Design Considerations for a Continuous Fetal Monitoring System Using Wearable Sensors and Machine Learning

by

Lucy Spicher

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Mechanical Engineering)
in the University of Michigan
2026

Doctoral Committee:

Professor Kathleen H. Sienko, Co-Chair
Associate Professor Xun Huan, Co-Chair
Clinical Associate Professor Carrie Bell
Professor Leia Stirling
Professor Jenna Wiens

Lucy Spicher

lspicher@umich.edu

ORCID iD: 0009-0008-1340-9282

© Lucy Spicher 2026

Dedication

For those who believed big dreams could come from small places.

Acknowledgements

Dr. Seuss once wrote, “You’re off to great places! Today is your day! Your mountain is waiting, so… get on your way!” While many mountains in my life still lie ahead, I would not have climbed to this peak without the guidance, brilliance, and unwavering support of the mentors, peers, friends, and family who accompanied me along the way.

First and foremost, I want to extend my deepest gratitude to my co-advisors, Dr. Kathleen Sienko and Dr. Xun Huan, for allowing me to take a risk with this project. Your mentorship has transformed how I write, think critically, and approach science, and your teachings will guide me throughout my career. I am equally indebted to Dr. Bell; this dissertation quite literally would not exist without you. Your clinical perspective, enthusiasm, and tenacity were invaluable, and I will deeply cherish the hours spent collecting and reviewing data and the friendship we built. To my committee members, Dr. Jenna Wiens and Dr. Leia Stirling, thank you for challenging me and putting the right tools in my academic toolbox.

This research was also made possible by the human element of science. I would like to acknowledge the 23 participants who volunteered to advance our collective knowledge of prenatal care, and the 58 engineers who shared their experiences to improve our understanding of engineering design. Additionally, this work was financially supported by the National Science Foundation (NSF) Graduate Research Fellowship Program (GRFP), the NSF Engineering Design and Systems Engineering (EDSE) grant program, the National Institutes of Health (NIH) T32 Acute and Critical Care in Engineering Training Program (ACCETP), and the Michigan Institute for Data and AI in Society (MIDAS).

To my lab mates, past and present: you taught me valuable lessons about research and life. Safa, Jamie, Grace, and Nick, thank you for taking me under your wing and modeling honesty and integrity. Jean, Emma, and Nadine, thank you for helping me grow as a scientist and cheering me on as I crossed the finish line.

I also owe a special thank you to the foundational mentors who encouraged me to shatter glass ceilings long before this dissertation took shape. Thank you to Beverly Miller, Gay Boyer, Tricia Schultz, and Chris Ruff for recognizing my early potential; to Cheryl Knobloch and Jess

Menold for shaping me into a confident leader; and to John Gershenson and Ben Savonen for taking a gamble on me and helping me realize the real-world impact of my work.

While my academic community built the scientist, my "village" sustained the person. To my Fit2Live community, thank you for being my “third place” and keeping me grounded throughout graduate school. To my friends, thank you for constantly reminding me that life is meant to be lived outside the lab. Karina, Rachel, Elizabeth, Laura, Emily, and Charlie—thank you for the Nancy Drew nights, Thursday trivia at Ashley’s, camping trips, long runs, Zingerman’s coffee breaks, and spontaneous adventures that helped me navigate adulthood. Amy, Liv, and Shan, thank you for teaching me never to take life too seriously. And Brynn, my best friend since preschool, thank you for always cheering me on and welcoming me home.

Finally, my deepest love goes to my family. To my parents, Jeff and Melinda, who made endless sacrifices so that I could succeed: thank you for trusting me, believing in me, and supporting me through every unknown. I certainly would not have made it through 22 years of education and three continents of academic adventure without you. I am incredibly proud to be the first in our family to earn a doctorate degree, and this milestone belongs to you, too.

And to my fiancé, Mickey: I could not have made it through the grueling final stretch of this journey without you. You have shown me unconditional love and partnership, championing my dreams as if they were your own. Thank you for welcoming me into your family, and for being my home. We have such an exciting future ahead of us.

Table of Contents

Dedication ii

Acknowledgements iii

List of Tables x

List of Figures xi

Abstract xiv

Chapter 1 Background, Motivation, and Dissertation Overview 1

1.1 Background and Motivation 1

1.1.1 State of Prenatal Care 1

1.1.2 Physiology of the Fetal Status 4

1.1.3 Self-Monitoring Fetal Movement 5

1.1.4 Fetal Monitoring in the Clinical Setting 6

1.1.5 Opportunities for Continuous Fetal Monitoring through Wearable Sensors 8

1.1.6 Contextual Factors in Medical Device Design 12

1.1.7 Considerations for Deployment of a Fetal Monitoring System 13

1.2 Dissertation Overview 13

1.2.1 Dissertation Aims 13

1.2.2 Chapter Overview 14

1.3 References 16

Chapter 2 Comparative Analysis of Machine Learning Approaches for Fetal Movement Detection with Linear Acceleration and Angular Rate Signals 24

2.1 Abstract 24

2.2 Introduction 25

2.3 Materials and Methods 27
2.3.1 Participants 27
2.3.2 Sensor System 28
2.3.3 Experimental Protocol 28
2.3.4 Signal Processing and Data Labeling 29
2.3.5 Feature-based Approach: Random Forest Model 31
2.3.6 Time Series Approach: BiLSTM Model 32
2.3.7 Time Frequency Approach: CNN Model 32
2.3.8 Model Training and Evaluation 34
2.3.9 Thresholding Analysis and Training Set Size Reduction 35
2.4 Results 36
2.4.1 Summary of the Results 36
2.4.2 Model Performance Across Sensor Types 36
2.4.3 Model Performance Across Data Representations 37
2.4.4 Thresholding Analysis and Training Set Size Reduction 39
2.5 Discussion 42
2.5.1 Model Performance Across Sensor Types 42
2.5.2 Model Performance Across Data Representations 42
2.5.3 Implications for Fetal Movement Detection 43
2.5.4 Limitations and Future Work 44
2.6 Conclusions 45
2.7 Appendix A: Participant Demographics 46
2.8 References 47
Chapter 3 Four Data-Driven Design Choices for IMU-Based Fetal Movement Detection: Sensitivity, Thresholding, Windowing, and Filtering 53
3.1 Abstract 53

3.2 Introduction 53
3.3 Materials and Methods 55
3.3.1 Data Collection & Categorization 55
3.3.2 Signal Processing & Feature Calculation 57
3.3.3 Statistical Analysis and Planned Comparisons 58
3.4 Results and Discussion 59
3.4.1 Context: Maternal Perception Performance 59
3.4.2 RMS Magnitude Analysis 61
3.4.3 Windowing: Quantifying Fetal Movement Duration 63
3.4.4 Peak Frequency Analysis 64
3.5 Conclusion 66
3.6 Acknowledgements 66
3.7 References 66
Chapter 4 Methodological Evaluation of Label-Source Selection and Evaluation Strategies on Reported Performance in Wearable Fetal Movement Detection 69
4.1 Abstract 69
4.2 Introduction 70
4.3 Background 72
4.4 Methods 74
4.4.1 Data Collection 74
4.4.2 Signal Processing & Event Annotation 76
4.4.3 Splitting of Training and Testing Sets 79
4.4.4 Fetal Movement Classification and Evaluation 80
4.5 Results 82
4.5.1 Dataset Contextualization 82
4.5.2 Reliability of Maternal Perception Annotations for Model Training 83

4.5.3 Random Sampling Inflates Performance 86
4.6 Discussion 87
4.7 Conclusion 92
4.8 References 92
Chapter 5 Contextual Factors Incorporated by Experienced Design Practitioners: A Study Across Three Design Domains 97
5.1 Abstract 97
5.2 Introduction 97
5.3 Background 99
5.4 Methods 102
5.4.1 Participants 103
5.4.2 Data Collection 107
5.4.3 Data Analysis 108
5.5 Findings 110
5.5.1 Participant-level classification of contextual factors incorporated during design 110
5.5.2 Detailed Comparison of contextual factor incorporation across design domains 114
5.5.3 Detailed Comparison of contextual factor incorporation across income settings 117
5.5.4 Incorporating two or more contextual factors into a single design decision 122
5.6 Discussion 125
5.7 Conclusion 130
5.8 Appendix A: Contextual Factor Cards 131
5.9 Appendix B: Example Questions From Interview Protocol 132
5.10 Appendix C: Selected Themes of Incorporation of Contextual Factors Common Across Domain and Income Settings 132
5.11 References 134
Chapter 6 Discussion, Implications, and Future Work 141

6.1 Chapter Summaries 141

6.2 Discussion 144

6.2.1 Controlled Benchmarking for Technical Development of FM Detection Algorithms 144

6.2.2 Role of Ultrasound to Promote Algorithm Explainability through Reduced Uncertainty in Event Labels 146

6.2.3 Impact of Subject-Specific Physiology on Model Generalizability 147

6.2.4 Incorporating Context into the Design and Development of FM Detection Algorithms 148

6.3 Implications 150

6.3.1 Implications for Technical Development: Toward Standardized and Rigorous Benchmarking 150

6.3.2 Implications for Clinical Practice: Balancing Accessibility with Clinical Integrity for Decision Support Tools 154

6.3.3 Operationalizing Fetal Monitoring Systems 156

6.4 Recommendations and Future Work 158

6.5 References 160

List of Tables

Table 1: Demographic information for the participant population included in studies 1 & 2 related to the ability to perceive fetal movement 46

Table 2: Features calculated within each window 58

Table 3: Participant performance metrics of maternal perception against ultrasound clinical standard and related demographics 60

Table 4: Pairwise comparisons between no FM and weak FM groups for RMS magnitude feature (*p*-values Tukey-adjusted, α=0.05) 62

Table 5: Pairwise comparisons between strong FM and artifact groups for RMS magnitude feature (*p*-values Tukey adjusted, α = 0.05) 63

Table 6: Pairwise comparisons between strong FM and artifact groups for peak frequency feature (p-values Tukey adjusted, α=0.05) 65

Table 7: Participant-level analysis of perceived fetal movement compared to the ultrasound clinical standard 83

Table 8: Event-level performance, unseen trials 84

Table 9: Event-level performance, unseen subjects/LOSO 86

Table 10: Participant Information 103

Table 11: Final codebook containing 29 contextual factors organized into 9 primary classifications 109

Table 12: Contextual factor usage counts by domain and context (unit of analysis: participants) 111

List of Figures

Figure 1: Fetal mortality rates in the United States, 1990-2021 Source: National Vital Statistics Reports (2023) 1

Figure 2: Fetal Mortality Rates by Race in the United States Source: National Vital Statistics Report (2023) 2

Figure 3: Maternity care access in the United States Source: March of Dimes (2024) 3

Figure 4: Overview of dissertation chapters 15

Figure 5: Position and alignment of four tri-axial IMUs on the maternal abdomen for fetal movement detection 28

Figure 6: Example of fetal movement in the time series data. Measurements from accelerometer and gyroscope sensors are displayed. Bright green markers indicate maternal perception toggle presses, while labeled fetal movement events are defined to be within the gray boxes 31

Figure 7: Model architecture for the time series approach 32

Figure 8: Example of fetal movement in the time–frequency spectrum 33

Figure 9: Model architecture for the time–frequency approach 34

Figure 10: ROC curves and AUROC values (in legend) comparing acceleration, angular rate, and combined datasets with different models 37

Figure 11: Performance metrics of RF, BiLSTM, and CNN models. Each histogram displays five metrics, accuracy (ACC), specificity (SPEC), sensitivity (SEN), positive predictive value (PPV), and F1 score, with bars indicating the model's performance on combined, angular rate, and acceleration datasets 37

Figure 12: Confusion matrices showing the performance of the RF, BiLSTM, and CNN models across three datasets: acceleration, angular rate, and combined. The models' classification performances are visualized in terms of true positives (top left), false positives (bottom left), false negatives (top right), and true negatives (bottom right). Each cell displays the percentage of the total number of samples, with the raw count shown in parentheses. RF matrices are shown in green, BiLSTM matrices in purple, and CNN matrices in blue. The combined datasets are positioned in the top row, acceleration in the second row, and angular rate in the third row. It is important to note that the models were evaluated using different window sizes:

RF and BiLSTM used 0.5 s windows, while CNN used 8 s windows. This difference in window sizes may affect the distribution of predicted outcomes and should be considered when interpreting the results..... 38

Figure 13: Performance of RF, BiLSTM, and CNN models across varying threshold values (0.1 to 0.9). Each line plot represents a different performance metric, illustrating how model performance changed with threshold adjustments..... 40

Figure 14: Performance of RF, BiLSTM, and CNN models across varying training set sizes (25–100%). Each line plot represents a different performance metric, highlighting how model performance changed with increasing training data..... 41

Figure 15: ROC curves and AUROC values (in legend) for all model types across varying training set sizes..... 41

Figure 16: Functional Alignment Method for orienting angular rate data to the anatomical frame..... 57

Figure 17: Each analytical class plotted for the z-axis of acceleration across all four IMUs. The top left plot is an example of a maternal cough..... 61

Figure 18: RMS magnitude feature (y-axis of angular rate data) used to define sensitivity floor and artifact rejection threshold..... 62

Figure 19: Histogram of FM duration used to inform window size..... 64

Figure 20: Peak frequency feature (y-axis of acceleration data) used to define filter cutoffs..... 65

Figure 21: Placement and orientation of abdominal IMUs used to capture abdominal deformation caused by fetal movement..... 75

Figure 22: Application of the functional alignment method to angular rate data..... 78

Figure 23: Train/test splits employed for model training and evaluation. Approach A depicts model predictions on unseen trials, where every participant has at least one trial in the training and testing sets. Approach B depicts model predictions on unseen subjects. Approach C depicts random sampling of windows..... 80

Figure 24: Strategy used to extract probability scores from model predictions at the event-level81

Figure 25: Receiver-operating characteristic curves associated with each model following a train/test split where every participant has at least one trial in the training and testing sets. The blue curve represents predictions resulting from the model trained with maternal perception (MP) labels, while the orange curve represents predictions resulting from the model trained with ultrasound (US) labels. Bootstrapped 95% confidence intervals are illustrated to show performance separation..... 84

Figure 26: Receiver-operating characteristic curves associated with each model follow a randomly sampled 8:2 train/test split. The blue curve represents predictions resulting from the model trained with maternal perception (MP) labels, while the orange curve represents predictions resulting from the model trained with ultrasound (US) labels. 87

Figure 27: Multi-factor incorporation of contextual factor classifications. Node size represents the frequency with which a classification was incorporated in combination with at least one other classification. Lines connecting classifications indicate co-incorporation, illustrating the breadth of joint incorporation across the dataset. 122

Figure 28: Cross-domain validation and extension of contextual factors in engineering design. Green factors are representative of core contextual factors that are generalizable across domains. Yellow factors are domain-specific contextual factors that are adaptable considerations within domains. Grey factors are contextual factors that may or may not be important considerations, depending on the domain, but are key peripheral explorations..... 125

Figure 29: Summary of findings and their implications for future work. Concepts outlined with solid lines illustrate contributions of this dissertation. Concepts outlined with dashed lines are potential next steps informed by the findings from this work. 144

Abstract

Reduced fetal movement is a critical indicator of fetal risk, yet current monitoring relies on subjective maternal perception or provides only a snapshot of fetal health. While wearable inertial measurement units (IMUs) and machine learning offer a promising solution for continuous monitoring, development is limited by inconsistent methodological benchmarks and a lack of model generalizability. For successful adoption, technical solutions must be tailored to the complex social, political, clinical, and consumer contexts in which they operate. This dissertation establishes a rigorous technical and contextual foundation for the development of wearable IMU-based fetal movement detection systems.

Two studies established evidence-based benchmarks for model training and signal processing. In the first study, IMU data and maternal perception annotations were collected from 23 pregnant participants while seated to inform sensor selection and model architecture. Acceleration, angular rate, and combined datasets were used to train random forest, bi-directional long short-term memory (BiLSTM), and convolutional neural network (CNN) architectures to classify fetal movement. Results showed that combining acceleration and angular rate datasets significantly improved detection performance, with feature-based random forest models achieving the highest AUROC (0.81). The second study incorporated domain knowledge through clinician-validated ultrasound data from a subset of 16 participants to characterize the fetal movement events. Using linear mixed-effect models, statistically significant differences in signal magnitude, frequency, and duration between fetal movement events and maternal artifacts were used to derive data-driven parameters, including artifact thresholds and optimal filter cutoffs.

The third study evaluated label sources and data-splitting strategies to identify limitations in model generalizability using the same 16 participants. Random forest classifiers trained with subjective maternal perception faced a distinct performance ceiling, while ultrasound-labeled events reduced uncertainty and improved accuracy. Furthermore, common random sampling methods, which overlook subject-level data correlation, overestimated model capability in this

study (AUROC = 0.75) compared to a Leave-One-Subject-Out approach (AUROC = 0.61). These results demonstrated that subject-specific physiology like BMI and placental position obscure generalizable signal characteristics. At a 5% false-positive rate, models achieved only 25% sensitivity, indicating that current system performance limits deployment to unsupervised settings for fetal health reassurance.

The fourth study addressed the contextual intersection of design and deployment, acknowledging that technical rigor alone is insufficient for clinical adoption. An existing contextual factor taxonomy was applied to 58 semi-structured interviews with designers across various domains and income settings. Qualitative coding and summative content analysis identified sociocultural, political, and institutional factors as core considerations for successful implementation. Results suggest that fetal monitoring systems must be integrated into existing clinical workflows and domestic routines, and developers must navigate specific regulatory constraints to ensure safety and clinical trust.

The technical benchmarks and validation strategies presented provide a roadmap for moving fetal movement detection from laboratory proof-of-concept to a reproducible clinical standard by identifying the physiological barriers to generalizability and the contextual factors required for successful adoption. Together, these contributions bring us closer to the development of contextually appropriate, clinically-adopted, and personalized tools capable of studying longitudinal fetal activity trends to improve antenatal outcomes and reduce the risk of stillbirth.

Chapter 1 Background, Motivation, and Dissertation Overview

1.1 Background and Motivation

1.1.1 State of Prenatal Care

1.1.1.1 Prenatal Care Statistics in the United States

In the United States, there are approximately 4 million births every year [1,2]. Over 98% of women who give birth seek prenatal care at some point during the 40-42 week gestational cycle, positioning prenatal care as one of the most widely used preventative healthcare measures [1,2]. Despite its widespread use, prenatal care in the United States faces persistent challenges related to access, equity, and birth outcomes. In 2021, more than 20,000 stillbirths (fetal death after 28 weeks of gestation) were reported [3], and an estimated 32% of these deaths could have been prevented through adequate prenatal care [4]. Yet, despite advancements in medical technology and shifts in population health, the stillbirth rate has remained relatively unchanged since the 1990s and has stabilized since 2003 (Figure 1) [3].

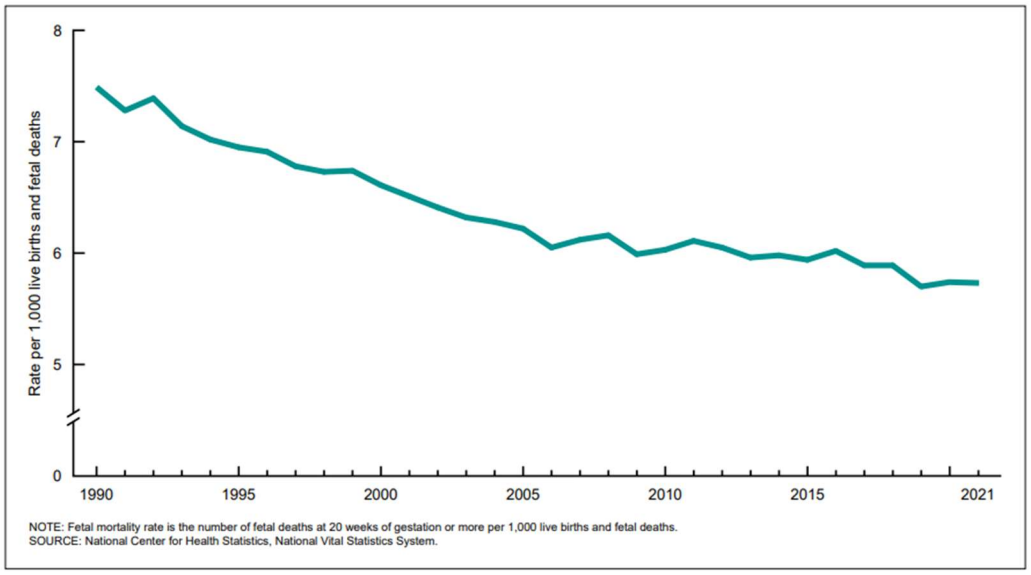


Figure 1: Fetal mortality rates in the United States, 1990-2021
Source: National Vital Statistics Reports (2023)

These outcomes disproportionately affect women of color. The preterm birth rate is 1.6 times higher for non-Hispanic Black women than for non-Hispanic White women [5], and the stillbirth rate is twice as high (9.89 vs. 4.85 per 1,000 births) (Figure 2) [3,6]. Having a stillbirth puts mothers at risk, too. Black mothers also face an increased risk of maternal death, with the likelihood rising as much as four-fold among those who have experienced a stillbirth [7]. The non-Hispanic Black maternal mortality is 2.5 times higher than that of non-Hispanic White women and 2.9 times higher than that of Hispanic women [8].

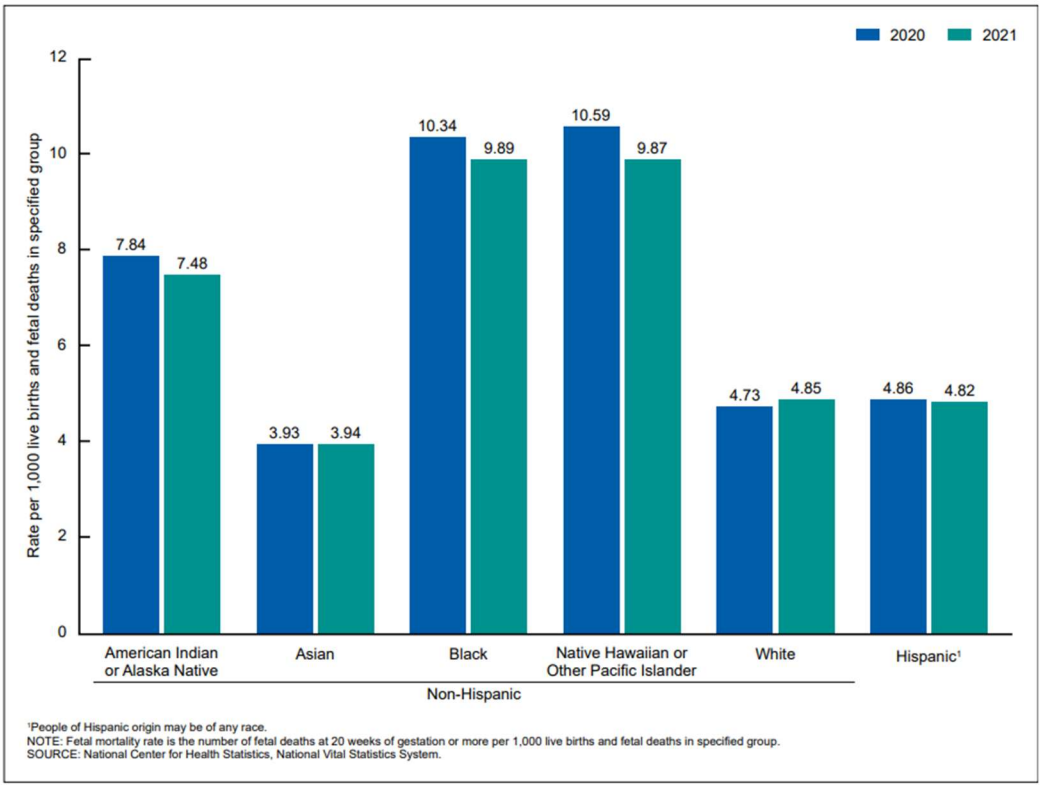


Figure 2: Fetal Mortality Rates by Race in the United States
Source: National Vital Statistics Report (2023)

Geographic access further compounds these disparities. In 35% of U.S. counties, classified as maternity care deserts (no birthing center or obstetric clinician) (Figure 3), patients often travel long distances for services, with an average drive time of 38 minutes (2.6 times longer than for women in full-access counties) [5]. Limited access is associated with a 13% higher risk of preterm birth and contributed to more than 10,000 excess preterm births between 2020 and 2022 [5].

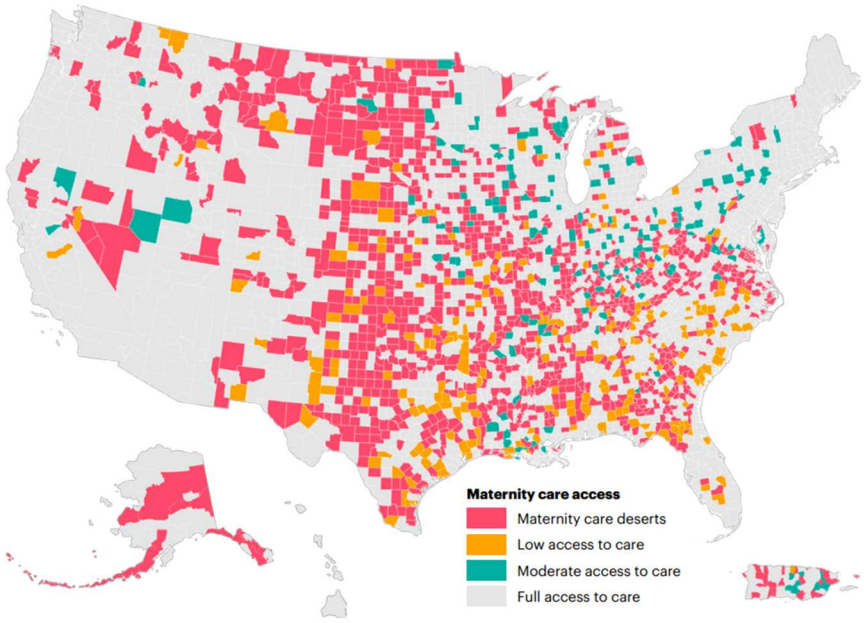


Figure 3: Maternity care access in the United States
Source: March of Dimes (2024)

1.1.1.2 Existing Model of Prenatal Care in the United States

These persistent disparities exist alongside a prenatal care delivery model that has remained largely unchanged for nearly 75 years [9–11]. The traditional model consists of 12–14 in-person visits: monthly appointments until 28 weeks of gestation, biweekly until 32 weeks, and weekly after 32 weeks until delivery [9,11,12]. This schedule is standard despite evidence that prenatal care is not a "one-size-fits-all" approach. Most visits focus on routine screening for vital signs and pregnancy symptoms, with only a few requiring critical services such as ultrasounds, vaccinations, or laboratory testing [11,12]. Accounting for travel, parking, and in-clinic procedures, adherence to the standard schedule equates to roughly 40 hours (an entire workweek) of missed employment or childcare [9]. For low-risk patients, more than 10 visits may be unnecessary and potentially harmful, increasing the likelihood of interventions such as additional ultrasounds or cesarean deliveries [13–15]. Reducing required visits from 14 to 10 could enable clinicians to care for 20 additional low-risk patients annually and avoid $3.9 billion per year from excess visits [16].

