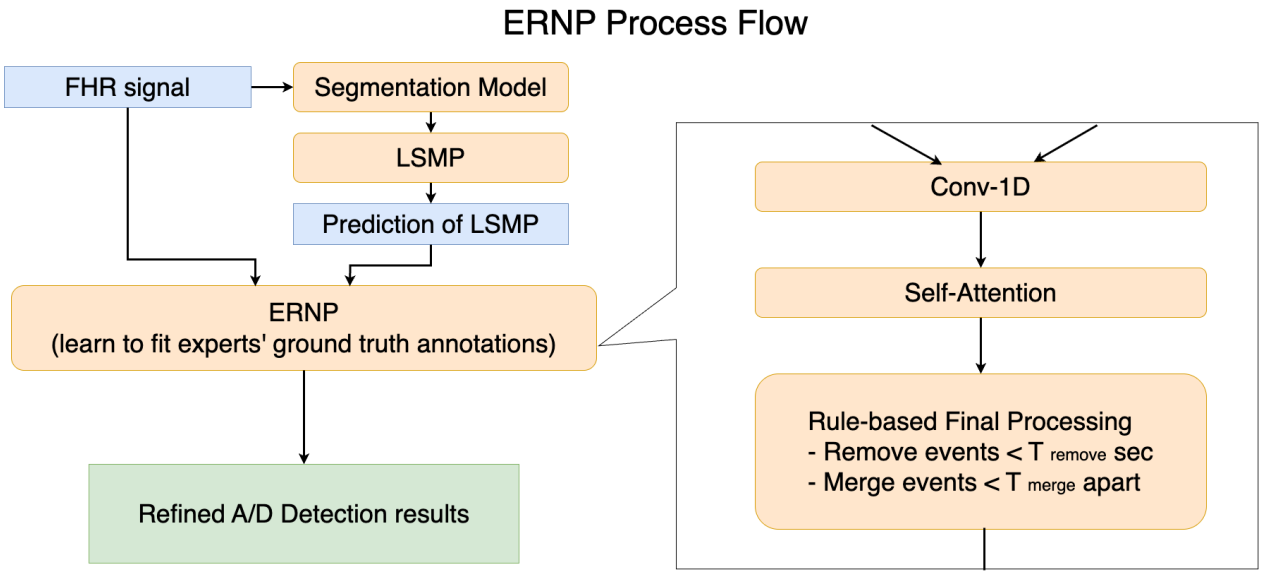

Figure 4.1 The proposed ERNP framework takes FHR signal and the corresponding LSMP results as input and refines them through Self-Attention-based event optimization, followed by a rule-based post-processing step to remove short events ($< T_{\text{remove}}$ sec) and merge adjacent events ($< T_{\text{merge}}$ sec apart). This hybrid approach ensures context-aware refinement, error suppression, and clinically meaningful outputs.

The process flow is illustrated in Figure 4.1. In practice, ERNP is implemented as a lightweight neural module consisting of the following components:

1. **Input Concatenation:** The LSMP output (predicted A/D mask) and the raw FHR signal are concatenated along the channel axis to form a input sequence with shape $[2, T]$, where T is the number of time steps.

2. **Conv-1D Layer:** A 1D convolutional layer with kernel size k extracts local temporal patterns across both channels.

3. **Self-Attention Layer:** One or more transformer encoder blocks with h attention heads model long-range dependencies and capture global contextual relationships between candidate A/D regions.

4. **Output Projection:** A linear layer maps the hidden representation back to a per-timestep A/D/None classification.

5. **Post-Refinement Filtering:** Two rule-based operations are applied:

   (a) Removing events shorter than a threshold T_{remove} (e.g., 3 seconds).

   (b) Merging adjacent events within T_{merge} seconds to reduce over-segmentation.

This hybrid approach—combining data-driven self-attention refinement with clinically guided rule-based heuristics—ensures that A/D detection is both robust and interpretable, bridging the gap between deep learning segmentation outputs and expert-level event annotations. The detailed structure of each module will be elaborated in the following subsections.

Choice of Input Signal. Although UC signals are available in public datasets, we do not utilize them in this study. Our focus is solely on acceleration and deceleration (A/D) event detection based on FHR signals. This decision is motivated by the fact that most existing clinical standards for the annotation of A/D events are based exclusively on FHR patterns, and the majority of prior automated methods also operate on FHR alone [35–38]. Therefore, we adopt a consistent input setup to facilitate fair comparisons, reproducibility, and compatibility with existing clinical practice.

4.2.1 Input

The input to ERNP consists of two temporally aligned one-dimensional sequences: the FHR signal and the output of LSMP. Both signals share the same length and are aligned point-by-point along the time axis.

The first input, FHR data, first undergoes a pre-processing pipeline (detailed in Section 4.3.2). Each FHR recording is then segmented using a sliding window of 1200 timestamps (corresponding to 20 minutes at a sampling rate of 1 Hz), with a stride of 300. We refer to each of these sliding window segments as a "sequence" for the purpose of model input and description.

For numerical stability and consistency across samples, each FHR sequence is normalized to the range $[-0.5, +0.5]$ using (4.1):

$$\mathrm{FHR}_{\mathrm{norm}} = \frac{\mathrm{FHR} - \mathrm{FHR}_{\mathrm{min}}}{\mathrm{FHR}_{\mathrm{max}} - \mathrm{FHR}_{\mathrm{min}}} - 0.5 \tag{4.1}$$

where $\mathrm{FHR}_{\mathrm{max}}$ and $\mathrm{FHR}_{\mathrm{min}}$ denote the maximum and minimum FHR values across the entire dataset. This normalization ensures that the signal range is centered around zero and compatible with the symbolic range used for the second input.

The second input, derived from LSMP output, is a symbolic representation encoded as a float-valued sequence with values $\{-0.5, 0, +0.5\}$, representing background, acceleration, and deceleration, respectively. Although represented as float values, they are interpreted as categorical class indicators rather than continuous numerical signals. The choice of this numerical range $[-0.5, +0.5]$ is intentional: it aligns the scale of the LSMP output with the normalized FHR signal, facilitating effective joint learning by the downstream convolutional and attention modules.

The two sequences are stacked as separate channels, resulting in a 2-channel input matrix of shape $[2, L]$, where $L = 1200$ denotes the length of each segment (i.e., 20 minutes of data). This dual-input design is inspired by the residual learning paradigm, where the model receives both the original signal and an initial coarse prediction (from LSMP) and learns to refine the residual difference between them.

4.2.2 Conv-1D module

The concatenated sequence is processed by a 1D convolutional layer, which enhances local feature extraction before passing the data to the self-attention module. A single Conv-1D kernel operates jointly on the two input channels. While the FHR signal and the LSMP labels are of different nature, the convolution allows the model to learn patterns across both sources by learning channel-wise weights. The Conv-1D module operates as follows:

1. Kernel Size: 21, which corresponds to a 21-second temporal window given the 1 Hz sampling rate. The input tensor has shape $[2, L]$, where the two channels correspond to the normalized FHR signal and the output from the prior LSMP. The Conv-1D layer applies a one-dimensional convolution over the time axis, allowing the network to capture local temporal patterns while jointly leveraging both input channels.

2. Output Channels: Set to be equal to the embedding dimension (d_{emb}) of the following self-attention module;

3. Padding: Applied to maintain sequence length, ensuring that the output tensor has the same temporal resolution as the input. Padding is set to 11.

After this transformation, the input tensor is converted into a higher-dimensional representation with shape $[d_{emb}, L]$, which is suitable for processing the following self-attention module.

This Conv-1D module serves as a feature extractor that transforms the raw input signals into a higher-dimensional representation, preparing them for the subsequent self-attention module. Its primary function is twofold: (1) to increase the embedding dimension so that the self-attention mechanism can operate on a more expressive feature space; and (2) to capture local temporal patterns and short-range dependencies that are relevant for FHR interpretation.

We apply Conv-1D before self-attention to ensure that the subsequent global modeling is grounded in locally meaningful representations, leading to more stable and interpretable refinements of A/D event predictions.

Importantly, ERNP leverages not only the raw FHR signal but also the coarse segmentation mask output from LSMP. This design reflects a residual-inspired strategy: instead of relearning A/D events from scratch, the model refines a preliminary prediction by modeling its alignment and deviation from the raw physiological signal. The LSMP input acts as a structured prior—albeit noisy—which guides the attention mechanism toward temporally relevant regions. This dual-input design improves robustness, especially in ambiguous cases where the raw signal alone may be insufficient for accurate event delineation.

4.2.3 Self-Attention module

The output of the previous Conv-1D module is processed by a multilayer self-attention module, which follows the Transformer encoder structure. This self-attention mechanism is designed to act as a refinement layer that can potentially correct misclassified events, filter out redundant predictions, and promote alignment with expert-annotated A/D labels. We hypothesize that by modeling global dependencies, self-attention can mitigate segmentation inconsistencies introduced in earlier stages. Experimental evidence in the third experiments of the experiment section supports its contribution to performance improvement. Instead of relying on local waveform variations, the self-attention mechanism in our method can capture global contextual dependencies, ensuring more reliable A/D detection. By leveraging global contextual relationships, self-attention mitigates common segmentation errors, such as false positive acceleration events in deceleration-dominant contexts and fragmented event detections.

The attention mechanism consists of the following steps:

1. **Linear projection:** The input tensor X with shape [Sequence Length, d_{model}] is first projected into three separate embeddings Q (Query), K (Key) and V (Value) in (4.2), where W_i^Q, W_i^K, W_i^V are three learnable projection matrices of the ith attention head for Q, K and V respectively. These projections allow the model to compute attention scores based on how different time steps relate to each other.

$$Q_i = XW_i^Q, \quad K_i = XW_i^K, \quad V_i = XW_i^V \tag{4.2}$$

2. **Scaled dot-product attention:** Attention weights are computed using (4.3). This mechanism determines how much each time step attends to other time steps, allowing the model to incorporate global information into local predictions.

$$Attention(Q_i, K_i, V_i) = softmax\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i \tag{4.3}$$

3. **Multi-head attention (MHA):** A common practice is to use multiple attention heads and concatenate the result (4.5). Each attention head learns distinct aspects of the temporal dependencies. Once the outputs from all attention heads are concatenated, they are linearly transformed back into the input feature size. Each attention head represents a scaled dot-product attention in (4.4).

$$head_i = Attention(Q_i, K_i, V_i) \tag{4.4}$$

$$MHA(Q, K, V) = Concat(head_1, head_2, ..., head_m) \tag{4.5}$$

4. **Add & norm (residual connection + layer normalization):** The multi-head attention output is added back to the original input tensor using a residual connection as shown in (4.6). This prevents gradient vanishing and improves training stability.

$$\tilde{X} = \text{LayerNorm}(X + \text{MultiHeadAttention}(X)) \tag{4.6}$$

5. **Feed-forward network (FFN) for feature transformation:** The refined feature representation is passed through a simple two-layer MLP with ReLU activation shown in (4.7). This allows further non-linear transformation of the feature space.

$$\text{FFN}(X) = \text{ReLU}(XW_1 + b_1)W_2 + b_2 \tag{4.7}$$

6. **Final add & norm (residual connection + layer normalization):** A second residual connection is applied to the FFN output to maintain feature stability as shown in (4.8).

$$X' = \text{LayerNorm}(\tilde{X} + \text{FFN}(\tilde{X})) \tag{4.8}$$

The final refined representation X' is then passed to the output layer. In the output layer, the processed feature tensor is projected back to a three-class segmentation output, representing the refined A/D event predictions. The self-attention module ensures that detected acceleration and deceleration events are contextually consistent with surrounding patterns, reducing false positive detections.

4.2.4 Loss functions

To address class imbalance and emphasize accurate segmentation across all categories, we employ the Generalised Dice Loss (GDL) [60] as our loss function. Unlike the standard Dice loss, which can be biased toward dominant classes, the Generalized Dice Loss incorporates a weighting scheme that assigns higher importance to underrepresented classes. This is particularly beneficial in our context, where the majority of timepoints belong to the background class .

Let $\hat{y} \in \mathbb{R}^{C \times N}$ denote the model's predicted class probabilities, where C is the number of classes, and N the sequence length. Let $\tilde{y} \in \{0, 1, 2\}^N$ be the ground truth class labels, where $0, 1$, and 2 represent the background class, the acceleration class, and the deceleration class respectively.

The ground truth $\tilde{y}$ is first converted to a one-hot representation $y \in \{0, 1\}^{C \times N}$. For each class c, a weight $w[c]$ is calculated based on the inverse squared frequency of that class in the ground truth in (4.9).

$$w[c] = \frac{1}{\left(\sum_{n=1}^{N} y[c, n]\right)^2 + \epsilon} \tag{4.9}$$

These weights are then used to compute the weighted overlap between predictions and targets. The Generalized Dice Loss is defined in (4.10):

$$\mathcal{L}\text{GDL} = 1 - \frac{2 \sum_{c=1}^{C} w[c] \sum_{n=1}^{N} \hat{y}[c, n] \cdot y[c, n]}{\sum_{c=1}^{C} w[c] \sum_{n=1}^{N} \left(\hat{y}[c, n]^2 + y[c, n]^2\right) + \epsilon} \tag{4.10}$$

Here, ϵ is a small constant, which we take $\epsilon = 10^{-6}$ in all the experiments, added to prevent division by zero. This formulation helps that the loss is sensitive to both over- and under-segmentation, and robust to class imbalance across different recordings. Particularly, it is suitable for fetal heart rate A/D segmentation, where event classes (A and D) are sparse compared to the background, and precise detection boundaries are critical for clinical interpretability.

It is worth noting that this formulation corresponds to the squared-norm variant of the Generalized Dice Loss, where the denominator uses $\hat{y}^2 + y^2$ instead of $\hat{y} + y$. Such a soft-Dice-style formulation provides smoother gradients and improved numerical stability in highly imbalanced segmentation scenarios, and our experimental results also confirm that this squared variant yields more stable convergence and better overall performance in the FHR A/D segmentation task.

4.2.5 Model training

Experimental results of hyperparameter tuning led to the following choice: the learning rate is 0.001; the number of epochs is 200; the batch size is 8; the dropout rate is 0.1. And we use the Adam optimizer [61], is a widely used optimizer in deep learning due to its efficiency and ability to handle sparse gradients.

4.2.6 Rule-based final processing

While the self-attention model refines A/D predictions, a final rule-based filtering processing ensures clinically meaningful event structures. This final processing is non-differentiable and is applied only during inference. It contains the following two steps: 1) Threshold-Based Event Filtering: Events shorter than T_{remove} seconds are eliminated as they are unlikely to be clinically relevant; 2) Event Merging: Adjacent events occurring within T_{merge} seconds of each other are merged to reduce over-segmentation errors. Both T_{remove} and T_{merge} are treated as tunable hyperparameters, and their optimal values are determined through experiments presented in section4.4.1.

4.3 Data sets

4.3.1 CULF-DB

We adopt a widely used open-access FHR dataset [58, 59] from the Catholic University of Lille France (CULF-DB) for model training and evaluation.

4.3.2 Data cleaning and pre-processing

We adopted a pre-processing pipeline inspired by prior state-of-the-art studies. However, due to the lack of publicly released pre-processing code or cleaned datasets, exact replication was not feasible. Instead, we designed a reproducible and robust pipeline grounded in domain knowledge and empirical observations. Specifically, the implemented steps include:

(1) Identification of invalid signal trajectories, where sudden jumps between adjacent timestamps exceed 40 bpm, which are treated as missing values; (2) Linear interpolation of missing values when the gap duration is shorter than 10 seconds; (3) Replacement of physiologically implausible values (e.g., zero values) using linear interpolation; (4) Manual exclusion of recordings with extreme artifacts or excessive missingness. Based on the expert-annotated "unreliable" segments provided in the public dataset, we identified and excluded five FHR recordings (IDs: 39, 40, 46, 56, 57) that exhibited pervasive signal corruption or were deemed clinically unusable.

This preprocessing approach ensures signal continuity while preserving clinically meaningful patterns. All steps are deterministic and reproducible, and we have released the complete data cleaning and pre-processing implementation in our public GitHub repository.

We down sampled the signal from 4 Hz to 1 Hz because according to medical knowledge, tiny waveforms have little effect on A/D recognition. Reducing the sampling rate can effectively decrease computational complexity and increase the speed of model training and inference. It can also provide a faster response speed in actual clinical scenarios in the future.

4.4 Experiments

To evaluate the robustness, effectiveness, and generalization of our proposed ERNP method, we design a series of experiments that assess its performance under different configurations and across multiple segmentation models. Specifically, we test ERNP in combination with several widely used *UNet-based* segmentation models, which serve as the first-stage predictors of candidate A/D regions.

As shown in the following subsections, our experiments aim to answer the following key research questions:

1. How do ERNP hyperparameters affect A/D event detection performance?
2. Can ERNP consistently improve A/D detection when applied to different segmentation models?

3. How do individual components of ERNP contribute to overall performance?

For all of the following experiments, we used 4-fold cross-validation.

To evaluate the performance of both the LSMP and our proposed ERNP across different models, we employ three evaluation metrics: F1-score (F1), positive predictive value (PPV), and sensitivity (Sen). The calculation of them are shown in the formula below.

$$F1 = 2 * \text{PPV} * \text{Sen}/(\text{PPV} + \text{Sen}) \tag{4.11}$$

$$\text{Sen} = \text{TP}/(\text{TP} + \text{FN}) \tag{4.12}$$

$$\text{PPV} = \text{TP}/(\text{TP} + \text{FP}) \tag{4.13}$$

The sensitivity (Sen) of A/D events is defined as the proportion of correctly detected A/D events relative to the total number of events identified by expert consensus. Similarly, the positive predictive value (PPV) is calculated as the proportion of correctly detected A/D events among all events detected by each method. Two detected events are considered to be in agreement when their segments overlap for at least 5 seconds. True positives (TP), false negatives (FN), and false positives (FP) represent the number of correctly detected events, missed events, and incorrectly detected events, respectively, in each FHR recording. We use the error reduction rate, which is represented by err.r(%), to measure the improvement of our method. The median values of the above metrics were calculated over a minimum of 10 independent runs.

4.4.1 Experiment 1: impact of ERNP hyperparameters

We conducted an extensive investigation of hyperparameter choice. To be more convincing and widely applicable, we need to tune the hyperparameters under ideal circumstances. In this controlled setting, we simulate an ideal LSMP output by deriving segmentation masks directly from ground-truth A/D labels. This allows us to benchmark the upper-bound performance of LSMP and analyze to what extent ERNP can further refine or surpass this baseline under optimal conditions.

We analyze the following three aspects:

1. Deep Learning Model Hyperparameters: The hyper-parameters in the Conv-1D module

and the self-attention module, including Conv-1D module's kernel size (k), the self-attention module's embedding dimension(d_{emb} or d), number of multi-attention heads (h) and number of hidden layers (l), and finally the choice of activation functions and dropout rates to balance expressiveness and regularization.

2. Event Duration Threshold (T_{remove}): The minimum event duration required for an A/D event to be considered valid. We evaluated different values, from 0 to 10 seconds at 1 second interval. And we analyzed their effect on precision-recall trade-offs.

3. Merging Threshold (T_{merge}): The maximum allowable time gap between two adjacent A/D events before they are merged. We experiment with values from 0 to 60 seconds at 5 second interval, to assess their impact on over-segmentation.

We conducted an extensive hyperparameter search by evaluating all combinations of the following parameters. For each configuration, we report the mean of the average F1 scores for acceleration and deceleration events. The hyperparameter ranges are as follows: $T_{\text{remove}} \in \{0, 1, 2, \ldots, 10\}$; $T_{\text{merge}} \in \{0, 5, 10, \ldots, 60\}$; kernel size $k \in \{7, 11, 21, 31, 41\}$; embedding dimension $d \in \{8, 16, 32, 64, 128, 256\}$; number of attention heads $h \in \{1, 2, 4, 8\}$; and number of attention layers $l \in \{1, 2, 3, 4\}$.