The COVID-19 pandemic prompted the first major shift to this model in decades. In March 2020, the American College of Obstetricians and Gynecologists (ACOG) endorsed both a reduced visit schedule and the integration of virtual care [17]. Many practices quickly adopted telemedicine, hybrid models, and streamlined schedules. While longitudinal data are not yet

available, early studies of virtual care and remote monitoring demonstrate positive maternal and fetal outcomes alongside high patient satisfaction. Notably, 85% of patients expressed interest in telemedicine contact between visits, and many reported confidence in performing remote monitoring tasks such as tracking weight, blood pressure, and fetal heart tones, activities that previously required in-person care [2].

1.1.2 Physiology of the Fetal Status

1.1.2.1 Warning Signs of an At-Risk Fetus

Not all adverse birth outcomes are preventable, but improving the quality and accessibility of prenatal care is critical to reducing preventable morbidity and mortality. Early signs of fetal compromise are often subtle and may be missed, making timely intervention difficult before serious complications arise. Monitoring the fetal status is essential for preventing fetal death and identifying cases that may benefit from interventions such as resuscitation or early delivery [18]. A fetus is considered at risk when oxygen availability is insufficient (hypoxemia) or when acid accumulates in the blood (acidemia) [18,19]. Three primary indicators of fetal well-being are sensitive to these conditions: fetal heart rate, movement, and tone [18,19]. These markers are physiologically interrelated, and the presence of fetal heart rate accelerations in association with fetal movement (reflecting adequate tone) is generally accepted as evidence of sufficient oxygenation to support the fetal nervous system [18]. Both fetal movement and fetal heart rate provide important insights into neurological, cardiovascular, and respiratory function [18].

1.1.2.2 Fetal Movement as an Indicator of Fetal Health

Fetal movement develops around the ninth week of gestation and can be perceived by the mother after approximately the 16th week [20]. The frequency of movements increases as gestation progresses, peaking between 28 and 34 weeks [21], and gradually declines thereafter due to improved motor coordination and reduced amniotic fluid volume resulting from increased fetal size [22]. Fetal movement is the simplest and most commonly assessed indicator of fetal wellbeing because it is the only one that can be directly perceived by the mother without clinical instrumentation.

Given its dynamic nature throughout gestation, consistent monitoring of fetal movement patterns is important. Regular fetal movement is a reassuring sign of fetal health [22,23], which

includes regular bouts of fetal hiccups—rhythmic, repetitive diaphragmatic movements that are widely recognized as a normal indicator of healthy central nervous system development [24,25]. Conversely, reduced fetal movement (RFM) is associated with complications like stillbirth, fetal growth restriction, congenital anomalies, and fetal hypoxia [26,27]. In fact, 30-55% of women who experience stillbirth report RFM in the preceding week [28,29], and those with RFM are up to five times more likely to experience stillbirth compared with the general pregnant population [30]. RFM is an adaptive response to an insufficient placental environment and serves as a key warning sign that timely intervention may be necessary [22,23,26].

1.1.3 Self-Monitoring Fetal Movement

1.1.3.1 Maternal Perception of Fetal Movement

Because fetal movement is the only indicator of fetal wellbeing that can be assessed by the mother without clinical instrumentation, maternal perception (or more commonly kick counting) is one of the oldest and most commonly used methods to monitor the fetal status [23]. Even without prompting by a clinician, many pregnant women use fetal movement patterns as a screening tool, with approximately 15% of women presenting during the third trimester with concerns about RFM [23]. Daily counting can help mothers detect changes in fetal activity and heighten vigilance for warning signs, but it also places a significant responsibility on the mother, particularly in the context of reduced prenatal visit schedules [23].

1.1.3.2 Limitations of Self-Monitoring

Despite its accessibility, maternal perception of fetal movement is inherently subjective, imprecise, and, for some, anxiety-inducing. Sensitivity compared with ultrasound detection ranges from 33% to 88% [22,31], and is influenced by posture, placental location, and the fetal sleep–wake cycle [18,23,28,31,32]. Movements are most often perceived when the mother is lying down, followed by sitting, and least when standing, likely reflecting differences in distraction rather than true movement frequency [22,32]. An anterior placenta may dampen movement perception [22,31], and fetal activity follows a diurnal pattern, with stronger and more frequent movements later in the day [20,31,33,34]. The correlation between actual fetal movements and maternal perception during normal daily activities is unknown but is expected to be significantly lower than when women are resting and focused during ultrasound examination [23]. Fetal movement counting requires sustained attention, as some movements are subtle.

Women who are busy or distracted often report an inaccurate reduction in movements [22]. Compliance is variable, with only 60% of women adhering to counting protocols and just 50% responding appropriately to alarm triggers [30]. The constant vigilance required can generate significant anxiety, and women presenting with RFM are at increased risk of obstetric intervention, particularly induction of labor and planned early-term birth [35].

1.1.3.3 Published Standards and Guidelines

Because fetal movement patterns change dynamically throughout gestation, establishing standardized thresholds for clinically significant RFM is challenging. While most guidelines advise clinical assessment when there is any perceived reduction in frequency or change in pattern [35], the wide range of normal variation makes quantification difficult [22]. Definitions of RFM in the literature range from complete absence of movement for 24 hours to fewer than 10 movements in 2 hours, with median counting times reported between 15 minutes and 12 hours [23]. No objective lower limit of normal has been determined, as both interindividual and intraindividual variation are substantial [22,23,30]. Consequently, while maternal perception remains a valuable screening tool, it is best interpreted in combination with other clinical assessments.

1.1.4 Fetal Monitoring in the Clinical Setting

1.1.4.1 Technologies for Fetal Surveillance

When RFM is reported, clinical assessment typically begins with the non-stress test (NST), which evaluates the relationship between fetal heart rate and movement over a 20-minute period [18]. A NST is considered reassuring if there are at least two accelerations of fetal heart rate within the observation window [18]. Although reactivity generally indicates fetal well-being, its absence is often attributable to the fetal sleep cycle but can also reflect early signs of compromise [18]. Non-reassuring results occur in approximately 15% of cases (and up to 50% before 28 weeks of gestation) [18].

Fetal heart rate is most often measured using cardiotocography (CTG), which employs two transducers: one to detect heart rate via reflected ultrasound waves, and another to measure uterine contractions through pressure changes at the abdominal surface [36]. While CTG is sensitive to fetal activity, it is also prone to signal noise, necessitating frequent adjustments [36].

This, combined with biological variability, contributes to the NST's high false-positive rate for non-reassuring results (75–90%) [18].

In some cases, clinicians supplement the NST with a biophysical profile (BPP), which integrates four ultrasound-based assessments: fetal breathing movements, gross body or limb movements, tone (flexion/extension of extremities), and amniotic fluid volume alongside the NST [18]. Each parameter is scored as 0 (absent/abnormal) or 2 (present/normal), with totals ≥8 considered reassuring and ≤4 abnormal [18].

The contraction stress test (CST) evaluates fetal heart rate in relation to uterine contractions utilizing the same CTG technology as the NST [18]. The CST is conducted to identify late decelerations that may occur during the transient reduction in oxygen supply caused by contractions [18]. A reassuring CST shows no late or significant variable decelerations, indicating adequate oxygenation [18]. Approximately 25% of patients with a non-reassuring NST will have a reassuring CST [18]. The CST should not be performed in cases of preterm labor, preterm rupture of membranes, prior extensive uterine surgery, or placenta previa [18]. As with the NST, high false-positive rates (8–57%) can limit specificity [37].

1.1.4.2 Interpretation of Clinical Testing

Non-reassuring NST results are generally followed by either a BPP or CST. If both are non-reassuring, early delivery is considered, although resuscitative measures such as maternal repositioning, fluid administration, or oxygen may be employed to prolong gestation when feasible [18]. When the NST is non-reassuring but the CST is reassuring, the absence of reactivity may be explained by factors such as prematurity, sleep cycle, or preexisting neurological conditions [18]. Reassuring results are typically considered valid for one week, with weekly (or more frequent) testing recommended for high-risk pregnancies [18]. If RFM is not persistent or recurrent, repeat testing is not necessary [18].

1.1.4.3 Limitations of Clinical Fetal Monitoring

While critical for assessing fetal and maternal status, current clinical monitoring provides only a "snapshot in time" that may not fully capture fetal well-being. Technologies such as CTG are susceptible to noise [36] and cannot readily be applied during activities of daily living. Interpretation requires specialized expertise and a clinical setting, increasing both time and cost burdens. High false-positive rates [18,37] further contribute to unnecessary interventions up to

and including early delivery. As a result, while existing clinical monitoring techniques yield valuable information, they are insufficient for continuous, accurate monitoring outside of clinical settings.

1.1.5 Opportunities for Continuous Fetal Monitoring through Wearable Sensors

1.1.5.1 History of Wearable Medical Devices for Continuous Monitoring

Wearable devices have entered both clinical and consumer spaces over the past two decades, enabled by advances in miniaturized electronics, wireless data transmission, and reductions in cost [38–40]. Early systems, such as the Holter monitor for ambulatory cardiac monitoring, demonstrated the value of collecting physiological signals outside the hospital [40,41]. Before the availability of wearable devices, long-term monitoring often required repeated or extended hospital visits, disrupting daily life and creating substantial financial burden [40]. The ability to monitor patients in their own environments has shifted the model of care, making it possible to collect larger volumes of data, provide clinicians with more continuous information, and reduce unnecessary admissions [40]. Continuous glucose monitors, for example, allow patients with diabetes to respond quickly to critical changes [42], while sleep monitoring systems capture longitudinal data that support the diagnosis and management of chronic conditions such as insomnia and narcolepsy [43]. These examples highlight the broader potential of wearable devices to improve patient outcomes, enable earlier diagnoses, and expand opportunities for data-driven decision-making [40]. Despite this promise, most commercially available wearables have been designed for consumer rather than clinical use. Although they can promote awareness and support behavior change, their data quality and accuracy often remain insufficient for medical decision-making [40]. As a result, the disruptive potential of wearable devices in healthcare has not yet been fully realized.

In the context of fetal monitoring, these limitations are especially evident. Despite decades of research, no existing technology enables continuous, accurate collection of fetal movement data outside the clinic. This gap has left the field without a foundational understanding of fetal movement patterns across gestation or evidence-based definitions of reduced fetal movement (RFM). Current clinical tools are highly susceptible to signal noise, limiting their use during activities of daily living (ADL) and restricting insight into how fetal movement responds to real-world environments. The absence of high-quality, longitudinal

datasets hinders three critical objectives: (1) establishing consensus on normal fetal movement patterns across gestation, (2) evaluating the potential of fetal movement monitoring for stillbirth risk management, and (3) improving clinical decision-making in cases of reported RFM. Addressing these gaps will require new wearable technologies capable of accurate, continuous monitoring in daily life, laying the foundation for the next generation of fetal surveillance.

1.1.5.2 Background on Inertial Measurement Units (IMUs)

Among the sensing technologies that have enabled wearable health monitoring, inertial sensors were quickly adopted by researchers to capture human movement patterns outside of laboratory settings [38,44]. Inertial measurement units (IMUs) typically include triaxial accelerometers and gyroscopes. Accelerometers measure specific force by detecting proof-mass displacement. Gyroscopes measure angular rate using Coriolis forces acting on a vibrating element [45]. Together, IMUs capture both linear and rotational motion signals that can be mapped to patterns of human movement. Paired with the evolution of signal processing algorithms, wearable IMU systems have moved from laboratory and ambulatory-based settings to unsupervised naturalistic environments, allowing researchers to learn new insights about human activity in real-world conditions [39]. These technologies have enabled the development of a variety of health applications including the assessment of frailty, detection of falls, and diagnosis of chronic movement disorders, and the promotion of healthy living [39]. Acceptance among consumers and clinicians alike has been driven by wearable IMUs cost-effective and unobtrusive manner, allowing the technology to evolve to the point where such systems are used interchangeably with conventional lab-based motion analysis systems and in routine clinical practice to monitor and assess several types of movement disorders [39]. Like other wearables, though, IMUs are known to impose several signal processing challenges, despite the value they provide [38].

1.1.5.3 Overview of Prior Work in Fetal Monitoring

In response to the gaps in current fetal monitoring technologies, there is growing interest in using wearable sensor-based systems to measure fetal movement outside the clinic. Wearable sensors are non-invasive and can enable objective, continuous data collection. These sensors are compact and low-cost and have the potential to facilitate long-term and at-home monitoring of fetal wellbeing. With longitudinal datasets, wearable sensor-based systems have the potential to

impact fetal and maternal health by informing clinical decision making and serve as an academic research tool to uncover critical information about population trends in fetal movement patterns.

Sensor Configurations and Modalities

Over the past decade, research on fetal movement detection has explored a wide range of sensor configurations, signal processing strategies, and machine learning models to identify fetal activity outside the clinic. Most studies have relied on arrays of abdominal accelerometers [46–54] (typically three to five sensors) placed to capture motion tangential to the maternal abdomen, where fetal movements transmit the greatest force. Some supplemented abdominal sensors with a reference on the chest or upper back to help distinguish fetal from maternal movement [46,48,49,51,54]. Only one study by Delay et al. used a single accelerometer [52].

Accelerometers have shown great promise for detecting fetal movement, but there is no consensus on the optimal configuration. Altini et al. is the only study to directly compare the number of sensors, reporting that two abdominal accelerometers achieved the best trade-off between accuracy and simplicity, with only marginal gains from adding more sensors [54]. A reference on the upper back improved performance, though a chest reference was not tested. Other sensor types including fiber-optic [55], acoustic [56], and piezoelectric sensors [57] have been investigated but generally underperformed compared to accelerometers.

Sensor fusion, which integrates multiple sensor types, has shown promise in related fields such as infant apnea detection [58], pulmonary monitoring [59], and dietary behavior tracking [60]. In fetal monitoring, only two studies have attempted fusion: Ghosh et al. combined accelerometers with other sensors (90% accuracy vs. maternal perception, but tested on just five participants) [57], and Alwis et al. combined accelerometers with gyroscopes via inertial measurement units (88% accuracy) [61]. While limited, these results suggest that multimodal approaches may improve fetal movement detection performance, though systematic comparisons of sensor trade-offs remain lacking.

Signal Processing Techniques

Wearable systems are prone to high levels of signal noise, with artifacts arising from maternal motion [47]. Even subtle sources such as respiration and heart rate can be detected in acceleration signals and resemble fetal movement, while activities like laughing, shifting

position, or talking introduce superimposed motion that is especially challenging to separate from true fetal activity [48,53].

To address signal artifacts, most studies begin with filtering methods such as infinite impulse response (IIR) [46–49,52,54] or Kalman [50] filters to remove baseline drift, electrical interference, and high-frequency noise. High-pass IIR filters are particularly common for removing low-frequency maternal respiration. Increasingly, researchers have leveraged the frequency domain to improve discrimination, applying Fourier transforms or time–frequency methods (e.g., short-time Fourier transform, wavelets) to capture both spectral and temporal features of fetal activity. For example, Mesbah et al. and Xu et al. applied continuous wavelet decomposition to identify and remove maternal heartbeat artifacts before detecting fetal movements. Other preprocessing methods include independent component analysis [48], thresholding [46,49,50,61], and statistical outlier removal [53]. Despite these efforts, fetal movement signals remain complex and variable, observed as short bursts of low-frequency activity embedded in noisy maternal signals, making detection inherently challenging.

Machine Learning Approaches

To capture the dynamic characteristics of fetal movement, signal processing techniques have been paired with machine learning models. Shallow classifiers such as support vector machines [49], random forests [54], and ensemble methods [48,53] remain common, while deep learning architectures, particularly convolutional neural networks (CNNs) [52] and CNN–LSTM hybrids [61], are becoming more popular for their ability to learn spatial–temporal patterns directly from raw signals.

Mesbah et al. and Xu et al. both compared multiple feature-based classifiers against ultrasound annotations. Mesbah achieved highest accuracy using a bagged random forest (89%), while Xu reported an extra trees classifier outperforming other classifiers (87% accuracy). However, no study has directly compared shallow models to deep networks, and deep learning remains underexplored in this domain. Reported performance varies widely across studies, influenced by sensor setup, preprocessing pipeline, and algorithm choice. Generally, deep learning with multimodal inputs and frequency-domain features yields higher performance, while simpler models often trade sensitivity for specificity [62,63].

Label Source and Definition

Performance comparisons are complicated by differences in label source and fetal movement definitions. Most studies use either maternal perception (button press) or ultrasound imaging for reference. Maternal perception is widely available but subjective and imprecise, with some mothers detecting only one-third of true fetal movements [31]. Ultrasound provides visual confirmation but can miss events outside the imaging plane or obscured by maternal motion [46]. Ultrasound also requires review by a trained clinician to provide signal annotations. Labeling strategies further increase variability: some studies mark precise movement onset and offset [46,49,61], while others allow ±30-second buffers [53]. These inconsistencies undermine comparability between studies and make it difficult to establish a reliable benchmark. Ultimately, neither maternal perception nor ultrasound provides a perfect reference, creating fundamental challenges for training and validating detection algorithms.

Limitations and Barriers to Deployment

Despite progress, the field remains fragmented. Wide variation in sensor configurations, preprocessing methods, and algorithm choices has produced an unstandardized and difficult-to-navigate research landscape [30,54]. In addition, studies often report different sets of performance metrics, which hinders direct comparison and further limits generalizability. Nearly all published studies have been conducted with stationary, lying-down participants, leaving the real-world effectiveness of current systems largely untested.

There is an unmet need for reliable fetal movement monitoring methods, given their crucial role in assessing fetal well-being and reducing adverse birth outcomes. However, progress toward commercial deployment remains limited by three major challenges: (1) difficulty separating fetal signals from maternal motion artifacts, (2) lack of a perfect clinical standard for evaluating algorithm performance, and (3) inconsistent study protocols and metrics that limit reproducibility and comparison across studies. Together, these challenges create barriers to deploying clinically meaningful and generalizable solutions for fetal monitoring.

1.1.6 Contextual Factors in Medical Device Design

While wearable sensors hold clear promise for continuous fetal monitoring, their success will ultimately depend on thoughtful integration into real-world care settings. Researchers note that despite the pace of innovation in medical device design, design processes often exclude

meaningful considerations of intended end users, stakeholders, and the broader communities and contexts in which they are situated [64]. Failure to incorporate meaningful contextual information has been shown to lead to many medical device failures [64–66]. Contextual factors are the elements of a technology's broad use context that could affect how the technology would be implemented and used in practice, including social, cultural, political, and economic factors [64,66,67]. Studies of engineering practice have identified that professional engineers investigate contextual factors and sociotechnical considerations [64,67–70], but there is limited guidance on what aspects of context to consider and how to incorporate contextual factors into design processes.

1.1.7 Considerations for Deployment of a Fetal Monitoring System

Unlike consumer devices, wearable medical technologies are used by diverse patient populations under the guidance of healthcare providers, often without the patient's voluntary choice [40]. As such, factors such as usability, comfort, cultural acceptability, and long-term compliance become as critical as technical performance. Devices must be designed to operate reliably during activities of daily life, minimize burden for expectant mothers, and provide clinicians with data that are interpretable and actionable. Beyond user experience, deployment requires attention to power management, data security, and decisions around local storage versus wireless transmission [40], each of which may vary depending on clinical context and resource setting. These design considerations highlight that the challenges of wearable fetal monitoring extend beyond algorithm development and sensor accuracy. Understanding how contextual factors shape medical device design will be essential to ensuring that new technologies are not only feasible in principle but also acceptable, sustainable, and impactful in practice.

1.2 Dissertation Overview

1.2.1 Dissertation Aims

This dissertation aims to advance the design and development of a wearable, continuous fetal monitoring system for real-world deployment through two complementary objectives: (1) the development and validation of algorithms for fetal movement detection using wearable sensors and (2) the investigation of contextual considerations in medical device design that are essential for deployment across diverse care settings.

1.2.2 Chapter Overview

This dissertation is organized into two parts, as outlined in Figure 4, aligned with these aims:

Part 1: Establish evidence-based benchmarks for signal processing and algorithm development

This part addresses the lack of standardized methods in existing literature for detecting fetal movement with wearable sensors. Data were collected from 23 third-trimester participants who wore four abdominal IMUs and one chest reference while reporting perceived fetal movements and receiving an ultrasound scan.

Study 1: *Comparative Analysis of Machine Learning Approaches for Fetal Movement Detection with Linear Acceleration and Angular Rate Signals*

This study evaluated whether combining accelerometer and gyroscope signals improves detection performance. Three model types—Random Forest (RF), BiLSTM, and CNN—were trained on hand-engineered features, time series, and spectrograms. Results showed that integrating acceleration and angular rate signals improved detection, with RF and BiLSTMs achieving the best overall performance, while RF showed the greatest capacity for fetal movement detection when limiting false reassurance of fetal health.

Study 2: *Four Data-Driven Design Choices for IMU-Based Fetal Movement Detection: Sensitivity, Thresholding, Windowing, and Filtering*

This study used clinician-validated ultrasound data to precisely define fetal movement events, enabling the characterization of IMU signals for data-driven signal processing and hardware design parameters to support future algorithm development. Linear Mixed-Effect Models (LMEMs) were applied to 1371 fetal movement events across four analytical classes (No fetal movement, Weak fetal movement, Strong fetal movement, Artifact) and the estimated marginal means (EMMs) of the groups were compared to quantify the presented design choices.

Study 3: *Methodological Evaluation of Label-Source Selection and Evaluation Strategies on Reported Performance in Wearable Fetal Movement Detection*

This study compared the performance of maternal perception and ultrasound labels for training machine learning models that classify fetal movement events from wearable IMU signals under realistic evaluation strategies. The results demonstrated that while maternal perception is a scalable training source, it introduces label noise that constrains model reliability.

Furthermore, the significant performance disparity between random sampling and participant-independent splitting underscores the necessity of rigorous validation to avoid false reassurance in clinical deployment.

Part 2: Define a contextual framework for design-decision making

This part addresses the lack of design support tools for incorporating contextual factors into the design of emerging technologies to support adoption alongside technical development.

Study 4: *Contextual Factors Incorporated by Experienced Design Practitioners: A Study Across Three Design Domains*

This study examined the validity of an existing contextual factor taxonomy across design domains and its ability to capture a shared structure of contextual reasoning in engineering design. Using qualitative data collected from 58 practitioners across engineering design domains and income settings, an existing taxonomy of contextual factors was applied and critically evaluated. Qualitative coding and summative content analysis showed that many contextual factor categories are generalizable across domains, suggesting the presence of a core, underlying structure in how designers account for context. At the same time, domain-specific patterns were identified that motivated nuanced design behaviors within the taxonomy. This work advances theoretical understanding of how context is incorporated into design processes and informs the development of more adaptable design methods, tools, and educational approaches.

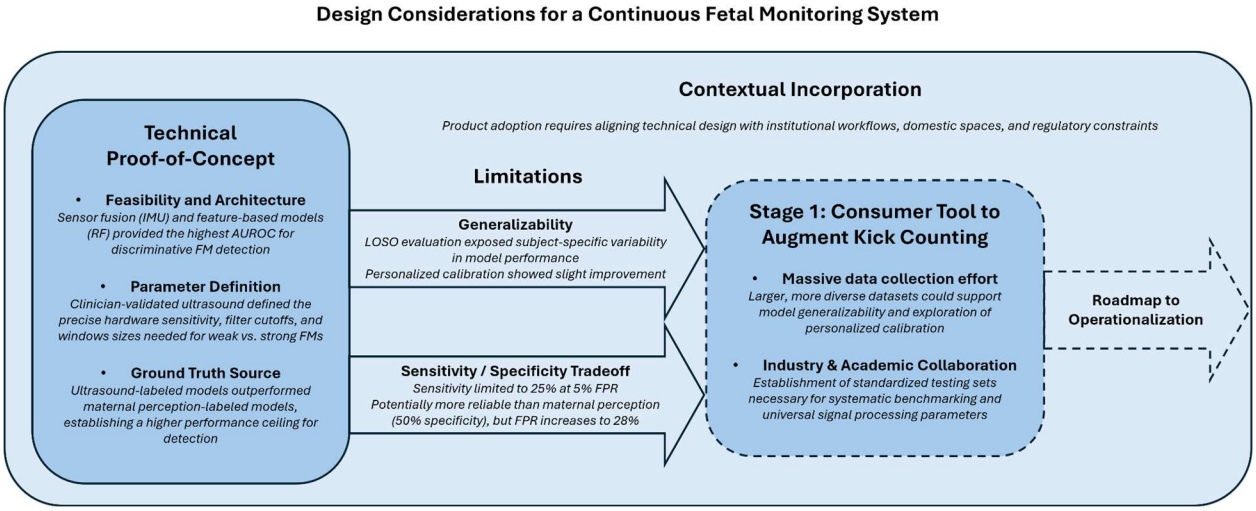


Figure 4: Overview of dissertation chapters

Together, these studies advance both the methodological rigor and validity of fetal movement detection systems and the contextual framework necessary to support future

development that supports adoption of wearable monitoring systems into clinical and consumer settings.

1.3 References

[1] Osterman, M. J. K., and Martin, J. A., 2016, *Timing and Adequacy of Prenatal Care in the United States, 2016*.

[2] Peahl, A. F., Novara, A., Heisler, M., Dalton, V. K., Moniz, M. H., and Smith, R. D., 2020, "Patient Preferences for Prenatal and Postpartum Care Delivery: A Survey of Postpartum Women," Obstetrics and Gynecology, **135**(5), pp. 1038–1046. https://doi.org/10.1097/AOG.0000000000003731.

[3] Gregory, E. C. W., Valenzuela, C. P., and Hoyert, D. L., 2020, *National Vital Statistics Reports Volume 71, Number 8 July 26, 2023*. [Online]. Available: https://www.cdc.gov/nchs/products/index.htm.

[4] Gardosi, J., Madurasinghe, V., Williams, M., Malik, A., and Francis, A., 2013, "Maternal and Fetal Risk Factors for Stillbirth: Population Based Study," BMJ (Online), **346**(7893), pp. 1–14. https://doi.org/10.1136/bmj.f108.

[5] Stoneburner, A., Lucas, R., Fontenot, J., Brigance, C., Jones, E., and DeMaria, A., 2024, *Nowhere to Go: Maternity Care Deserts Across the US*. [Online]. Available: https://www.marchofdimes.org/maternity-care-deserts-report. [Accessed: 16-Sep-2025].

[6] MacDorman, M. F., and Gregory, E. C. W., 2015, "Fetal and Perinatal Mortality: United States, 2013," National Vital Statistics Reports, **64**(8), pp. 1–24.

[7] Wall-Wieler, E., Carmichael, S., Gibbs, R., Lyell, D., Girsen, A., Yasser, Y., and Butwick, A., 2019, "Severe Maternal Morbidity Among Stillbirth and Live Birth Deliveries in California," Obstetrics & Gynecology, **134**(1).

[8] Hoyert, D., 2023, *Maternal Mortality Rates in the United States, 2021*, Atlanta, Georgia. https://doi.org/10.15620/cdc:124678.

[9] Peahl, A. F., Gourevitch, R. A., Luo, E. M., Fryer, K. E., Moniz, M. H., Dalton, V. K., Fendrick, A. M., and Shah, N., 2020, "Right-Sizing Prenatal Care to Meet Patients' Needs and Improve Maternity Care Value," Obstetrics and Gynecology, **135**(5), pp. 1027–1037. https://doi.org/10.1097/AOG.0000000000003820.

[10] Alexander, G. R., and Kotelchuck, M., 2001, *Assessing the Role and Effectiveness of Prenatal Care: History, Challenges, and Directions for Future Research*.

[11] Peahl, A. F., Smith, R. D., and Moniz, M. H., 2020, "Prenatal Care Redesign: Creating Flexible Maternity Care Models through Virtual Care," Am. J. Obstet. Gynecol., **223**(3), pp. 389.e1-389.e10. https://doi.org/10.1016/j.ajog.2020.05.029.

[12] Kilpatrick, S., Papile, L.-A., Macones, G., and editors, 2017, *Guidelines for Perinatal Care*, Elk Grove Village, IL/Washington D.C. .

[13] Carter, E. B., Tuuli, M. G., Caughey, A. B., Odibo, A. O., Macones, G. A., and Cahill, A. G., 2016, "Number of Prenatal Visits and Pregnancy Outcomes in Low-Risk Women," Journal of Perinatology, **36**(3), pp. 178–181. https://doi.org/10.1038/jp.2015.183.

[14] Dowswell T, Carroli G, Duley L, Gates S, Gülmezoglu AM, Khan-Neelofur D, and Piaggio G, 2015, " Alternative versus Standard Packages of Antenatal Care for Low-risk Pregnancy," Cochrane database of systematic reviews, **7**.

[15] Butler Tobah, Y. S., LeBlanc, A., Branda, M. E., Inselman, J. W., Morris, M. A., Ridgeway, J. L., Finnie, D. M., Theiler, R., Torbenson, V. E., Brodrick, E. M., Meylor de Mooij, M., Gostout, B., and Famuyide, A., 2019, "Randomized Comparison of a Reduced-Visit Prenatal Care Model Enhanced with Remote Monitoring," Am. J. Obstet. Gynecol., **221**(6), pp. 638.e1-638.e8. https://doi.org/10.1016/j.ajog.2019.06.034.

[16] Truven Health Analytics, 2013, *The Cost of Having a Baby in the United States*. [Online]. Available: https://www.catalyze.org/product/2013-cost-baby-united-states/. [Accessed: 16-Sep-2025].

[17] Peahl, A. F., and Howell, J. D., 2021, "The Evolution of Prenatal Care Delivery Guidelines in the United States," Am. J. Obstet. Gynecol., **224**(4), pp. 339–347. https://doi.org/10.1016/j.ajog.2020.12.016.

[18] Walton, J. R., and Peaceman, A. M., 2012, "Identification, Assessment and Management of Fetal Compromise," Clin. Perinatol., **39**(4), pp. 753–768. https://doi.org/10.1016/j.clp.2012.09.001.

[19] Manning, F. A., and Platt, L. D., 1979, "Maternal Hypoxemia and Fetal Breathing Movements," Obstetrics & Gynecology, **53**(6), pp. 758–760.

[20] Raynes-Greenow, C. H., Gordon, A., Li, Q., and Hyett, J. A., 2013, "A Cross-Sectional Study of Maternal Perception of Fetal Movements and Antenatal Advice in a General Pregnant Population, Using a Qualitative Framework," BMC Pregnancy Childbirth, **13**. https://doi.org/10.1186/1471-2393-13-32.

[21] Mangesi, L., Hofmeyr, G. J., Smith, V., and Smyth, R. M. D., 2015, "Fetal Movement Counting for Assessment of Fetal Wellbeing," Cochrane Database of Systematic Reviews, **2015**(10). https://doi.org/10.1002/14651858.CD004909.pub3.

[22] Olesen, A. G., and Svare, J. A., 2004, "Decreased Fetal Movements: Background, Assessment, and Clinical Management," Acta Obstet. Gynecol. Scand., **83**(9), pp. 818–826. https://doi.org/10.1111/j.0001-6349.2004.00603.x.

[23] Heazell, A. E. P., and Frøen, J. F., 2008, "Methods of Fetal Movement Counting and the Detection of Fetal Compromise," J. Obstet. Gynaecol. (Lahore)., **28**(2), pp. 147–154. https://doi.org/10.1080/01443610801912618.

[24] Thompson, J. M. D., Wilson, J., Bradford, B. F., Li, M., Cronin, R. S., Gordon, A., Raynes-Greenow, C. H., Stacey, T., Cullling, V. M., Askie, L. M., O'Brien, L. M., Mitchell, E. A., McCowan, L. M. E., and Heazell, A. E. P., 2021, "A Better Understanding of the Association between Maternal Perception of Foetal Movements and Late Stillbirth—Findings from an Individual Participant Data Meta-Analysis," BMC Med., **19**(1). https://doi.org/10.1186/s12916-021-02140-z.

[25] Heazell, A. E. P., Budd, J., Li, M., Cronin, R., Bradford, B., McCowan, L. M. E., Mitchell, E. A., Stacey, T., Martin, B., Roberts, D., and Thompson, J. M. D., 2018, "Alterations in Maternally Perceived Fetal Movement and Their Association with Late Stillbirth: Findings from the Midland and North of England Stillbirth Case-Control Study," BMJ Open, **8**(7). https://doi.org/10.1136/bmjopen-2017-020031.

[26] Scala, C., Bhide, A., Familiari, A., Pagani, G., Khalil, A., Papageorghiou, A., and Thilaganathan, B., 2015, "Number of Episodes of Reduced Fetal Movement at Term: Association with Adverse Perinatal Outcome," Am. J. Obstet. Gynecol., **213**(5), pp. 678.e1-678.e6. https://doi.org/10.1016/j.ajog.2015.07.015.

[27] Ryo, E., Kamata, H., and Yatsuki, K., 2025, "Correlation between Perinatal Abnormalities and Decreased Fetal Movement as Counted by a Fetal Movement Acceleration Measurement Recorder: A Prospective Cohort Study," International Journal of Gynecology and Obstetrics, **169**(2), pp. 656–662. https://doi.org/10.1002/ijgo.16101.

[28] Dutton, P. J., Warrander, L. K., Roberts, S. A., Bernatavicius, G., Byrd, L. M., Gaze, D., Kroll, J., Jones, R. L., Sibley, C. P., Frøen, J. F., and Heazell, A. E. P., 2012, "Predictors of Poor Perinatal Outcome Following Maternal Perception of Reduced Fetal Movements -

a Prospective Cohort Study," PLoS One, **7**(7). https://doi.org/10.1371/journal.pone.0039784.

[29] Bekiou, A., and Gourounti, K., 2020, "Reduced Fetal Movements and Perinatal Mortality," Mater. Sociomed., **32**(3), pp. 227–234. https://doi.org/10.5455/msm.2020.32.227-234.

[30] Patel, S., Conway, A. E., Adjei, T., Abati, I., Dhawan, S., Yu, Z., Vaidyanathan, R., and Lees, C., 2025, "Is It Possible to Monitor Fetal Movements with a Wearable Device? A Review of Novel Technologies," European Journal of Obstetrics and Gynecology and Reproductive Biology, **305**, pp. 329–338. https://doi.org/10.1016/j.ejogrb.2024.12.011.

[31] Hijazi, Z. R., Callan, S. E., and East, C. E., 2010, "Maternal Perception of Foetal Movement Compared with Movement Detected by Real-Time Ultrasound: An Exploratory Study," Australian and New Zealand Journal of Obstetrics and Gynaecology, **50**(2), pp. 144–147. https://doi.org/10.1111/j.1479-828X.2009.01122.x.

[32] Minors, D. S., and Waterhouse, J. M., 1979, "The Effect of Maternal Posture, Meals and Time of Day on Fetal Movements," BJOG, **86**(9), pp. 717–723. https://doi.org/10.1111/j.1471-0528.1979.tb11273.x.

[33] Bradford, B. F., Cronin, R. S., McCowan, L. M. E., McKinlay, C. J. D., Mitchell, E. A., and Thompson, J. M. D., 2019, "Association between Maternally Perceived Quality and Pattern of Fetal Movements and Late Stillbirth," Sci. Rep., **9**(1). https://doi.org/10.1038/s41598-019-46323-4.

[34] Ozkaya, E., Baser, E., Cinar, M., Korkmaz, V., and Kucukozkan, T., 2012, "Does Diurnal Rhythm Have an Impact on Fetal Biophysical Profile?," The Journal of Maternal-Fetal & Neonatal Medicine, **25**(4), pp. 335–338. https://doi.org/http://dx.doi.org/10.3109/14767058.2011.576721.

[35] Turner, J. M., Cincotta, R., Chua, J., Gardener, G., Petersen, S., Thomas, J., Lee-Tannock, A., and Kumar, S., 2023, "Decreased Fetal Movements—the Utility of Ultrasound to Identify Infants at Risk and Prevent Stillbirth Is Poor," Am. J. Obstet. Gynecol. MFM, **5**(2). https://doi.org/10.1016/j.ajogmf.2022.100782.

[36] Park, S., Nguyen, T., George, T. M., Mayo, O. A., Mutersbaugh, J. E., Luo, G., Romero, R., and Gandjbakhche, A. H., 2025, "From Traditional to Cutting-Edge: A Review of

Fetal Well-Being Assessment Techniques," IEEE Trans. Biomed. Eng., **72**(5), pp. 1542–1552. https://doi.org/10.1109/TBME.2024.3512414.

[37] Lagrew, D., 1995, "The Contraction Stress Test," Clin. Obstet. Gynecol., **38**(1), pp. 11–25.

[38] Seel, T., Kok, M., and McGinnis, R. S., 2020, "Inertial Sensors—Applications and Challenges in a Nutshell," Sensors (Switzerland), **20**(21), pp. 1–5. https://doi.org/10.3390/s20216221.

[39] Picerno, P., Iosa, M., D'Souza, C., Benedetti, M. G., Paolucci, S., and Morone, G., 2021, "Wearable Inertial Sensors for Human Movement Analysis: A Five-Year Update," Expert Rev. Med. Devices, **18**(sup1), pp. 79–94. https://doi.org/10.1080/17434440.2021.1988849.

[40] Rodriguez-Villegas E, Iranmanesh S, and Imtiaz SA, 2018, "Wearable Medical Devices: High-Level System Design Considerations and Tradeoffs," IEEE Solid-State Circuits Magazine, pp. 43–52.

[41] Galli, A., Ambrosini, F., and Lombardi, F., 2016, "Holter Monitoring and Loop Recorders: From Research to Clinical Practice," Arrhythm. Electrophysiol. Rev., **5**(2), pp. 136–143. https://doi.org/10.15420/AER.2016.17.2.

[42] Vaddiraju, S., Burgess, D. J., Tomazos, I., Jain, F. C., and Papadimitrakopoulos, F., 2010, *Technologies for Continuous Glucose Monitoring: Current Problems and Future Promises*. [Online]. Available: www.journalofdst.org.

[43] Vitazkova, D., Kosnacova, H., Turonova, D., Foltan, E., Jagelka, M., Berki, M., Micjan, M., Kokavec, O., Gerhat, F., and Vavrinsky, E., 2025, "Transforming Sleep Monitoring: Review of Wearable and Remote Devices Advancing Home Polysomnography and Their Role in Predicting Neurological Disorders," Biosensors (Basel)., **15**(2). https://doi.org/10.3390/bios15020117.

[44] Dobkin, B. H., and Dorsch, A., 2011, "The Promise of MHealth: Daily Activity Monitoring and Outcome Assessments by Wearable Sensors," Neurorehabil. Neural Repair, **25**(9), pp. 788–798. https://doi.org/10.1177/1545968311425908.

[45] Groves, P. D., 2015, "Navigation Using Inertial Sensors," IEEE Aerospace and Electronic Systems Magazine, **30**(2), pp. 42–69. https://doi.org/10.1109/MAES.2014.130191.

[46] Boashash, B., Khlif, M. S., Ben-Jabeur, T., East, C. E., and Colditz, P. B., 2014, "Passive Detection of Accelerometer-Recorded Fetal Movements Using a Time-Frequency Signal Processing Approach," Digital Signal Processing: A Review Journal, **25**(1), pp. 134–155. https://doi.org/10.1016/j.dsp.2013.10.002.

[47] Zhao, X., Zeng, X., Koehl, L., Tartare, G., and Jonckheere, J. De, 2020, "A Wearable System for In-Home and Long-Term Assessment of Fetal Movement," IRBM, (41), pp. 205–211.

[48] Mesbah, M., Khlif, M. S., Layeghy, S., East, C. E., Dong, S., Brodtmann, A., Colditz, P. B., and Boashash, B., 2021, "Automatic Fetal Movement Recognition from Multi-Channel Accelerometry Data," Comput. Methods Programs Biomed., **210**.

[49] Layeghy, S., Azemi, G., Colditz, P., and Boashash, B., 2014, "Non-Invasive Monitoring of Fetal Movements Using Time-Frequency Features of Accelerometry," IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP).

[50] Liang, S., Peng, J., and Xu, Y., 2021, "Passive Fetal Movement Signal Detection System Based on Intelligent Sensing Technology," J. Healthc. Eng., **2021**. https://doi.org/10.1155/2021/1745292.

[51] Khlif, M. S., Boashash, B., Layeghy, S., Ben-Jabeur, T., Mesbah, M., East, C., and Colditz, P., 2011, "Time-Frequency Characterization of Tri-Axial Accelerometer Data for Fetal Movement Detection," IEEE International Symposium on Signal Processing and Information Technology, ISSPIT 2011, pp. 466–471. https://doi.org/10.1109/ISSPIT.2011.6151607.

[52] Delay, U. H., Nawarathne, B. M. T. M., Dissanayake, D. W. S. V. B., Ekanayake, M. P. B., Godaliyadda, G. M. R. I., Wijayakulasooriya, J. V., and Rathnayake, R. M. C. J., 2020, "Non Invasive Wearable Device for Fetal Movement Detection," 2020 IEEE 15th International Conference on Industrial and Information Systems, ICIIS 2020 - Proceedings, (978), pp. 285–290. https://doi.org/10.1109/ICIIS51140.2020.9342662.

[53] Xu, J., Zhao, C., Ding, B., Gu, X., Zeng, W., Qiu, L., Yu, H., Shen, Y., and Liu, H., 2022, "Fetal Movement Detection by Wearable Accelerometer Duo Based on Machine Learning," IEEE Sens. J., **22**(12), pp. 11526–11534. https://doi.org/10.1109/JSEN.2022.3172451.

[54] Altini, M., Mullan, P., Rooijakkers, M., Gradl, S., Penders, J., Geusens, N., Grieten, L., and Eskofier, B., 2016, "Detection of Fetal Kicks Using Body-Worn Accelerometers During Pregnancy: Trade-Offs Between Sensors Number and Positioning," IEEE EMBC.

[55] Abeywardena, C. L., Vanheusden, F. J., Walker, K. F., Arm, R., and Zhang, Q., 2020, "Fetal Movement Counting Using Optical Fibre Sensors," Sensors, **21**(48).

[56] Lai, J., Woodward, R., Alexandrov, Y., ain Munnee, Q., Lees, C. C., Vaidyanathan, R., and Nowlan, N. C., 2018, "Performance of a Wearable Acoustic System for Fetal Movement Discrimination," PLoS One, **13**(5). https://doi.org/10.1371/journal.pone.0195728.

[57] Ghosh, A. K., Catelli, D. S., Wilson, S., Nowlan, N. C., and Vaidyanathan, R., 2024, "Multi-Modal Detection of Fetal Movements Using a Wearable Monitor," Information Fusion, **103**. https://doi.org/10.1016/j.inffus.2023.102124.

[58] Cohen, G., and de Chazal, P., 2015, "Automated Detection of Sleep Apnea in Infants: A Multi-Modal Approach," Comput. Biol. Med., **63**, pp. 118–123. https://doi.org/10.1016/j.compbiomed.2015.05.007.

[59] Nemati, E., Zhang, S., Ahmed, T., Rahman, M. M., Kuang, J., and Gao, A., 2021, "CoughBuddy: Multi-Modal Cough Event Detection Using Earbuds Platform," 2021 IEEE 17th International Conference on Wearable and Implantable Body Sensor Networks, BSN 2021, **2021-Janua**, pp. 0–3. https://doi.org/10.1109/BSN51625.2021.9507017.

[60] Lotfi, R., Eskicioglu, R., Tzanetakis, G., and Irani, P., 2020, "A Comparison between Audio and IMU Data to Detect Chewing Events Based on an Earable D," 11th Augmented Human Interntional Conference.

[61] Alwis, P., Thilakasiri, I., Liyanage, S., Nanayakkara, R., Godaliyadda, R., Ekanayake, M., Rathnayake, C., Wijayakulasooriya, J., and Herath, V., 2024, "Application of an LSTM-Based Channel Attention and Classification Mechanism in Fetal Movement Monitoring." https://doi.org/10.36227/techrxiv.24630642.v3.

[62] Nazir, Z., Zarymkanov, T., and Park, J.-G., 2022, "A Machine Learning Model Selection Considering Tradeoffs between Accuracy and Interpretability," Advances in Science, Technology and Engineering Systems Journal, **7**(4), pp. 72–78.

[63] Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., and Farhan, L., 2021, "Review of Deep

Learning: Concepts, CNN Architectures, Challenges, Applications, Future Directions," J. Big Data, **8**(1). https://doi.org/10.1186/s40537-021-00444-8.

[64] Rodriguez, N. M., Burleson, G., Linnes, J. C., and Sienko, K. H., 2025, "Thinking Beyond the Device: An Overview of Human-and Equity-Centered Approaches for Health Technology Design," Annual Review of Biomedical Engineering Downloaded from www.annualreviews.org. Guest, **18**, p. 27. https://doi.org/10.1146/annurev-bioeng-081922.

[65] Fisher, M., and Johansen, E., 2020, "Human-Centered Design for Medical Devices and Diagnostics in Global Health," Global Health Innovations, **3**(1).

[66] Aranda-Jan, C., Jagtap, S., and Moultrie, J., 2016, "Towards A Framework for Holistic Contextual Design for Low-Resource Settings," International Journal of Design, **10**(3), pp. 43–63. [Online]. Available: www.ijdesign.org.

[67] Burleson, G., Wojciechowski, L., Toyama, K., and Sienko, K. H., 2024, "Characterizing the Use of Contextual Factors in Engineering Design: An Exploration of Global Health Designer Practice," Design Science, **10**. https://doi.org/10.1017/dsj.2024.28.

[68] Atman, C., Borgford-Parnell, J., Deibel, K., Kang, A., Ng, W. H., Kilgore, D., and Turns, J., 2009, *Matters of Context in Design*, CRC Press.

[69] Aranda-Jan, C. B., Jagtap, S., and Moultrie, J., 2016, "Towards A Framework for Holistic Contextual Design for Low-Resource Settings," International Journal of Design, **10**(3). [Online]. Available: www.ijdesign.org.

[70] Mosyjowski, E., Espinoza von Bischhoffshausen, J., Lattuca, L., and Daly, S., 2020, "Student and Practitioner Approaches to Systems Thinking: Integrating Technical and Contextual Considerations," *ASEE Annual Conference*, American Society of Engineering Education.

[71] Mesbah, M., Khlif, M. S., East, C., Smeathers, J., Colditz, P., and Boashash, B., 2011, "Accelerometer-Based Fetal Movement Detection," Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, pp. 7877–7880. https://doi.org/10.1109/IEMBS.2011.6091942.

Chapter 2 Comparative Analysis of Machine Learning Approaches for Fetal Movement Detection with Linear Acceleration and Angular Rate Signals[1]

2.1 Abstract

Reduced fetal movement (RFM) can indicate that a fetus is at risk, but current monitoring methods provide only a "snapshot in time" of fetal health and require trained clinicians in clinical settings. To improve antenatal care, there is a need for continuous, objective fetal movement monitoring systems. Wearable sensors, like inertial measurement units (IMUs), offer a promising data-driven solution, but distinguishing fetal movements from maternal movements remains challenging. The potential benefits of using linear acceleration and angular rate data for fetal movement detection have not been fully explored. In this study, machine learning models were developed using linear acceleration and angular rate data from twenty-three participants who wore four abdominal IMUs and one chest reference while indicating perceived fetal movements with a handheld button. Random forest (RF), bi-directional long short-term memory (BiLSTM), and convolutional neural network (CNN) models were trained using hand-engineered features, time series data, and time–frequency spectrograms, respectively. Models were evaluated against a standardized testing set to ensure a fair comparison across data representations. The results showed that combining accelerometer and gyroscope data improved detection performance across all models compared to either one alone. Simpler models (RF and BiLSTM) significantly outperformed CNN in detection performance and data efficiency. While RF demonstrated the highest capacity for reducing the false-positive rate to clinically acceptable levels (less than 10%), this tradeoff necessitated a significant trade-off, with sensitivity dropping to approximately 20%. These findings suggest that while simpler, data-efficient models are currently the superior choice for proof-of-concept monitoring, further development is required for deployment to clinical or consumer settings.

[1] Adapted from the published work: Spicher, L., Bell, C., Sienko, K. H., & Huan, X. (2025). Comparative analysis of machine learning approaches for fetal movement detection with Linear acceleration and angular rate signals. Sensors, 25(9), 2944.

2.2 Introduction

Prenatal care is essential for maternal and fetal health, with regular monitoring used to identify an at-risk fetus and improve birth outcomes. Key indicators of fetal wellbeing include fetal heart rate, movement, and muscle tone [1]. Among these indicators, maternal perception of fetal movement stands out as one of the oldest and most common methods of monitoring fetal health [2]. Reduced fetal movement (RFM) can signal complications: stillbirth, fetal growth restriction, congenital anomalies, and maternal hemorrhage [3,4]. An estimated 25% of pregnancies in which mothers report RFM result in poor perinatal outcomes [3]. However, maternal perception inherently is subjective, influenced by factors such as activity level, placental position, and fetal sleep–wake cycles, and lacks standard guidelines for optimal number of movements [1,3,5–7]. Clinicians initiate antenatal testing where there is a report of RFM by completing a non-stress test. Fetal heart rate and uterine activity are monitored for 20 minutes for fetal heart rate accelerations or decelerations with or without contractions. Non-stress tests have significant false positive rates, resulting in additional testing to assess fetal wellbeing [1,8]. The contraction stress test provides information about fetal reactivity with uterine contractions and the biophysical profile measures fetal breathing, movement, tone, and amniotic fluid via ultrasound [1]. All these tests require trained clinicians and clinical settings and provide only a "snapshot in time" of fetal health, which may yield inconclusive results [1,8]. Despite the utility of clinical fetal monitoring, these methods can prompt cautious but sometimes unnecessary interventions—including early delivery, hospitalization, or more frequent monitoring [1,2,4,7,8]. Continuous, objective fetal movement monitoring could reduce reliance on subjective perception and isolated assessments, offering a more comprehensive understanding of fetal wellbeing and improving antenatal care.

Wearable sensors offer a promising solution for continuous, objective fetal movement monitoring. However, distinguishing fetal movements from superimposed maternal movements in wearable sensor data continues to be a challenge [9–12]. Accelerometers have shown potential in controlled research environments when study participants are stationary. For instance, Khlif et al. achieved moderate accuracy (59%), sensitivity (76%), and specificity (55%) using a thresholding algorithm with three abdominal accelerometers [13]. Mesbah et al. applied signal preprocessing to remove maternal artifacts, achieving 95.8% accuracy but observing a sharp drop to 87.6% when maternal artifacts were included [9]. Delay et al. explored multiclass

classification (e.g., fetal movement, maternal laughter, and respiratory motion), achieving an 86% true positive rate and 7% false positive rate with a single accelerometer [14]. Xu et al. achieved 86.6% accuracy and 84.2% F1 score using two accelerometers but relied on synthetic minority oversampling [15]. These examples highlight the complexity of fetal movement detection, particularly the difficulty of distinguishing fetal movements from maternal motion—a challenge that becomes even greater in real-world settings outside controlled environments. Furthermore, differences in study populations and methodologies—such as gestational age, baseline fetal activity, preprocessing techniques, model architecture, and data representations—affect the results reported [16]. Comparing studies based on these metrics alone is not applicable to a broad patient population, as success reflects conditions and datasets specific to the study, and algorithms may not be generalizable. Understanding these nuances is crucial for interpreting research in this field and appreciating the variability inherent in fetal movement detection approaches.

Although evaluating the effectiveness of fetal movement detection approaches presents significant challenges, isolating and detecting fetal movements becomes particularly difficult when superimposed maternal movements are captured by wearable sensors [9–12]. In related biomedical applications where superimposed movements are also a concern (e.g., monitoring of sleep apnea in preterm infants, pulmonary conditions, and dietary behaviors), inertial measurement units (IMUs) have shown promise for accurate event detection [17–20]. Traditional accelerometer-based approaches detect fetal movement by identifying isolated changes in acceleration caused by abdominal deformation. When maternal movement occurs, a mixture of signals from both fetal and maternal motion are captured [9–12]. Angular rate gyroscopes, although rarely used to date in fetal movement detection, measure rotational motion and could help distinguish between the superimposed movements [21,22]. During maternal movement, the torso can be approximated as a rigid body, leading to similar gyroscope readings across sensors [23,24]. Thus, combining accelerometers with gyroscopes may detect localized abdominal deformation, potentially allowing for the identification of fetal movement in the presence of maternal motion.