We have three heat maps (Figure 4.2, 4.3 and 4.4) showing the results of different choices of two hyperparameters (T_{remove} and T_{merge}, k and d, h and l) averaged over all other parameters. To account for the variability induced by random initialization, each hyperparameter configuration was evaluated over multiple independent runs with different random seeds. For each heatmap cell, the reported value corresponds to the mean F1-score across all runs and across the remaining hyperparameters marginalized in the corresponding figure. In addition, the standard deviation (std) and 95% confidence interval (CI) were computed for each cell based on the number of runs contributing to that configuration.

For clarity and readability, only the mean values are visualized in the heatmaps, while the overall variability statistics (std and CI ranges) are reported separately.

Then, under the ideal condition, a set of optimal hyperparameters is selected based on these hyperparameter experiments, and the comprehensive results are compared with LSMP to demonstrate the improvement of ERNP.

Analysis of ERNP's hyperparameters:

From the heat map in Figure 4.2 we can see that, the best performance is shown when T_{remove} is 6 seconds. In contrast, LSMP removes all events shorter than 15 seconds. Based on this comparison, we surmise that the ERNP model outputs less fragmented results than

Figure 4.2 Mean F1-score heatmap as a function of T_{remove} and T_{merge}, averaged over all remaining hyperparameters and multiple independent runs. Each cell represents the mean across runs; standard deviation and 95% confidence interval statistics are summarized in the text.

Table 4.1 Performances of LSMP and ERNP across Acceleration, Deceleration, and Average A/D in ideal condition. Error rate (err.r) is computed only for the F1 score, thus Sen and Spe columns are marked as N/A.

	Acceleration			**Deceleration**			**Average A/D**		
	Sen(%)	**PPV(%)**	**F1(%)**	**Sen(%)**	**PPV(%)**	**F1(%)**	**Sen(%)**	**PPV(%)**	**F1(%)**
LSMP	82.11	53.35	64.68	76.39	75.57	75.98	79.25	64.46	70.33
ERNP	76.08	67.09	**70.98**	80.22	79.89	**79.79**	77.62	73.37	**75.34**
err.r(%)	N/A	N/A	17.84	N/A	N/A	15.86	N/A	N/A	16.89

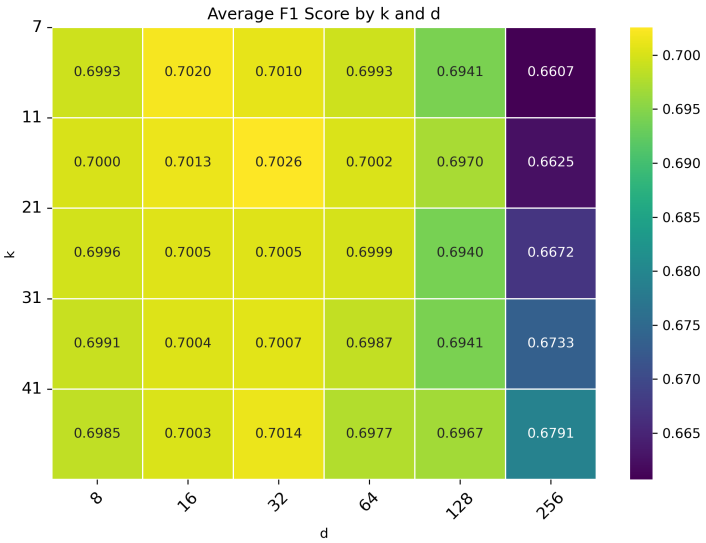

Figure 4.3 Mean F1-score heatmap as a function of kernel size k and embedding dimension d, evaluated using the optimal (T_{remove} and T_{merge}) selected from Figure 4.2 and averaged over multiple independent runs. Variability across runs is quantified using standard deviation and 95% confidence intervals.

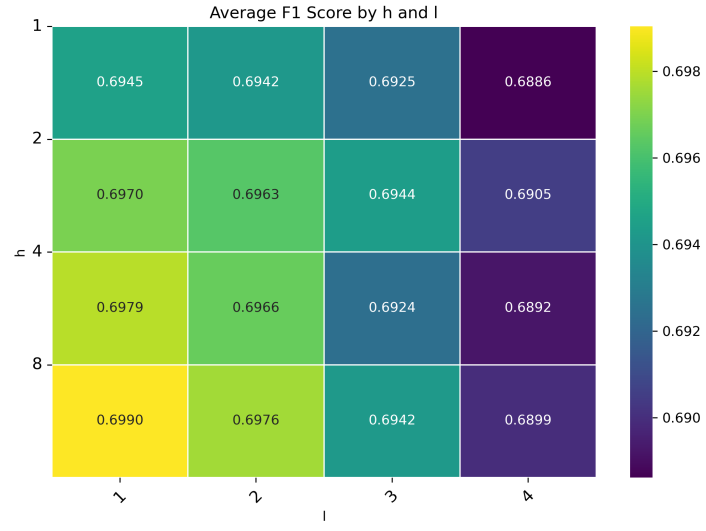

Figure 4.4 Mean F1-score heatmap as a function of the number of attention heads h and layers l, evaluated using the optimal (T_{remove}, T_{merge}, k, d) selected from Figures 4.2 and 4.3 and averaged over multiple independent runs. For clarity, only mean values are visualized.

the Segmentation model, which is why the threshold for removing fragmented events can be lowered. Interestingly, we observe that performance degrades as T_{merge} increases, with the best results obtained when $T_{\text{merge}} = 0$. This indicates that no merging of adjacent events yields the best performance. This trend contrasts with traditional pipelines, where merging is essential to address over-fragmentation. We surmise that ERNP mitigates event fragmentation more effectively than baseline methods, thereby eliminating the need for post-merge correction.

According to the heat map in Figure 4.3, the choice of kernel size (k) has a relatively minor effect on model performance, with F1 score differences within 0.5%. Among the tested values, $k = 11$ yields the highest F1 score, suggesting that a convolutional kernel roughly covering an 11-second receptive field may be sufficient for initial local pattern extraction. However, this effect is marginal and may fall within the range of random variation, particularly when $d = 32$. Therefore, we refrain from making definitive claims about the optimal kernel size based on this experiment alone. The embedding dimension (d) of the self-attention module must be carefully selected. When d is 32, the model achieves the best performance. From the same heat map shows that, when d is greater than 128, the model performance is significantly reduced. We suspect that this is because when d is too large, the generalization ability of the model will be significantly reduced, and it is more prone to overfitting.

From the heat map in Figure 4.4 we can see that the value of multi-attention heads (h) is in the range of 1 to 8, the higher the better. However, the benefit of increasing h is not obvious, and the difference is less than 0.5% at most. The smaller the number of hidden layers (l) of the self-attention module, the better. However, increasing it does not significantly reduce the model performance. The maximum F1 score difference in the range of 1 to 4 does not exceed 1%.

Across the three hyperparameter heatmaps, the median standard deviation of the F1-score across runs remains below 0.01, while the median 95% confidence interval half-width is below 0.015. Even in the worst case, the observed variability remains limited (maximum standard deviation below 0.03), indicating that the performance trends observed in Figures 4.2,4.3 and 4.4 are stable with respect to random initialization.

From the above analysis, we believe that the self-attention mechanism enhances contextual learning. But, excessive model complexity can lead to overfitting and diminish returns. Based on the experimental results, we found the best combination to be: $\{T_{\text{remove}} = 6,\ T_{\text{merge}} = 0,\ k = 11,\ d = 32,\ h = 8,\ l = 1\}$. These values will also represent the best hyperparameter combination of ERNP and is adopted for all subsequent experiments.

Analysis of ERNP's improvements:

Table 4.1 demonstrates that under ideal conditions—where the LSMP's input is derived directly from ground-truth labels—ERNP still yields substantial improvements for both acceleration and deceleration detection. Specifically, ERNP reduces the error rate by 17.84% for acceleration and 15.86% for deceleration, resulting in an average error reduction of 16.89% across both classes.

This result is particularly noteworthy because it suggests that ERNP is capable of refining even the best-case outputs of LSMP, effectively pushing the upper bound of what traditional post-processing methods can achieve. In other words, ERNP does not merely correct noise or recover from segmentation mistakes—it enhances the representation and decision quality beyond what is possible through rule-based heuristics alone. This indicates that the refinement capabilities of ERNP are not limited by the imperfections of first-stage predictions but are instead inherently powerful, even when initialized with near-perfect segmentation.

These findings confirm that ERNP is not just a corrective post-processor but a principled refinement module that can raise the ceiling of A/D detection accuracy in clinical FHR analysis.

4.4.2 Experiment 2: Performance comparison under different SOTA deep learning segmentation models

To demonstrate ERNP's applicability beyond the ideal conditions scenario, we evaluate it across multiple state-of-the-art deep learning segmentation models. Unlike the idealized setting where LSMP receives ground-truth A/D annotations as input, here all inputs to LSMP are derived from actual model predictions. This setting better reflects real-world usage, allowing us to assess whether ERNP consistently improves A/D detection performance across different segmentation models. We also aim to identify the best-performing combination of segmentation model with ERNP for this task.

We replicate existing SOTA deep learning segmentation models in this field, and refer to each model by their original publication reference. The detailed parameters of the model structure are all referenced in the respective reference. For the sake of reproducibility, we open source our implementation of each segmentation model code in our GitHub repository, along with the preprocessing modules for full reproducibility of the current work.

It is worth mentioning that almost every SOTA model uses ensemble methods with different kernel sizes, but they also show through experiments that the performance improvement from the ensemble method is limited. For example, as reported by Mujun Liu et al. [35], ensemble models combining four convolutional kernel sizes (21, 31, 61, 81) showed marginal gains in

the average A/D F1-score—less than 1% compared to the best single-kernel selection—with overlapping confidence intervals suggesting statistically insignificant improvement. However, this came at the cost of approximately 4 times both inference and training time. In this article, since we compare performance across model architectures, no ensemble method is used. All SOTA models have shown through experiments that 21 is the best kernel size when the sample rate is uniformly fixed to 1 Hz, so this article uses this configuration.

To evaluate the adaptability of the proposed ERNP across different architectural paradigms, we selected representative and relevant A/D segmentation backbones from recent literature and our own implementations:

1. **CNN-based segmentation models:** including UNet-derived architectures without Transformer-style self-attention, such as CTGNet [38] and EMAU-Net [35]. We also include ETCNN [36], which incorporates lightweight attention modules but remains primarily convolutional in design.
2. **Attention-augmented segmentation models:** to further examine ERNP's behavior when integrated into architectures with residual learning and attention mechanisms, we implemented an internal attention-based baseline that shares the same convolutional backbone structure as ERNP but without the refinement stage. This enables a controlled assessment of ERNP's contribution under similar architectural conditions.

The chosen models cover both conventional convolutional designs and attention-augmented variants, ensuring a diverse yet relevant evaluation set for testing ERNP's generalization ability.

For each segmentation model, we measured: 1) F1, Sensitivity and PPV before and after ERNP integration, and 2) Error reduction rate (%) achieved by using ERNP. Each model is evaluated under LSMP and ERNP. The corresponding results are presented in Table 4.2.

Analysis:

ERNP consistently improves F1-scores across all tested segmentation models, highlighting its robustness and generalizability, inspired by the residual learning paradigm. On average, it achieves an error reduction rate of approximately 15.24%.

The most substantial improvement, with an error reduction of 19.95%, is observed when ERNP is applied to the $CTGNet$ segmentation model, suggesting that ERNP is particularly effective in complementing this architecture.

Notably, even the weakest-performing baseline model benefits significantly from ERNP integration, underscoring its ability to correct diverse segmentation errors regardless of the initial

Table 4.2 Comparison of LSMP and ERNP refinement strategies across multiple segmentation models for A/D detection. Results are reported on acceleration, deceleration, and average A/D performance. Error rate (err.r) is computed only for the F1 score, thus Sen and Spe columns are marked as N/A.

Segmentation Model	**Acceleration**			**Deceleration**			**Average A/D**		
	Sen(%)	**Spe(%)**	**F1(%)**	**Sen(%)**	**Spe(%)**	**F1(%)**	**Sen(%)**	**Spe(%)**	**F1(%)**
CTGNet + LSMP	69.68	48.42	57.13	67.12	69.89	68.48	68.40	59.16	62.81
EMAUNet + LSMP	73.26	47.43	57.58	62.42	70.80	68.19	69.51	59.12	62.88
ETCNN + LSMP	74.01	52.21	61.23	69.47	69.31	69.39	71.74	60.76	65.31
Attention-augmented + LSMP	72.13	48.04	57.67	69.59	70.97	70.27	70.86	59.50	63.08
CTGNet + ERNP	72.32	**61.27**	**66.05**	**74.91**	**72.73**	**73.31**	**73.52**	**67.02**	**70.23**
EMAUNet + ERNP	**72.88**	59.62	65.50	73.67	72.19	72.85	73.18	65.97	69.49
ETCNN + ERNP	73.26	57.26	64.28	77.26	66.54	71.50	75.26	61.90	67.89
Attention-augmented + ERNP	71.75	59.70	65.33	74.66	70.81	72.91	73.42	65.29	68.89
CTGNet err.r(%)	N/A	N/A	**20.81**	N/A	N/A	**15.32**	N/A	N/A	**19.95**
EMAUNet err.r(%)	N/A	N/A	18.67	N/A	N/A	14.65	N/A	N/A	17.81
ETCNN err.r(%)	N/A	N/A	7.87	N/A	N/A	6.89	N/A	N/A	7.44
Attention-augmented err.r(%)	N/A	N/A	18.10	N/A	N/A	8.88	N/A	N/A	15.74

model quality.

4.4.3 Experiment 3: Ablation study

To further investigate the contributions of individual ERNP components, we conduct an ablation study by incrementally removing specific components:

1. LSMP only: The baseline segmentation output with conventional post-processing;
2. LSMP + rule-based filtering (LSMP + Rule): Only applying the rule-based final processing without the self-attention refinement;
3. LSMP + ERNP (full pipeline): Our complete method, integrating both self-attention and rule-based final processing.

Table 4.3 F1-score (%) results of ablation study. Where err.r(%) represents the error reduction rate.

Segmentation Model	**LSMP**	**LSMP + Rule**	**LSMP + ERNP**	**err.r(%) of ERNP**
ground-truth labels	70.33	73.54	74.74	14.86
CTGNet	62.81	65.86	**70.23**	**19.95**
EMA UNet	62.88	66.67	69.49	17.81
ETCNN	**65.31**	**67.85**	67.89	7.44
Attention-augmented	63.08	67.15	68.89	15.74

Analysis:

Table 4.3 shows that, the full ERNP pipeline consistently and significantly outperforms all ablated variants across models and metrics. Removal of the entire ERNP module, including the self-attention, convolutional, and final processing layers, significantly degrades performance, highlighting the importance of dedicated refinement beyond initial segmentation.

Rule-based final processing yields measurable gains. However, without a learned refinement mechanism, it is insufficient to recover the full performance gap, suggesting limitations in addressing subtle segmentation errors.

4.5 Discussion

4.5.1 Advantages of proposed ERNP

The advantages of the proposed ERNP approach can be summarized as follow:

Bridges the gap between LSMP and expert annotations: The training objective of ERNP is to match expert-labeled A/D events, whereas no existing LSMP paradigm solutions directly learns from expert annotations at its final output stage. Following the paradigm of residual learning, by refining LSMP outputs based on real-world expert annotations, ERNP improves alignment with clinical decision-making.

Enhances any existing LSMP-based pipeline: ERNP functions as an independent refinement component, making it easy to integrate into any existing deep learning pipeline without modifying model architectures. This flexibility ensures broad applicability across different A/D detection frameworks. The only condition is that the input of ERNP must contain the output of LSMP.

Ensures alignment with medical guidelines: ERNP takes as input the post-processing output derived from widely validated and medical-guideline-based LSMP paradigm methods. This design supports the model incorporates and adheres to clinical standards for A/D detection, preserving clinical consistency.

Eliminates the need for baseline annotations: Unlike many existing methods that depend on baseline annotations for model training or event detection, ERNP operates without requiring baseline labels in either training or inference. This significantly reduces the annotation burden on medical experts, accelerating AI-driven research in fetal heart rate A/D detection.

Bridges data-driven and rule-based processing: ERNP combines self-attention-based event refinement with rule-based heuristics, ensuring that detections remain both data-driven and clinically meaningful. The learned model corrects inconsistencies, while the rule-based step enforces critical domain-specific constraints.

4.5.2 Why did we choose to improve rather than replace existing postprocessing methods?

Rather than redesign the current LSMP paradigm, our goal was to enhance and generalize it, ensuring seamless integration, broad applicability, and computational efficiency while achieving substantial performance gains. Below are several key reasons behind this decision:

Preserving model-specific optimizations. Many existing LSMP paradigm methods are highly optimized for specific deep learning models. Replacing them entirely could risk losing domain-specific refinements that researchers have carefully tuned over time. Instead of discarding these optimizations, ERNP builds upon them, leveraging their strengths while mitigating their weaknesses.

Addressing specific weaknesses without reinventing the wheel. Current state-of-the-art LSMP-based methods follow a segmentation + heuristic post-processing paradigm, where the post-processing step significantly impacts final performance. Instead of redefining the entire framework, our approach targets and improves the most problematic part—the event filtering and merging process—where the refinement based on the self-attention mechanism provides significant added value.

Enhancing reproducibility and benchmarking. Many existing methods are not open-source, leading to challenges in reproducing results. By keeping the original method intact and adding a separate refinement step, we provide a transparent and easily comparable improvement, ensuring fair benchmarking and reproducibility.

Preserving medical standard practices and interpretability. Existing LSMP pipelines often follow gold-standard annotation practices used by medical experts when identifying A/D events, where the reference baseline plays a crucial role. Replacement of the entire method could disrupt this clinically established framework, potentially reducing the interpretability of the final results. By enhancing rather than replacing the LSMP paradigm pipeline, we ensure that critical clinical decision-making steps remain intact, while improving overall detection accuracy. This approach maintains both scientific validity and practical usability in real-world medical applications.

4.5.3 Discussion of self-attention model

Unlike LSMP, which classifies A/D events based purely on local fluctuations, self-attention learns to model long-range dependencies between events, helping that detected accelerations are contextually consistent. Clinically, isolated acceleration events in a predominantly decelerating pattern are unlikely. By capturing relationships between neighboring events,

self-attention suppresses false accelerations in such contexts, ensuring better alignment with expert-annotated ground truth.

Moreover, by learning relationships between different detected events, self-attention helps in refining predictions where segmentation models may over-segment or miss critical patterns. For instance, in cases where multiple decelerations occur in close succession, self-attention prevents the incorrect classification of minor fluctuations as acceleration, a common failure mode of LSMP.

4.5.4 Clinical Implications, Data Limitations, and Future Directions

While the proposed ERNP framework demonstrates substantial improvement in acceleration/deceleration event detection using FHR-only data, several factors limit the direct translation of these results into clinical practice.

First, dataset characteristics impose inherent constraints. The widely used CTU-UHB dataset contains many recordings from the last 90 minutes before delivery, often during the second stage of labor, where signal dropouts and discontinuities are frequent. Without contextual maternal information, many of these segments are clinically uninterpretable, and the experienced obstetricians in our team estimate that only 20–30% of the data are of practical diagnostic value. The FHRMA dataset, although of higher overall quality, also exhibits limitations. Specifically, its deceleration annotations collapse all clinically distinct subtypes (e.g., early, late, variable, and prolonged decelerations) into a single class, despite well-established physiological and prognostic differences among these patterns. According to expert consensus, such oversimplification reduces the dataset's educational and clinical utility, especially for complex subtypes such as variable decelerations with multiple morphological variants.