Only one study to date has reported using both accelerometer and gyroscopic data obtained from IMUs for fetal movement detection [25]. The recent preprint by Alwis et al. demonstrated that a fusion deep learning model combining a convolutional neural network

(CNN) and long short-term member (LSTM) layers achieved 88% accuracy, 0.86 sensitivity, and 0.91 specificity using spectrogram-based time–frequency representations [25]. While this study highlighted the potential of IMUs for fetal movement detection, it did not address the distinct contributions of acceleration versus angular rate data or evaluate their potential combined benefits relative to an accelerometer-only approach. Furthermore, Alwis et al. relied exclusively on spectrogram features and a specific fusion model architecture, leaving questions about how alternative data representations and machine learning models may perform.

Given the variability across study methodologies, comparing how different data representations affect performance is crucial for advancing fetal movement detection approaches. For example, hand-engineered features capture domain-specific patterns, time series characterize temporal dependencies, and time–frequency spectrograms offer insights into frequency and time dynamics simultaneously [9,14,25]. Comparing these approaches provides a valuable understanding of the strengths and limitations of each method and helps standardize evaluations in the field.

This study explores fetal movement detection using acceleration and angular rate signals from wearable IMUs. The primary contributions of this work are as follows: (1) a comparative analysis of accelerometer and angular rate gyroscope performance in detecting fetal movement, (2) an assessment of their combined performance to enhance detection accuracy, and (3) an evaluation of various machine learning models trained on three distinct data representations—hand-engineered features, time series data, and time–frequency spectrogram representations. Additionally, this study examines the effect of decision thresholds for binary classification on model performance and the effect of training set size on model robustness. This analysis aims to provide a clearer understanding of each approach's benefits and limitations, addressing gaps in current models and offering insights into how more efficient, deployable solutions can be developed for real-world use.

2.3 Materials and Methods

2.3.1 Participants

Twenty-three participants were recruited for a single data collection session at the University of Michigan. Participants were selected based on the following criteria: between 18 and 49 years old, currently carrying a singleton pregnancy, and between 24 and 32 weeks of

gestational age. Appendix A provides additional demographic information about the participant population. Exclusion criteria included a diagnosis of gestational diabetes, hypertension, or any known fetal disease or physical abnormalities, to minimize confounding factors that could affect movement patterns. Written informed consent was obtained from all participants, and the study protocol was reviewed and approved by the University of Michigan Institutional Review Board (HUM00204999).

2.3.2 Sensor System

Data were collected using four tri-axial IMUs (Opal, APDM Inc., Portland, OR, USA) positioned around the participant's umbilicus with a medical-grade adhesive. The x-axis of each IMU was aligned with gravity and the z-axis was perpendicular to the abdomen to ensure proper and consistent measurements of angular rates (Figure 5). To help distinguish between fetal and maternal movement, a chest IMU was added as a reference [9,13]. By approximating the torso as a rigid body, the angular rates measured by the chest and abdominal IMUs were assumed to be similar during maternal movement. Since the chest IMU was not expected to capture isolated abdominal deformations, it was considered useful for distinguishing maternal movement from fetal movement. The IMUs collected synchronized acceleration and angular rate data at a sampling rate of 128 Hz.

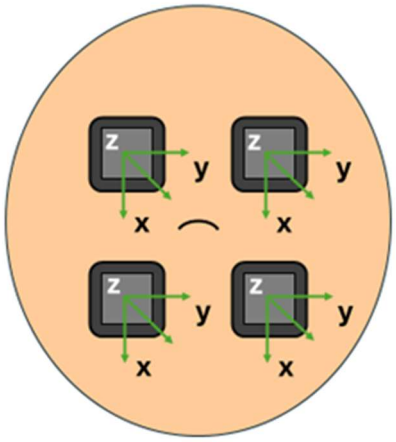


Figure 5: Position and alignment of four tri-axial IMUs on the maternal abdomen for fetal movement detection

2.3.3 Experimental Protocol

At the start of each session, participants' heights and weights were recorded, and the sensor system was fitted. Calibration data were collected by having participants perform a series of specified movements, which were later used to align the axes of the angular rate data. While

accelerometer-only approaches have detected fetal movement in prior studies by capturing changes in acceleration tangential to the abdomen [10,13,14], approaches that use angular rate data have the potential to distinguish fetal movement from maternal motion but require an estimation of the sensors' orientation [26]. Therefore, a Functional Alignment Method was employed to determine the anatomical axes relative to the IMU fixed frame of reference [26]. Since the IMU sensors were placed on the abdomen and the abdomen was assumed to move as a rigid body with rotation about the hip, the hip axis was used as a reference point to estimate IMU orientation. At the start of data collection, participants began from a standing position and were instructed to lean forward, bending at the hip, to a comfortable degree. This motion was repeated three times, and the calibration data from these hip-hinging movements were used to align the IMU axes (details related to signal processing are in Section 2.3.4).

Following calibration, participants were seated in a chair. Each participant held a unique IMU which functioned as a handheld toggle. When fetal movement was perceived, participants pressed the button on this IMU to mark the event for subsequent movement detection analysis. Data collection was conducted in 10–15 min intervals to balance practical constraints with capturing meaningful fetal movement data. Since fetal movements often occur in clusters [27,28], shorter recording periods allowed for natural fluctuations while preventing participant fatigue and loss of focus, which could affect the accuracy of self-reported movements. Additionally, these intervals provided opportunities for participants to engage in behaviors that stimulate fetal movement, such as rubbing the abdomen or consuming a snack, increasing the likelihood of detecting activity [28]. While maternal perception captures only about 40% of true movements [6], simply increasing the recording duration would not necessarily yield more detected events. Instead, shorter trials helped sustain participant engagement, minimize fatigue-related artifacts, and effectively extend total observation time across the session. Each participant completed 2–3 trials, generating a total of 49 trials.

2.3.4 Signal Processing and Data Labeling

After data collection, the synchronized files from the IMUs were downloaded and reviewed for signal quality. Data from two participants were excluded due to signal acquisition issues. Raw acceleration and angular rate signals from each IMU were processed through a zero-phase bandpass filter (1–20 Hz) to reduce drift and noise [16].

For angular rate data, a Functional Alignment Method was applied to estimate each sensor's orientation to improve differentiation between maternal and fetal movement signals [26]. During static standing, the global gravity vector was identified to establish a reference for the alignment of the vertical axis. Using rotation about the frontal axis, from the hip-hinging movements, the frontal and sagittal axes were identified. Each sensor's x-axis was aligned with global gravity, the y-axis with the participant's frontal axis, and the z-axis with the sagittal axis. This alignment allowed the angular rate data to capture deviations from expected rigid body motion—a key feature for distinguishing localized abdominal deformations caused by fetal movements from larger maternal movements. By aligning the axes this way, rotational changes inconsistent with whole-body movement patterns could be identified as potential fetal movements.

Axis alignment was not performed on acceleration data to preserve tangential fetal movement information on the z-axis, which is critical for capturing small, localized changes in abdominal acceleration that may be otherwise lost through axis realignment. Preserving the raw acceleration signal ensured that subtle changes in movement direction and magnitude—key indicators of fetal movement—were retained for analysis.

This combination of angular rate alignment and unaltered acceleration data leveraged each signal's strengths: angular rate data for detecting deviations from maternal rigid body motion and acceleration data for capturing localized movement dynamics. An example of the time series measurements is provided in Figure 6.

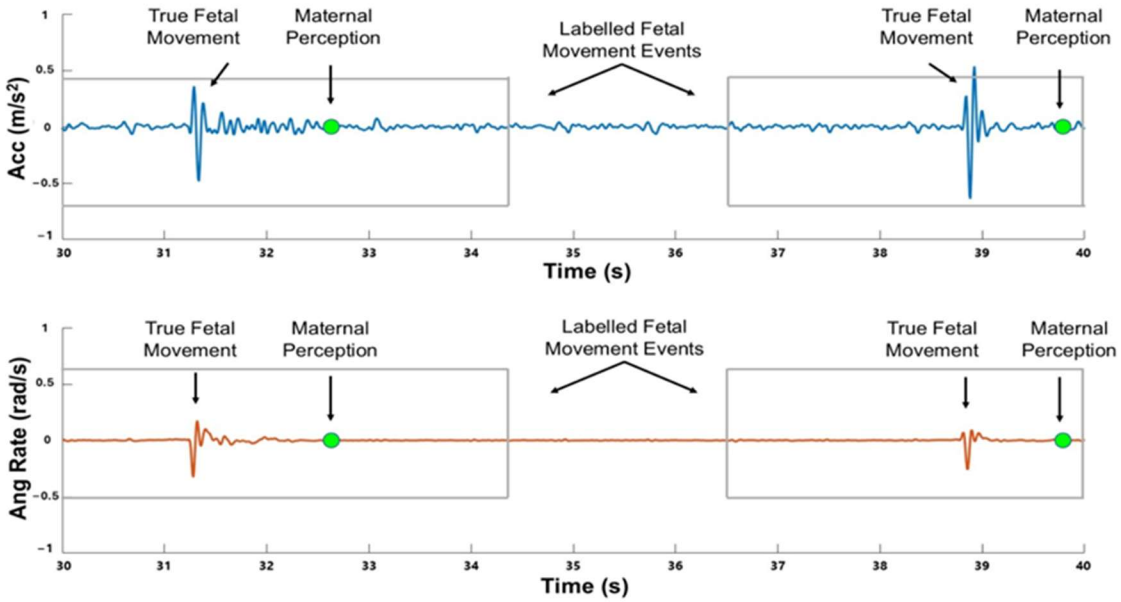


Figure 6: Example of fetal movement in the time series data.
Measurements from accelerometer and gyroscope sensors are displayed. Bright green markers indicate maternal perception toggle presses, while labeled fetal movement events are defined to be within the gray boxes

Fetal movement events were defined 3.5 s prior to and 1.5 s after each recorded press of the handheld toggle to account for participant reaction time and fetal movement duration [29] (see Figure 6 for an example). The full time series data were then segmented by a window of fixed size. Binary labels were assigned to each window, where windows containing more than 10% overlap with event intervals were labeled as 1 and all others as 0. This 10% threshold, determined through visual inspection, effectively captured short fetal movements [30]. After labeling, approximately 11% of windows were positively labeled. To mitigate data imbalance in the training set, two-thirds of negatively labeled windows were discarded to achieve an approximate 1:3 positive-to-negative ratio with 26% positive training windows [16].

2.3.5 Feature-based Approach: Random Forest Model

For the first data representation, all data trials were segmented into 0.5 s windows with no overlap. Statistical and correlation features were calculated for each axis and sensor, including the mean, standard deviation, and range. In addition, the cross-correlation between axes and the cross-correlation with the reference sensor for each axis were computed, as described in our previous study [30]. This process resulted in 75 acceleration features and 75 angular rate features per window. The features of each window were then used to train a random forest (RF) model. The model architecture consisted of an RF classifier with hyperparameters

optimized via a random search based on the F1 score evaluated on the validation set. The final output was determined by aggregating the predictions of all trees using majority voting, based on a chosen threshold, to classify samples into {0,1} [30].

2.3.6 Time Series Approach: BiLSTM Model

In contrast to the feature-based approach used for the first data representation, the second data representation used the time series data. Using information from the time domain maintains the natural temporal dynamics and memory of activity history, capturing transient accelerations and rotations that could indicate movement events. Preserving the time series structure may improve sensitivity to short, low-frequency bursts, characteristic of fetal movement [32].

For the second data representation, all data trials were again segmented into 0.5 s windows with no overlap. These windows were used directly as inputs into a machine learning model. The time series inputs were then used to train a bi-directional long short-term memory (BiLSTM) model, a type of recurrent neural network and variant to LSTM that is designed for capturing temporal dependencies from past and future contexts simultaneously [33]. Given that fetal movements often follow patterns or brief bursts, BiLSTM's memory capability enables it to recognize and retain relevant time dependencies while ignoring noise or less relevant temporal artifacts.

The model architecture presented in Figure 7 included a single BiLSTM layer with 64 units regularized with 0.3 dropout and 0.3 recurrent dropout. The output from this layer was passed to a fully connected dense layer with sigmoid activation to produce a scalar value, which was rounded to {0,1} following a thresholding rule for binary classification. The model was trained using the Adam optimizer with binary cross-entropy loss.

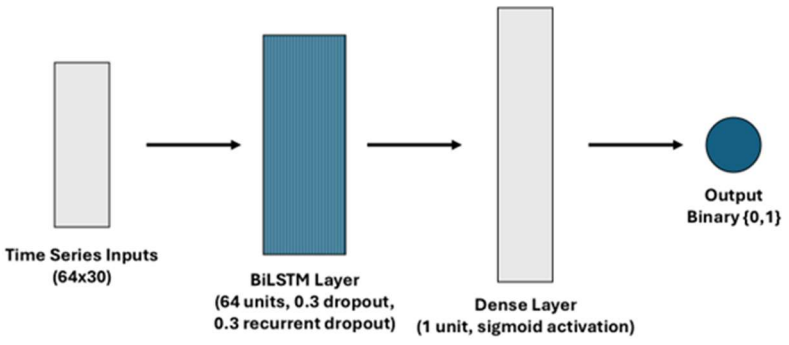


Figure 7: Model architecture for the time series approach

2.3.7 Time Frequency Approach: CNN Model

A time–frequency data representation was used for the third data representation to enable the model to leverage frequency-specific features, which may correspond to fetal movement rhythms or recurrent patterns that are not easily seen in the time domain [34]. Frequency–domain analysis has proven useful in other biomedical applications, particularly in distinguishing movement from background noise, and may improve model robustness against maternal motion artifacts [35]. Spectrograms also allow the model to focus on frequency components associated with fetal movement, improving sensitivity to movement events in noisy environments.

For the third data representation, all data trials were segmented into overlapping 8 s windows with a 1 s stride. Although shorter windows (e.g., 0.5 s) are suitable for capturing fine-grained temporal patterns in feature-based or time series approaches [16], a longer window was necessary to balance the tradeoff between time and frequency resolution in the time–frequency representation [25]. Specifically, a 0.5 s window provided better time resolution, but poor frequency resolution, limiting the ability to capture meaningful frequency–domain patterns. Conversely, the 8 s window al-lowed for improved frequency resolution, which is crucial for distinguishing fetal movement patterns from maternal motion in the spectral domain. Within each 8 s window, the short-time Fourier transform (STFT) was computed using a sliding Hanning window of 16 samples with a stride length of 1 [25]. The magnitude of the STFT is known as the spectrogram time–frequency representation of the signal (Figure 8). A magnitude spectrogram was constructed for each window, and these spectrograms were used as inputs into a machine learning model.

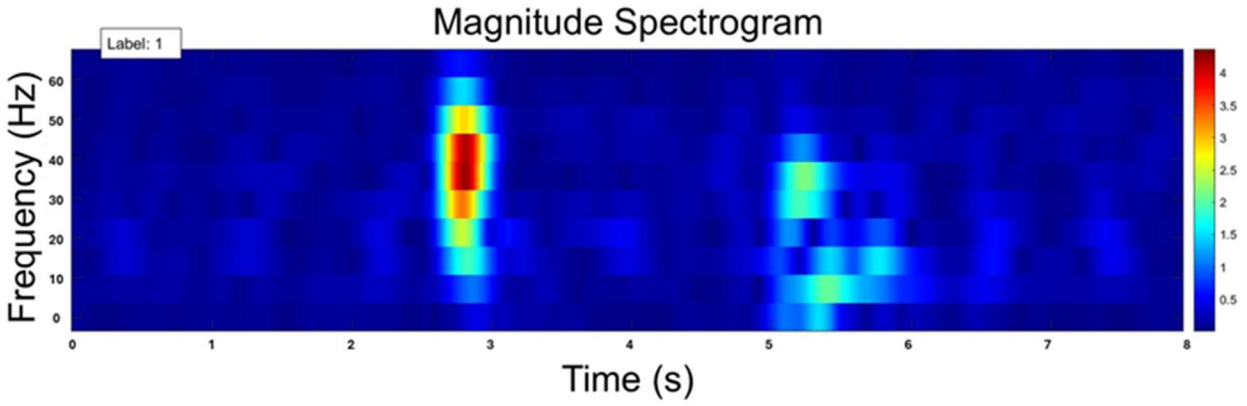


Figure 8: Example of fetal movement in the time–frequency spectrum

Magnitude spectrograms were used to train a CNN. CNNs are effective for pattern recognition in image data, as they excel in identifying localized features and patterns [36]. By

applying a CNN to frequency–domain representations, the model can learn frequency-specific movement characteristics while ignoring irrelevant noise, making it ideal for movement detection in an environment with variable maternal activity.

The model architecture, presented in Figure 9, consisted of an initial two-dimensional (2D) convolutional layer with 32 filters, each of size (3,3), and ReLu activation. This layer was followed by a (2,2) max-pooling layer to downsample the feature map. A second convolutional layer with 16 filters, each of size (3,3), and ReLU activation was applied to further refine the extracted features. The output of the second convolutional layer was then flattened and passed to a dense layer with 16 units and ReLu activation, a dropout layer with rate 0.25, and a final output layer with sigmoid activation to produce a scalar value, which was rounded to {0,1} following a thresholding rule for binary classification. The model was trained using the Adam optimizer and binary cross-entropy loss.

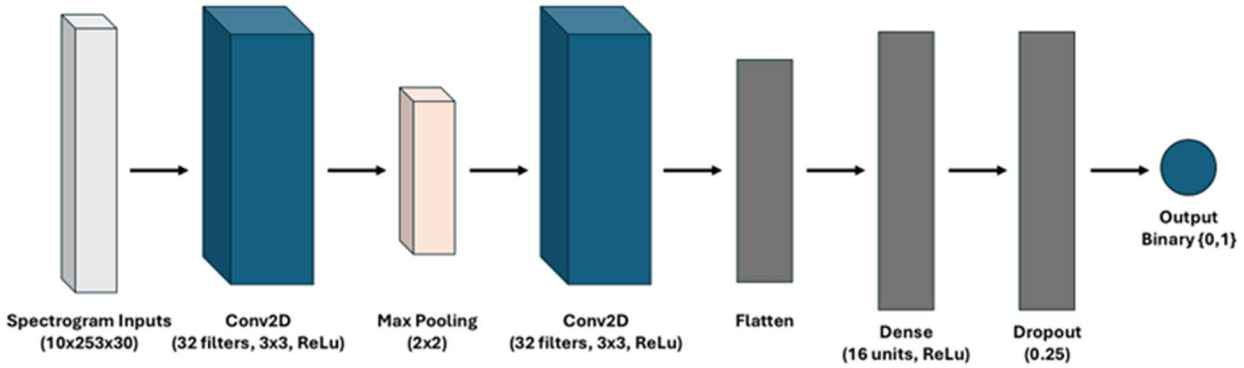


Figure 9: Model architecture for the time–frequency approach

2.3.8 Model Training and Evaluation

From the full dataset, six trials were randomly selected for testing, six for validation, and the remaining thirty-seven for training. This division ensured that performance could be assessed on data unseen to the model, allowing for an evaluation of model generalization across different scenarios. Three dataset configurations were explored: (1) acceleration data only, (2) angular rate data only, and (3) a combination of both data types. For evaluation, the area under the receiver operating characteristic curve (AUROC), accuracy, sensitivity, specificity, positive predictive value (PPV), and F1 score were calculated to present a comprehensive view of model performance. Confusion matrices were also generated for each combination to provide a visual representation of classification performance. To maintain consistency across varying temporal resolutions, the predictions from the RF and BiLSTM models were aggregated from their

original 0.5s windows into 8s overlapping windows with a 1s stride by taking the maximum prediction probability across all sub-windows to determine the final classification for the 8s interval, matching the evaluation framework of the CNN. This alignment ensured that the reported performance metrics reflected a standardized assessment across all model types.

To ensure robust and consistent evaluation, cross-validation was employed with a randomized repeated split approach [31]. Specifically, 10 sets of validation and testing splits were pre-selected prior to model training, with each split generated independently and without replacement to maintain diversity across trials. Model performance was then evaluated on the validation and testing sets to reflect its predictive capability on unseen data scenarios. This Monte Carlo cross-validation approach ensured that each model's performance was averaged across 10 different splits, providing a robust estimate of generalization [31]. Furthermore, using the same validation and testing splits for all data representations ensured a fair and direct comparison of model performance across different dataset configurations. Other cross-validation approaches, such as k-fold, could also be employed. To assess the statistical significance of differences between models, p-values were calculated using paired statistical tests across the 10 validation and testing splits.

2.3.9 Thresholding Analysis and Training Set Size Reduction

In addition to the primary model evaluations, two exploratory analyses, thresholding and training set size reduction, were conducted to further understand model performance. Setting a good threshold for the machine learning models is an important part of creating an effective binary classification model. A standard 0.5 threshold is often not the best choice, and a smaller threshold (towards 0) tailors the model to predict more positives, while a larger threshold (towards 1) tailors the model to predict more negatives. This effect can also be quite sensitive at times, leading to fragile models. Therefore, a thresholding analysis was performed to vary the threshold from 0.1 to 0.9 in increments of 0.1.

Another consideration for data-driven modeling is whether the training set is sufficiently rich. This consideration is non-trivial and depends on the architecture and complexity of the models being used, the nonlinearity of the underlying data input–output relationship, and the settings and demographics being targeted. Nonetheless, sensitivity analysis can be performed by varying the training data size and providing insights for understanding model limitations and future data collection. The three models in this study were evaluated for robustness and

sensitivity to data availability by training and testing them at 100%, 75%, 50%, and 25% of the full training set size.

2.4 Results

2.4.1 Summary of the Results

The results presented in this study demonstrate the comparative performance for detecting fetal movement when using only acceleration data, only angular rate data, and combining both. By evaluating three distinct data representations and models—hand-engineered features classified with RF, time series data analyzed with BiLSTM, and time–frequency spectrograms analyzed with CNN—this study highlights the strengths and limitations of each approach. Across all model types, combining acceleration and angular rate data improved performance for all metrics. Additionally, RF and BiLSTM models outperformed CNN, with RF showing the greatest capacity to reduce false positives but at a significant tradeoff with sensitivity.

2.4.2 Model Performance Across Sensor Types

Mean receiver operating characteristic (ROC) curves, averaged across 10 testing splits, were generated for binary classification under varying thresholds (Figure 10), with the corresponding mean area under ROC (AUROC) values shown in the legend. The hand-engineered features trained with RF achieved the highest AUROC of 0.81 for the combined dataset, indicating “very good” performance [37] in distinguishing fetal movement from non-movement. Single-mode datasets performed significantly lower than the combined dataset (acceleration: p-value = 0.005; angular rate: p-value = 0.002) with an AUROC of 0.76. Time series data trained with BiLSTM also yielded “very good” performance [37] with an AUROC of 0.80 for the combined dataset. The single-mode datasets scored 0.78 and 0.74, respectively (acceleration: p-value = 0.3; angular rate: p-value = 0.1). Time-frequency spectrograms analyzed with CNN achieved “good” performance with an AUROC of 0.75 for the combined dataset. The single model datasets achieved lower AUROC scores of 0.73 and 0.74, respectively (acceleration: p-value = 0.1; angular rate: p-value = 0.8).

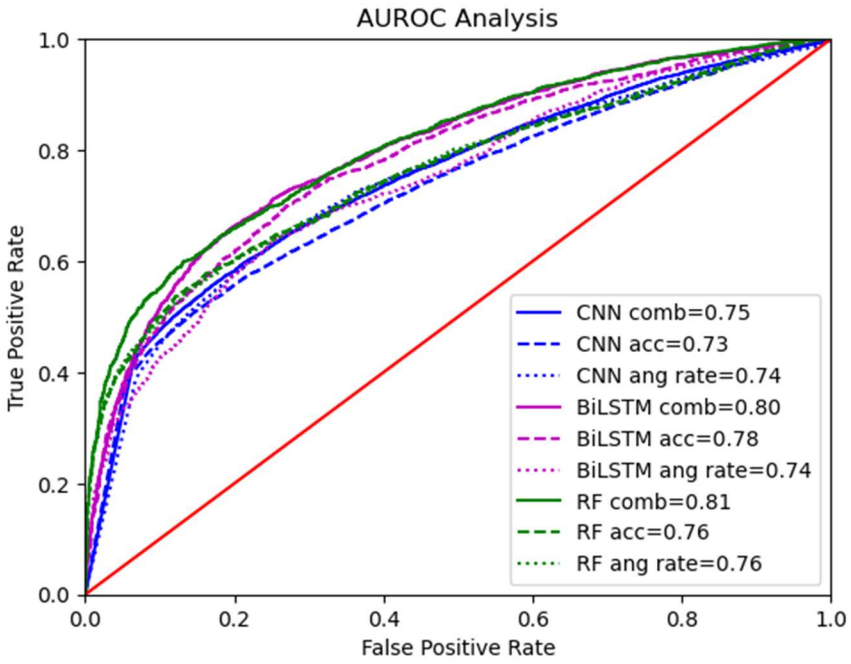


Figure 10: ROC curves and AUROC values (in legend) comparing acceleration, angular rate, and combined datasets with different models

2.4.3 Model Performance Across Data Representations

Accuracy, sensitivity, PPV, and F1 score, averaged over 10 testing splits and standardized at a specificity of 0.75, are presented in Figure 11 for each dataset and model. Confusion matrices, presented in Figure 12, provide further insight into these results by showing the distribution of true positives, false positives, false negatives, and true negatives across models and datasets.

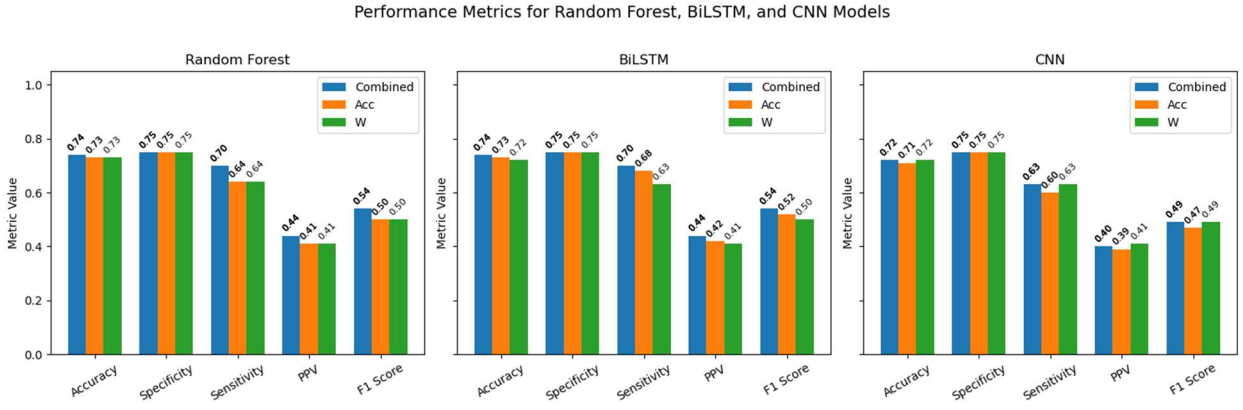


Figure 11: Performance metrics of RF, BiLSTM, and CNN models. Each histogram displays five metrics, accuracy (ACC), specificity (SPEC), sensitivity (SEN), positive predictive value (PPV), and F1 score, with bars indicating the model's performance on combined, angular rate, and acceleration datasets

For the hand-engineered features classified with RF, the combined dataset achieved increased performance for all metrics. Metrics with significant increases included sensitivity from 0.64 to 0.70 (acceleration: p-value = 0.003; angular rate: p-value = 0.006), F1 score from 0.50 to 0.54 (acceleration: p-value = 0.006; angular rate: p-value = 0.06), and AUROC from 0.76 to 0.81 (acceleration: p-value = 0.005; angular rate: p-value = 0.002). Furthermore, the confusion matrices, shown in green, illustrate that the combined dataset exhibits more true positives (top left) and fewer false negatives (top right) compared to the single-mode datasets.

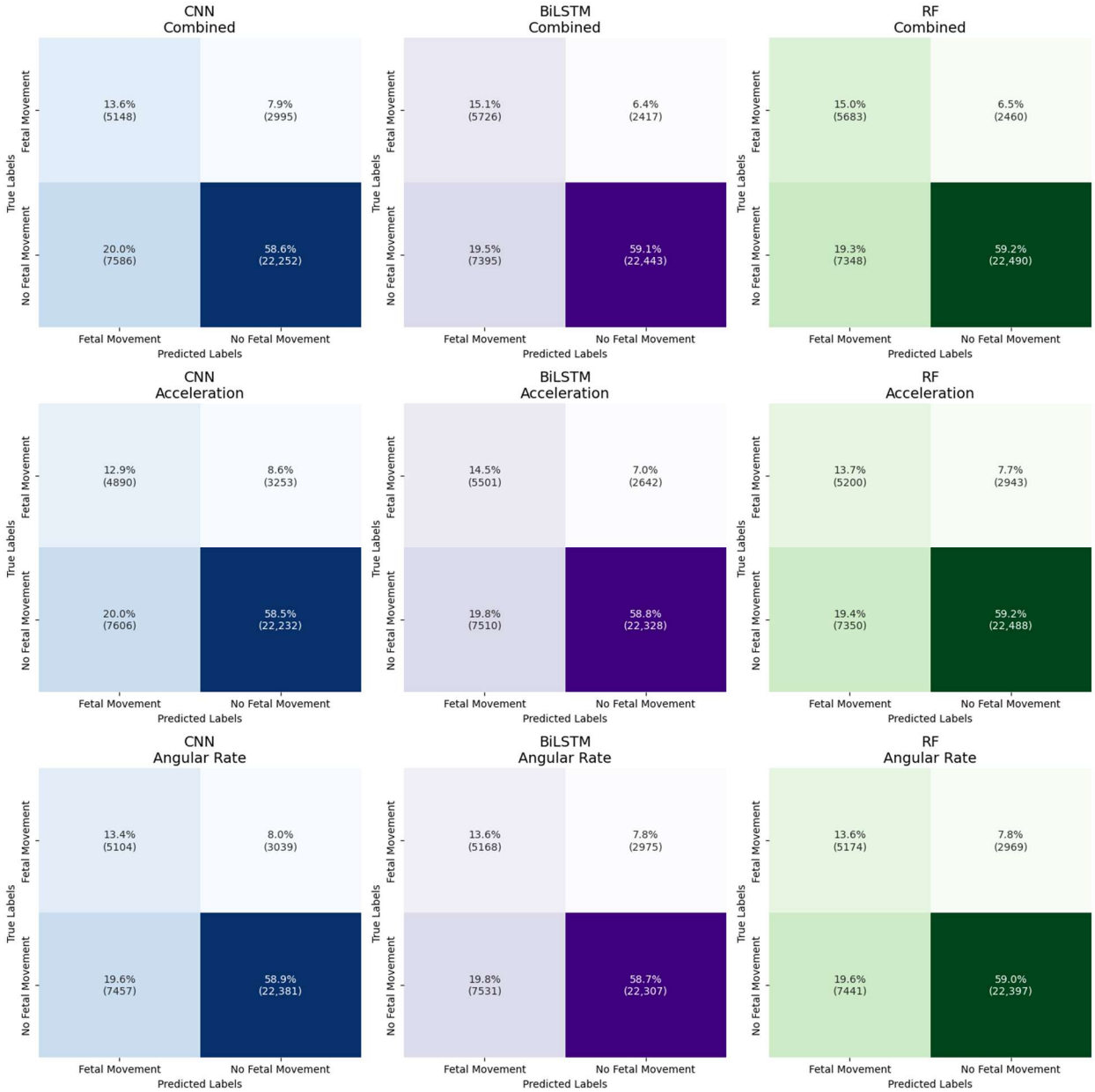


Figure 12: Confusion matrices showing the performance of the RF, BiLSTM, and CNN models across three datasets: acceleration, angular rate, and combined. The models' classification performances are visualized in terms

of true positives (top left), false positives (bottom left), false negatives (top right), and true negatives (bottom right). Each cell displays the percentage of the total number of samples, with the raw count shown in parentheses. RF matrices are shown in green, BiLSTM matrices in purple, and CNN matrices in blue. The combined datasets are positioned in the top row, acceleration in the second row, and angular rate in the third row. It is important to note that the models were evaluated using different window sizes: RF and BiLSTM used 0.5 s windows, while CNN used 8 s windows. This difference in window sizes may affect the distribution of predicted outcomes and should be considered when interpreting the results.

The time series representations analyzed with BiLSTM resulted in the same trend: performance increased for all metrics with the combined dataset. However, these performance increases were not significant (p-value > 0.05) across the 10 testing splits. The confusion matrices, shown in purple, demonstrate similar trends to the RF model, highlighting the ability of the combined dataset to capture more true positives and fewer false negatives. When comparing data representations, the BiLSTM and RF models achieved nearly identical performance. The RF model achieved a higher AUROC compared to BiLSTM (0.81 and 0.80, respectively), but this difference was not significant (p-value = 0.2). Similar trends are observed when comparing the confusion matrices.

Finally, for the time–frequency spectrograms analyzed with CNN, the combined dataset demonstrated the best performance in terms of AUROC. However, the increase in AUROC from 0.73 (acceleration: p-value = 0.1) and 0.74 (angular rate: p-value = 0.8) to 0.75 was not significant. The angular rate and combined datasets outperformed the acceleration dataset across all other performance metrics. However, these performance increases were not significant (p-value > 0.05) across the 10 testing splits. When comparing data representations, the RF and BiLSTM models outperformed CNN across all performance metrics. Metrics with significant differences included sensitivity of 0.63 compared to 0.70 (RF: p-value = 0.03, BiLSTM: p-value = 0.05), F1 score of 0.49 compared to 0.54 (RF: p-value = 0.009, BiLSTM: p-value = 0.005), and AUROC of 0.75 compared to 0.81 (BiLSTM: p-value = 0.0002) and 0.80 (RF: p-value = 2 × 10−5). The confusion matrices, shown in blue, demonstrate a slightly more limited ability to capture true positives paired with an increased proportion of false negative and false positive predictions.

2.4.4 Thresholding Analysis and Training Set Size Reduction

Performance metrics resulting from varying the threshold for binary classification from 0.1 to 0.9 for each model type, trained and tested on the combined dataset, are reported in Figure 13. When evaluating RF across varying thresholds, accuracy, specificity, and PPV followed a

sigmoidal trend, where performance linearly increased with plateaus at low and high thresholds. Sensitivity linearly decreased at increasing thresholds. The F1 score peaked between thresholds of 0.5 and 0.6. Notably, when the RF threshold was tuned to achieve a clinically acceptable false-positive rate of less than 10% (specificity > 0.90), the model's sensitivity dropped to approximately 0.20. For BiLSTM, the trends were relatively similar, but performance metrics had less variation across thresholds. Specifically, the F1 score remained relatively stable at 0.40 regardless of the threshold. Additionally, the maximum achievable performance was more limited. While RF was able to achieve specificity near 1.0 at higher thresholds, BiLSTM achieved a maximum specificity of approximately 0.80. Similarly, PPV peaked near 0.40 for BiLSTM, while it peaked near 0.90 for RF. Finally, the performance metrics for CNN were notable stable across all thresholds. Like BiLSTM, the maximum achievable performance was more limited than RF. Specificity peaked near 0.75 and PPV near 0.40. The F1 score remained stable around 0.45. This analysis highlights how model performance can be highly dependent on threshold choice. Overall, RF demonstrated the highest achievable performance in terms of specificity and PPV.

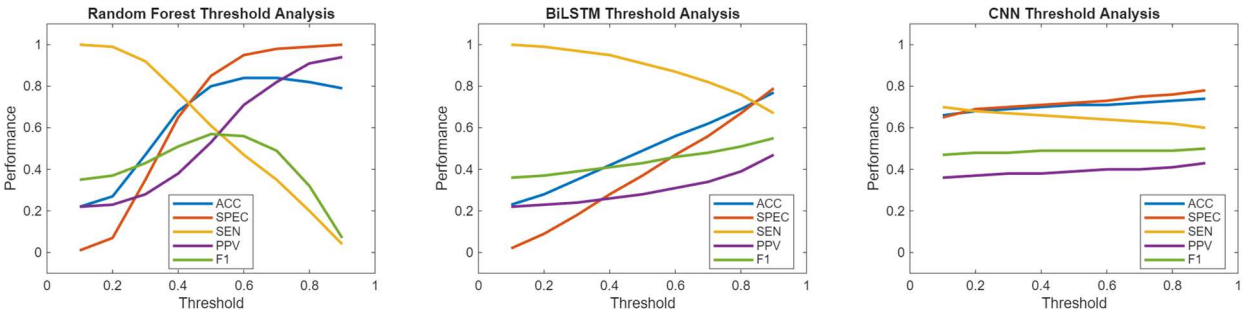


Figure 13: Performance of RF, BiLSTM, and CNN models across varying threshold values (0.1 to 0.9). Each line plot represents a different performance metric, illustrating how model performance changed with threshold adjustments

Performance metrics and ROC curves resulting from training and testing each model type on the combined dataset at 100%, 75%, 50%, and 25% of the training set size are reported in Figures 14 and 15, respectively.

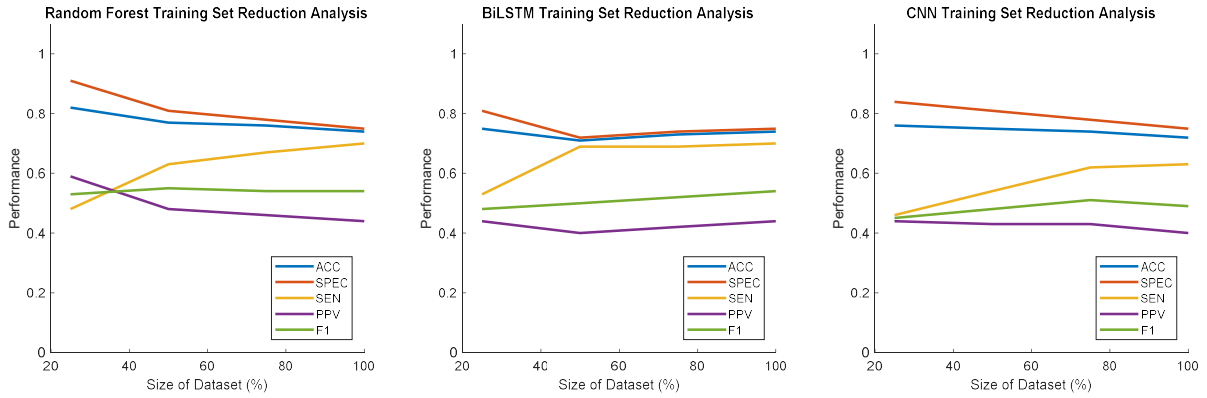


Figure 14: Performance of RF, BiLSTM, and CNN models across varying training set sizes (25–100%). Each line plot represents a different performance metric, highlighting how model performance changed with increasing training data

For RF, AUROC increased the most from 25% to 50% of training set size. This trend aligns with the other performance metrics, as performance stabilizes above 50% of training set size. For BiLSTM, AUROC steadily increases with training set size, and like RF, performance stabilizes above 50% of training set size. For CNN, AUROC and the other performance metrics stabilize above 75% of the training set size. Across all three models, sensitivity and AUROC benefit most from increased training set size.

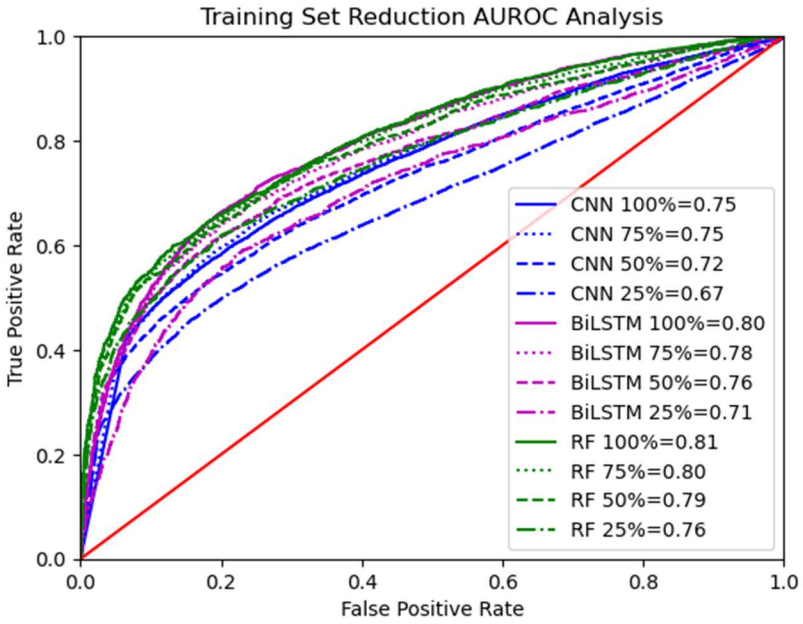


Figure 15: ROC curves and AUROC values (in legend) for all model types across varying training set sizes

2.5 Discussion

This study trained and tested machine learning models to detect fetal movement using acceleration and angular rate signals from wearable IMUs. Acceleration data, angular rate data, and combined datasets were used to train on hand-engineered features classified with RF, time series data analyzed with BiLSTM, and time–frequency spectrograms analyzed with CNN. Combining acceleration and angular rate data improved all performance metrics for RF and BiLSTM. Angular rate and combined datasets achieved similar performance for CNN and outperformed the acceleration dataset for all metrics. When thresholds were chosen to standardize specificity, RF and BiLSTM demonstrated almost identical performance across all metrics, significantly outperforming CNN. RF demonstrated the highest achievable performance capacity depending on the threshold chosen, highlighting the benefits of feature-based approaches for proof-of-concept datasets.

2.5.1 Model Performance Across Sensor Types

While single-mode sensors were able to learn discriminative patterns related to fetal movement, combining acceleration and angular rate data consistently improved performance across all performance metrics and model architectures. Angular rate data achieved comparable performance to the combined dataset with CNN. Across all data types and models, the increased performance of the combined dataset suggested that the complementary information resulting from the combination of acceleration and angular rate datasets had discriminative power over either dataset on its own. The improvements observed with the combined dataset were driven by an increase in true positives and a decrease in false negatives across all model types. These improvements suggest that combining acceleration and angular rate data enhances the models' ability to accurately identify fetal movement events, improving the reliability of these methods for clinical use. Reducing false positive outputs prevents false reassurance of fetal wellbeing, while fewer false negatives can help avoid unnecessary medical interventions.

2.5.2 Model Performance Across Data Representations

Across all data representations, models successfully learned discriminative and generalizable patterns of fetal movement, achieving "good" or "very good" performance [37]. However, when evaluating performance at a standardized specificity, hand-engineered features (RF) and time-series data (BiLSTM) yielded comparable results, while the time–frequency

approach (CNN) resulted in significantly lower AUROC, sensitivity, and F1 scores. This performance gap is likely driven by the relationship between model complexity and dataset size. Training set size analysis revealed that the CNN required 75% of the total data to achieve performance gains, whereas the RF and BiLSTM models reached a performance plateau at only 50%. These findings suggest that for proof-of-concept studies with smaller populations, simpler architectures are more data-efficient and less prone to the noise-related artifacts that hinder deep learning models like CNN.

Beyond data efficiency, the temporal resolution of the models played a role in signal clarity. The improved performance of RF and BiLSTM may stem from the alignment between their 0.5s windows and the transient nature of fetal movement. In contrast, the 8s window required to maintain an appropriate frequency resolution for CNN likely diluted the signal-to-noise ratio, as brief fetal movements occupied a smaller fraction of the analysis window. In this comparison, the hand-engineered features used for training RF more effectively described the physical dynamics of abdominal deformation than the complex spectral patterns the CNN attempted to learn from noisy, longer-duration windows.

Finally, the models exhibited distinct behaviors regarding classification thresholds, representing a trade-off between performance capacity and operational stability. The RF model demonstrated the highest capacity for peak performance, achieving specificity values above 75%. However, performance metrics were highly dependent on precise threshold tuning. In contrast, the BiLSTM and CNN metrics were more stable across thresholds, offering more predictable performance that may be easier to deploy across diverse operational settings without intensive recalibration. Ultimately, the high performance and data efficiency of the RF and BiLSTM models indicate that simpler architectures may offer better performance for fetal movement detection in proof-of-concept studies.

2.5.3 Implications for Fetal Movement Detection

When considering the deployment of a fetal movement detection system in real-world settings, the tradeoffs between interpretability, stability, and model performance must be considered [38–40]. While combining acceleration and angular rate signals from IMUs was shown to be consistently more effective than individual sensors, the choice of model depends on the specific application and data available. RF excels at simpler tasks and offers high performance capacity and interpretability through binary decision trees, making it suitable for

environments where immediate, explainable feedback is necessary [41]. However, the stability of BiLSTM across thresholds suggests it may be more robust to varying user environments without requiring intensive threshold recalibration. While CNN has the potential to excel at complex tasks in larger datasets, it is currently more computationally demanding and requires a larger volume of data to surpass the performance of simpler models [42]. For smaller datasets or consumer-level wearables where power consumption and ease of use are constraints to adoption, RF and BiLSTM provide a more practical solution [4,39,43].

While all models achieved "good" overall performance, the requirements for real-world clinical deployment impose strict constraints on error types, particularly false positives. In the context of fetal monitoring, a high false-positive rate (e.g., 20–25%) could provide false reassurance to a mother during periods of fetal distress. The RF model was uniquely capable of reducing the false-positive rate (FPR) to below 10% but at a significant tradeoff with sensitivity, which dropped to approximately 20%. This significant tradeoff implies that models must sacrifice a majority of detected events to achieve a clinically safe false positive rate. This sensitivity trade-off is further complicated by the nature of the clinical standard used in this study. Maternal perception, while a standard clinical metric, is only 33–88% sensitive to movements confirmed via ultrasound [4, 46]. Consequently, a model that is only 20% sensitive to maternally perceived events is, in reality, capturing an even smaller fraction of total fetal activity. This discrepancy suggests that while simpler models like RF provide the best available capacity for noise rejection, when current performance levels are tuned for clinical safety, these models may not yet be sufficient for standalone diagnostic use. In conclusion, further advancements in sensor fusion and the selection of data-efficient models may provide a reliable solution for both clinical and consumer fetal monitoring.

2.5.4 Limitations and Future Work

While this study provided valuable insights into data-driven fetal movement detection using wearable IMUs, there are several important limitations and areas for future research before implementation in clinical or real-world settings is possible. Although machine learning models demonstrated strong performance across various sensor types and data representations, the clinical standard in the current study was based on maternal perception, which is subjective and has intrinsic human error [4,6]. Counting perceived fetal movements requires attention and varies with maternal activity and fetal sleep cycles, so consequently some fetal movement may not be

perceived. Maternal anxiety can negatively affect maternal perception, resulting in patients seeking care for inadequate movement [4,44,45]. Validating these models for clinical use requires comparison with ultrasound, which provides a detailed view of fetal activity [4,46]. Future work should aim to conduct prospective studies where fetal movement predicted from models using wearable IMU data is compared directly with real-time ultrasound observations to assess its true accuracy and reliability.

Another important consideration is the gestational age of the participant population. Participants in this study were recruited between 24 and 32 weeks of gestational age, as mothers can typically perceive fetal movement by the 20th week and there is an increased risk of labor after 32 weeks [47]. While the frequency of fetal movement is highest between 28 and 34 weeks, the risk of stillbirth increases with gestational age being highest after 41 weeks [48]. Future studies should consider prioritizing the recruitment of participants between 28 and 32 weeks to capture a greater number of fetal movements.

For a fetal movement detection system to be effective in real-world settings, it must be robust enough to handle a wide range of maternal activities of daily living (ADLs) without sacrificing performance [43]. Current models were all trained in con-trolled environments, but real-world use presents additional challenges, such as increased noise from maternal body movement, varying sensor placements, inconsistent data quality, and the lack of reliable data labeling, whether through maternal perception or ultrasound [4]. Future research should focus on testing these models in more diverse settings, such as during physical activity, sleep, and other routine behaviors, to evaluate their performance and robustness under real-world conditions. In addition, the development of adaptive algorithms that can adjust for such variability could be a promising direction for improving system reliability in non-clinical environments.

2.6 Conclusions

In conclusion, this study demonstrated that combining acceleration and angular rate data significantly improves fetal movement detection across all data representations and model types. RF and BiLSTM achieved better overall performance than CNN, likely due to model complexity, dataset size, and temporal resolution. These findings suggest that shorter window sizes may be better suited for capturing the transient dynamics of fetal movement than the longer windows required for spectral analysis. Furthermore, while RF demonstrated the highest performance capacity through threshold tuning, BiLSTM offers greater operational stability.

Ultimately, these findings highlight the advantage of using combined datasets of acceleration and angular rates signals with simpler model architectures for proof-of-concept studies. This approach provides a robust foundation for developing continuous fetal monitoring systems that can reliably distinguish fetal activity, moving closer toward a clinical tool for improving prenatal outcomes, provided that the trade-offs between false-positive reduction and detection sensitivity can be further optimized.

2.7 Appendix A: Participant Demographics

This section contains demographic data applicable to the participant population in which IMU signals were acquired. S09 and S16 were excluded from the analysis due to signal acquisition issues reported in Section 2.2.4. Gestational age (GA) refers to the number of weeks that have passed since the first day of the participant's last menstrual period. Only the week was reported, e.g., 32 weeks, and 5 days is recorded as 32 weeks. Parity refers to the number of times a woman has given birth, where mothers in their first pregnancy are nulliparous and those with previous pregnancies are parity ≥ 1. GA and parity were self-reported by participants. Body mass index (BMI) is calculated based on the height and weight of the participant taken on the day of participation. A pre-pregnancy BMI was not collected. A BMI ≥ 30 is classified as obese. BMI was calculated according to Equation 1. GA, parity, and BMI can impact a participant's ability to perceive fetal movement [4] and may impact the maternal perception labels applied to this analysis.

$$\text{BMI} = \frac{\text{weight (kg)}}{\text{height (m)}^2} \quad \text{(Eqn. 1)}$$

Fetal movement follows a diurnal pattern, with stronger and more frequent movements generally occurring later in the day [28,47,49]. As a result, the time of day may influence both a participant's ability to perceive fetal movement and the frequency and intensity of movements that occur. To account for this potential effect, session timing was recorded and categorized as follows: "morning" sessions began between 7 and 10 a.m., "afternoon" sessions began between 2 and 4 p.m., and "evening" sessions began at or after 5 p.m.

Table 1: Demographic information for the participant population included in studies 1 & 2 related to the ability to perceive fetal movement

Participant Identifier	GA (in Weeks)	Parity	BMI	Time of Day of Data Collection

S01	32	Parity ≥ 1	37.2	Evening
S02	28	Nulliparous	38.7	Morning
S03	32	Nulliparous	31.7	Afternoon
S04	32	Parity ≥ 1	36.4	Evening
S05	29	Nulliparous	28.2	Morning
S06	31	Nulliparous	28.8	Afternoon
S07	26	Parity ≥ 1	36.3	Afternoon
S08	26	Nulliparous	25.7	Evening
S10	25	Parity ≥ 1	27.7	Morning
S11	31	Nulliparous	25.0	Evening
S12	28	Nulliparous	23.5	Evening
S13	24	Not reported	27.1	Evening
S14	26	Parity ≥ 1	35.0	Evening
S15	28	Nulliparous	25.0	Evening
S17	32	Nulliparous	39.8	Evening
S18	27	Not reported	30.3	Evening
S19	25	Parity ≥ 1	27.1	Evening
S20	25	Parity ≥ 1	30.5	Evening
S21	30	Parity ≥ 1	35.9	Evening
S22	24	Parity ≥ 1	24.8	Evening
S23	27	Nulliparous	25.8	Morning

2.8 References

[1] Walton, J.R.; Peaceman, A.M. Identification, Assessment and Management of Fetal Compromise. Clin. Perinatol. 2012, 39, 753–768.

[2] Frøen, J.F.; Frøen, J.F. A Kick from Within-Fetal Movement Counting and the Cancelled Progress in Antenatal Care Introduction and History. J. Perinat. Med. 2004, 32, 13–24.

[3] Dutton, P.J.; Warrander, L.K.; Roberts, S.A.; Bernatavicius, G.; Byrd, L.M.; Gaze, D.; Kroll, J.; Jones, R.L.; Sibley, C.P.; Frøen, J.F.; et al. Predictors of Poor Perinatal Outcome Following Maternal Perception of Reduced Fetal Movements—A Prospective Cohort Study. PLoS ONE 2012, 7, e39784. https://doi.org/10.1371/journal.pone.0039784.

[4] Brown, R.; Johnstone, E.D.; Heazell, A.E.P. Professionals' Views of Fetal-Monitoring Support the Development of Devices to Provide Objective Longer-Term Assessment of Fetal Wellbeing. J. Matern.-Fetal Neonatal Med. 2016, 29, 1680–1686. https://doi.org/10.3109/14767058.2015.1059808.