Second, clinical interpretation of FHR patterns is inherently context-dependent. Obstetricians typically encounter three categories of morphological patterns: (1) guideline-defined typical patterns that are unambiguous; (2) atypical but distinctive patterns that are rare yet clearly recognizable to trained practitioners; and (3) morphologically ambiguous patterns, for which interpretation varies even among experts and often requires consideration of labor stage, temporal context, and concurrent clinical data. This reflects the need for semantic-level analysis in automated CTG interpretation, moving beyond waveform morphology toward contextual reasoning.

Third, annotation quality and consistency remain critical. We have developed a guideline-based annotation manual and implemented a multi-level review system—similar to triple-review frameworks used in large-scale image datasets—to improve label reliability.

Looking forward, several research directions emerge. The modular nature of ERNP allows straightforward extension to multi-modal inputs, including UC signals as well as quantifiable patient-specific clinical variables with proven medical relevance. Equally important is the development of explainable and clinically understandable AI, incorporating visualization of learned feature clusters, transparent training procedures, and explicit clinical benchmarking. Large language models offer a promising avenue for generating human-interpretable rationales for model predictions. Finally, the creation of new tasks and datasets—particularly those addressing fine-grained classification of deceleration subtypes—will be essential for pushing the scientific and clinical boundaries of automated CTG analysis.

4.6 Conclusion

In this paper, we propose Event Refinement via Neural Processing (ERNP), a novel method that significantly improves A/D detection performance. Our contributions include: 1. Introducing a new residual inspired architecture that consistently outperforms standard methods across various deep learning segmentation models; 2. Providing an open-source implementation of pre-processing, post-processing, and state-of-the-art model replication, solving reproducibility issues; 3. Benchmarking state-of-the-art deep learning models under ERNP, identifying the best-performing configurations for fetal heart rate analysis. Our results demonstrate that ERNP is a robust and generalizable post-processing method that can enhance any A/D detection model, making it the best current approach in the field. In future work, we plan to extend ERNP to multi-modal setups that integrate uterine contraction signals, potentially enhancing clinical relevance and predictive performance.

CHAPTER 5 MULTI-MODAL MULTI-SCALE DEEP CONVOLUTIONAL NEURAL NETWORKS FOR RECOGNIZING ACCELERATION AND DECELERATION GRAPHS IN INTRAPARTUM FETAL CONTINUOUS CARDIOTOCOGRAPHY

5.1 Introduction

In this study, we emphasize feature-level recognition and leverages all three available signal modalities in Fetal-CTG: FHR, UC, and FM. In particular, the FM channel is often overlooked in existing studies due to its relatively sparse and noisy nature. However, through a systematic evaluation of multiple fusion strategies, we demonstrate that incorporating FHR alongside FM and UC yields the best performance in our framework. This highlights the complementary roles of FM and UC in A/D recognition when properly integrated via multi-stream fusion.

Designing a machine learning model for Fetal-CTG A/D detection presents several challenges. Developing datasets is time-consuming and labor-intensive, and annotation is difficult due to the circular definitions of baseline and A/D in medical guidelines [11–13]. Additionally, the data for different Fetal-CTG features is imbalanced, with the deceleration class accounting for less than 10% of the overall data.

Given the challenges of automated fetal-CTG analysis, our aim is to develop a deep-learning abnormal detection model accurately recognizing A/D in multi-modal Fetal-CTG data, from different signal sources. As deep learning models require accurately-annotated data for training, a large Fetal-CTG dataset with high-quality signal data and A/D annotations was established, which have been annotated and cross reviewed by senior obstetricians.

In summary, we propose and experimentally validate novel algorithms for A/D recognition based on convolutional neural networks. The adapted ResNet architectures process multiple 1D signals or channels from fetal-CTG records. These signals are jointly used for A/D recognition. We also develop and compare multi-scale fusion strategies to better exploit this information. Compared with the state-of-the-art method [16], our approach significantly improved F1-score by 13.75% for acceleration and 15.47% for deceleration. Unlike Studies 1 and 3, this chapter formulates A/D recognition as a sequence-level classification task.

5.2 Methods

The pipeline of our multi-modal 1D-CNN method is shown in Figure 5.1. In this pipeline, the multi-modal fetal-CTG data are filtered, segmented and pre-processed before fed into the CNN model. These pre-processing operations are discussed in section 4.3.2. The three data channels are then encoded at two different scales, using large-scale and small-scale receptive fields, and the 6 resulting streams as fed as input to separate 1D residual CNNs. The outputs of the CNNs are then combined for multi-stream fusion and then passed to fully-connected layers which serve as a classifier to predict the probabilities of the three A/D classes (acceleration, deceleration and background).

The proposed A/D detector effectively solves the mutual reference problem of medical guidelines where recognizing A/D requires to have the baseline which, in turn, requires the exclusion of A/D and baseline-variability regions.

5.2.1 1D Residual Convolutional Neural Network

A CNN is a type of neural network that employs multiple layers of shared weights to represent convolution filters. These filters combine local features to model complex visual patterns in images. The 1D CNN, a variant of CNN, is commonly utilized in signal classification tasks and medical imaging. It is specifically designed to process 1D signals.

For all the convolutional layer in our 1D convolutional model, the output feature value Z_{ij} calculated from the input feature value X_i at the time-axis location i in the j^{th} feature map can be formulated as in equation 5.1.

$$Z_{ij} = W_j^{\mathrm{T}} X_i + b_j \tag{5.1}$$

Where Z_{ij} is the output feature value at the time-axis location i in the j-th feature map, X_i is the input feature value at the time-axis location i, W_j is the shared weight coefficients of the j^{th} filter, and b_j is the bias weight coefficient of the j^{th} filter.

We empirically set the kernel size of the 1D convolution to 5 and its stride to 2. Moreover, we used the Rectified Linear Unit (ReLU) as activation function and apply max-pooling after this activation to reduce feature dimensionality. After the last convolutional block, a flatten operation is used to reshape the 2D feature maps into a single vector that can be processed by the fully-connected layers. Our model has three fully-connected layers with 1024, 256 and 3 neurons, respectively. The number of neurons in the last fully-connected layer is equal to number of prediction classes. For the loss function, we used Focal loss [62] for balancing the

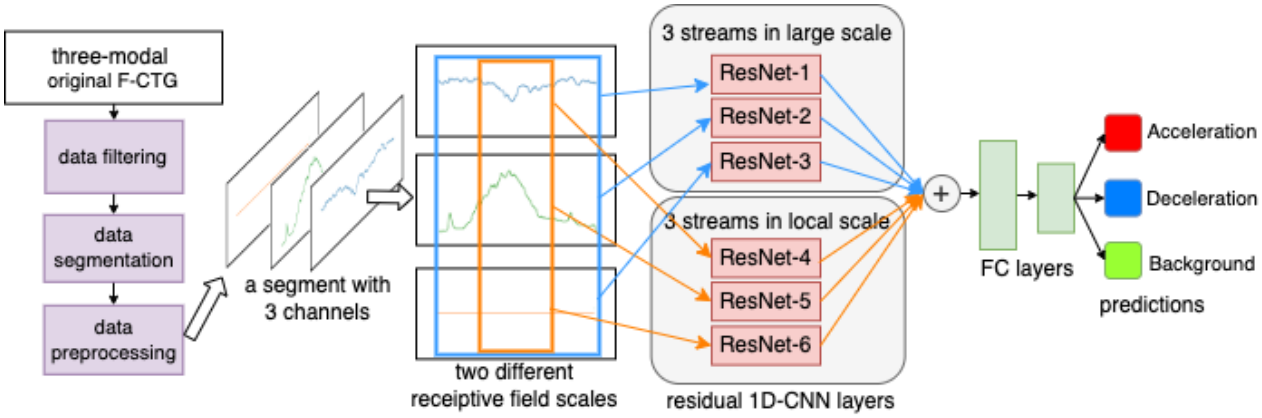

Figure 5.1 The structure of the proposed scheme.

training difficulties of these three classes. We found that the focusing parameter $\gamma = 2$ of Focal loss works best in our experiments.

A residual module inspired by residual neural network (ResNet) [63] was implemented in this study, based on the idea of predicting residuals using shortcut connections. When building ultra-deep networks, the straight stacking of numerous original residual learning modules will result in a parameter explosion, and also may cause a gradient dissipation problem, which would prevent convergence during network training and result in disadvantaged prediction performance. To address this issue, the ResNet model was modified significantly from a vanilla CNN structure, through identity mapping produced by the skip connection technique. This ResNet structure aims to minimize the number of parameters of the deep network without sacrificing accuracy [64]. Recent research works have proven the efficacy of ResNet structure on medical image tasks with 1D-CNN, such as ECG [65] and chest X-Ray [66]. As our network structure is deep, the use of residual blocks (ResBlock) improves the performance by handling residual error with residual blocks without causing vanishing or explosion of the gradient.

The core architecture of our residual module, the Residual Block, is illustrated in Figure 5.2. As in the previous model, the input x is fed to a convolution with a ReLU activation to obtain transformed features $F(x)$. However, the input of the block is then added to these features to compute the final output, $F(x) + x$.

The number of residual blocks is an important design choice for our architecture. A CNN with too few residual blocks lacks the capacity to model complex patterns, whereas using too many may lead to overfitting. Based on cross-validation experiments comparing networks with 1, 2, and 3 residual blocks shown in section 5.4.3, we empirically determined that the

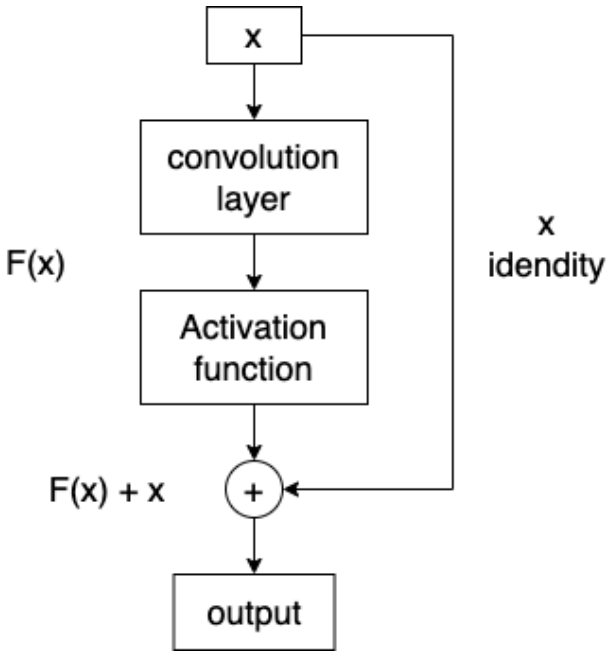

Figure 5.2 The structure of the Residual Block.

architecture with two residual blocks strikes the optimal balance between model capacity and generalization performance.

5.2.2 Multi-stream fusion strategies

Multi-stream fusion strategies are widely used in multi-modal tasks, as reviewed in the study [67]. The basic design is to feed multiple data streams to separate feature extractors so that they can learn customized patterns. Then, a concatenate operation is employed to merge the streams into a single output tensor, which is the input of the subsequent pattern recognition modules. We designed and evaluated four different multi-stream fusion approaches, which are illustrated in Figure 5.3, named based on where the fusion occurs: input-fusion, early-fusion, intermediate-fusion and late-fusion.

Input-fusion

The input-fusion strategy, illustrated in Figure 5.3 (a), represents the simplest approach. Analogously to how CNNs combine information from RGB channels in image recognition tasks, our model fuses the Fetal-CTG signal from the FHR, UC, and FM channels as distinct components of the input tensor. The combined tensor was then fed into the unique convolutional block,

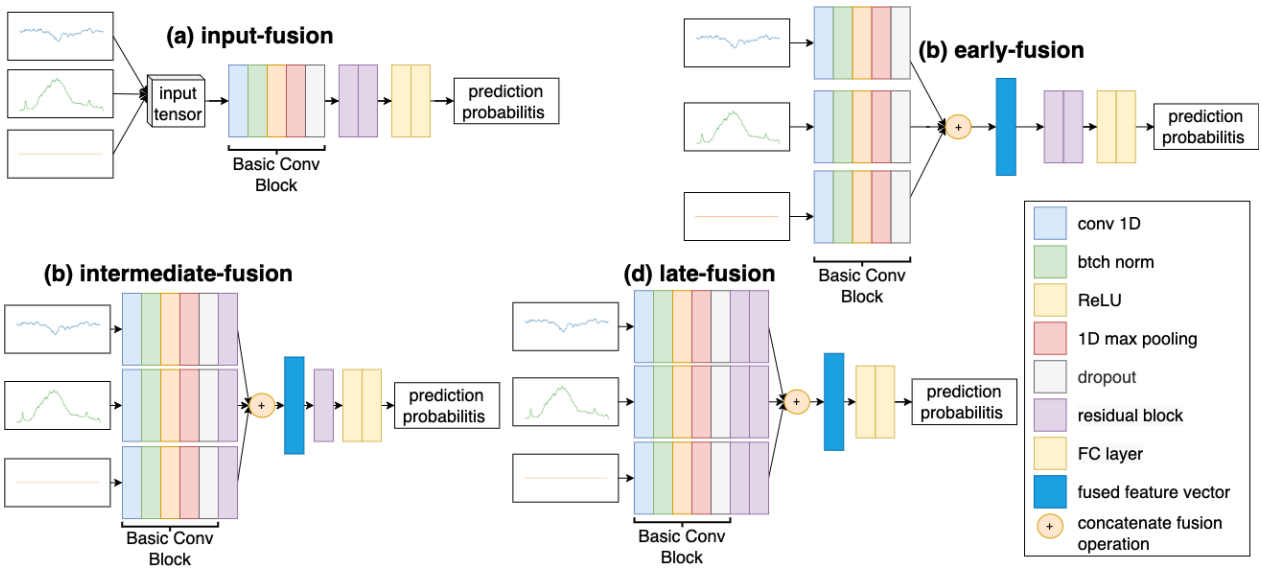

Figure 5.3 Architecture of the four fusion strategies: (a) input-fusion, (b) early-fusion, (c) intermediate-fusion, and (d) late-fusion.

Early-fusion

The second approach, depicted in Figure 5.3 (b), merges the three signal channels after the first convolutional block. The resulting tensor is then fed into the first residual block. In this approach, individual convolutional blocks function as modality-specific feature extractors applied to separate channels.

Intermediate-fusion

In the intermediate-fusion strategy, the three channels are combined into one tensor after the first residual block. The output tensor then enters the second residual block. Thus, each signal channel has its convolutional and residual block to extract modality-specific features. Refer to Figure 5.3 (c) for the architecture.

Late-fusion

The late-fusion strategy combines the three signal channels after the second residual block, feeding the output tensor to a fully-connected layer. This architecture, with each feature extractor having its convolutional block and two residual blocks, is displayed in Figure 5.3 (d).

5.2.3 Multi-stream fusion for a larger receptive field

To better exploit the target segment, we also explored a multi-scale fusion strategy with larger input receptive fields. The main idea of this strategy is to include the data around the target segment as a new stream with a wider receptive field. Since the original and surrounding streams have different scales, they cannot be directly aligned in a single time-axis stream. Thus, these multi-scale signals were treated as different streams before data fusion, as shown in Figure 5.1.

To enlarge the stream scale, the width of the input time window was extended by once or twice its size to the left and right, resulting in a total length of 3 or 5 times the original width. For each multi-stream model, there are two streams representing the target data segment and the surrounding data segment, respectively. It is important to note that the labels of the target segments are not impacted by the surrounding information, and the large-scale streams only provide a greater receptive field.

5.2.4 Training process

For the training phase, weights of the CNN model are optimized through back-propagation using Focal loss. The loss on the training and validation sets are used to decide the number of training epochs. If the gap between the training and the validation losses increases, overfitting occurs. When this happens, the training is stopped and the weights are saved. The model's accuracy is then calculated on testing set. All models are trained with 65 epochs, the Adam optimizer, and a batch size of 64. The learning rate was initialized to 0.1 and reduced by 0.001 every 15 epochs.

5.3 Data sets

We used our private large scale dataset, the 7AH-SYSU dataset.

In our dataset, the three measured signals are fetal heart-rate (FHR), uterine contractions (UC) and fetal-movement (FM). The sampling rate for each recording is 8 frames per second. The first pre-processing is to preliminary exclude obvious interference. We followed the standard filtering technique in this domain [16, 68], which exclude FHR channels containing a gap of 25 bpm between two successive frames.

A sliding window with 512 frames length and 25 frames stride length was employed to divide the data into segments. For each segment, we calculate the proportion of the total count of A/D classes. If the proportion of A/D is less than or equal to 30%, we label the segment as a

background class. Otherwise, we label the segment as the class with the highest proportion within the segment. Following is the detailed calculation. Let W be the window size for segmenting the data. For each segment i, we calculate as the proportion of non-background classes as follows equation 5.2:

$$P_i = \frac{\sum C_i}{W} \tag{5.2}$$

where C_i represents the count of non-background classes in segment i, P_i is the proportion of non-background classes.

If $P_i \leq 0.3$, the segment is labeled as the background class; otherwise, the segment is labeled as one of acceleration or deceleration who is with the highest proportion within the segment as following equation 5.3:

$$class_i = \begin{cases} 0 & if P_i \leq 0.3 \\ argmax_j(C_{ij}) & otherwise \end{cases} \tag{5.3}$$

where C_{ij} is the count of class j in segment i, and $j \in$ {Acceleration, Deceleration}

The number of samples in the background class is much larger than the other two classes. To re-balance the background and A/D data points, we use a re-sampling technique. The background class is down-sampled to the same amount as acceleration class. Then the deceleration class is over-sampled to the same amount as acceleration class. The amount of A remains the same.

Based on the domain knowledge of clinicians, the small high-frequency signals are mostly noise generated by equipment. In actual clinical practice, doctors pay more attention to the low-frequency overall trend change of the signal. Therefore, a widely used low pass filter, Savitzky-Golay filter [69] was used to denoise and smooth the time series data. The window-length of the filter was set to 25 and the order of the filter's polynomial was set to 3. Figure 5.4 compares the same data segment with and without smoothing, each of the three curves representing a different fetal-CTG channel.

The original data has three channels, two of which are integer values and one binary. They need to be standardized and normalized to ensure the stability of the model training. We standardize the data of each channel separately with statistical mean and standard deviation, and then normalize it to the $[-0.5, 0.5]$ range.

After data standardization and normalization, we used data augmentation techniques to increase and diversify the data for training, a common practice to help reduce overfitting.

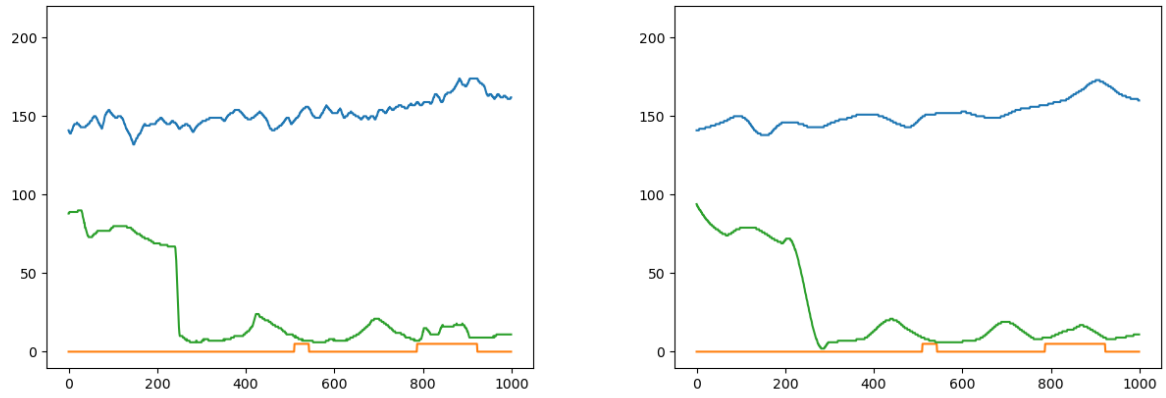

Figure 5.4 Example of a Fetal-CTG data segment with (left) and without (right) smoothing processing. The X-axis represents the time axis of Fetal-CTG (frame). The Y-axis represents fetal heart rate (bpm). The blue, green and orange curves are FHR, UC and FM respectively.