[5] Minors, D.S.; Waterhouse, J.M. The Effect of Maternal Posture, Meals and Time of Day on Fetal Movements. BJOG Int. J. Obstet. Gynaecol. 1979, 86, 717–723. https://doi.org/10.1111/j.1471-0528.1979.tb11273.x.

[6] Hijazi, Z.R.; Callan, S.E.; East, C.E. Maternal Perception of Foetal Movement Compared with Movement Detected by Real-Time Ultrasound: An Exploratory Study. Aust. N. Z. J. Obstet. Gynaecol. 2010, 50, 144–147. https://doi.org/10.1111/j.1479-828X.2009.01122.x.

[7] Heazell, A.E.P.; Frøen, J.F. Methods of Fetal Movement Counting and the Detection of Fetal Compromise. J. Obstet. Gynaecol. 2008, 28, 147–154. https://doi.org/10.1080/01443610801912618.

[8] Heazell, A.E.P.; Sumathi, G.M.; Bhatti, N.R. What Investigation Is Appropriate Following Maternal Perception of Reduced Fetal Movements? J. Obstet. Gynaecol. 2005, 25, 648–650. https://doi.org/10.1080/01443610500278303.

[9] Mesbah, M.; Khlif, M.S.; Layeghy, S.; East, C.E.; Dong, S.; Brodtmann, A.; Colditz, P.B.; Boashash, B. Automatic Fetal Movement Recognition from Multi-Channel Accelerometry Data. Comput. Methods Programs Biomed. 2021, 210, 106377. https://doi.org/10.1016/j.cmpb.2021.106377.

[10] Layeghy, S.; Azemi, G.; Colditz, P.; Boashash, B. Non-Invasive Monitoring of Fetal Movements Using Time-Frequency Features of Accelerometry. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4 May 2014; IEEE: New York, NY, USA, 2014; pp. 4379–4383.

[11] Liang, S.; Peng, J.; Xu, Y. Passive Fetal Movement Signal Detection System Based on Intelligent Sensing Technology. J. Healthc. Eng. 2021, 2021, 1745292. https://doi.org/10.1155/2021/1745292.

[12] Zhao, X.; Zeng, X.; Koehl, L.; Tartare, G.; De Jonckheere, J. A Wearable System for In-Home and Long-Term Assessment of Fetal Movement. IRBM 2020, 41, 205–211.

[13] Khlif, M.S.; Boashash, B.; Layeghy, S.; Ben-Jabeur, T.; Mesbah, M.; East, C.; Colditz, P. Time-Frequency Characterization of Tri-Axial Accelerometer Data for Fetal Movement Detection. In Proceedings of the 2011 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Bilbao, Spain, 14 December 2011; IEEE: New York, NY, USA, 2011; pp. 466–471. https://doi.org/10.1109/ISSPIT.2011.6151607.

[14] Delay, U.; Nawarathne, T.; Dissanayake, S.; Gunarathne, S.; Withanage, T.; Godaliyadda, R.; Rathnayake, C.; Ekanayake, P.; Wijayakulasooriya, J. Novel Non-

Invasive in-House Fabricated Wearable System with a Hybrid Algorithm for Fetal Movement Recognition. PLoS ONE 2021, 16, e0254560. https://doi.org/10.1371/journal.pone.0254560.

[15] Xu, J.; Zhao, C.; Ding, B.; Gu, X.; Zeng, W.; Qiu, L.; Yu, H.; Shen, Y.; Liu, H. Fetal Movement Detection by Wearable Accel-erometer Duo Based on Machine Learning. IEEE Sens. J. 2022, 22, 11526–11534. https://doi.org/10.1109/JSEN.2022.3172451.

[16] Altini, M.; Mullan, P.; Rooijakkers, M.; Gradl, S.; Penders, J.; Geusens, N.; Grieten, L.; Eskofier, B. Detection of Fetal Kicks Using Body-Worn Accelerometers During Pregnancy: Trade-Offs Between Sensors Number and Positioning. In Proceedings of the 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL, USA, 16 August 2016; IEEE: New York, NY, USA, 2016; pp. 5319–5322.

[17] Nemati, E.; Zhang, S.; Ahmed, T.; Rahman, M.M.; Kuang, J.; Gao, A. CoughBuddy: Multi-Modal Cough Event Detection Using Earbuds Platform. In Proceedings of the 2021 IEEE 17th International Conference on Wearable and Implantable Body Sensor Networks (BSN), Virtual, 27 July 2021; IEEE: New York, NY, USA, 2021; pp. 1–4. https://doi.org/10.1109/BSN51625.2021.9507017.

[18] Lotfi, R.; Eskicioglu, R.; Tzanetakis, G.; Irani, P. A Comparison between Audio and IMU Data to Detect Chewing Events Based on an Earable Device. In Proceedings of the 11th Augmented Human International Conference, Winnipeg, MB, Canada, 27 May 2020; pp. 1–8.

[19] Gomes, D.; Sousa, I. Real-Time Drink Trigger Detection in Free-Living Conditions Using Intertial Sensors. Sensors 2019, 19, 2145.

[20] Senyurek, V.Y.; Imtiaz, M.H.; Belsare, P.; Tiffany, S.; Sazonov, E. A CNN-LSTM Neural Network for Recognition of Puffing in Smoking Episodes Using Wearable Sensors. Biomed. Eng. Lett. 2020, 10, 195–203.

[21] Luinge, H.J.; Veltink, P.H. Measuring Orientation of Human Body Segments Using Miniature Gyroscopes and Accelerometers. Med. Biol. Eng. Comput. 2005, 43, 273–282.

[22] Varkey, J.P.; Pompili, D.; Walls, T.A. Human Motion Recognition Using a Wireless Sensor-Based Wearable System. Pers. Ubiquitous Comput. 2012, 16, 897–910. https://doi.org/10.1007/s00779-011-0455-4.

[23] Kot, A.; Nawrocka, A. Modeling of Human Balance as an Inverted Pendulum. In Proceedings of the 2014 15th International Carpathian Control Conference (ICCC), Velke Karlovice, Czech Republic, 28–30 May 2014; IEEE: New York, NY, USA, 2014; pp. 254–257. https://doi.org/10.1109/CarpathianCC.2014.6843607.

[24] Noamani, A.; Vette, A.H.; Rouhani, H. Nonlinear Response of Human Trunk Musculature Explains Neuromuscular Stabiliza-tion Mechanisms in Sitting Posture. J. Neural Eng. 2022, 19, 026045. https://doi.org/10.1088/1741-2552/ac63ed.

[25] Alwis, P.; Thilakasiri, I.; Liyanage, S.; Nanayakkara, R.; Godaliyadda, R.; Ekanayake, M.; Rathnayake, C.; Wijayakulasooriya, J.; Herath, V. Application of an LSTM-Based Channel Attention and Classification Mechanism in Fetal Movement Monitoring. TechRxiv 2024, https://doi.org/10.36227/techrxiv.24630642.v1.

[26] Vitali, R.V.; Perkins, N.C. Determining Anatomical Frames via Inertial Motion Capture: A Survey of Methods. J. Biomech. 2020, 106, 109832.

[27] Bradford, B.F.; Cronin, R.S.; McCowan, L.M.E.; McKinlay, C.J.D.; Mitchell, E.A.; Thompson, J.M.D. Association between Maternally Perceived Quality and Pattern of Fetal Movements and Late Stillbirth. Sci. Rep. 2019, 9, 9815. https://doi.org/10.1038/s41598-019-46323-4.

[28] Bradford, B.F.; Cronin, R.S.; McKinlay, C.J.D.; Thompson, J.M.D.; Mitchell, E.A.; Stone, P.R.; McCowan, L.M.E. A Diurnal Fetal Movement Pattern: Findings from a Cross-Sectional Study of Maternally Perceived Fetal Movements in the Third Trimester of Pregnancy. PLoS ONE 2019, 14, e0217583. https://doi.org/10.1371/journal.pone.0217583.

[29] Lai, J.; Woodward, R.; Alexandrov, Y.; ain Munnee, Q.; Lees, C.C.; Vaidyanathan, R.; Nowlan, N.C. Performance of a Wearable Acoustic System for Fetal Movement Discrimination. PLoS ONE 2018, 13, e0195728. https://doi.org/10.1371/journal.pone.0195728.

[30] Spicher, L.; Bell, C.; Huan, X.; Sienko, K. Exploring Random Forest Machine Learning for Fetal Movement Detection Using Abdominal Acceleration and Angular Rate Data. In Proceedings of the 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL, USA, 15 July 2024; IEEE: New York, NY, USA, 2024I; pp. 1–4. https://doi.org/10.1109/EMBC53108.2024.10782456.

[31] Shan, G. Monte Carlo Cross-Validation for a Study with Binary Outcome and Limited Sample Size. BMC Med. Inform. Decis. Mak. 2022, 22, 270. https://doi.org/10.1186/s12911-022-02016-z.

[32] Fadili, Y.; El Yamani, Y.; Kilani, J.; El Kamoun, N.; Baddi, Y.; Bensalah, F. An Enhancing Timeseries Anomaly Detection Using LSTM and Bi-LSTM Architectures. In Proceedings of the 2024 11th International Conference on Wireless Networks and Mobile Communications (WINCOM), Leeds, UK, 23 July 2024; IEEE: New York, NY, USA, 2024; pp. 1–6.

[33] Graves, A.; Fernández, S.; Schmidhuber, J. Bidirectional LSTM Networks for Improved Phoneme Classification and Recogni-tion. In Proceedings from the International Conference on Artificial Neural Networks (ICANN), Warsaw, Poland, 11 September 2005; Springer: Berlin/Heidelberg, Germany, 2005; pp. 799–804.

[34] Addison, P.; Walker, J.; Guido, R. Time—Frequency Analysis of Biosignals. In IEEE Engineering in Medicine and Biology Magazine; IEEE: New York, NY, USA, 2009; Volume 28, pp. 14–29.

[35] David, M.; Hirsch, M.; Karin, J.; Toledo, E.; Akselrod, S.; An, A.S. An Estimate of Fetal Autonomic State by Time-Frequency Analysis of Fetal Heart Rate Variability. J. Appl. Physiol. 2007, 102, 1057–1064. https://doi.org/10.1152/japplphysiol.00114.2006.

[36] Chauhan, R.; Ghanshala, K.K.; Joshi, R.C. Convolutional Neural Network (CNN) for Image Detection and Recognition. In Proceedings of the 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), Jalandhar, India, 15 December 2018; IEEE: New York, NY, USA, 2018; pp. 278–282.

[37] Šimundić, A.-M. Measures of Diagnostic Accuracy: Basic Definitions. ejifcc 2009, 19, 203.

[38] Rudin, C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. Nat. Mach. Intell. 2019, 1, 206–215.

[39] Tonekaboni, S.; Joshi, S.; Mccradden, M.D.; Goldenberg, A. What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. In Proceedings of the Machine Learning for Healthcare Conference, Ann Arbor, Michigan, 9-10 August 2019; PMLR; pp. 359–380.

[40] Stiglic, G.; Kocbek, P.; Fijacko, N.; Zitnik, M.; Verbert, K.; Cilar, L. Interpretability of Machine Learning-Based Prediction Models in Healthcare. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 2020, 10, e1379.

[41] Nazir, Z.; Zarymkanov, T.; Park, J.-G. A Machine Learning Model Selection Considering Tradeoffs between Accuracy and Interpretability. In Proceedings of the 2021 13th International Conference on Information Technology and Electrical Engi-neering (ICITEE), Virtual, 14 October 2021; IEEE: New York, NY, USA, 2021; pp. 63–68.

[42] Alzubaidi, L.; Zhang, J.; Humaidi, A.J.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaría, J.; Fadhel, M.A.; Al-Amidie, M.; Farhan, L. Review of Deep Learning: Concepts, CNN Architectures, Challenges, Applications, Future Directions. J. Big Data 2021, 8, 53. https://doi.org/10.1186/s40537-021-00444-8.

[43] Rodriguez-Villegas, E.; Iranmanesh, S.; Imtiaz, S.A. Wearable Medical Devices: High-Level System Design Considerations and Tradeoffs. IEEE Solid-State Circuits Mag. 2018, 10, 43–52. https://doi.org/10.1109/MSSC.2018.2867247.

[44] Peahl, A.F.; Gourevitch, R.A.; Luo, E.M.; Fryer, K.E.; Moniz, M.H.; Dalton, V.K.; Fendrick, A.M.; Shah, N. Right-Sizing Prenatal Care to Meet Patients' Needs and Improve Maternity Care Value. Obstet. Gynecol. 2020, 135, 1027–1037.

[45] Carter, E.B.; Tuuli, M.G.; Caughey, A.B.; Odibo, A.O.; Macones, G.A.; Cahill, A.G. Number of Prenatal Visits and Pregnancy Outcomes in Low-Risk Women. J. Perinatol. 2016, 36, 178–181. https://doi.org/10.1038/jp.2015.183.

[46] Besinger, R.E.; Johnson, T.R.B. Doppler Recordings of Fetal Movement: Clinical Correlation with Real-Time Ultrasound. Obstet. Gynecol. 1989, 74, 277–280.

[47] Raynes-Greenow, C.H.; Gordon, A.; Li, Q.; Hyett, J.A. A Cross-Sectional Study of Maternal Perception of Fetal Movements and Antenatal Advice in a General Pregnant Population, Using a Qualitative Framework. BMC Pregnancy Childbirth 2013, 13, 32. https://doi.org/10.1186/1471-2393-13-32.

[48] Mangesi, L.; Hofmeyr, G.J.; Smith, V.; Smyth, R.M.D. Fetal Movement Counting for Assessment of Fetal Wellbeing. Cochrane Database Syst. Rev. 2015, 10, CD004909.

[49] Ozkaya, E.; Baser, E.; Cinar, M.; Korkmaz, V.; Kucukozkan, T. Does Diurnal Rhythm Have an Impact on Fetal Biophysical Profile? J. Matern.-Fetal Neonatal Med. 2012, 25, 335–338. https://doi.org/10.3109/14767058.2011.576721.

Chapter 3 Four Data-Driven Design Choices for IMU-Based Fetal Movement Detection: Sensitivity, Thresholding, Windowing, and Filtering[2]

3.1 Abstract

Prenatal care requires objective, continuous monitoring, yet current fetal movement (FM) detection methods rely on subjective maternal perception or intermittent clinical testing. Laboratory-based FM detection systems lack data-driven standards for signal preparation and algorithm design. This study presents a methodological approach for characterizing inertial measurement unit (IMU) signals and applies the approach to a unique, clinician-validated dataset of eighteen pregnant participants to provide data-driven insights for objective FM monitoring under stationary conditions. Specifically, suggested guidance for four design choices was presented: (1) The minimum hardware sensitivity required to reliably capture weak FM, (2) the signal magnitude threshold to reliably capture strong FM while rejecting artifacts from maternal movement, (3) the maximum window size for data analysis based on observed FM durations, and (4) the cutoff frequencies for filter design to reject signal noise and artifacts. Linear Mixed-Effect Models (LMEMs) were applied to 1371 FM events across four analytical classes (No FM, Weak FM, Strong FM, Artifact) and the estimated marginal means (EMMs) of the groups were compared to quantify the presented design choices. The methodology presented acts as a foundational step toward standardizing the signal preparation and algorithm design of future fetal movement detection systems.

3.2 Introduction

In the United States, over 20,000 stillbirths are reported annually [1], yet nearly 32% may be preventable through improved prenatal care [2]. Early signs of fetal compromise are often subtle, making timely intervention difficult. Reductions in fetal movement (FM) can indicate complications [3], but current reliance on subjective maternal perception [4] and intermittent

[2] Adapted from the published work: Spicher, L., Bell, C., Huan, X., & Sienko, K. (2026, April). Four Data-Driven Design Choices for IMU-Based Fetal Movement Detection: Sensitivity, Thresholding, Windowing, and Filtering. In 2026 Design of Medical Devices Conference.

clinical assessment provides only a “snapshot in time” of the fetal status [5]. A continuous, objective FM monitoring system is essential to bridge this gap, providing the longitudinal data necessary for personalized risk assessment. By enabling data collection during daily activities, wearable systems can identify changes in FM patterns over time. Similar to how the Holter monitor transformed arrhythmia detection through continuous ECG monitoring [5], a wearable FM monitor offers a proactive path to care that currently does not exist for fetal health.

Accelerometer-based systems have shown promise for FM monitoring. For example, Mesbah et al. achieved high performance compared to ultrasound-validated FMs (AUROC = 0.98, SEN = 0.95, and PPV = 0.92) using three abdominal accelerometers and one reference chest sensor paired with a Random Forest classifier [6]. Zhao et al. also achieved high performance (SPEC = 0.99, SEN = 0.77) but compared perception-validated FMs and used two abdominal accelerometers paired with a fuzzy ARTMAP classifier [7]. However, these two studies are difficult to compare directly because they used different sensor configurations, distinct machine learning classifiers, and different label sources to define FM events. Despite different design choices, both emphasized the difficulty of overcoming artifacts caused by maternal movement, which is the primary obstacle to progress in the field of fetal movement detection.

This methodological ambiguity reflects a common and ongoing challenge in mature physiological signal processing fields. For example, while continuous ECG monitoring has been deployed in real-world settings, experimental studies to improve signal quality and reject artifacts are difficult to compare, particularly across data preprocessing and segmentation methods; researchers are advocating for consensus methods and data-driven standards for signal preparation [8]. In fetal monitoring studies, a similar lack of consensus is apparent. Researchers report hardware specifications, artifact thresholding methods, and frequency cutoffs for filtering that range widely without formal analysis to justify the numerical choices [6], [7], [9]. Many different window sizes are chosen for data segmentation [6], [7], [9], but without rationale for why they were chosen. To our knowledge, no prior work has systematically contextualized or characterized FM data from IMUs, leaving a critical gap in data-driven standards for signal preparation and algorithm design choices.

The objective of this study, therefore, was to present a rigorous, data-driven methodology for characterizing IMU signals of FM, and to apply this approach to a unique, clinician validated

dataset to provide data-driven insights and guidance for the engineering of objective FM monitoring technology under stationary conditions. After contextualizing the dataset by quantifying the precision and recall of maternal perception against the ultrasound clinical standard to provide a reference for comparing future datasets, this study established a data-driven methodology for quantifying four key design choices: the suggested minimum hardware sensitivity and artifact rejection threshold based on root mean square (RMS) magnitude, cutoff frequencies for filter design based on differences in peak frequency between fetal and maternal movement artifacts, and maximum window size for data analysis based on the observed range of FM duration.

3.3 Materials and Methods

Eighteen participants between 18 and 49 years old were recruited for a single data collection session at the University of Michigan. All participants were between 24 and 32 weeks of gestational age and carrying a singleton pregnancy with no known fetal disease or physical abnormalities. Participants with a diagnosis of gestational diabetes or hypertension were excluded. The study was approved by the University of Michigan Institutional Review Board (HUM00204999).

3.3.1 Data Collection & Categorization

During data collection, participants were fitted with an array of four abdominal IMUs (Opal, APDM Inc., Portland, OR, USA) centered around the umbilicus and attached with adhesive tape. Synchronized acceleration and angular rate data were collected at a sampling rate of 128 Hz. A series of three hip-hinging movements were conducted first to calibrate the IMUs. Participants were then seated in a chair with their feet flat on the floor, and data were collected in 10–15-minute trials. Participants were instructed to remain stationary during data collection. Each participant completed 2-3 trials, resulting in 35 total trials in the dataset. The study team recorded any maternal movement events with a timestamp.

While seated, two measurements were collected to later validate FMs: maternal perception annotations and ultrasound recordings. Participants received an additional handheld IMU with a pressable button to indicate perceived FM and were instructed to press the button at the onset of FM. Additionally, an ultrasound scan was administered by a board-certified OB/GYN throughout the duration of each trial. Before data collection, a clear visual of the fetus

was obtained, and if the view was obscured during data collection, the ultrasound probe was moved to regain a visual. Recordings of the ultrasound scan were collected when FM was identified. These recordings were post-processed and time synchronized with the IMU data. The ultrasound recordings were reviewed by the OB/GYN after data collection to provide refined annotations. Data from two of the eighteen participants were excluded due to signal acquisition issues upon review.

The ultrasound recordings were taken as ground-truth reference data. The onset (T_{Start}) and offset (T_{End}) of any FM were manually determined by visual inspection of the ultrasound recordings and used to index windows of FM. These FM events were then categorized by the board-certified OB/GYN into two clinically relevant groups, "strong" and "weak", based on subjective assessment of uterine wall contact and full-body displacement. Weak FMs were those that did not contact the uterine wall and were not expected to be perceived by the participant [4]. Examples include flexion/extension of the hands/feet, twitches, or non-prolonged rolling. Strong FMs were those that did contact the uterine wall and were expected to be perceived. Examples include fetal kicks resulting in full body movement or long, sustained rolling/stretching [4]. Because the frequency and intensity of FM was highly variable between participants, distinctions between "weak" and "strong" FMs were influenced by the behavior of the individual fetus for each participant. Differentiating these strong and weak FMs is critical, as the goal of a clinically viable device is to reliably detect strong FM that mothers would likely perceive, while establishing the necessary sensitivity floor to capture subtle (but clinically relevant) weak FM that mothers often miss.

Maternal movement events identified by the study team were timestamped and used to index windows labeled as "Artifact." Examples of maternal movement include talking, coughing, deep breathing, and gross body movement. For this initial analysis, any FM detected via ultrasound during a maternal movement event was labeled as "Artifact." This conservative labeling approach was intentionally adopted to isolate fetal signals under stationary conditions and to prevent maternal motion from confounding the characterization of the FM signal itself. By restricting this analysis to stationary data, we aimed to establish a clear baseline understanding of system performance before addressing the more complex challenge of disentangling concurrent maternal and fetal movements. All remaining windows were labeled as "No FM" and therefore

did not contain any maternal or fetal movement. All segmented data was grouped into the four final analytical classes: “Strong FM,” “Weak FM,” “Artifact,” and “No FM”.

Then, maternal perception data was compared against the ultrasound clinical standard. Each FM event identified via ultrasound was also marked as “perceived” or “missed” based on the maternal perception annotations. “Perceived” FMs were those for which the maternal perception button was pressed within 5 seconds of the event. All other FM events were marked as “missed.” Maternal perception annotations that did not align with FM events identified by ultrasound were marked as “false positive”. Though maternal perception of FM is subjective, it is also possible that false positives may indicate FM outside of the plane of view of the ultrasound.

3.3.2 Signal Processing & Feature Calculation

The raw tri-axial acceleration and angular rate signals from each IMU were processed through a zero-phase bandpass filter (1-20 Hz) for denoising [10]. Using the IMU signals during calibration, the Functional Alignment Method [11] was applied to orient the angular rate data to the same frame of reference: each IMU’s x-axis was aligned with global gravity (identified during static standing), the y-axis with the participant’s frontal axis, and the z-axis with the sagittal axis. The frontal and sagittal axes were identified via rotation about the hip axis (Figure 16). The acceleration data was not oriented to retain the strength of the acceleration signals during motion tangential to the abdomen.

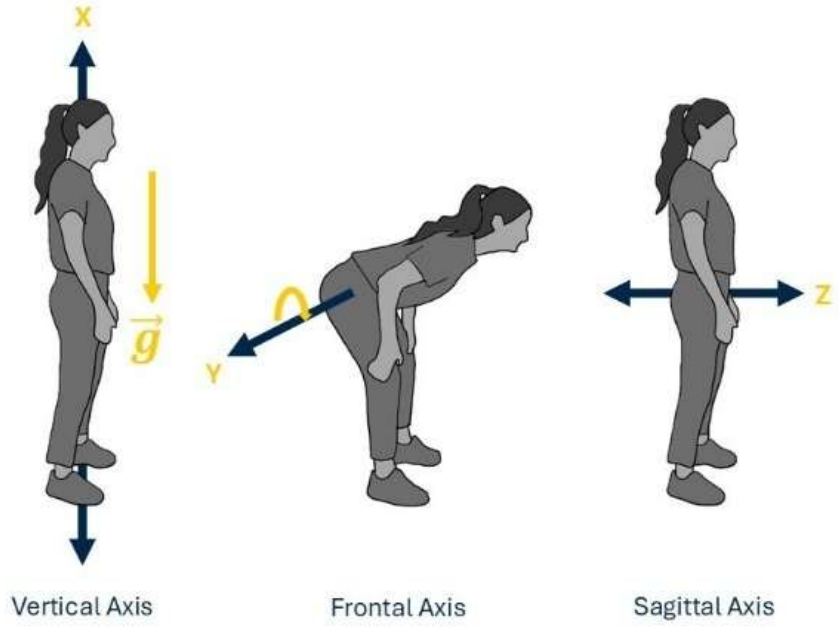


Figure 16: Functional Alignment Method for orienting angular rate data to the anatomical frame

After preprocessing, the IMU signals were segmented based on the timestamps defined by each analytical class. Feature calculation was performed on the segmented IMU data across all four IMUs and all six channels (3-axis acceleration and 3-axis angular rate). Within each variable-duration window, a series of comprehensive statistical features were calculated to characterize the signal in the time and frequency domains. The features are defined in Table 2.

Table 2: Features calculated within each window

Feature Name	**Calculation**	**Rationale**
Duration	Ultrasound-validated onset and offset of FM, $T_{end} - T_{start}$	Quantifies duration of FM events. Used to inform windowing strategies.
Root Mean Square (RMS)	Signal intensity calculated as the square root of the mean of square values within a window.	Selected as the primary metric for magnitude boundaries, as it is a fundamental measure of the FM signal's energy content [71]. Used to derive minimum hardware sensitivity and maximum artifact threshold.
Peak Frequency	Frequency corresponding to the maximum power spectral density value within a window of unfiltered data.	Characterizes the dominant frequency of distinct physical events Used to inform filter cutoff frequencies for effective noise mitigation.

Within each window, the IMU exhibiting the largest feature magnitude was retained, while the remaining three were discarded. This selection was performed independently on a window-by-window basis, meaning the retained IMU could differ between windows. The purpose of this approach was to emphasize the strongest available signal (i.e., the highest signal-to-noise ratio) for the subsequent statistical analysis.

3.3.3 Statistical Analysis and Planned Comparisons

The performance metrics of maternal perception against the ultrasound clinical standard were calculated, specifically reporting precision (positive predictive value) and recall (sensitivity) to provide clinical context for the dataset.

The analysis was designed to rigorously test three pairwise comparisons for data-driven signal preparation and algorithm design: (1) Sensitivity Floor: Comparison between no FM and weak FM based on RMS features; (2) Artifact Threshold: Comparison between strong FM and artifacts based on RMS features; and (3) Filter Cutoffs: Comparison between strong FM and

artifacts based on peak frequency features. Because multiple events (e.g., strong FMs, weak FMs, and artifacts) were identified across several trials for the same participant, the effect of the analytical class on each feature was assessed with Linear Mixed-Effect Models (LMEMs) to properly account for the variance attributed to the individual subject and trial. For each feature, an independent LMEM was computed across all six axes (three-axis acceleration and three-axis angular rate):

$$Feature \sim Group + (1|Particpant) + (1|Trial{:}\,Participant),$$

where the fixed effect was the group (e.g., "No FM", "weak FM", "strong FM", or "artifact") and the random effects were the random intercept for participant and the random intercept for trial nested within participant. This approach controlled for inherent individual differences between trials and participants. An overall comparison of means was conducted on the resulting LMEM to determine the statistical significance of the group factor. Pairwise comparisons between the estimated marginal means (EMMs) of the four groups were then calculated and adjusted using the Tukey method for multiple comparisons. Significance was defined at the $p <$ 0.05 level.

The statistically significant mean differences derived from these focused comparisons were then used to quantify data-driven choices for hardware sensitivity, algorithm thresholds, and filter design. Furthermore, the distribution of all FM durations was quantified to establish guidance for windowing strategies in subsequent algorithm design.

3.4 Results and Discussion

3.4.1 Context: Maternal Perception Performance

1371 FM events were identified by ultrasound. A total of 534 FMs were perceived by participants, resulting in a recall (sensitivity) of 39% across the dataset (compared to 33-88% reported in literature [4]). Participants indicated perceived FM 91 times that were not observed by ultrasound, for a precision (positive predictive value) of 85%. A participant's ability to perceive FM was influenced by factors such as gestational age, body mass index (BMI), and placental location [3], [4]. A participant-level analysis of the performance metrics of maternal perception against the ultrasound clinical standard is available in Table 3.

Table 3: Participant performance metrics of maternal perception against ultrasound clinical standard and related demographics

Subject #	**FMs**	**Precision**	**Recall**	**GA**	**BMI**	**Placenta Location**
1	29	1.00	0.62	31	28.8	Anterior
2	50	0.86	0.12	26	36.3	Anterior
3	91	1.00	0.19	26	25.7	Anterior
4	Excluded due to signal acquisition issues					
5	58	0.91	0.36	25	27.7	Anterior
6	17	0.50	0.12	31	25.0	Anterior
7	116	1.00	0.16	28	23.5	Posterior
8	43	0.48	0.22	24	27.1	Posterior
9	148	0.80	0.53	26	35.0	Anterior
10	98	0.95	0.18	28	25.0	Anterior
11	Excluded due to signal acquisition issues					
12	91	0.71	0.40	32	39.8	Anterior
13	85	1.00	0.25	27	30.3	Anterior
14	82	0.97	0.41	25	27.1	Anterior
15	105	0.78	0.44	25	30.5	Posterior
16	120	0.89	0.55	30	35.9	Anterior
17	185	0.96	0.69	24	24.8	Posterior
18	53	0.90	0.34	27	25.8	Anterior
TOTAL	**1371**	**0.87 [0.48,1.00]**	**0.39 [0.12,0.69]**			

Of the 1371 FM events identified by ultrasound, 552 (40%) were categorized as “strong FM” and 821 (60%) as “weak FM”. Participants perceived 59% of strong FMs, but only 26% of weak FMs, supporting the assertion that subtle weak FM is less likely to be perceived. Given that strong FMs are more readily perceived, detection technologies must reliably capture them.

Furthermore, the minimum signal magnitude for detection (sensitivity floor) can be quantified to reliably detect weak FMs, which appear much smaller in amplitude and can be difficult to distinguish from the ambient noise of no FM as seen in Figure 17. The sensitivity floor should allow the IMU sensors to capture most of the weak FM signals, acknowledging that some of these subtle FM events may still fall below IMU detection thresholds.

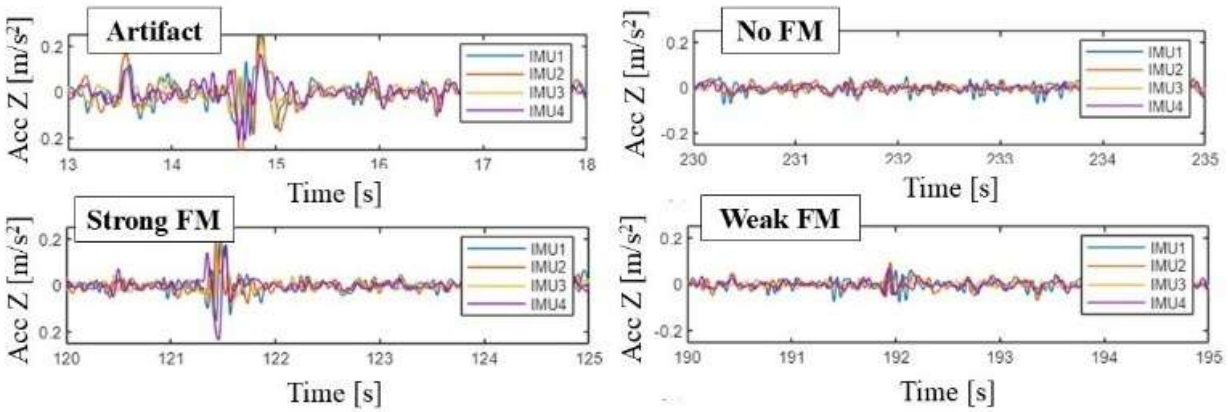


Figure 17: Each analytical class plotted for the z-axis of acceleration across all four IMUs. The top left plot is an example of a maternal cough.

Review of study team annotations identified a total of 49 artifact events. These artifacts included gross maternal movement, deep breathing, coughing, laughing, and talking. While gross maternal movements may be much larger in magnitude than FM signals, other artifacts like the maternal cough shown in Figure 17 can resemble strong FMs in magnitude, making these events difficult to differentiate from actual FM. False positive predictions of FM caused by artifacts are dangerous because they give patients and clinicians false reassurance of fetal wellbeing. Though false negative predictions caused by a lack of separation of weak FM from the no FM baseline signal may increase the incidence of additional evaluation and intervention, objectively detecting most FMs can still help identify true reductions in FM and prompt timely intervention. In risk-averse systems, a conservative approach that tolerates under-prediction of FMs is preferable, as it facilitates additional assessment of fetal well-being and reduces the likelihood of missing clinically relevant complications.

3.4.2 RMS Magnitude Analysis

A comparison of RMS magnitude between groups was conducted to establish thresholds for FM detection while reducing the potential of false positive predictions (Figure 18). LMEMs applied to the RMS magnitude demonstrated a significant effect of group (p<0.001) across the three axes of acceleration and three axes of angular rate, accounting for random effects of participant and trial. The resulting estimated marginal means (EMMs) were used to define the hardware sensitivity floor (Section 3.4.2.1) and the artifact rejection threshold (Section 3.4.2.2).

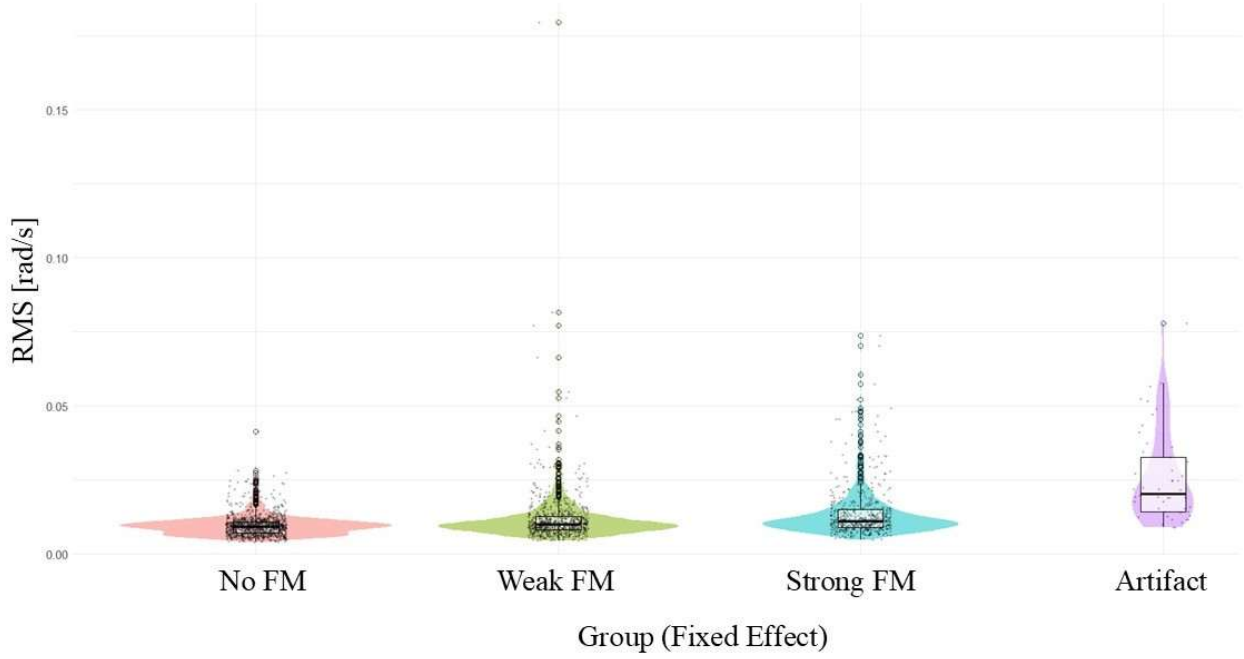


Figure 18: RMS magnitude feature (y-axis of angular rate data) used to define sensitivity floor and artifact rejection threshold

3.4.2.1 Sensitivity Floor: Defining Minimum Hardware Requirements

First, RMS was analyzed to identify significant differences between no FM and weak FM windows to quantify the sensitivity floor for hardware selection. Pairwise comparisons showed statistically significant separation for all three axes of acceleration and three axes of angular rate (Table 4). The significant differences at the lower detection limit confirmed the ability of both acceleration and angular rate IMU signals to reliably detect non-perceived FM. Setting an RMS noise floor significantly lower than 0.0246 m/s^2 and 0.00821 rad/s for hardware selection may help avoid missing clinically valuable weak FM events.

Table 4: Pairwise comparisons between no FM and weak FM groups for RMS magnitude feature (*p*-values Tukey-adjusted, α=0.05)

Signal Comparison	**EMM (No FM)**	**EMM (Weak FM)**	***p*-value**	**Cohen's d**
Acc X-axis	0.0342 m/s^2	0.0380 m/s^2	<0.0001	-0.273
Acc Y-axis	0.0222 m/s^2	0.0246 m/s^2	<0.0001	-0.248
Acc Z-axis	0.0254 m/s^2	0.0294 m/s^2	<0.0001	-0.378
Gyro X-axis	0.00651 rad/s	0.00821 rad/s	<0.0001	-0.332
Gyro Y-axis	0.00957 rad/s	0.01160 rad/s	<0.0001	-0.333
Gyro Z-axis	0.00710 rad/s	0.00871 rad/s	<0.0001	-0.277

3.4.2.2 Artifact Rejection: Defining Algorithm Threshold

RMS was also evaluated to identify significant differences between artifact and strong FM windows to quantify a magnitude threshold for mitigating false positives. Pairwise comparisons showed statistically significant separation for all three axes of acceleration and three axes of angular rate (Table 5). The significant difference between groups confirmed the ability to filter out higher magnitude maternal movement through artifact thresholding. For example, applying magnitude-based thresholds of 0.061 m/s^2 and 0.021 rad/s, set at one and a half times the mean strong FM signal, could support effective artifact rejection near the upper detection limit.

Table 5: Pairwise comparisons between strong FM and artifact groups for RMS magnitude feature (*p*-values Tukey adjusted, $\alpha = 0.05$)

Signal Comparison	**EMM (Strong FM)**	**EMM (Artifact)**	***p*-value**	**Cohen's d**
Acc X-axis	0.0405 m/s^2	0.0706 m/s^2	<0.0001	-0.900
Acc Y-axis	0.0264 m/s^2	0.0595 m/s^2	<0.0001	-0.907
Acc Z-axis	0.0354 m/s^2	0.0579 m/s^2	<0.0001	-0.840
Gyro X-axis	0.01037 rad/s	0.01775 rad/s	0.0001	-0.817
Gyro Y-axis	0.01394 rad/s	0.02436 rad/s	<0.0001	-0.965
Gyro Z-axis	0.01141 rad/s	0.01844 rad/s	0.0003	-0.780

3.4.3 Windowing: Quantifying Fetal Movement Duration

FM events exhibited a wide physiological range (0.5s to 95s) with a significant right-tail skew in duration. The 80th and 95th percentiles were 4s and 12s, respectively (Figure 19). Prior studies identified FMs between 1s and 96s, with most occurring between 1s and 4s [13], though longer FMs are more likely to be perceived [4]. In this dataset, strong FMs had a 95th percentile of 17s, which is significantly longer than the 95th percentile of 7s for weak FMs. Based on these distributions, a window size ≤ 12s captures 95% of the observed FM events. Though a universal standard remains elusive, choosing a window size within this range necessitates a tradeoff: smaller windows (e.g., ≤ 4s) offer high granularity but may over-segment events and miss broader temporal trends [14], while larger windows (e.g., approaching 12s) capture overall patterns but risk summarizing out finer variations [10]. The choice is further influenced by the desired level of temporal information, computational constraints, and the architecture of the data-driven detection model. Reported window sizes for FM detection algorithms range from 0.5s to

avoid averaging out features within the window [10] to 8s to balance the tradeoffs between time and frequency resolution [14]. Despite the lack of a standardized window size, evaluating the duration of FM for a specific population provides critical, data-driven insight for informing algorithm design across multiple stages: (1) dynamic windowing functions, (2) postprocessing of model predictions, (3) evaluation of algorithm performance, and (4) hardware sampling rate. The final choice of window size depends on the objective of the study and what is known about the physiological signal used to capture events.

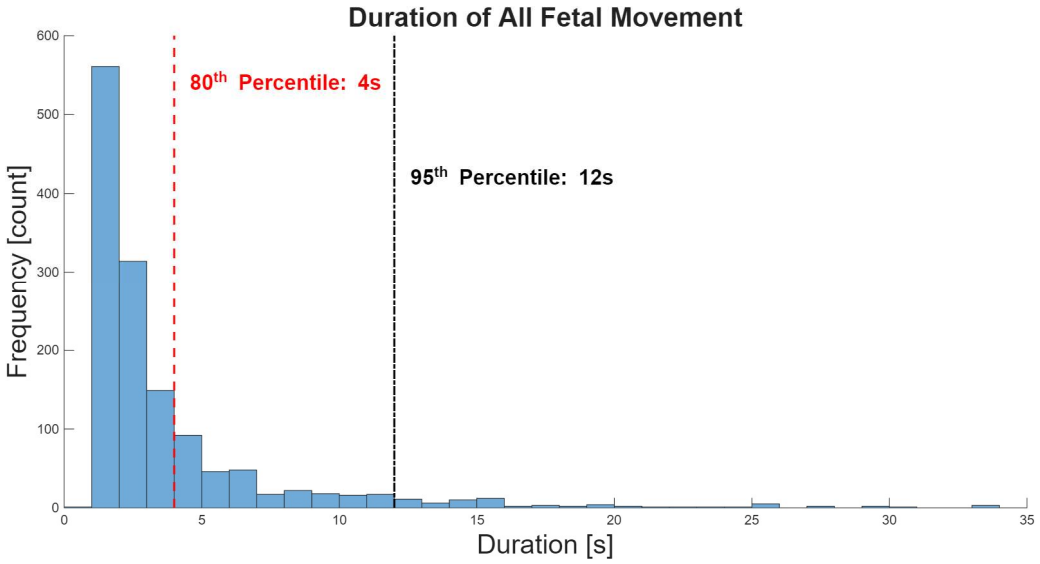


Figure 19: Histogram of FM duration used to inform window size

3.4.4 Peak Frequency Analysis

Though a magnitude threshold can be used to eliminate IMU signal noise, artifacts like maternal laughter and coughing can resemble FM in magnitude [9]. Frequency filter design can be considered for further noise mitigation for FM event detection. LMEMs applied to the peak frequency demonstrated a significant fixed effect of group (p<0.0001) across all three axes of acceleration and three axes of angular rate. The spectral energy of strong FM was concentrated between 4 and 10 Hz with an upper limit of 23 Hz. Artifacts showed spectral energy between 1.5 and 7 Hz with an upper limit of 15 Hz (Figure 20).

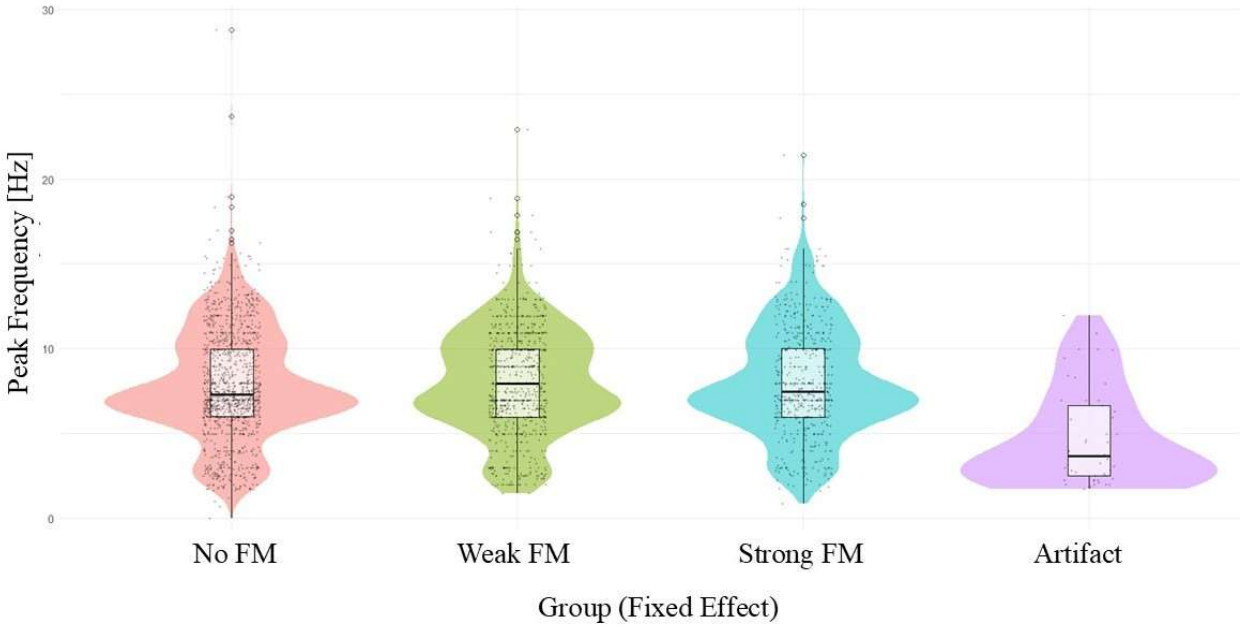


Figure 20: Peak frequency feature (y-axis of acceleration data) used to define filter cutoffs

3.4.4.1 Filter Cutoffs: Defining Noise Limits

Pairwise comparison of artifact and strong FM peak frequencies showed statistically significant separation for all three axes of acceleration and three axes of angular rate (Table 6). Maternal respiration is typically below 1 Hz [6], [15], and maternal movement has been identified between 1 and 10 Hz [15]. Other artifacts like maternal heartbeat have been reported between 1 and 4 Hz. Based on this analysis of peak frequency, a lower cutoff frequency of 4.0 Hz may remove most of the artifact energy that lies between 1.5 Hz and 4.0 Hz. An upper cutoff frequency of 24 Hz would retain the entire spectral content of strong FM (up to 23 Hz) while filtering out high frequency noise. These cutoff frequencies result in a final, recommended bandpass filter of 4.0–24 Hz.

Table 6: Pairwise comparisons between strong FM and artifact groups for peak frequency feature (p-values Tukey adjusted, α=0.05)

Signal Comparison	**EMM (Strong FM)**	**EMM (Artifact)**	***p*-value**	**Cohen's d**
Acc X-axis	7.89 Hz	6.84 Hz	0.0001	0.567
Acc Y-axis	7.90 Hz	5.17 Hz	<0.0001	1.08
Acc Z-axis	8.62 Hz	5.85 Hz	<0.0001	0.695
Gyro X-axis Gyro Y-axis	8.02 Hz	4.83 Hz	<0.0001	1.06
Gyro Z-axis	7.11 Hz	3.76 Hz	<0.0001	0.949

	7.31 Hz	4.29 Hz	<0.001	0.953

3.5 Conclusion

The lack of standardized, data-driven methodology for FM detection remains a barrier to real-world deployment, mirroring historical challenges in mature fields like ECG. This study partially addressed that gap by employing a rigorous methodology to guide objective, quantitative design choices for IMU-based FM detection under stationary conditions: defining the IMU sensitivity floor significantly lower than 0.0246 m/s^2 (or 0.00821 rad/s) to reliably capture weak FMs, setting a magnitude threshold of 0.061 m/s^2 (or 0.021 rad/s) for effective artifact rejection, defining a 4.0-24.0 Hz bandpass filter for noise mitigation, and setting a window size ≤12s to guide algorithm choices. These signal characterizations provide data-driven reference points for future design choices like event-triggered data acquisition, dynamic sensor activation, and adaptive windowing strategies based on detected movement modes. Future work with datasets incorporating maternal movement is required to accurately characterize FM signals for a deployable FM monitoring system. Ultimately, this initial analysis of IMU signal characterization is the first step toward consensus methods for rigorous comparison and validation of emerging technologies, enabling the clinical trust and technical reliability required to transition fetal monitoring from intermittent snapshots to a continuous, objective standard of care.

3.6 Acknowledgements

This research was funded by the Michigan Institute for Data and AI in Society (MIDAS) through the Propelling Original Data Science program. Statistical advice was provided by Abner Heredia Bustos through the University of Michigan Consulting for Statistics, Computing and Analytics Research (CSCAR) program.

3.7 References

[1] E. C. W. Gregory, C. P. Valenzuela, and D. L. Hoyert, “National Vital Statistics Reports,” National Vital Statistics Reports, vol. 71, no. 4, 2022.

[2] J. Gardosi, V. Madurasinghe, M. Williams, A. Malik, and A. Francis, “Maternal and fetal risk factors for stillbirth: Population based study,” BMJ (Online), vol. 346, no. 7893, pp. 1–14, 2013.

[3] J. R. Walton and A. M. Peaceman, “Identification, assessment and management of fetal compromise,” Clin. Perinatol., vol. 39, no. 4, pp. 753–768, 2012.

[4] Z. R. Hijazi, S. E. Callan, and C. E. East, “Maternal perception of foetal movement compared with movement detected by real-time ultrasound: An exploratory study,” Australian and New Zealand Journal of Obstetrics and Gynaecology, vol. 50, no. 2, pp. 144–147, Apr. 2010.

[5] E. Rodriguez-Villegas, S. Iranmanesh, and S. A. Imtiaz, “Wearable Medical Devices: High-Level System Design Considerations and Tradeoffs,” IEEE Solid-State Circuits Magazine, vol. 10, no. 4, pp. 43–52, Sep. 2018.

[6] M. Mesbah et al., “Automatic fetal movement recognition from multi-channel accelerometry data,” Comput. Methods Programs Biomed., vol. 210, 2021.

[7] X. Zhao, X. Zeng, L. Koehl, G. Tartare, and J. De Jonckheere, “A Wearable System for In-Home and Long-Term Assessment of Fetal Movement,” IRBM, no. 41, pp. 205–211, 2020.

[8] Y. Ansari, O. Mourad, K. Qaraqe, and E. Serpedin, “Deep learning for ECG Arrhythmia detection and classification: an overview of progress for period 2017– 2023,” 2023, Frontiers Media SA.

[9] B. Boashash, M. S. Khlif, T. Ben-Jabeur, C. E. East, and P. B. Colditz, “Passive detection of accelerometerrecorded fetal movements using a time-frequency signal processing approach,” Digital Signal Processing: A Review Journal, vol. 25, no. 1, pp. 134–155, 2014.

[10] M. Altini et al., “Detection of Fetal Kicks Using BodyWorn Accelerometers During Pregnancy: Trade-offs Between Sensors Number and Positioning,” IEEE EMBC, 2016.

[11] X. Robert-Lachaine, H. Mecheri, C. Larue, and A. Plamondon, “Accuracy and repeatability of single-pose calibration of inertial measurement units for whole-body motion analysis,” Gait Posture, vol. 54, pp. 80–86, May 2017.

[12] M. Mesbah, M. S. Khlif, C. East, J. Smeathers, P. Colditz, and B. Boashash, “Accelerometer-based fetal movement detection,” Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, pp. 7877–7880, 2011.

[13] J. ten Hof et al., “Quantitative analysis of fetal general movements: methodological considerations,” Early Hum. Dev., vol. 56, pp. 57–73, 1999, [Online].

[14] P. Alwis et al., “Application of an LSTM-Based Channel Attention and Classification Mechanism in Fetal Movement Monitoring,” Sep. 10, 2024.

[15] A. K. Ghosh, D. S. Catelli, S. Wilson, N. C. Nowlan, and R. Vaidyanathan, “Multi-modal detection of fetal movements using a wearable monitor,” Information Fusion, vol. 103, Mar. 2024.

Chapter 4 Methodological Evaluation of Label-Source Selection and Evaluation Strategies on Reported Performance in Wearable Fetal Movement Detection

4.1 Abstract

Continuous fetal movement (FM) monitoring has the potential to transform prenatal care through the identification of at-risk fetuses. To address the lack of standardization in existing literature, this study compared the performance of maternal perception and ultrasound labels for training machine learning models that classify FM events from wearable inertial measurement unit (IMU) signals under realistic evaluation strategies with respect to ultrasound labels. Data were collected from eighteen pregnant participants [mean gestational age: 27 weeks (SD: 2.6); mean BMI: 29.3 (SD: 4.9)] who wore four abdominal IMUs and one chest reference sensor for 10-15 minute stationary, data collection trials. Maternal perception annotations were collected using a handheld push button, and ultrasound recordings were captured and reviewed by a board-certified OB/GYN to identify FM events. Random forest machine learning classifiers were trained using time-domain statistical features extracted from the multi-node IMU array labeled with two clinical standard annotations: maternal perception and ultrasound. Model training was conducted such that (1) every participant had at least one trial in the training and testing sets (personalized) and (2) a leave-one-subject-out approach (LOSO). Model training was also conducted with random sampling for comparison to prior work. Under the personalized approach, models trained with maternal perception labels achieved an area under the receiving operating characteristic (AUROC) curve of 0.61 (95% CI 0.57-0.65) compared to models trained with ultrasound labels, which achieved an AUROC of 0.65 (95% CI 0.61-0.69). At a false positive rate of 5%, both models achieved similar sensitivity to FM events (24-25%). At a sensitivity of 50%, models trained with ultrasound labels demonstrated greater specificity (72% compared to 65%). While the personalized approach showed slight improvements over LOSO, random sampling led to inflated performance. These results demonstrate that while maternal perception is a scalable training source, it introduces label noise that constrains model reliability. Furthermore, the significant performance disparity between random sampling and participant-independent splitting underscores the necessity of rigorous validation to avoid false reassurance

in clinical deployment. Though this wearable system shows promise for FM detection, the limited sensitivity to meaningfully reduce false alarms limits deployment for clinical reassurance. This work establishes a necessary methodological framework for benchmarking the true capabilities and limitations of wearable fetal monitoring systems.