We adopt three widely used signal processing data augmentation approaches, Gaussian noise augmentation, vertical random panning augmentation and random cropping augmentation. The Gaussian noise augmentation is to add Gaussian-distribution based random noise values over the origin data values. The standard deviation of the Gaussian distribution we used is 0.1. The vertical random panning augmentation is to add a random value over all the values of each data point. The overall signal looks like it is raising or lowering. The range of the random value is from -0.1 to 0.1. The random cropping augmentation is to compress or stretch the input data by a random range on the time axis. If compression processing is performed, the compressed data will be randomly filled with 0 before and after, respectively, to resize back to the length of the input data. If stretching is performed, the stretched data is cropped at random positions to the length of the input data. In our experiments, we randomly compress or stretch 10% of the length of the input data. These data augmentation approaches are only applied on the training set, not on the validation or testing set.

5.4 Experiments

5.4.1 4-fold cross-validation and evaluation

Sensitivity, specificity, accuracy and F1-score of the models are evaluated.

For all experiments, we use inter-patient 4-fold cross-validation to obtain unbiased measures of performance. In each cross-validation fold, we use the data corresponding to 75% of patients for training (90%) and validation (10%), and the remaining 25% of the data for

testing. Table 5.1 shows the number of data cases in the training, validation and testing sets in each cross-validation fold.

5.4.2 Experiments on the number of feature maps

The number of feature maps of each convolutional layer is a critical hyperparameter, as it directly impacts the size of the model and its learning capacity. It is important to note that larger does not necessarily mean better. While the number of feature maps increases, both the training and inference times will augment. Moreover, having excessive capacity for the amount of available data can lead to overfitting, which can ultimately harm the performance of the model.

For the set up of this experiment, we consider the input-fusion architecture with three residual blocks. For simplicity, the same number of feature maps is used in all convolution layers.

Table 5.2 shows the results of this experiment. We find that, as the number of feature maps increases (up to 256), performance also improves. The best performance, corresponding to a F1-score of 82.19%, is achieved when using 256 feature maps. This model size is optimal, as it provides a good balance between training time and performance. It is neither too small, which may result in poor performance, nor too large, which would be time-consuming to train. We thus keep this setting for all following experiments.

For drilling deeper into the accuracy of each class, the confusion matrix of this model for a sampled cross validation fold is presented in Figure 5.5. The Acceleration class exhibits a near-perfect recognition accuracy of 98.65%, while the Background class has the lowest recognition accuracy at 78.38%. Moreover, the incorrect detection between the Background class and Deceleration class is significantly higher than the other class combinations. However, the significant accuracy differences between classes are consistently underwhelming.

5.4.3 Experiments of multi-stream fusion strategies

Next, we investigate the performance of various fusion strategies with the same network architecture. We explore the performance of different multi-stream fusion strategies with a varying number of residual blocks and fusing depths. We compare these fusion strategies with the same network architecture for each single-stream signal method but without fusing. By doing so, we can estimate the specific improvement brought by the fusion strategies.

The performance of these approaches with one, two, and three residual blocks is shown in Table 5.3. Where for the setting with three residual blocks, reported in the third part of Table 5.3, intermediate-fusion-1 performs the fusion operation between the first and second

Fold	Training	Validation	Testing
1	5,118	954	561
2	4,635	1,170	789
3	5,226	921	486
4	4,920	627	1,086

Table 5.1 Number of data cases in training, validation and testing set in 4-fold cross-validation.

# feature maps	Sen (%)	Spe (%)	Acc (%)	F1 (%)
8	80.92	80.18	85.13	79.94
16	80.13	79.28	84.45	79.04
32	82.30	81.59	86.19	81.58
64	80.97	80.57	85.42	80.39
128	82.01	81.64	86.23	81.55
256	**82.72**	**82.32**	**86.73**	**82.19**
512	82.39	81.96	86.46	81.81

Table 5.2 F1-score results of different feature map numbers for ResNet 1D-CNN models.

residual blocks, while intermediate-fusion-2 places it between the second and third residual blocks.

Among the selections of a single signal, we see that the FHR signal gives the best performance, and the UC signal contributes the least. The observations are consistent with those of the clinician. The FHR signal is the primary signal channel for interpretation, whereas the UC and FM signals play supporting roles. It is almost impossible to detect Acceleration or Deceleration using the UC and FM signals independently.

Most importantly, our fusion approaches have been proven to improve model performance. In experiments involving one, two and three ResBlock architectures, the best fusion strategy improved the F1-score by 1.74%, 1.14%, and 0.86%, respectively, compared to using only the FHR signal. On average, the fusion strategies improved the F1-score by 1.25% compared to only the FHR signal channel method.

As we increase the number of residual blocks, the intermediate-fusion and late-fusion approaches which have similar accuracy and F1-scores outperform the input-fusion and early-fusion strategies. This is an intriguing finding, as it suggests that feature fusion with a higher degree of abstraction are more effective, which supports the idea that deeper neural network architectures can lead to better performance. Additionally, this finding implies that it is more effective to perform the fusion operation at feature vector layers with higher degrees

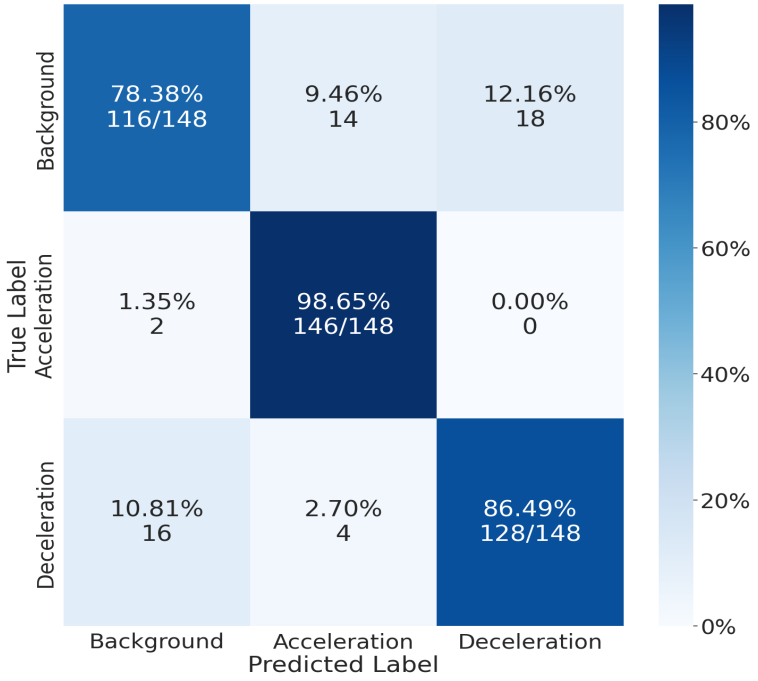

Figure 5.5 Confusion matrix of a representative cross-validation fold for the best-performing model, configured with 256 feature maps, three residual blocks, and the input-fusion strategy. This result is obtained from the experiment described in Section 5.4.2.

of abstraction, rather than at earlier layers with more primitive features. This finding is inspiring for the application of deep learning algorithms in the medical field and our future research.

Curves showing the loss values and accuracy of this model at each training epoch are given in Figures 5.6 and 5.7, respectively. After the fifteenth and the thirty epoch, three curves all converged more stable because of learning rate decay. The test curves are always close to the validation curves, which means that our validation set can well reflect the objective experimental results. The training accuracy curve quickly approaches around 95%, while the verification curve and test curve basically stay at around 80% after 30 epochs. This shows that the overfitting problem of the current model is still challenging.

As a conclusion of this experiment, among these fusion approaches, the two-residual-block intermediate-fusion achieved the best performance with F1-score of 83.83%, accuracy of 88.00%, sensitivity of 84.47%, and specificity of 83.98%. This F1-score result is 1.79% better than the previous experiment, which directly uses the input-fusion approach. Based on these results, we use the two-residual-block intermediate-fusion approach in subsequent experiments.

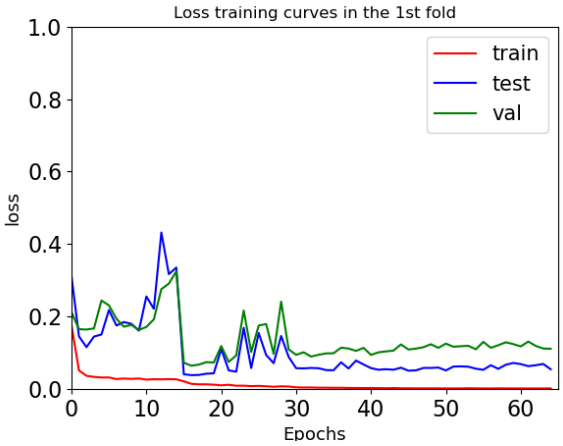

Figure 5.6 Training (red), validation (blue), and testing (green) loss of the best ResNet 1D-CNN model with 256 feature maps during training. The curves correspond to a representative training run and are provided to illustrate the optimization dynamics and convergence behavior.

5.4.4 Experiments of the larger input receptive field approache

In table 5.4, we compare the performance of the previously introduced multi-stream fusion strategies with 3× and 5× larger input receptive fields. Only the 3× scale shows improvements compared to the original 1× scale. Furthermore, the 3× scale approach yields the highest F1-score of 84.62%. We can observe that the main advantage of the 3× scale approach is its higher sensitivity, which is 1.11% higher than the original 1× scale approach. In terms of specificity and accuracy, both approaches demonstrate similar performance.

These results suggest that the 5× scale approach, which involves aggressive downsampling along the time axis, may lead to the loss of high-frequency local information due to excessive compression of the input signal. By combining information from the original scale and the 3× scale inputs, the proposed multi-stream fusion approach effectively preserves both precise local features and broader contextual information, thereby enhancing overall classification accuracy.

For drilling deeper into the accuracy of each class, the confusion matrix of this model for the third cross validation fold is shown in Figure 5.8. The high false negative rate of the Deceleration class is the primary challenge. The confusion matrix indicates that 14.6% of Decelerations are incorrectly classified as the Background class, which is the highest false negative rate in the matrix. We believe that is because of the nature difficulty and complexity of the signal waveform in the Deceleration class. Additionally, the Deceleration class exhibits high intra-class heterogeneity. To address this issue, we plan to develop a customized data

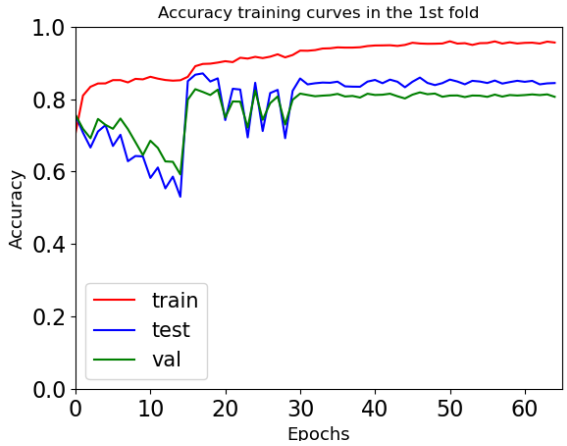

Figure 5.7 Training (red), validation (blue), and testing (green) accuracy of the best ResNet 1D-CNN model with 256 feature maps during training. This figure shows a representative run and is intended for qualitative analysis of training dynamics.

augmentation method for the Deceleration class, which will be the focus of our next phase of research.

Curves showing the loss values and accuracy of this model at each training epoch are given in Figures 5.9 and 5.10, respectively. Compared to not using the multi-scale approach, we found that the overfitting problem is alleviated when using it. In particular, when comparing Figure 5.6 and Figure 5.9, we observe that the test curve for the experiment using the multi-scale approach is closer to the training curve. While the test and validation curves in Figure 5.9 exhibit larger fluctuations than the previous experiment, they quickly become stable after the first learning rate decay at the fifteenth epoch. Overall, these figures indicate that second half training process with the 3-time-scale model is easier to converge, more accurate, and more stable.

5.4.5 Comparative experiments with state-of-the-art

We compared our method to the current state-of-the-art (SOTA) presented in the paper [16] for A/D recognition in Fetal-CTG. We meticulously re-implemented, fine-tuned, and evaluated the SOTA method on our dataset.

The state-of-the-art method re-implementation details

We re-implemented the SOTA method on our dataset to enable a direct and objective comparison with our proposed model. Since the performance of our method and the SOTA

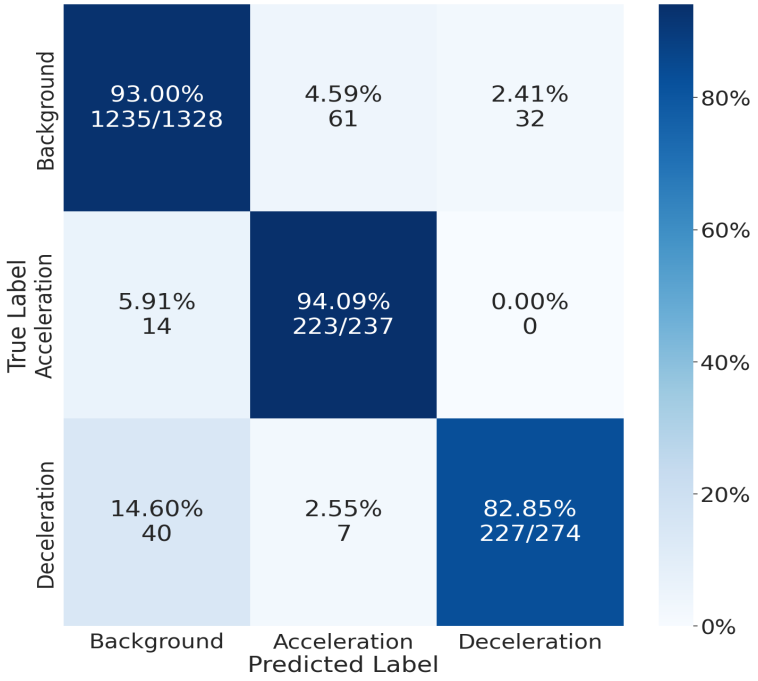

Figure 5.8 The confusion matrix of a sampled (the third) cross validation for the 3-time-scale model, whose performance is the best in this paper.

method was evaluated on different datasets, a direct comparison is not possible. Unfortunately, we did not have access to the private dataset used in the SOTA paper, hence the need for re-implementation. It is worth noting that re-implementation is a commonly accepted comparison method in the field, and numerous papers have used it to compare the accuracy of various previous methods.

We meticulously followed the method outlined in the SOTA paper to ensure a fair comparison with their approach. Specifically, the SOTA method firstly employed the Empirical Mode Decomposition (EMD) approach [70] to estimate the baseline from the FHR. Then, An acceleration or deceleration in a sliding window was detected using a threshold of 15 beats per minute (bpm). During the EMD process, the index of Intrinsic Mode Functions (IMFs) and Standard Deviation (SD) threshold are adjustable parameters. Through a rigorous parameter tuning process, we ultimately determined that the optimal parameters for our dataset were an index of IMFs equal to 2 and a SD threshold of 1×10^{-8}. The A/D detection rules were also fixed in the SOTA method, so we strictly adhered to the rules outlined in the paper, including the FHR threshold and the too-short A/D removal rule. After careful parameter fine-tuning, the re-implemented SOTA method achieved its best possible performance on our dataset.

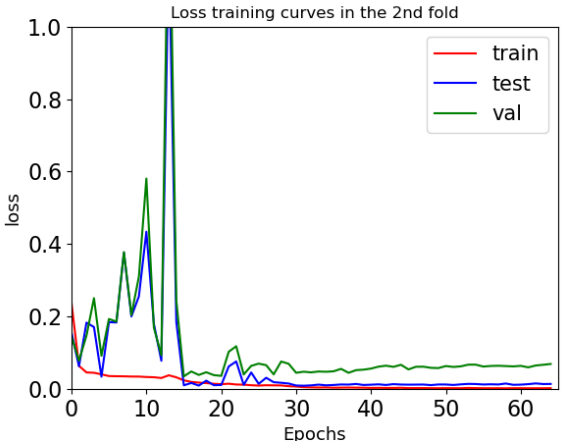

Figure 5.9 Training (red), validation (blue), and testing (green) loss of the proposed three-time-scale model during training. The curves correspond to a representative fold and illustrate the convergence behavior of the model.

Comparative experiments on the three-classification task

In Table 5.5, we present a comparison of the performance of our proposed method and the SOTA method on the testing dataset. Our method outperforms the state-of-the-art approach, with significantly better average results, as demonstrated by an F1-score of 84.62%, accuracy of 87.92%, sensitivity of 85.58%, and specificity of 83.98%. Notably, our method achieves a 21.54% higher F1-score and a 24.92% higher accuracy than the SOTA method.

We would like to emphasize that the performance of our re-implemented SOTA method is relatively lower than what was reported in their original paper. Specifically, the reported SOTA performance in the paper was an F1-score of 71.50%, sensitivity of 90.00%, and specificity of 67.50%. However, even when compared to the reported SOTA performance, our proposed method achieved a significantly better F1-score, which is 13.12% higher. We take pride in our objective data collection and labeling process, which is completely isolated from the algorithm development process. As a result, we have a high degree of confidence in the objectivity and reliability of the test results that we have obtained.

To further investigate and compare the recognition accuracy of our proposed model and the SOTA method for acceleration and deceleration, we performed one-versus-all binary classification experiments. The details of these experiments will be discussed in the following subsection 5.4.5.

Figure 5.10 Training (red), validation (blue), and testing (green) accuracy of the proposed three-time-scale model during training. Only a representative training run is shown for clarity.

Comparative experiments on the one-versus-all binary classification task

We conducted one-versus-all binary classification experiments to compare the recognition accuracy of the models for A/D. Comparison with the state-of-the-art methods are shown in Table 5.6. For both the A/D classification tasks, our model achieves significantly higher F1-score, accuracy, sensitivity, and specificity than the state-of-the-art method. Regarding the acceleration classification task, our method achieves better results with an F1-score of 88.08%, accuracy of 89.62%, sensitivity of 84.47%, and specificity of 88.52%, outperforming the state-of-the-art by 13.75% in F1-score and 15.21% in accuracy. For the deceleration classification task, our method achieves better results with an F1-score of 87.69%, accuracy of 87.11%, sensitivity of 88.52%, and specificity of 87.11%, outperforming the state-of-the-art by 15.47% in F1-score and 14.54% in accuracy.

It is worth highlighting that the SOTA method's F1-score for Deceleration is 2.11% lower than that for Acceleration. In contrast, our model's F1-score for Deceleration is only 0.39% lower than that for Acceleration. This demonstrates that our model's discrimination of Acceleration and Deceleration is lower, while the SOTA method's identification ability for Deceleration is significantly lower than its ability to identify Acceleration.

Our model shows significant superiority over the SOTA method in both the three-class classification tasks and the one-versus-all binary classification task. This is due to the fact that the SOTA method's A/D judgment relies on designed rules, which lack the necessary specificity and meticulousness. To examine the generality of our model, a comparative experiment on

Approach	**Sen (%)**	**Spe (%)**	**Acc (%)**	**F1 (%)**
One ResBlock:				
only FHR signal	80.83	80.45	85.33	80.30
only UC signal	45.26	47.32	60.49	44.22
only FM signal	51.37	50.90	63.18	49.99
Input-fusion	**82.39**	82.06	86.55	**82.04**
Early-fusion	82.32	**82.20**	**86.65**	80.50
Late-fusion	79.74	80.49	85.37	80.32
Two ResBlock:				
only FHR signal	83.43	83.03	87.27	82.84
only UC signal	44.94	46.00	59.50	43.01
only FM signal	53.60	53.34	65.01	52.62
Input-fusion	83.09	82.61	86.99	82.53
Early-fusion	82.87	82.65	86.98	82.54
Intermediate-fusion	**84.47**	83.98	**88.00**	**83.98**
Late-fusion	84.26	**84.00**	**88.00**	83.91
Three ResBlock:				
only FHR signal	83.57	83.12	87.34	82.97
only UC signal	43.63	46.01	59.51	43.49
only FM signal	54.42	53.57	65.18	52.64
Input-fusion	82.87	82.32	86.73	82.19
Early-fusion	82.29	81.88	86.40	81.78
Intermediate-fusion 1	83.10	82.70	87.03	82.65
Intermediate-fusion 2	83.10	82.70	87.65	83.43
Late-fusion	**84.29**	**83.93**	**87.96**	**83.83**

Table 5.3 Performance of single-signal approaches and fusion strategies with one, two and three residual block.

scale times	**Sen (%)**	**Spe (%)**	**Acc (%)**	**F1 (%)**
1× scale	84.47	**83.98**	**88.00**	83.98
3× scale	**85.58**	83.93	87.92	**84.62**
5× scale	81.77	74.38	80.92	77.03

Table 5.4 Average accuracy and F1-score of different surrounding scales.

CTU-UHB dataset [24] is provided in the following subsection5.4.6.

Methods	**Sen (%)**	**Spe (%)**	**Acc (%)**	**F1 (%)**
state-of-the-art	63.28	63.37	63.00	63.08
Ours	**85.58**	**83.93**	**87.92**	**84.62**

Table 5.5 Average accuracy and F1-score results of our presented method and the state-of-the-art method.

Methods	**Sen (%)**	**Spe (%)**	**Acc (%)**	**F1 (%)**
Acceleration:				
state-of-the-art	74.69	74.40	74.41	74.33
Ours	**84.47**	**88.52**	**89.62**	**88.08**
Deceleration:				
state-of-the-art	73.78	72.57	72.57	72.22
Ours	**88.52**	**87.11**	**87.11**	**87.69**

Table 5.6 Average accuracy and F1-score results of our single-class acceleration and deceleration classification results versus state-of-the-art results.

5.4.6 Generalization on the Public CTU-UHB Dataset

To further assess the generalization capability of our model, we conducted a comparative experiment on the publicly available CTU-UHB dataset [24], using the expert-annotated labels from the work [71]. Unlike our in-house dataset, the labeled CTU-UHB dataset only includes two channels: FHR and UC. To accommodate our model structure, we set the FM channel input to zero during both training and inference.

Table 5.7 presents the overall multi-class classification results. Our method achieves an F1-score of 65.91% and an accuracy of 74.25%, significantly outperforming the state-of-the-art rule-based method, which achieves an F1-score of 57.18% and an accuracy of 60.41%. This corresponds to a relative improvement of approximately 9% in F1-score and 14% in accuracy.

We also evaluated the single-class (one-vs-all) performance for acceleration and deceleration detection tasks. The results are presented in Table 5.8. For acceleration detection, our model achieved an F1-score of 75.48%, which is 18.21% higher than the state-of-the-art method. For deceleration detection, our model slightly outperforms the state-of-the-art with an F1-score of 76.42% versus 71.34%. These results further validate the robustness and generalizability of our method across datasets and clinical contexts.

Overall, our model demonstrates strong cross-dataset generalization and confirms its capability to outperform existing methods even when adapted to different signal modalities and

Methods	**Sen** (%)	**Spe** (%)	**Acc** (%)	**F1** (%)
State-of-the-art	57.05	57.31	60.41	57.18
Ours	**66.17**	**65.65**	**74.25**	**65.91**

Table 5.7 Average accuracy and F1-score results of our method and the state-of-the-art method on the labeled CTU-UHB dataset.

Methods	**Sen** (%)	**Spe** (%)	**Acc** (%)	**F1** (%)
Acceleration:				
State-of-the-art	42.07	89.69	71.09	57.27
Ours	**77.93**	**74.60**	**74.60**	**75.48**
Dcceleration:				
State-of-the-art	78.02	65.71	71.54	71.34
Ours	**78.54**	**75.78**	**75.78**	**76.42**

Table 5.8 Binary classification results for acceleration and deceleration detection on the labeled CTU-UHB dataset.

clinical annotation standards.

5.5 Discussion

Regarding Fetal-CTG AAMs, while end-to-end diagnostic approaches are now considered the mainstream technical direction, we believe that a guideline-aligned independent full-feature classifier, which can recognize clinical features including Baseline, A/D, Variability, and Sinusoidal Pattern, still has significant clinical value. Because it can greatly help clinicians avoid neglecting key feature signals during busy daily routine works while improving their feature interpretation consistency simultaneously. Among all these clinical features, the specificity of A/D is poor, leading to significant inter- and intra-observer variability in clinical practice. The A/D detection methods have not improved substantially for a long time. In this study, our experimental results significantly outperform the state-of-the-art algorithm. The same method can also be implemented for variability detection, and subsequently, the Baseline and Sinusoidal Pattern can be computerized. This study can be considered as the key foundational method research of the full-feature classifier.

Regarding the prospects of the dataset, we only needed to annotate 667 Fetal-CTG samples to achieve good accuracy and interpretation effect. This not only demonstrates the effectiveness of deep 1D-CNN in dealing with the A/D detection task, but also serves as a good reference

for the amount of data that must be prepared for future improvements to this A/D detection model and the development of classifiers for other similar Fetal-CTG features.

We understand that each sub-feature of A/D has its potential clinical interpretation, but it is difficult to collect and annotate enough data for them due to their naturally low frequency (especially variable deceleration and prolonged deceleration). Therefore, for the early-stage research that aims to test the capacity of our customized method, we decided to combine all sub-features of deceleration as a single deceleration class in the current dataset version. To our best knowledge, the related works are mainly based on single hospital's dataset, as different hospitals have various devices based on different standards for collect CTG/FHR data collection, and it is difficult to align/reconcile all the data.

In the next stage of research, we plan to upgrade our dataset to collect enough data and annotations for necessary Fetal-CTG features and sub-features. We will also develop an ensemble algorithm that combines state-of-the-art deep learning and machine learning technology to explore the full potential of Fetal-CTG full-feature automatic analysis.

5.6 Conclusion

We proposed a multi-modal, multi-scale 1D-CNN method based on ResNet architectures to solve the task of Fetal-CTG A/D recognition. To train this data-driven CNN model, we built a dataset with A/D annotations that were cross-reviewed by highly qualified obstetricians. The experimental results of our model outperformed previous state-of-the-art methods. Our next steps include upgrading the dataset with necessary features' data collection and animation, and completing the Fetal-CTG full-feature automated classification. Our vision is to develop an automatic analysis of Fetal-CTG and extend it to a decision support system for clinicians.

CHAPTER 6 FNO-AUGUNET: A FOURIER-NEURAL-OPERATOR-AUGMENTED 1D-UNET FOR FETAL HEART RATE ACCELERATION AND DECELERATION DETECTION

6.1 Introduction

6.1.1 Background

FHR monitoring remains a cornerstone of intrapartum care, and acceleration/deceleration patterns are widely recognized as key indicators of fetal well-being. As detailed in earlier chapters, these events reflect autonomic regulation and oxygenation status, providing clinicians with actionable insight into potential distress. Building on this physiological foundation, this chapter focuses on improving the algorithmic modeling of such patterns. Specifically, we aim to capture both the localized morphology and the global temporal context of FHR signals through a dual-stream representation, where the auxiliary Fourier Neural Operators (FNO) [72] stream models global dependencies in the frequency domain, extending beyond conventional single-branch architectures.

6.1.2 Limitation of Existing Methods

State-of-the-art models based on 1D-UNet [35–39] have demonstrated strong performance in modeling localized temporal patterns of FHR signals. However, their convolutional receptive fields remain inherently limited, making it difficult to capture long-range temporal dependencies. Conversely, models that emphasize global or frequency-domain representations tend to overlook fine-grained morphological variations critical for identifying transient events. As a result, existing single-stream architectures struggle to jointly model both local detail and global context. This limitation motivates the design of a dual-stream local–global architecture, in which a FNO serves as a global augmentation stream to complement the 1D-UNet's localized feature extraction.

6.1.3 Motivation for Fourier Neural Operators

Modeling the global temporal behavior of FHR signals requires mechanisms that can capture long-range dependencies beyond the limited receptive fields of convolutional filters. In this context, spectral representations provide a complementary perspective to the raw time-domain signal, as they implicitly encode periodicities, oscillatory dynamics, and correlations

that extend over large temporal scales.

FNO are particularly well suited for such global modeling because they perform learnable transformations in the frequency domain via the Fourier transform [73, 74]. By parameterizing a subset of Fourier modes with complex-valued weights, FNO can efficiently capture global dependencies that conventional convolutional or recurrent encoders approximate only through deep stacking. Here, each Fourier mode corresponds to a specific temporal frequency component of the signal, and learning over these modes allows the network to directly model interactions across all time scales. These modes are obtained by applying the Fourier transform to the input representation, decomposing the signal into sinusoidal basis functions of different frequencies, among which only a limited set of dominant modes are retained and parameterized. Furthermore, the use of the Fast Fourier Transform (FFT) yields favorable computational efficiency and parameter economy compared with deep convolutional networks. Building on these properties, we employ an FNO stream as a global feature augmentation pathway to enhance the contextual modeling capability of the 1D-UNet.

We therefore propose a 1D-UNet with an FNO-Augmented Stream, hereafter referred to as **FNO-AugUNet**, where the 1D-UNet serves as the main branch for local morphological feature extraction, while the auxiliary FNO stream enriches global temporal representations directly from the same time-domain input.

6.1.4 Contributions

In this chapter, we advance the state of FHR A/D detection by introducing a dual-stream local–global framework and systematically validating its effectiveness. Our main contributions are summarized as follows:

1. We propose a dual-stream architecture that integrates a 1D-UNet for local morphological modeling with an FNO-based global augmentation stream operating directly on the same FHR input.

2. To the best of our knowledge, this is the first work to introduce the Fourier Neural Operator method to fetal CTG–related recognition tasks, demonstrating its effectiveness in improving overall recognition performance.

3. We evaluate the proposed FNO-AugUNet against state-of-the-art methods on both a benchmark public dataset and a large-scale private dataset, demonstrating consistent improvements in event-level and time-step-wise F1 performance.

Together, these contributions highlight the potential of dual-stream local–global modeling, which is implemented via an FNO-augmented 1D-UNet, to deliver more reliable and interpretable decision support in intrapartum fetal monitoring.

6.2 Method

6.2.1 Architecture of FNO-AugUNet

The overall architecture of the proposed FNO-AugUNet framework is illustrated in Figure 6.1. The model processes the FHR signal through two parallel streams operating on the same input:

- **1D-UNet Stream.** The raw FHR sequence is directly fed into a 1D-UNet, which captures local morphological patterns, baseline variations, and transient acceleration/deceleration features in the temporal domain. This design follows the paradigm adopted by recent state-of-the-art FHR detection models that employ 1D-UNet backbones for time-domain feature extraction.
- **FNO Stream.** In parallel, the FHR sequence is processed by FNO layers. This stream performs spectral transformations internally to capture global temporal dependencies and long-range correlations that extend beyond the convolutional receptive field of the UNet.

The feature maps from both streams are temporally aligned and fused through a summation fusion module. The fused representation then produces per-timestep predictions of acceleration and deceleration events.

This dual-stream design leverages the complementary strengths of the two pathways: the 1D-UNet Stream provides fine-grained temporal sensitivity, while the FNO Stream enhances global contextual modeling. Together, they yield a unified representation capable of robust and interpretable fetal event detection.

6.2.2 The FNO Stream

Unlike convolutional encoders, whose receptive fields are limited to local neighborhoods determined by kernel size and network depth, the FNO Stream captures long-range temporal dependencies by performing spectral transformations within the network. This enables global information exchange across distant time points, which conventional convolutions can only approximate through deep stacking.

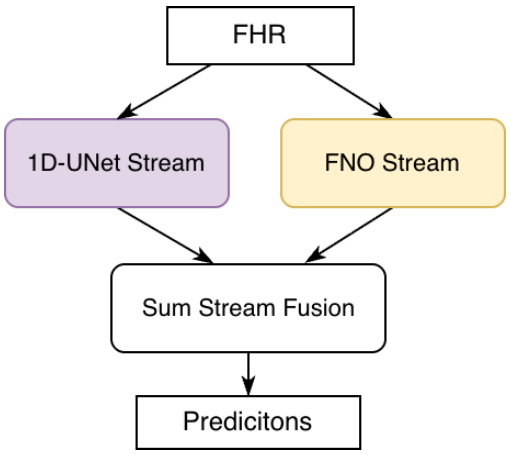

Figure 6.1 Architecture of the proposed FNO-AugUNet, a dual-stream local–global framework that integrates a 1D-UNet for time-domain local feature extraction and an FNO-based stream for implicit frequency-domain global modeling.

The architecture of the FNO Stream, which serves as the global modeling branch of the proposed dual-stream framework, is illustrated in Figure 6.2. Given the raw FHR signal $X \in \mathbb{R}^{C_{\text{signal}} \times T}$, where C_{signal} is the number of input channels (set to 1 in our implementation, as only the FHR signal is used) and T is the temporal length, the signal is directly processed by a sequence of FNO layers. Each FNO layer applies a Fast Fourier Transform (FFT) to convert the temporal representation into the frequency domain, parameterizes a subset of Fourier modes with learnable complex weights, and transforms them back to the temporal domain via an inverse FFT (IFFT). Through this process, the FNO Stream can effectively model global temporal correlations.

FNO blocks. Each FNO layer consists of an FNO block followed by a series of residual-style operations. The input of each layer is first passed into its internal FNO block, which performs spectral convolution with truncated Fourier modes, allowing efficient global modeling of temporal dependencies.

Given the input $Z^{(l)} \in \mathbb{R}^{C \times T}$, the Fourier transform, denoted by $\mathcal{F}(\cdot)$, is applied along the temporal dimension to obtain the frequency-domain representation:

$$\hat{Z}^{(l)} = \mathcal{F}(Z^{(l)}) \in \mathbb{C}^{C' \times T} \tag{6.1}$$

where $Z^{(l)}$ represents the input to the l-th FNO layer. At the first layer (where $l = 1$), $Z^{(1)} \in \mathbb{R}^{1 \times T}$ corresponds to the raw FHR signal, while for deeper layers (where $l > 1$), it corresponds to the output of the previous FNO layer, i.e., $Z^{(l)} = h^{(l-1)}$. For simplicity, the layer index (l) is omitted in the subsequent equations.

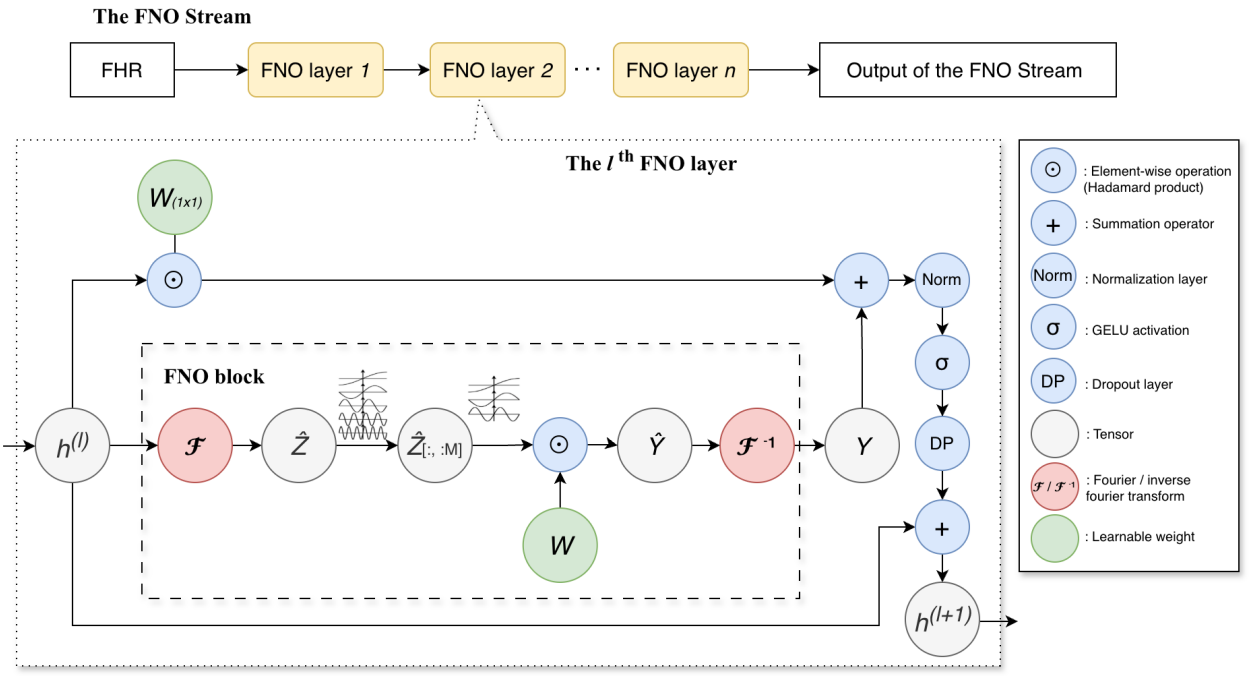

Figure 6.2 The structure of the FNO stream.

Only the lowest M Fourier modes are retained to capture the dominant temporal patterns while suppressing high-frequency noise. The spectral convolution is then defined as:

$$\hat{Y}_{[:,m]} = W_{[:,m]} \odot \hat{Z}_{[:,m]}, \quad m = 1, \ldots, M \tag{6.2}$$

where $W_{[:,m]}$ denotes the complex-valued learnable coefficients for the m-th mode, and $\odot$ represents element-wise complex multiplication. Let $\hat{Y} \in \mathbb{C}^{C' \times T}$ be the zero-padded matrix that places $\{\hat{Y}_{[:,m]}\}_{m=1}^{M}$ at the retained frequencies and zeros elsewhere.

Finally, the inverse Fourier transform, which is represented as $\mathcal{F}^{-1}()$ in the equation, reconstructs the output $\hat{Y}$ in the temporal domain:

$$Y = \mathcal{F}^{-1}(\hat{Y}) \in \mathbb{R}^{C' \times T} \tag{6.3}$$

After the inverse Fourier transform, the spectral output is denoted as $Y^{(l)}$, serves as the input to the subsequent residual and nonlinear transformation.

Although the operation described above involves an element-wise multiplication in the frequency domain rather than a conventional sliding-window convolution in the time domain, it is referred to as a spectral convolution because of the convolution theorem [75, 76]: a con-

volution in the temporal domain is equivalent to a pointwise multiplication in the frequency domain. By learning complex-valued coefficients $W_{[:,m]}$ for each retained Fourier mode, the FNO block effectively performs a learnable global convolution over the entire sequence.

This operation enables the encoder to efficiently model long-range dependencies through spectral representations while maintaining linear computational complexity with respect to sequence length. In our experiments, two stacked FNO layers and $M = 96$ retained modes provided the most stable performance.

Residual and nonlinear transformation. Each FNO block is combined with a point-wise 1×1 convolution ($W_{1\times1}$), normalization ($Norm$), and nonlinearity in a residual fashion:

$$h^{(l)} = h^{(l-1)} + \sigma\Big(\mathrm{Norm}(Y^{(l)} + W_{1\times1}h^{(l-1)})\Big) \tag{6.4}$$

Here, $h^{(l-1)} \in \mathbb{R}^{C'\times T}$ represents the temporal representation propagated from the previous FNO layer. The term $Y^{(l)} \in \mathbb{R}^{C'\times T}$ corresponds to the output of the FNO block within the same l-th FNO layer, which encodes global dependencies by aggregating information across the entire temporal sequence. $\sigma(\cdot)$ denotes GELU activation and dropout regularization. A stack of L such blocks forms the FNO stream.

This residual formulation enables stable training and facilitates the flow of frequency-domain information across layers. By preserving the identity mapping of the input through skip connections, the network mitigates vanishing gradient issues and supports deeper spectral modeling. The inclusion of point-wise convolution and normalization further enhances representational flexibility by allowing channel-wise mixing and adaptive feature scaling before the nonlinear transformation. Together, these design choices ensure that each FNO block not only captures global frequency-domain correlations but also refines them through localized, learnable adjustments in the temporal domain. Consequently, the resulting FNO stream achieves a balanced representation—retaining the efficiency of spectral convolution while preserving the fine-grained temporal dynamics essential for accurate fetal heart-rate pattern analysis.

Output formulation. The encoder output is mapped to class logits by a final 1×1 convolution: $\ell_{\mathrm{spec}} = W_{\mathrm{head}}h^{(l)} \in \mathbb{R}^{N_{\mathrm{cls}}\times T}$, where N_{cls} represents the number of classes. These logits represent per-timestep probabilities of acceleration/deceleration events in the spectral stream and are aligned with the outputs from the time-domain UNet for subsequent fusion.

Comparison to conventional encoders. Conventional convolutional encoders, such as 1D-UNet, rely on localized kernels and hierarchical downsampling to capture temporal dependencies. Their receptive field is inherently limited by the kernel size, so long-range tem-

poral interactions must be approximated by stacking multiple layers, which increases model depth and parameter count. In contrast, the proposed FNO encoder operates directly on the time-domain signal while implicitly performing spectral transformations to capture global temporal dependencies. Instead of sliding local kernels in time, it learns complex-valued filters applied to truncated Fourier modes, allowing each mode to aggregate information across the entire temporal sequence. This design enables efficient global modeling of temporal dynamics with significantly fewer layers and parameters, while maintaining linear computational complexity. For quasi-periodic FHR signals, such global frequency-domain representations are particularly effective in capturing long-term rhythmic patterns that conventional CNNs often fail to model.

6.2.3 The 1D-UNet Stream

The time-domain stream adopts the standard one-dimensional U-Net architecture, which has been widely established as a state-of-the-art baseline for biomedical sequence segmentation tasks, especially for the A / D detection task [35–39]. To ensure a fair comparison and isolate the effect of our proposed frequency-domain module, we employ the most basic configuration of the 1D-UNet without additional modifications.

Concretely, the model consists of an encoder–decoder hierarchy with symmetric skip connections. The encoder progressively down-samples the raw FHR sequence through convolutional layers with kernel size $k = 21$ and stride $s = 11$, followed by max-pooling layers of size 2. The number of channels doubles at each downsampling stage (in our case they are 32, 64, 128 and 256), enabling multi-scale representation of temporal features. The decoder mirrors this structure, performing upsampling by transposed convolutions and concatenating the corresponding encoder features via skip connections. All convolutional layers are followed by Batch Normalization, ReLU activations and a dropout layer. A final 1×1 convolutional layer produces per-timestep class logits: $\ell_{\text{time}} \in \mathbb{R}^{N_{\text{cls}} \times T}$, where N_{cls} represents the number of classes.

This architecture has been extensively validated in prior work as a baseline method for time-domain FHR analysis, capable of capturing local morphology, baseline shifts, and transient acceleration/deceleration events. In this proposed framework, it serves as the time-domain branch against which the proposed frequency-domain encoder is integrated and compared.

6.2.4 Summation Stream Fusion

Firstly, to ensure proper synchronization between the 1D-UNet and FNO streams, both outputs are temporally aligned on a shared timeline so that their feature representations correspond to the same temporal axis. Then, to combine the complementary information extracted from these two streams, we design a late-fusion-style logit summation fusion module that operates on the per-timestep logits.

Let $\ell_{\text{time}} \in \mathbb{R}^{N_{\text{cls}} \times T}$ and $\ell_{\text{spec}} \in \mathbb{R}^{N_{\text{cls}} \times T}$ denote the class logits produced by the 1D-UNet and the FNO stream, respectively. The two outputs are directly summed:

$$\ell_{\text{fused}} = \ell_{\text{time}} + \ell_{\text{spec}} \tag{6.5}$$

The fused logits ℓ_{fused} are directly provided to the loss function without an explicit softmax. Consistent with Section 4.2.4, the same GDL loss formulation is employed here to supervise the fused output, ensuring coherence in the optimization objectives across all experiments. This design preserves numerical stability and enables end-to-end optimization of the fused representation.

6.2.5 Model Training Details

The training procedure of the proposed dual-stream framework largely follows the configuration established in Chapter 4, with minor adjustments for stability. We adopt the cross-entropy loss function, which has been widely used in segmentation-style classification tasks due to its numerical stability and suitability for multi-class prediction. In our experiments, this choice consistently yielded the best performance among the tested alternatives, including Dice and F1-optimized losses.

For optimization, we employ the Adam optimizer [61], chosen for its efficiency and robustness in handling sparse and noisy gradients. The learning rate is set to 1×10^{-3}, with a batch size of 64 and training for 100 epochs. To improve generalization and mitigate overfitting, a dropout rate of 0.1 and 0.7 are applied within the 1D-CNN stream and the FNO stream, respectively. These values were determined empirically through validation experiments as the optimal configuration for balancing performance and stability. No additional weight decay is used, as dropout alone provided sufficient regularization in our experiments.

Early stopping was further employed based on the validation event-level F1 score (defined in Section 6.3.2) to prevent over-training and reduce variance across runs. This training protocol was applied consistently across all baseline and proposed models to ensure fairness

in comparison.

6.3 Experiments

6.3.1 Dataset and Data Pre-Processing

In this study, we evaluate the proposed FNO-AugUNet on two datasets: the publicly available CULF-DB and our private 7AH-SYSU dataset. Both datasets, including their annotation of acceleration and deceleration events, have been introduced in detail in earlier chapters. For consistency, we adopt the same annotation protocol and follow the identical train/validation/test split strategy previously described. This ensures fair comparison with prior models presented in this thesis.

Also, for the input FHR signal pre-processing (including resampling, artifact clipping, gap imputation, and per-recording normalization) has already been described in detail in Section 4.3.2 and is applied here without modification.

6.3.2 Evaluation Metrics

All reported results represent the median value of at least ten independent training runs. This multi-run evaluation strategy reduces randomness-induced bias and ensures that the observed improvements reflect consistent trends rather than chance variation. For each metric, the reported median value was computed within the 95% confidence interval (CI_{95}) of the repeated runs, following a robust estimation procedure to mitigate the influence of outliers.

We evaluate model performance using two complementary categories of F1-based metrics. Detailed definitions and computation procedures have been introduced in earlier chapters and are adopted here without modification. In this chapter, we report only F1 scores to avoid redundancy.

1. **Event-level F1.** Reported as *Event-level F1(%)* in the result tables, this metric evaluates the agreement between predicted and ground-truth A/D events.

Each predicted probability sequence from the model is first converted into discrete events through a standardized *post-processing pipeline*. Specifically:

- **Thresholding and removal of short segments:** Consecutive time steps predicted as the same class (acceleration or deceleration) are grouped into a preliminary event. Segments shorter than 3 s are discarded to suppress spurious detections.
- **Event merging:** For each class, two neighboring events are merged into a single event

if the temporal gap between them is less than 15 s. This prevents fragmented predictions of the same physiological episode.

Event definition: Each final event is represented by its onset time t_{start} and offset time t_{end}, where t_{start} corresponds to the first time step labeled as the given class, and t_{end} marks the last consecutive time step. The duration of an event is therefore $\Delta t = t_{\text{end}} - t_{\text{start}}$.

Two events (predicted and ground-truth) are considered to be in agreement if their time segments overlap for at least 5 s. Based on this matching rule, the number of true positives (TP), false negatives (FN), and false positives (FP) are counted across all fetal heart rate (FHR) recordings.

The event-level sensitivity (**Sen**) and positive predictive value (**PPV**) are then defined as:

$$\text{Sen} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \qquad \text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{6.6}$$

and the F1-score is computed as their harmonic mean:

$$\text{F1} = \frac{2 \times \text{PPV} \times \text{Sen}}{\text{PPV} + \text{Sen}}. \tag{6.7}$$

This event-based evaluation reflects clinical practice, where obstetricians interpret A/D patterns as discrete temporal episodes rather than time-step-wise fluctuations, ensuring that performance metrics align with clinical interpretability.

It is worth noting that, in rare cases, a predicted event may overlap with a ground-truth event by more than the predefined threshold (5 s), yet exhibit considerable temporal misalignment elsewhere (e.g., partial overlap at only one end). Such cases could theoretically exploit the overlap rule of the event-level metric. However, empirical inspection of all test recordings indicates that these situations occur only sporadically and have a negligible impact on the overall fairness and objectivity of the reported results.

This event-based criterion is the most widely adopted evaluation protocol in current state-of-the-art studies, despite its inherent limitations. We therefore employ it as the primary performance metric to ensure fair and direct comparison with existing methods. Nevertheless, to provide a more fine-grained and temporally precise assessment of model behavior, we further report the *time-step-wise F1* metric.

2. **Time-step-wise F1.** Reported as *Time-step-wise F1(%)* in the result tables, this metric evaluates the segmentation model's raw outputs at each temporal sample before any post-processing. In contrast to the event-level F1, which measures the correctness of clinically

meaningful acceleration and deceleration episodes after heuristic grouping, the time-step-wise F1 directly reflects how accurately the model distinguishes instantaneous physiological states at the frame level.

3. **Rationale for dual-level evaluation.** Existing studies on A/D recognition in fetal heart rate (FHR) analysis almost exclusively report event-level metrics—computed after merging and filtering predicted segments—because these events correspond more closely to obstetric interpretation. However, such post-processing inevitably conceals the temporal precision of the model's predictions. Small but consistent offsets between predicted and ground-truth event boundaries may exert little influence on event-level F1, yet they reveal systematic early or delayed responses in the model's detection behavior.

To address this limitation, we introduce a complementary time-step-wise F1 metric to quantify the model's fine-grained temporal precision and recall prior to heuristic smoothing. This comparison between event-level and time-step-wise performance provides a more complete understanding of model behavior:

- It disentangles clinical detection accuracy (event-level) from raw segmentation fidelity (time-step-wise);
- It highlights whether observed improvements originate from better temporal localization or from post-processing effects;
- It offers insight into overfitting patterns. For example, models with inflated event-level F1 but poor frame-level agreement often rely excessively on post-processing.

Together, these two complementary evaluations yield a balanced view of both the clinical interpretability and intrinsic segmentation quality of each model.

4. **Error reduction rate (%)**. To highlight the relative improvement over baselines, we further report the percentage error reduction rate (ERR), computed from F1 scores as:

$$\mathrm{ERR}(\%) = \frac{\mathrm{F1}_{\mathrm{model}} - \mathrm{F1}_{\mathrm{baseline}}}{1 - \mathrm{F1}_{\mathrm{baseline}}} \times 100\%. \tag{6.8}$$

All reported results are based on at least ten independent runs per experimental setting.

6.3.3 Model Hyperparameter Fine-Tuning

To identify an effective configuration of the proposed FNO-based stream, we conducted a series of fine-tuning experiments focusing on key architectural hyperparameters, including

the number of FNO layers (depth) and the number of retained Fourier modes (M). These parameters directly control the model's capacity to capture global temporal dependencies and its computational efficiency.

We systematically evaluated different combinations of stacked FNO layers and retained Fourier modes. Specifically, the network was tested with one to four FNO layers and with the number of retained Fourier modes M ranging from 16 to 256. The empirical results indicate that a configuration with two FNO layers and $M = 96$ consistently provides the best trade-off between accuracy, stability, and computational cost. Deeper configurations (three or more layers) yielded marginal improvements but increased overfitting risk and training variance, while shallower variants underfit the data. Therefore, this configuration was adopted as the default setting in all subsequent experiments.

6.3.4 Experiment 1: FNO-AugUNet vs. 1D-UNet

The first experiment evaluates the benefit of the FNO stream by comparing the proposed FNO-AugUNet with the 1D-UNet baseline. The 1D-UNet is widely used in FHR analysis and serves as a strong baseline, as it captures morphology, baseline shifts, and transient temporal patterns. Both models were trained and evaluated under identical settings on the public CULF-DB dataset to ensure a fair comparison.

Table 6.1 summarizes the results of Experiment 1 on the CULF-DB dataset. Compared with the time-domain-only 1D-UNet baseline, our proposed FNO-AugUNet achieves substantial improvements on both evaluation metrics. Specifically, the event-level F1 score increases from 63.90% to 68.71%, corresponding to a relative error-rate reduction of 13.32%, while the time-step-wise F1 rises from 60.60% to 63.80%, yielding a 8.12% reduction in time-step-level misclassification errors. These consistent gains across metrics confirm that integrating frequency-domain information significantly enhances the model's ability to delineate A/D patterns. Notably, for deceleration events, the event-level F1 shows a remarkable relative error-rate reduction of 20.22%, underscoring the FNO stream's strong advantage in capturing long-range dependencies and complex deceleration patterns.

The performance improvement can be attributed to the complementary nature of the two streams. While the time-domain 1D-UNet stream captures local morphological cues and short-term dynamics in the FHR signal, the FNO stream provides a global frequency perspective that stabilizes detection under nonstationary noise and baseline fluctuations. Overall, these results demonstrate the effectiveness and robustness of the proposed dual-stream FNO-AugUNet, validating its advantage over conventional single-stream architectures.

Table 6.1 Experiment 1 on CULF-DB: FNO-AugUNet vs. baseline 1D-UNet.

Model	Acceleration	Deceleration	Average A/D	
	Event-level F1(%)	Event-level F1(%)	Event-level F1(%)	Time-step-wise F1(%)
1D-UNet (baseline)	57.48	70.28	63.90	60.60
Our FNO-AugUNet	**61.23**	**76.29**	**68.71**	**63.80**
Our FNO-AugUNet ERR(%)	8.82	20.22	13.32	8.12

6.3.5 Experiment 2: FNO-only Stream vs. FNO-AugUNet

The second experiment examines the effectiveness of relying exclusively on the FNO streams. The performance of the only FNO stream is compared against that of our FNO-AugUNet to assess the necessity of integrating spectral and temporal representations.

Table 6.2 presents the results of Experiment 2 on the CULF-DB dataset. The FNO-only stream yields substantially lower scores, with an average event-level F1 of 39.92% and a time-step-wise F1 of 48.74%. This pronounced degradation confirms that frequency-domain information alone cannot reliably characterize the nonstationary morphology of FHR patterns. The FNO stream can effectively capture global periodicities and energy distributions but lose crucial temporal dependencies and instantaneous transitions, which can be recognized by the UNet style structure, that are indispensable for accurate A/D recognition.

In contrast, the dual-stream model achieves markedly higher performance—68.71% in event-level F1 and 63.80% in time-step-wise F1—representing nearly a 50% improvement over the FNO-only baseline. These gains demonstrate that spectral features become truly informative only when complemented by the fine-grained temporal structure learned from the raw FHR signal. The dual-stream formulation thus enables a synergistic interaction between the spectral and temporal encoders: the spectral branch provides global frequency context and noise robustness, while the temporal branch preserves local waveform dynamics and phase continuity. Together, they yield a more physiologically faithful representation of fetal cardiac responses, underscoring the necessity and effectiveness of integrating both streams in intrapartum CTG analysis.

6.3.6 Experiment 3: Comparison with SOTA Models

In the final experiment, we benchmark our FNO-AugUNet against several state-of-the-art models for fetal heart rate analysis, namely CTG-UNet, EMA-UNet, and ET-CNN. All baseline implementations were reproduced following the descriptions in their original publications to ensure fairness and consistency. Experiments were conducted on both the public CULF-DB

Table 6.2 Experiment 2 on CULF-DB: comparison between frequency-domain-only FNO and our FNO-AugUNet.

Model	**Acceleration**	**Deceleration**	**Average A/D**	
	Event-level F1(%)	**Event-level F1(%)**	**Event-level F1(%)**	**Time-step-wise F1(%)**
FNO-only Stream	38.57	42.73	39.92	48.74
Our FNO-AugUNet	**61.23**	**76.29**	**68.71**	**63.80**

Table 6.3 Experiment 3: comparison with SOTA models on CULF-DB and 7AH-SYSU datasets. ERR(%) is computed as the relative error-rate reduction of our dual-stream model compared to the best-performing SOTA baseline on each dataset.

CULF-DB				
Model	**A. Event-level F1(%)**	**D. Event-level F1(%)**	**Average A/D: Event-level F1(%)**	**Time-step-wise F1(%)**
CTG-UNet	57.13	68.48	62.81	60.60
EMA-UNet	57.58	68.19	62.88	59.36
ET-CNN	**61.23**	69.39	65.31	53.58
Our dual-stream model	**61.23**	**76.29**	**68.71**	**63.80**
Our model ERR(%)	0.00	22.54	9.80	8.12
7AH-SYSU				
Model	**A. Event-level F1(%)**	**D. Event-level F1(%)**	**Average A/D: Event-level F1(%)**	**Time-step-wise F1(%)**
CTG-UNet	68.71	65.47	66.99	64.60
EMA-UNet	69.94	66.74	68.23	**65.23**
ET-CNN	64.33	63.84	64.08	57.10
Our dual-stream model	**71.28**	**67.24**	**69.65**	64.53
Our model ERR(%)	4.46	1.50	4.47	-2.01

dataset and our large-scale private 7AH-SYSU dataset, enabling evaluation across datasets with different sizes and distributions.

Table 6.3 summarizes the quantitative results. The last row of Table 6.3 reports the error-rate reduction for each metric, calculated relative to the best-performing baseline rather than a fixed reference. This dynamic definition ensures that the ERR reflects genuine performance gains beyond the strongest available alternative in each experimental context. Across both datasets, our dual-stream model consistently achieves the highest event-level F1 scores among all competitors.

In particular, the model yields marked improvements in deceleration detection, which is heavily based on modeling long-range temporal dependencies—precisely the strength of our spectral pathway within the FNO encoder. In the public CULF-DB, the proposed model improves the average A/D event-level F1 by 9.80% over the strongest baseline (ET-CNN), while also significantly reducing the relative error rate by 22.54% in deceleration detection. On the larger and more diverse 7AH-SYSU dataset, similar trends are observed: the dual-

stream network surpasses all existing architectures with a 4.47% improvement in the average event-level F1, demonstrating stable generalization across different data distributions.

Although the improvement in deceleration detection is particularly pronounced, the performance on acceleration (A) events remains competitive. In the CULF-DB dataset, the acceleration event-level F1 score of our model matches that of ET-CNN, indicating that the FNO-augmented stream preserves the model's sensitivity to short-duration patterns, which is a characteristic characteristic feature of accelerations. Moreover, when evaluated on the large-scale 7AH-SYSU dataset, our model achieves the highest F1 for acceleration detection among all SOTA methods. This improvement in the larger dataset implies that the frequency–temporal fusion mechanism can capture stable and transferable representations of acceleration morphology, which become more evident under broader data distributions.

Despite achieving superior event-level metrics, our model exhibits a marginally lower time-step-wise F1 score on the 7AH-SYSU dataset compared with EMA-UNet. This observation can be attributed to the inherent trade-off between event-level precision and frame-level continuity in temporal segmentation. The dual-stream architecture, by emphasizing global spectral dependencies and cross-scale temporal fusion, tends to produce smoother and more temporally coherent predictions, which may slightly reduce local sensitivity at the precise onset and offset boundaries of events. In contrast, models such as EMA-UNet rely more heavily on localized convolutional receptive fields and moving-average mechanisms, which favor frame-wise accuracy but are less effective at capturing long-range contextual cues. Furthermore, the 7AH-SYSU dataset exhibits higher intra-class variability than CULF-DB, making frame-level alignment more challenging even when overall event detection remains robust. Nevertheless, the consistent improvement in event-level F1—which aligns more closely with clinical annotation granularity—suggests that our approach delivers better physiological interpretability and practical reliability in real-world fetal monitoring scenarios.

Taken together, these results validate the robustness, scalability, and clinical applicability of the proposed FNO-AugUNet with time–frequency fusion framework, showing that the integration of global spectral modeling and temporal feature learning leads to consistent benefits across datasets of varying sizes and acquisition environments.

6.3.7 Summary of Experimental Findings

Across three experiments, the proposed FNO-AugUNet consistently demonstrates advantages over single-stream baselines and state-of-the-art alternatives.

Experiment 1 shows that integrating a FNO based frequency-domain branch into a stan-

dard 1D-UNet significantly improves both event-level and time-step-wise F1 scores in both acceleration and deceleration recognition, confirming the value of spectral information.

Experiment 2 indicates that an FNO stream alone is insufficient for robust detection, as the FNO layers under-perform compared with temporal baselines. This finding highlights the necessity of combining temporal and spectral cues rather than relying on either in isolation.

Finally, Experiment 3 establishes that our dual-stream model surpasses strong state-of-the-art baselines across both the public CULF-DB dataset and the large-scale private 7AH-SYSU dataset. The observed improvements are particularly notable for deceleration detection, a clinically critical and challenging task.

Overall, the experiments confirm that the dual-stream design with the help of the FNO method achieves robust and generalizable improvements by jointly leveraging temporal morphology and spectral long-range dependencies. The results provide strong empirical support for the proposed architecture and motivate its use as a reliable solution for fetal heart rate acceleration and deceleration detection.

6.4 Discussion

6.4.1 Advantages of the Dual-Stream Approach

The proposed dual-stream architecture effectively leverages the complementary strengths of temporal and spectral representations. The time-domain 1D-UNet captures morphological features, baseline shifts, and transient signal patterns with high fidelity, while the frequency-domain FNO stream excels at modeling long-range dependencies through global spectral representations. By fusing these two streams, the framework achieves both improved accuracy and greater calibration stability compared to single-stream baselines. This demonstrates that temporal morphology and global spectral context are not redundant but mutually reinforcing, leading to more reliable detection of acceleration and deceleration events. Furthermore, the spectral branch based on FNO achieves these gains with fewer parameters and higher efficiency compared to conventional convolutional encoders, highlighting its practical scalability.

6.4.2 Limitations

Despite its advantages, the proposed FNO-AugUNet has several limitations. First, while the introduction of the FNO stream substantially enhances global contextual modeling, the frequency-domain representation inevitably abstracts away certain fine-grained temporal nu-

ances. Because the Fourier transform captures periodic structures in a compact spectral form, transient and localized signal variations—such as small, short-lived fluctuations in the FHR trace—may become less pronounced after transformation. As a result, the model's sensitivity to subtle local changes could be somewhat reduced, particularly for weak or borderline acceleration and deceleration events. Future extensions could mitigate this limitation by incorporating hybrid representations that jointly preserve temporal locality and frequency-domain globality, for example through wavelet-based or adaptive spectral encoding schemes.

Second, domain shift between datasets with different recording conditions remains a non-trivial challenge. Despite the strong cross-validation performance observed on the CULF-DB and private datasets, variations in sampling rate, signal quality, or annotation criteria across institutions may still degrade generalization. This limitation highlights the importance of future work on domain adaptation and robustness analysis, especially under heterogeneous clinical environments where fetal monitoring equipment and annotation practices differ. Investigating model adaptation strategies—such as self-supervised pretraining, adversarial domain alignment, or calibration-based transfer—could further enhance the reliability and clinical applicability of the proposed framework.

6.4.3 Clinical Implications

From a clinical perspective, automatic detection of acceleration and deceleration events has the potential to substantially reduce clinician workload and mitigate inter-observer variability. Importantly, spectral representations provide an additional layer of interpretability: Fourier-mode saliency maps reveal which frequency components are emphasized by the model. Combined with conventional time-domain saliency analysis, this dual interpretability may increase clinicians' trust and confidence in adopting AI-assisted decision support tools. Ultimately, such systems could facilitate more consistent fetal monitoring in both high-resource and resource-limited settings.

6.4.4 Future Work

Future research will explore several promising directions. First, integration of uterine contraction (UC) signals with FHR analysis may enable more comprehensive monitoring of maternal–fetal interactions. Second, semi-supervised or self-supervised pretraining strategies may improve robustness under limited annotated data. Finally, incorporating differentiable post-processing modules such as the ERNP framework could yield refined event delineation and closer alignment with clinical annotation protocols.

6.5 Conclusion

6.5.1 Summary of Findings

This chapter introduced FNO-AugUNet, a dual-stream time–frequency fusion network that combines a time-domain 1D-UNet with an auxiliary FNO stream. Through extensive experiments, we demonstrated that this architecture outperforms strong single-stream baselines as well as existing state-of-the-art models. The results confirm that the FNO based neural network provides a powerful mechanism for spectral modeling and that its integration with temporal morphology analysis enables accurate and robust detection of acceleration and deceleration events.

6.5.2 Contributions to the Thesis

The proposed FNO-AugUNet contributes to the broader thesis narrative on AI for fetal monitoring by demonstrating the value of jointly leveraging time- and frequency-domain information. Methodologically, the chapter establishes that FNO-based encoders offer a scalable and efficient alternative to conventional models. Empirically, the evaluation across both public and private datasets validates the robustness and generalizability of the approach. Collectively, these findings advance the state of the art in automatic fetal heart rate event detection.

6.5.3 Outlook

Looking ahead, the proposed design lays the foundation for clinically deployable AI solutions in obstetrics. By emphasizing both accuracy and interpretability, the framework aligns with the dual goals of improving clinical outcomes and supporting clinician decision-making. Future integration with additional modalities such as UC signals, combined with scalable pretraining and advanced post-processing, may further enhance clinical reliability. Ultimately, the dual-stream paradigm represents a step toward building trustworthy, data-driven monitoring systems capable of supporting maternal and neonatal care at scale.

CHAPTER 7 CONCLUSION

7.1 Summary of Works

7.1.1 Summarize the findings and significance of the three studies.

This thesis presented three complementary studies addressing the challenge of acceleration and deceleration event detection in fetal cardiotocography, with a particular focus on fetal heart rate signals.

The first study proposed the Event ERNP framework, a residual-inspired and self-attention-based refinement module. ERNP improves upon the widely adopted LSMP post-processing pipeline by explicitly learning to correct coarse segmentation outputs using both the raw FHR signal and LSMP's outputs. Experimental results on public and private datasets demonstrated that ERNP consistently enhanced detection accuracy across different backbone segmentation models. This work contributes a reproducible, lightweight, and model-agnostic refinement strategy that can be flexibly integrated into state-of-the-art pipelines.

The second study introduced a multi-modal multi-scale residual convolutional network designed to incorporate FHR, UC, and FM signals. By combining multi-stream fusion strategies and multi-scale receptive fields, this architecture effectively captured temporal dependencies across signals and scales. The results showed substantial improvements over single-modal and single-scale baselines, achieving state-of-the-art performance in both accuracy and F1-score, and validating the clinical importance of integrating UC and FM information in addition to FHR.

The third study introduced a dual-stream time–frequency fusion model based on Fourier Neural Operators. This approach performs joint modeling of the temporal and spectral domains of the FHR signal. The FNO layer captures global frequency structure and noise-resilient patterns, while the temporal UNet focuses on local morphological details and transient changes. Empirical results showed that the proposed dual-stream design achieved an error reduction rate of approximately 8% over the single-stream 1D-UNet baseline and improved the event-level F1 by around 10% compared with state-of-the-art methods on public datasets. Notably, the model yielded substantial gains in deceleration detection, with the event-level F1 and the error reduction rate reaching improvements of about 22%. These findings confirm that frequency-domain representations are not independently sufficient but become highly informative when integrated with temporal features through late-fusion-style stream fusion approaches.

Together, the three studies highlight post-processing refinement, multi-signal fusion, and front-end architectural innovation as crucial and complementary directions for achieving robust and generalizable A/D detection.

7.1.2 The potential of clinical integration

The methods developed in this thesis have strong implications for clinical adoption. By providing event-level detection of accelerations and decelerations, our proposed methods can produce interpretable outputs that align with existing NICHD/ACOG guidelines. ERNP, in particular, offers a transparent plug-in refinement step that can be easily attached to current automated systems without requiring changes in workflow. The multi-modal framework further demonstrates the value of integrating heterogeneous obstetric signals, paving the way toward comprehensive intrapartum monitoring systems.

If deployed in practice, these approaches could reduce inter-observer variability, support timely clinical decision-making, and lower unnecessary interventions. Importantly, the reproducibility and open-source availability of our implementations help ensure transparency, which is a prerequisite for clinical validation and regulatory approval. The FNO based dual-stream architecture is particularly promising for deployment in clinical monitoring systems, as its frequency-domain branch can mitigate sensor noise and signal dropout, while the temporal branch preserves real-time responsiveness for event-level alerts.

7.1.3 Implications for the generalization of medical time series

Beyond the domain of FHR analysis, the contributions of this thesis provide generalizable insights for medical time-series modeling. ERNP demonstrates that residual refinement of heuristic pipelines can bridge the gap between non-differentiable clinical rules and deep learning outputs, a principle that may be extended to ECG, EEG, or other biosignals where rule-based interpretation remains dominant. Similarly, the multi-modal multi-scale design highlights the importance of modeling signals across both multiple channels and multiple temporal resolutions, a pattern applicable to diverse healthcare contexts such as sleep staging, seizure detection, or ICU monitoring. Furthermore, the FNO-based study underscores the value of incorporating frequency-domain representations to capture global temporal dependencies and noise-resilient patterns, suggesting a promising direction for improving robustness and interpretability in complex medical time-series analysis.

In this sense, the thesis not only addresses a specific clinical problem in obstetrics but also contributes methodological innovations with broader relevance to the field of medical AI.

7.2 Philosophical and Social Perspective: AI for Good

The works presented in this thesis are not only technical contributions but also part of a broader movement often referred to as "AI for Good [77, 78]." This notion, endorsed by the United Nations, highlights the role of artificial intelligence in tackling some of humanity's most pressing challenges—healthcare, poverty reduction, and environmental sustainability [79–82]. Within this perspective, the central question is not only what AI can achieve but also how AI should be designed and deployed to maximize societal benefit while minimizing potential harm.

In the context of maternal and perinatal care, AI systems for fetal heart rate monitoring exemplify this dual responsibility. On the one hand, the automation of acceleration and deceleration detection addresses tangible clinical needs: it reduces the burden on over-stretched clinicians, mitigates human interpretation variability, and offers decision support in environments where obstetric expertise is scarce. Such improvements have the potential to enhance the fairness and accessibility of healthcare, allowing more women and infants in both developed and resource-limited regions to benefit from more consistent standards of care.

On the other hand, the "good" that AI brings must be critically examined. Clinical deployment of AI technologies carries the risk of over-reliance, misinterpretation of algorithmic outputs, or inequities arising from biased data sources. Therefore, the pursuit of AI for Good requires not only algorithmic innovation but also ethical governance, stakeholder involvement, and long-term monitoring of social impact. Ensuring transparency and reproducibility—as emphasized in this thesis by open-sourcing code and detailed methodological reporting—is one way to strengthen trustworthiness. Another is to engage clinicians, patients, and policymakers in the co-design of systems so that the solutions remain aligned with actual human needs rather than abstract technical benchmarks.

Finally, the reflection on AI for Good invites us to consider broader philosophical questions: What does "good" mean in the rapidly evolving intersection of technology and medicine? How can we ensure that the benefits of AI are distributed equitably, rather than concentrated in well-resourced healthcare systems? And how can we design AI systems that empower, rather than replace, human judgment? These questions remain open, but they are essential for guiding the next generation of research and development.

In conclusion, the research presented here contributes to AI for Good by demonstrating how advanced deep learning methods can be responsibly applied to maternal healthcare. By balancing innovation with reproducibility, interpretability, and social accountability, it aims to set an example of how technical research can meaningfully translate into real-world

benefits.

7.3 Limitations

Although this paper proposes several innovative models and experimental designs and achieves good performance on multiple public and private datasets, it still has the following limitations and requires further improvement and expansion in subsequent research.

7.3.1 Limitations of data sources

Despite the use of both public and private datasets, several limitations arise from the nature of the data. Public datasets such as CTU-UHB and CULF-DB are relatively small in scale, contain limited clinical diversity, and often suffer from annotation inconsistencies due to differences in labeling protocols. Although our private dataset is larger and has been annotated and iteratively reviewed by experienced clinicians, it is still confined to a single institutional setting, which may limit cross-hospital generalizability.

Although experiments were conducted on both public datasets (CULF-DB, CTU-UHB) and a private clinical dataset, all datasets predominantly originate from similar clinical environments and fetal monitoring protocols. This may limit the generalizability of the proposed models across hospitals with different annotation standards, sampling frequencies, or fetal monitoring devices. Future work should incorporate data from a broader range of populations and clinical contexts, including low-resource settings, to better assess model robustness and fairness.

The ground-truth labels for acceleration and deceleration events are based on human annotations, which are inherently subjective and sometimes inconsistent across annotators. Although care was taken to use high-quality annotations and follow clinical guidelines, no inter-annotator agreement statistics were available. The potential presence of missed labels, boundary uncertainty, or interpretational ambiguity may affect training and evaluation. Further work could explore probabilistic labeling strategies or unsupervised label refinement techniques that leverage intrinsic signal patterns to improve annotation quality.

Furthermore, due to the complexity and uncontrollability of actual clinical situations, fetal monitoring signals are often noisy, incomplete, or subject to device-specific artifacts, making it difficult to ensure fully representative coverage of real-world conditions. The multi-modal model relies on both FHR and UC signals. In practice, UC channels may be missing or corrupted due to hardware failure, maternal movement, or clinical decisions. The current model does not include explicit mechanisms to handle missing modalities. This limits the

applicability of the model in real-time clinical deployment. These factors may introduce bias and reduce the robustness of the proposed methods when deployed in different clinical environments. Future research could integrate modality dropout techniques or imputation networks to ensure robustness under incomplete input scenarios.

7.3.2 Limitations of the method design

The methodological innovations proposed in this thesis also come with inherent constraints.

For the first study, our proposed ERNP relies on prior LSMP-based masks, which means that its effectiveness may be partially bounded by the quality of upstream segmentation. Although ERNP was designed as a differentiable post-processing layer and conceptually supports end-to-end learning, in this study, it was trained in a two-stage pipeline rather than jointly with the base detection model. This decision was made to allow fair comparisons with other post-processing baselines. However, further work is needed to explore the benefits of true end-to-end optimization and feedback from the post-processor to the detection backbone.

In the second study, multimodal fusion strategies have not fully explored all possible architectures. For example, advanced mechanisms such as Transformer fusion and cross-attention were not included in the comparison. Future research on adaptive fusion is needed.

The FNO-based study also presents certain limitations. Although the Fourier Neural Operator effectively models long-range temporal dependencies and demonstrates strong noise resilience, its reliance on frequency-domain representations introduces several challenges. First, the truncation of high-frequency modes, though beneficial for stability, may lead to the loss of clinically relevant fine-grained temporal details. Second, the current implementation assumes a fixed Fourier basis and uniform sampling, which may not fully capture nonstationary or irregularly sampled patterns commonly found in real-world clinical recordings. Future work could investigate hybrid spectral–temporal operators or adaptive frequency selection mechanisms to achieve a more balanced representation between global and local dynamics.

7.3.3 Practical deployment considerations are still limited

From a practical standpoint, several challenges remain before real-world deployment. First, the computational demand of training and inference, while manageable in a research setting, may pose difficulties for real-time monitoring on resource-constrained hospital equipment. Second, issues of data privacy, security, and regulatory approval are non-trivial barriers to clinical integration. Finally, the gap between algorithmic evaluation and actual clinical workflow must be carefully bridged: clinicians require seamless integration with existing

monitoring systems, intuitive interfaces, and clear guidelines for interpretation. Without these, even accurate models may struggle to gain acceptance in practice.

This dissertation focused on methodological contributions and offline evaluation metrics. The latency, memory, and computational requirements of the proposed models, especially ERNP and multistream CNNs, were not tested in real-time clinical settings. While the multimodal architecture demonstrates performance gains, it increases computational complexity and requires synchronized multi-channel data (FHR, UC, FM), which are not always available in clinical practice. Deployment on edge devices (e.g., fetal monitors or mobile units in remote clinics) may require additional pruning, quantization, or distillation strategies.

And, all experiments remain in the offline validation phase, lacking interactive system evaluation with physician feedback, making it impossible to measure clinical friendliness and decision support effectiveness. Although the detection of A/D events is a foundational step in FHR interpretation, the models do not yet offer full diagnostic explanations or risk scoring that clinicians can act on directly. No integration with downstream decision-support systems was performed. The interpretation of predicted events and their correlation with neonatal outcomes remains an open direction, especially for translating model predictions into actionable clinical insights.

7.3.4 Limitations of the evaluation system

Although the thesis employed cross-validation, multiple datasets, and comparative baselines, limitations in the evaluation framework must be acknowledged. The majority of experiments were retrospective, relying on pre-recorded data rather than prospective clinical trials. Metrics such as F1-score, sensitivity, and PPV capture event-level performance but cannot fully reflect the broader impact on clinical decision-making, such as reduction in unnecessary interventions or improvements in neonatal outcomes. In addition, inter-observer variability in annotations remains an unresolved factor that may affect the validity of reported performance.

Currently, event-level evaluation indicators, such as F1-score, are used primarily. Indicators based on clinical risk levels, such as the unequal weighting of misdiagnosis and missed diagnosis evaluation mechanism, have not yet been built or used. This may cause the model to prioritize technically optimal metrics over clinically meaningful performance, particularly in balancing safety (avoiding false alarms) and sensitivity (detecting all potential abnormalities).

7.3.5 Summary

In summary, the limitations of this thesis can be grouped into four categories: data availability and representativeness, methodological assumptions, practical deployment barriers, and evaluation constraints. These limitations do not undermine the contributions of the work, but rather highlight the complexity of developing clinically reliable AI systems. Acknowledging these boundaries provides a realistic perspective on the current state of research and naturally motivates the future directions discussed in Section 7.4.

Given the above limitations, subsequent research can conduct more in-depth exploration in areas such as data diversity expansion, model structure optimization, clinical deployment verification, and multidimensional evaluation mechanisms, thereby promoting intelligent fetal heart monitoring technology to higher practicality and clinical credibility.

7.4 Future Research

Building upon the findings and acknowledging the limitations discussed above, several research directions emerge for the continued advancement of automated fetal monitoring and medical time-series analysis.

7.4.1 Methodological innovations

From a methodological perspective, several avenues remain open. First, integrating differentiable post-processing mechanisms could further unify training and inference, thereby overcoming the current separation between model prediction and heuristic refinement. Achieving this end-to-end differentiability would allow models to directly optimize for event-level accuracy rather than relying on non-trainable rules applied after inference. Future work may extend the dual-stream framework into a fully differentiable time–frequency learning paradigm, integrating spectral transformations such as continuous wavelets or learnable Fourier filters within end-to-end training to further enhance interpretability and domain adaptation.

Second, exploring more advanced architectures may yield further performance improvements. Promising directions include transformer-based temporal models that can capture long-range dependencies in FHR signals, hybrid networks that combine frequency- and time-domain representations to better characterize oscillatory patterns, and physics-informed neural networks that incorporate physiological priors into the learning process. These approaches hold potential to improve both accuracy and robustness under challenging signal conditions.

Third, interpretability must remain a central focus. Beyond widely used techniques such

as attention visualization, future work should consider more explanatory mechanisms. For instance, counterfactual reasoning can be employed to answer questions such as: 'What minimal change to the input signal would alter the model's decision?' In the context of FHR analysis, this could mean identifying how much shorter a deceleration event would need to be for the model to classify it as normal, or how baseline shifts influence predictions. Such explanations are more aligned with clinicians' reasoning processes, as they highlight decision boundaries and provide actionable insights. Similarly, rule-based explanations and hybrid human–AI interpretability frameworks may further bridge the gap between algorithmic predictions and clinical trust.

7.4.2 Data expansion and standardization

Future work should prioritize the development of larger, more diverse, and internationally standardized datasets. Multi-center collaborations can help mitigate institutional biases and ensure that models generalize across populations, devices, and clinical practices. In addition, standardized annotation protocols—ideally with consensus across multiple clinicians—would reduce subjectivity and enhance reproducibility. Leveraging semi-supervised or active learning approaches may also help reduce the annotation burden while improving label quality.

7.4.3 Toward real-world deployment

A crucial step for future research is the translation from retrospective validation to prospective clinical trials. Embedding AI systems in actual labor and delivery units would allow researchers to assess not only technical accuracy but also workflow integration, clinician acceptance, and impact on maternal–fetal outcomes. Beyond technical validation, regulatory approval processes, such as clearance from the U.S. Food and Drug Administration (FDA) or CE marking in Europe, represent essential milestones for medical AI deployment. Meeting these requirements demands rigorous clinical evidence, standardized reporting, and transparent risk assessment.

Another important aspect of deployment is workflow integration and human–AI collaboration. AI systems should not replace clinicians but act as assistive tools that enhance decision-making. Designing intuitive user interfaces, alert mechanisms, and interpretable outputs is therefore critical for fostering clinician trust and ensuring safe adoption. In practice, models must be evaluated for how they change clinician behavior, reduce cognitive workload, and support shared accountability between human experts and algorithms.

This future direction also involves optimization for low-resource settings, such as developing

lightweight versions of the models that can run on portable monitoring devices or mobile platforms. This could expand access to high-quality obstetric care in underserved regions, where specialist availability is limited. Parallel attention should be given to privacy-preserving techniques (e.g., federated learning, differential privacy) to ensure compliance with healthcare regulations and protect patient confidentiality during data collection and model training.

In summary, real-world deployment of automated fetal monitoring systems requires not only algorithmic excellence but also regulatory clearance, clinician–AI collaboration, and privacy-preserving safeguards. Addressing these aspects will be essential to ensure that AI systems are not only effective in controlled research settings but also trustworthy, safe, and impactful in clinical practice.

7.4.4 Generalization to broader medical time series

Finally, the approaches proposed in this thesis—residual refinement of heuristic pipelines and multi-modal multi-scale modeling—can be extended to other biomedical signal processing domains. Potential applications include arrhythmia detection in ECG, seizure onset prediction in EEG, and monitoring of ICU vital signs. Investigating how these methods generalize across domains will help establish a unified framework for interpretable and clinically reliable time-series analysis.

7.4.5 Summary

In conclusion, while this thesis has demonstrated tangible progress in automated acceleration/deceleration detection and multi-modal fetal monitoring, it also opens a pathway toward broader innovations in medical AI. The ultimate goal of future research should not only be technical excellence but also clinical relevance, ethical responsibility, and global accessibility, ensuring that AI truly fulfills its promise as a technology for good.

REFERENCES

[1] W. G. A. by the Guidelines Review Committee *et al.*, "Who recommendations: intrapartum care for a positive childbirth experience," *Geneva: World Health Organization Copyright© World Health Organization*, vol. 2018, 2018.

[2] H. Campbell *et al.*, "Exploring the potential cost-effectiveness of a new computerised decision support tool for identifying fetal compromise during monitored term labours: an early health economic model," *Cost Effectiveness and Resource Allocation*, vol. 22, no. 1, p. 72, 2024.

[3] S. M. Vijgen *et al.*, "Cost-effectiveness of cardiotocography plus st analysis of the fetal electrocardiogram compared with cardiotocography only," *Acta obstetricia et gynecologica Scandinavica*, vol. 90, no. 7, pp. 772–778, 2011.

[4] M. I. Evans *et al.*, "Improving the interpretation of electronic fetal monitoring: the fetal reserve index," *American Journal of Obstetrics & Gynecology*, vol. 228, no. 5, pp. S1129–S1143, 2023.

[5] K. B. Kozhimannil *et al.*, "The rural obstetric workforce in us hospitals: challenges and opportunities," *The Journal of Rural Health*, vol. 31, no. 4, pp. 365–372, 2015.

[6] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature medicine*, vol. 25, no. 1, pp. 44–56, 2019.

[7] X. Liu *et al.*, "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis," *The lancet digital health*, vol. 1, no. 6, pp. e271–e297, 2019.

[8] A. M. Oprescu *et al.*, "Artificial intelligence in pregnancy: A scoping review," *IEEE Access*, vol. 8, pp. 181 450–181 484, 2020.

[9] P. Giaxi *et al.*, "Artificial intelligence and machine learning: an updated systematic review of their role in obstetrics and midwifery," *Cureus*, vol. 17, no. 3, 2025.

[10] J. L. Aeberhard *et al.*, "Introducing artificial intelligence in interpretation of foetal cardiotocography: medical dataset curation and preliminary coding—an interdisciplinary project," *Methods and protocols*, vol. 7, no. 1, p. 5, 2024.

[11] J. T. Parer and T. King, "Fetal heart rate monitoring: is it salvageable?" *American journal of obstetrics and gynecology*, vol. 182, no. 4, pp. 982–987, 2000.

[12] S. Santo *et al.*, "Agreement and accuracy using the figo, acog and nice cardiotocography interpretation guidelines," *Acta obstetricia et gynecologica Scandinavica*, vol. 96, no. 2, pp. 166–175, 2017.

[13] H. Yang *et al.*, "Expert consensus on the application of electronic fetal heart rate monitoring," *Chinese Journal of Perinatal Medicine*, vol. 18, no. 7, pp. 486–490, 2015.

[14] J. Spilka *et al.*, "Analysis of obstetricians' decision making on ctg recordings," *Journal of biomedical informatics*, vol. 51, pp. 72–79, 2014.

[15] A. G. Cahill *et al.*, "A prospective cohort study of fetal heart rate monitoring: deceleration area is predictive of fetal acidemia," *American journal of obstetrics and gynecology*, vol. 218, no. 5, pp. 523–e1, 2018.

[16] A. H. de l'Aulnoit *et al.*, "Automated fetal heart rate analysis for baseline determination and acceleration/deceleration detection: A comparison of 11 methods versus expert consensus," *Biomedical Signal Processing and Control*, vol. 49, pp. 113–123, 2019.

[17] P. Brocklehurst, "A study of an intelligent system to support decision making in the management of labour using the cardiotocograph–the infant study protocol," *BMC pregnancy and childbirth*, vol. 16, no. 1, pp. 1–15, 2016.

[18] P. Brocklehurst *et al.*, "Computerised interpretation of fetal heart rate during labour (infant): a randomised controlled trial," *The Lancet*, vol. 389, no. 10080, pp. 1719–1729, 2017.

[19] P. J. Steer *et al.*, "Computerised analysis of intrapartum fetal heart rate patterns and adverse outcomes in the infant trial," *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. 126, no. 11, pp. 1354–1361, 2019.

[20] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[21] J. Ogasawara *et al.*, "Deep neural network-based classification of cardiotocograms outperformed conventional algorithms," *Scientific reports*, vol. 11, no. 1, pp. 1–9, 2021.

[22] Y. Kong *et al.*, "Deep gaussian mixture model on multiple interpretable features of fetal heart rate for pregnancy wellness," in *Advances in Knowledge Discovery and Data*

Mining: 25th Pacific-Asia Conference, PAKDD 2021, Virtual Event, May 11–14, 2021, Proceedings, Part I. Springer, 2021, pp. 238–250.

[23] Z. Zhao *et al.*, "Computer-aided diagnosis system of fetal hypoxia incorporating recurrence plot with convolutional neural network," *Frontiers in physiology*, vol. 10, p. 255, 2019.

[24] J. Spilka *et al.*, "Automatic evaluation of fhr recordings from ctu-uhb ctg database," in *International Conference on Information Technology in Bio-and Medical Informatics*. Springer, 2013, pp. 47–61.

[25] I. Nunes and D. Ayres-de Campos, "Computer analysis of foetal monitoring signals," *Best Practice & Research Clinical Obstetrics & Gynaecology*, vol. 30, pp. 68–78, 2016.

[26] K. R. Greene, "Intelligent fetal heart rate computer systems in intrapartum surveillance," *Current Opinion in Obstetrics and Gynecology*, vol. 8, no. 2, pp. 123–128, 1996.

[27] S. B. Thacker, D. F. Stroup, and H. B. Peterson, "Efficacy and safety of intrapartum electronic fetal monitoring: an update," *Obstetrics & Gynecology*, vol. 86, no. 4, pp. 613–620, 1995.

[28] V. Chudáček *et al.*, "Open access intrapartum ctg database," *BMC pregnancy and childbirth*, vol. 14, no. 1, pp. 1–12, 2014.

[29] H. Liang and Y. Lu, "A cnn-rnn unified framework for intrapartum cardiotocograph classification," *Computer Methods and Programs in Biomedicine*, vol. 229, p. 107300, 2023.

[30] I. Ben M'Barek, G. Jauvion, and P.-F. Ceccaldi, "Computerized cardiotocography analysis during labor–a state-of-the-art review," *Acta Obstetricia et Gynecologica Scandinavica*, vol. 102, no. 2, pp. 130–137, 2023.

[31] N. Alharbi *et al.*, "Fetal hypoxia detection using machine learning: A narrative review," *AI*, vol. 5, no. 2, pp. 516–532, 2024.

[32] R. D. Keith and K. R. Greene, "4 development, evaluation and validation of an intelligent system for the management of labour," *Bailliere's clinical obstetrics and gynaecology*, vol. 8, no. 3, pp. 583–605, 1994.

[33] M. Liu *et al.*, "Baseline/acceleration/deceleration determination of fetal heart rate signals using a novel ensemble lcresu-net," *Expert Systems with Applications*, vol. 218, p. 119610, 2023.

[34] ——, "Automated fetal heart rate analysis for baseline determination using emau-net," *Information Sciences*, vol. 644, p. 119281, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025523008666

[35] ——, "Automated fetal heart rate analysis for baseline determination using emau-net," *Information Sciences*, vol. 644, p. 119281, 2023.

[36] Q. Wu *et al.*, "Etcnn: an ensemble transformer-convolutional neural network for automatic analysis of fetal heart rate," *Biomedical Signal Processing and Control*, vol. 96, p. 106629, 2024.

[37] M. Wang *et al.*, "Automated analysis of fetal heart rate baseline/acceleration/deceleration using mtu-net3+ model," *Biomedical Engineering Letters*, vol. 14, no. 5, pp. 1037–1048, 2024.

[38] M. Zhong *et al.*, "Ctgnet: automatic analysis of fetal heart rate from cardiotocograph using artificial intelligence," *Maternal-Fetal Medicine*, vol. 4, no. 2, pp. 103–112, 2022.

[39] M. Liu *et al.*, "Baseline/acceleration/deceleration determination of fetal heart rate signals using a novel ensemble lcresu-net," *Expert Systems with Applications*, vol. 218, p. 119610, 2023.

[40] M. Shafiq and Z. Gu, "Deep residual learning for image recognition: A survey," *Applied sciences*, vol. 12, no. 18, p. 8972, 2022.

[41] H. Dang *et al.*, "A deep biometric recognition and diagnosis network with residual learning for arrhythmia screening using electrocardiogram recordings," *IEEE Access*, vol. 8, pp. 153 436–153 454, 2020.

[42] W. Jifara *et al.*, "Medical image denoising using convolutional neural network: a residual learning approach," *The Journal of Supercomputing*, vol. 75, pp. 704–718, 2019.

[43] R. Pardasani *et al.*, "Development of a novel artificial intelligence algorithm for interpreting fetal heart rate and uterine activity data in cardiotocography," *Frontiers in Digital Health*, vol. 7, p. 1638424, 2025.

[44] X. Qin *et al.*, "Fetalet: interpretable fetal heart rate anomaly detection via shapelet learning," *Complex & Intelligent Systems*, vol. 11, no. 11, pp. 1–16, 2025.

[45] I. Sato *et al.*, "Comparison and verification of detection accuracy for late deceleration with and without uterine contractions signals using convolutional neural networks," *Frontiers in Physiology*, vol. 16, p. 1525266, 2025.

[46] J. Tolladay, M. Tome, and A. Georgieva, "A deep learning method for locating fetal heart rate decelerations during labour using crowd-sourced data," *Expert Systems with Applications*, vol. 255, p. 124609, 2024.

[47] A. Alashbi *et al.*, "Advancing fetal heart rate monitoring: Ai-based approach for baseline detection and classification," in *2024 11th International Conference on Wireless Networks and Mobile Communications (WINCOM)*. IEEE, 2024, pp. 1–5.

[48] T. C. Lwin *et al.*, "Advanced predictive analytics for fetal heart rate variability using digital twin integration," *Sensors*, vol. 25, no. 5, p. 1469, 2025.

[49] Z. Yu *et al.*, "Ctggan: Reliable fetal heart rate signal generation using gans," in *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024, pp. 1–8.

[50] Y. Deng *et al.*, "A time-series progressive generative adversarial network for improving imbalanced fetal heart rate signal classification," *Applied Intelligence*, vol. 55, no. 18, p. 1135, 2025.

[51] Y. Zhang *et al.*, "Fhrgan: Generative adversarial networks for synthetic fetal heart rate signal generation in low-resource settings," *Information Sciences*, vol. 594, pp. 136–150, 2022.

[52] X. Li *et al.*, "Fhrdiff: Leveraging diffusion models for conditional fetal heart rate signal generation," in *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2024, pp. 2148–2155.

[53] W. Jatmiko, "Generative adversarial networks for unbalanced fetal heart rate signal classification," *Accessed: Mar*, vol. 19, 2024.

[54] D. C. Nguyen *et al.*, "Federated learning for smart healthcare: A survey," *ACM Computing Surveys (Csur)*, vol. 55, no. 3, pp. 1–37, 2022.

[55] S. Garst, J. Dekker, and M. Reinders, "A comprehensive experimental comparison between federated and centralized learning," *Database*, vol. 2025, p. baaf016, 2025.

[56] M. U. Nasir *et al.*, "Federated machine learning based fetal health prediction empowered with bio-signal cardiotocography." *Computers, Materials & Continua*, vol. 78, no. 3, 2024.

[57] V. D. Janapa *et al.*, "Federated learning for pregnancy care: Smartwatch and mobile app for fetal monitoring and promoting normal deliveries," *Journal of Electronic Materials*, vol. 54, no. 11, pp. 9524–9536, 2025.

[58] S. Boudet *et al.*, "A fetal heart rate morphological analysis toolbox for matlab," *SoftwareX*, vol. 11, p. 100428, 2020.

[59] ——, "Fetal heart rate signal dataset for training morphological analysis methods and evaluating them against an expert consensus," 2024.

[60] C. H. Sudre *et al.*, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *International Workshop on Deep Learning in Medical Image Analysis*. Springer, 2017, pp. 240–248.

[61] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[62] T.-Y. Lin *et al.*, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[63] K. He *et al.*, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[64] W. Xu, Y.-L. Fu, and D. Zhu, "Resnet and its application to medical image processing: Research progress and challenges," *Computer Methods and Programs in Biomedicine*, vol. 240, p. 107660, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260723003255

[65] J. Xie, S. Stavrakis, and B. Yao, "Automated identification of atrial fibrillation from single-lead ecgs using multi-branching resnet," *Frontiers in Physiology*, vol. 15, p. 1362185, 2024.

[66] J.-X. Wu *et al.*, "Chest x-ray image analysis with combining 2d and 1d convolutional neural network based classifier for rapid cardiomegaly screening," *IEEE Access*, vol. 10, pp. 47 824–47 836, 2022.

[67] Y. Li *et al.*, "A review of deep learning-based information fusion techniques for multimodal medical image classification," *Computers in Biology and Medicine*, vol. 177, p. 108635, 2024.

[68] D. Ayres-de Campos *et al.*, "Sisporto 2.0: a program for automated analysis of cardiotocograms," *Journal of Maternal-Fetal Medicine*, vol. 9, no. 5, pp. 311–318, 2000.

[69] R. W. Schafer, "What is a savitzky-golay filter?[lecture notes]," *IEEE Signal processing magazine*, vol. 28, no. 4, pp. 111–117, 2011.

[70] Y. Lu and S. Wei, "Nonlinear baseline estimation of fhr signal using empirical mode decomposition," in *2012 IEEE 11th International Conference on Signal Processing*, vol. 3. IEEE, 2012, pp. 1645–1649.

[71] S. Romagnoli *et al.*, "Annotation dataset of the cardiotocographic recordings constituting the "ctu-chb intra-partum ctg database"," *Data in Brief*, vol. 31, p. 105690, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352340920305849

[72] Z. Li *et al.*, "Fourier neural operator for parametric partial differential equations," *arXiv preprint arXiv:2010.08895*, 2020.

[73] N. Liu, S. Jafarzadeh, and Y. Yu, "Domain agnostic fourier neural operators," *Advances in Neural Information Processing Systems*, vol. 36, pp. 47 438–47 450, 2023.

[74] A. Tran *et al.*, "Factorized fourier neural operators," *arXiv preprint arXiv:2111.13802*, 2021.

[75] A. V. Oppenheim, *Discrete-time signal processing*. Pearson Education India, 1999.

[76] R. Bracewell and P. B. Kahn, "The fourier transform and its applications," *American Journal of Physics*, vol. 34, no. 8, pp. 712–712, 1966.

[77] A. M. Corrêa Harcus, "Ai for good global summit," 2018.

[78] N. Tomašev *et al.*, "Ai for social good: unlocking the opportunity for positive impact," *Nature Communications*, vol. 11, no. 1, p. 2468, 2020.

[79] G. Ramos, "Ethics of artificial intelligence," 2023.

[80] L. Floridi *et al.*, "Ai4people—an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations," *Minds and machines*, vol. 28, no. 4, pp. 689–707, 2018.

[81] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of ai ethics guidelines," *Nature machine intelligence*, vol. 1, no. 9, pp. 389–399, 2019.

[82] R. Vinuesa *et al.*, "The role of artificial intelligence in achieving the sustainable development goals," *Nature communications*, vol. 11, no. 1, p. 233, 2020.