4.2 Introduction

During pregnancy, intermittent fetal monitoring is enacted to evaluate fetal health and prevent fetal death through early identification of an at-risk fetus that may benefit from clinical intervention [1]. In the United States, there are nearly 20,000 reported stillbirths (defined as fetal death after 28 weeks of gestation) each year, yet many of these fetal deaths have an unknown cause [2]. Other areas of medicine have benefitted from advancements in technology and shifts in population health, but prenatal care continues to experience an unchanging and stagnant stillbirth rate [2]. The presence of fetal movement is generally accepted as evidence of a healthy fetus, indicating proper oxygenation to support the fetal nervous system [1]. Fetal movement develops around the 9th week of gestation and can be perceived by the mother by approximately the 16th week [3]. Reductions in fetal movement are associated with negative outcomes at birth [4,5], with the likelihood of stillbirth increasing five-fold for those who have experienced reduced fetal movement [6]. Reduced fetal movement, therefore, serves as a key warning sign that intervention may be necessary [4,7,8].

Maternal perception of fetal movement, also known as “kick counting”, is often described as the first line of defense for identification of an at-risk fetus [1]. The two main methods of kick counting are the Sadovsky method, where mothers count the number of perceived fetal movements in a specified amount of time [9], and the Cardiff Count to Ten method, which relies on the amount of time it takes for mothers to perceive ten movements [10]. Definitions of what qualifies as “reduced fetal movement” vary widely [6-8], and a mother’s ability to perceive movement can be limited by factors such as body mass index (BMI), placental location, and even the time of day [3,11-13]; mothers perceive 33-88% of fetal movement that actually occurs [7,14]. When reductions in fetal movement are perceived, clinicians initiate a non-stress test (NST) where fetal heart rate and uterine activity are monitored for 20 minutes for fetal heart rate accelerations or decelerations with or without contractions [1,15]. The NST can be complemented by the biophysical profile (BPP), which measures fetal breathing, movement, tone, and amniotic fluid volume via ultrasound [1]. Clinical assessment is the best existing option

for objective prenatal care but these tests require trained clinicians and clinical settings, provide only a "snapshot in time" of the fetal status, and can prompt cautious but unnecessary intervention including early delivery, hospitalization, or more frequent monitoring [1,8,11,15,16].

Wearable devices offer an alternative to traditional long-term monitoring that requires repeated or extended clinical visits that disrupt daily life and create substantial financial burden [17]. In the context of fetal monitoring, accelerometers, acoustic sensors, and piezoelectric sensors have shown promise for the continuous detection of fetal movement [18-20]. Though an optimal sensor setup has yet to be agreed upon, pairing wearable sensors with machine learning models has shown encouraging prior results under varying evaluation conditions for detecting fetal movement. For example, Xu et al. [21] reported sensitivity and precision scores greater than 80% using two accelerometers paired with an extra trees classifier.

However, a lack of standardization across the literature creates a significant challenge when evaluating methodological choices across existing studies. Each existing study has used a unique combination of study cohort, sensor setup, ML classifier, event label source, class distinctions, window segmentation size, class balance, and train/test split. For example, Delay et al. [22] was the only study to use raw time-series data as the input to a deep learning model. Others relied on hand-engineered features. Xu et al. [21] and Mesbah et al. [23] used ultrasound annotations to label training data and report performance, while Zhao et al. [24] and Altini et al. [18] used maternal perception. Even the way fetal movement was defined varied between studies. Delay et al. [22] only considered limb movements, while Xu et al. [21] and Mesbah et al. [23] included all movements. Additionally, the standards for reporting performance metrics were inconsistent, with variable class balancing strategies and a lack of reported threshold values. Because of these methodological variations, it is difficult to benchmark and evaluate the benefits of any single approach because performance may be as much a function of experimental design as sensor efficacy. This lack of clarity necessitates studies that control for these variables to understand the underlying drivers of model performance. Therefore, the goal of this study is to use controlled benchmarking on a common test set to evaluate two methodological choices that drive performance at a high level during model training: choice of clinical standard used to label training data and data splitting approach. By holding other methodological variables constant,

these specific variables were isolated to determine their individual impact on model performance.

4.3 Background

The research landscape for fetal movement detection remains difficult to navigate due to wide variations in experimental design. While a range of sensor types have been tested, accelerometers are most commonly used, as they have shown high performance metrics when distinguishing fetal movements [21,23,24]. Multimodal systems that pair accelerometers with other sensor types have also been appropriately benchmarked to demonstrate increased performance over single-mode sensor systems [20,25]. Additionally, feature-based model architectures are most commonly used, likely due to their simplicity and stability with sample sizes reflective of early-stage proof-of-concept studies. Though deep learning architectures show promise for capturing complex, transient FM characteristics [22] (Delay et al.), these models require significantly larger sample sizes that are more reflective of population-level studies intended for system deployment in later development stages [25]. Though a standardized sensor setup and optimal architecture are not yet defined, existing literature generally supports the use of IMU-based systems paired with feature-based classifiers for evaluative studies.

However, two variables driving model performance still remain unclear: the clinical standard used to label training data and the strategies used for data splitting. When training and evaluating model performance, event definition has a significant impact on the difficulty of the classification task and the interpretation of the model outputs. While some FM detection models have been developed using maternal perception as the clinical standard [18,24,26], others have relied on ultrasound [19,21,23]. Researchers have used these label sources rather interchangeably despite distinct tradeoffs for real-world deployment: maternal perception can be cheaply and widely collected to offer large, supervised datasets for model training, while ultrasound labeling is resource-intensive, requiring clinical resources and extensive clinician time for annotation, resulting in much smaller and more expensive datasets. For training and evaluating models, ultrasound-defined events offer greater clinical precision, while maternal perception acts as a more practical but potentially noisy alternative.

The label source used for defining FM events is further complicated by varying definitions of what constitutes a “movement” and how it is annotated. For example, Mesbah et al. [23] achieved 92% specificity and 95% sensitivity using ultrasound-labeled events, while

Zhao et al. [24] achieved 99% specificity and 77% sensitivity using maternal perception-labeled events for model training and evaluation. Though these performance metrics appear to be similar, they are highly dependent on how a FM event is defined by the label-source. While Mesbah et al. [23] achieved high performance for classifying precise, clinically meaningful FM, Zhao et al. [24] achieved high performance for classifying the FMs mothers are already able to perceive. Adding label buffers to account for human reaction delays [19,21,25] and excluding certain types of fetal movement [22] further complicates the interpretation of model performance. The subjectivity and ambiguity of using maternal perception to train and evaluate ML models has been cited as a limitation [24], but no studies have systematically compared maternal perception and ultrasound as training label sources for fetal movement classification using ultrasound during evaluation.

When benchmark datasets are not available, as is the case with FM, isolated testing sets within individual studies can be used to compare methods [27,28]. However, how these testing sets are defined highlights another significant discrepancy in the field. Most existing studies have randomly sampled windows of data into training and testing sets, with variable ratios and data balancing strategies [21-24]. While random sampling is common in machine learning, it fails to account for data correlation within trials or subjects. This approach does not reflect realistic deployment scenarios and can inflate performance estimates by allowing highly correlated data from the same participant or trial to appear in both training and testing sets, thereby limiting generalizability of the model. To date, only one study used a leave-one-subject-out (LOSO) approach [18] to mitigate this issue. Accounting for subject-specific signal characteristics through data splitting approaches with LOSO promotes participant-level generalization, deployment realism, and prevention of data leakage. Without generalizable data splitting protocols, the clinical benefits of different sensor setups or classifiers remain difficult to evaluate fairly.

Therefore, the goal of this study is to systematically evaluate these methodological tradeoffs by comparing maternal perception and ultrasound as clinical standards for model training under generalizable data splitting approaches. Using a consistent inertial measurement unit (IMUs) sensor setup and random forest classifier, models were trained on subjective, easily-scalable maternal perception labels to investigate their ability to predict objective, ultrasound-verified events in realistic deployment scenarios.

4.4 Methods

There are two clinical standards for capturing fetal movement: maternal perception and ultrasound observation. Models for detecting fetal movement events can be developed and evaluated with maternal perception and ultrasound observation annotations to ensure clinical reliability upon deployment. The methodology for this study consists of four key components: data collection, signal processing and event annotation, splitting of training and testing sets, and fetal movement classification and evaluation. Distinct models were developed using both maternal perception and ultrasound event annotations and their performance was compared against the clinical standard ultrasound during the evaluation phase. This comparative analysis aimed to evaluate the reliability of maternal perception as a clinical standard for model development and discuss the tradeoffs of the proposed methodology for classification of fetal movement events.

4.4.1 Data Collection

Eighteen participants between 18 and 49 years old were recruited for a single data collection session at the University of Michigan. All participants were between 24 and 32 weeks gestational age and carrying a singleton pregnancy with no known fetal disease or abnormalities. Participants with a diagnosis of gestational diabetes or hypertension were excluded. All participants provided written informed consent. The study was reviewed by the University of Michigan Institutional Review Board (HUM00204999). This population sample size is appropriate for an investigation focused on early-stage methodological validation. The study aims to evaluate the consistency of the analytical pipeline and the reliability of subjective maternal labels under realistic data-splitting constraints, rather than seeking to provide a deployment-ready population estimate.

Data were collected with a system of wearable IMUs (Opal, APDM Inc., Portland, OR, USA). Four IMUs were attached to the surface of the abdomen to capture fetal movement and one additional IMU was attached to the chest to capture maternal movement, since the chest would be isolated from abdominal deformation caused by fetal movement. All IMUs were attached with medical-grade adhesive tape. The abdominal IMUs were centered around the umbilicus, with one IMU consistently placed in each quadrant to reliably capture fetal movement independent of fetal position (see Fig. 21 for IMU placement). The chest IMU was centered on the top section of the participant's sternum (e.g., manubrium) to reliably capture maternal heart

rate, respiration, and trunk movement while avoiding rotational noise. All IMUs were oriented such that the x-axis was aligned with gravity and the z-axis was perpendicular to the abdominal wall (or chest surface). Intentional sensor placement ensured proper and consistent measurement of angular rate signals. All IMUs collected synchronized tri-axial acceleration and angular rate at a sampling rate of 128 Hz.

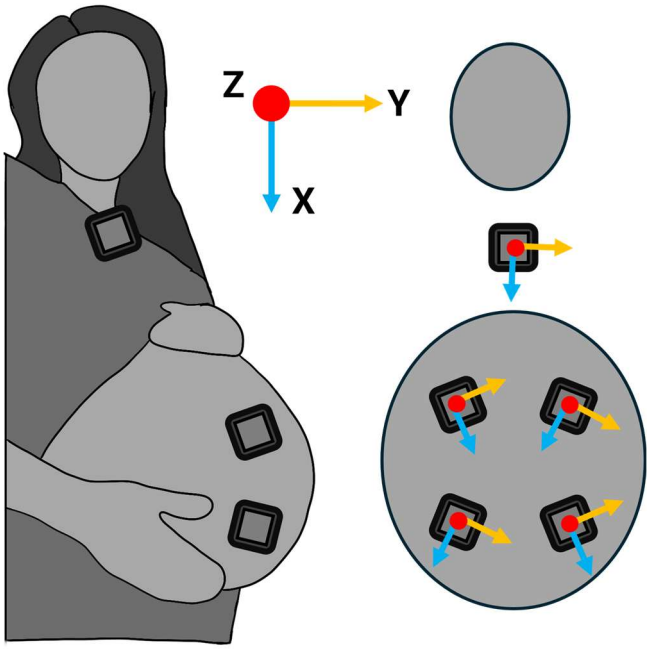


Figure 21: Placement and orientation of abdominal IMUs used to capture abdominal deformation caused by fetal movement

During data collection sessions, the height and weight of each participant was recorded. These measurements were used to calculate body mass index (BMI), which can indicate the level of adipose tissue between the fetus and outer abdominal wall and may affect the degree to which the abdomen deforms during fetal movement and the sensors' ability to capture fetal movement signals [29]. Then, the array of IMUs were fitted to the participant according to the placement strategy described above. Prior to data collection, participants were guided through a series of calibration movements, which were later utilized to estimate the orientation of the IMU sensors and align them within the same frame of reference. While most prior studies have relied on acceleration signals tangential to the abdomen, angular rate data may help isolate fetal movement signals in the presence of maternal motion, as fetal movement can be localized to a single sensor within the array, while maternal movement is global to the participant's body system [25]. A

functional alignment method was employed to determine the participant's anatomical axes relative to the IMU fixed frame of reference [30]. Because the sensors are placed on the abdomen, closest to the hip axis, and maternal motion can be simplified to rigid body motion with rotation about the hip axis, the hip axis was used as the reference axis to estimate orientation. At the start of data collection, participants were instructed to begin from a static standing position followed by a series of three rotations about the hip, leaning forward to a comfortable degree before returning to static standing. This motion was repeated three times, and the calibration data from these hip-hinging movements were used to employ the Functional Alignment Method and align the IMU axes of the sensor array. Details related to signal processing of the calibration data are located below in Section 4.4.2.

After performing the calibration movements, participants were seated in a chair and asked to remain upright with both feet flat on the floor. Participants were instructed to remain as stationary as possible throughout the duration of approximately 10-minute data collection trials. During data collection, participants were given an additional handheld IMU to indicate perceived fetal movement. The handheld IMU featured a trigger button that participants were instructed to press at the onset and continuously throughout perceived fetal movement. The annotations from the maternal perception trigger button were time-synchronized with the IMU array.

To capture fetal movement via ultrasound observation, an ultrasound scan was administered by a board-certified OB/GYN throughout the duration of each trial. Before the start of data collection, a clear visual of the fetus was obtained, prioritizing the trunk and/or limbs. If the visual was obscured during the trial, the ultrasound probe was repositioned. The space between the top two IMUs or the bottom two IMUs was prioritized to avoid interference with the sensor array and optimize the visual plane. Movement of the ultrasound probe was limited to the best possible extent to avoid introducing additional environmental noise to the sensor system. Recordings of the ultrasound scan were collected when fetal movement was observed. Data from participants was recorded for one to three trials, depending on the time available within the single data collection session.

4.4.2 Signal Processing & Event Annotation

After data collection, the synchronized files from the IMU array were downloaded and reviewed for signal quality. The ultrasound recordings were post-processed and time-synchronized with the data from the IMU array. Data from two participants were excluded due to

signal acquisition issues upon review, one related to the sensor array and one related to the ultrasound recordings. Data from 16 participants were used for analysis. The ultrasound recordings were reviewed by the OB/GYN to provide refined annotations. Both maternal perception and ultrasound observation annotations were post-processed and used for labeling the IMU data.

Raw acceleration and angular rate data from each IMU were processed through a zero-phase bandpass filter (4.0-24 Hz) to reduce signal drift and artifacts caused by maternal respiration at low frequency ranges and noise from environmental and physiological sources at high frequency ranges. The ultrasound annotations were analyzed in a prior study [31] to define the frequency range of fetal movement between 4.0-24 Hz.

After filtering, a functional alignment method was applied to the angular rate data to estimate each sensor's orientation and align the array into a local reference frame [30]. Axis alignment was not performed on the acceleration data to preserve tangential fetal movement information on the z-axis, which is critical for capturing small, localized changes in abdominal acceleration that may otherwise be diluted between axes through axis realignment. First, the global gravity vector was defined by the x-axis of the chest IMU during the static standing portion of the calibration movement to establish a reference for alignment of the vertical axis. Each sensor's x-axis was aligned with global gravity. Then, rotation of the chest IMU about the frontal axis during the hip-hinging movements was used to identify the frontal and sagittal axes. Each sensor's y-axis was aligned with the frontal axis, and the z-axis was aligned with the sagittal axis (Fig. 22). Axis alignment was applied independently for each trial. Rotational motion caused by fetal movement should then be localized to a single sensor within the array, while rotational motion caused by the maternal body system should be correlated across the array. In effect, the comparison of rotational motion signals between nodes of the IMU array should support differentiation between fetal and maternal movement.

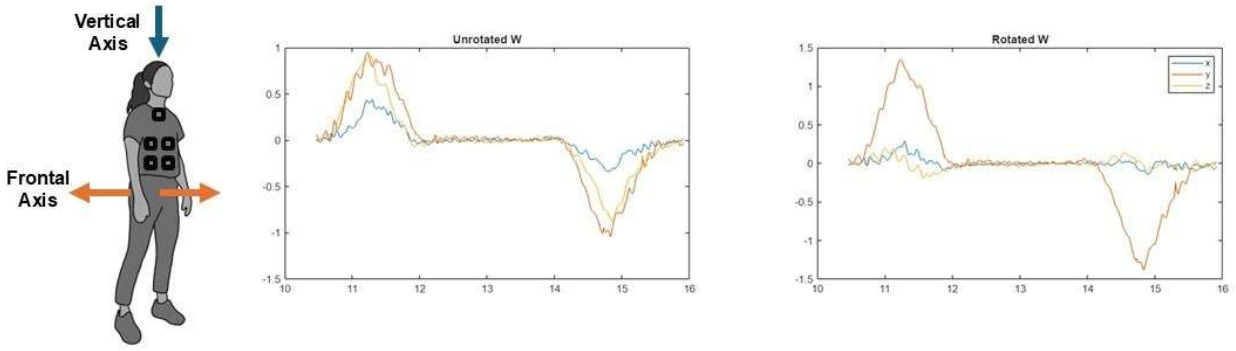


Figure 22: Application of the functional alignment method to angular rate data

After filtering and alignment, all data from each trial was segmented into 0.5s non-overlapping windows. The short duration of this window size is matched to the minimum resolution of the ultrasound annotations and allows the capture of transient signal variations caused by fetal movement while avoiding an averaging out of the signal in a longer time window [18,25]. Within these 0.5s windows, statistical features were calculated including the mean, standard deviation, and range of the signal in addition to the correlation between axes of each sensor and the correlation between each sensor and the reference chest sensor [18,25]. These features were calculated across tri-axial acceleration and angular rate signals for the four abdominal IMUs and the chest IMU resulting in a total of 150 features per window.

Each window was given a binary label of 1 or 0 for the presence or absence of fetal movement, respectively. Maternal perception annotations were dilated 3-seconds prior to and 1-second after the trigger was pressed to account for reaction delay and duration of the fetal movement event. This 4-second window was chosen based on a prior analysis of FM duration in this dataset, where the 80th percentile of FM duration was established as 4 seconds [31]. Mothers were instructed to continue pressing the trigger to indicate perceived fetal movement throughout its duration, so consecutive and overlapping maternal perception annotations were counted as one fetal movement event. The ultrasound observations were annotated with precise timestamps for the start and end of each fetal movement event and were applied directly to the dataset. After definition of FM events, the agreement between maternal perception- and ultrasound-defined events was calculated at the event-level for comparison to prior studies and to provide context for the study results. Ultrasound-defined events were considered the clinical standard for this comparison. Maternal perception-defined events that overlapped with ultrasound-defined events were considered true positives. All other ultrasound-defined events were considered false

negatives. Maternal perception-defined events that did not overlap with any ultrasound-defined events were considered false positives. Though maternal perception of fetal movement is subjective, it is important to note that false positive events may indicate that fetal movement occurred outside of the plane of view of the ultrasound probe. Nonetheless, we use ultrasound-defined events as the clinical standard.

4.4.3 Splitting of Training and Testing Sets

Only one study has conducted model training and evaluation using a leave-one-subject-out (LOSO) approach. All other studies have randomly sampled windows from the full dataset for training and testing sets, which represents an unrealistic setting: generally, data from a given trial (or even an event) is not available to predict other portions of that same trial for the same participant. This approach also reduces the difficulty of the prediction task for the ML model, since learned patterns can be extrapolated within the trial itself, rather than generalized across trials or across subjects. Some studies applied signal processing techniques to the dataset before splitting, resulting in data leakage. Randomly sampling windows does not account for inter-subject or inter-trial dependencies that may exist and splitting the data this way may result in an overly positive representation of model capability. Therefore, model training and evaluation for this study followed three different approaches to the train/test split: (A) splitting the dataset such that every participant has at least one trial in the training and testing sets, (B) LOSO, and © random sampling. It is critical to consider how the data are split during model development to ensure real-world conditions are being met.

Approach A: In the case that the dataset is split such that every participant has at least one trial in the training and testing sets (but the data from those trials is contained before splitting), it represents a situation, for example, that a pregnant participant completes an initial 30–60 min monitoring session at home or in the clinic. After that, the system can monitor fetal movement in future sessions without re-training. To employ this training and testing split, each participant was required to have at least two data trials. Thirteen of the 16 participants met this criterion and were included in model training and testing. Five of the 16 participants had data for three trials. The additional trial from these five participants was held out during training in a validation set for hyperparameter tuning. One trial from each participant was randomly selected for training, one trial for testing, and any additional trials were included in the validation set.

Approach B: In the case that a LOSO splitting approach is employed, it represents a scenario, for instance, where a new participant comes to the clinic, puts on sensors, and the system immediately estimates fetal movement without any prior calibration. There was no minimum data requirement to employ this approach, so all 16 participants were included in model training and testing. All trials from one participant were left out for testing and all remaining trials were included in the training set. This process was repeated for all participants. To avoid generalization to any specific subject during training, the best performing hyperparameters from the previous models were used for the LOSO model architecture.

Approach C: For comparison with prior work in the literature, models were also run with a random 8:2 train/test split across all time series. Using this approach, portions of the same trial from the same participant may be part of both the training and testing sets. As mentioned earlier, this scenario is unrealistic because data from a given trial is generally not available for both training and testing.

The train/test splits employed in this study are summarized in Fig. 23.

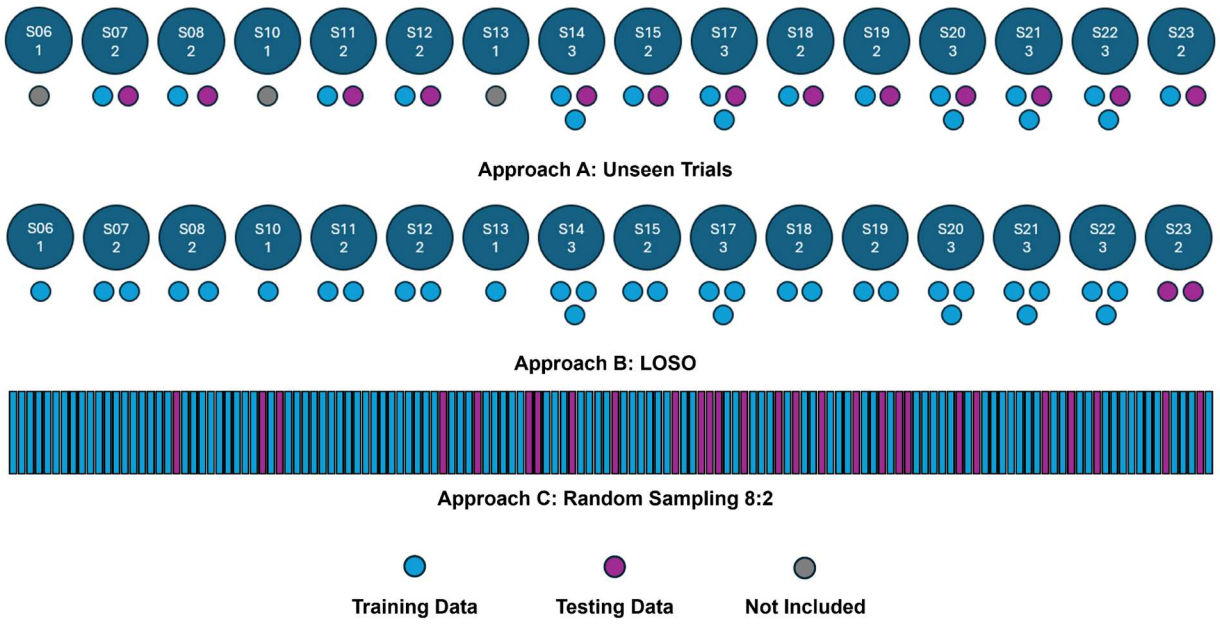


Figure 23: Train/test splits employed for model training and evaluation. Approach A depicts model predictions on unseen trials, where every participant has at least one trial in the training and testing sets. Approach B depicts model predictions on unseen subjects. Approach C depicts random sampling of windows.

4.4.4 Fetal Movement Classification and Evaluation

The 150 features calculated within each window (Section 4.4.2) were used to train random forest (RF) models for binary classification. RF was selected over more complex deep

learning architectures given the dataset size. The model architecture consisted of an RF classifier with hyperparameters optimized via a random search (using RandomizedSearchCV) based on the F1 score evaluated on the validation set. Because the task was binary classification, entropy was used as the criterion for measuring the quality of the split. Additionally, the class weights were balanced inversely proportional to the class frequencies in the input data to account for imbalances between the minority and majority classes within the training set. The final output was the probability score determined by aggregating the predictions of all trees using majority voting. Two models were independently trained: one with maternal perception labels and one with ultrasound labels. Predictions were made on held-out test data.

Because fetal movement events vary in duration and are the clinically relevant unit of analysis, performance was calculated at the event-level, rather than on a window-by-window basis. Using ultrasound-defined fetal movement events as the clinical standard, the maximum probability score predicted by the model within each event was used to evaluate the performance of the model. The maximum probability of the rest periods, defined between any two ultrasound-defined fetal movement events, was also defined to calculate performance of the models (Fig. 24). These probability scores were used to plot receiver operating characteristic (ROC) curves, and the area under the ROC curves (AUROC) was calculated. Model evaluation was performed based on mean AUROC scores on held-out test data. Bootstrapped 95% confidence intervals (CI) were also reported. Performance between models trained with maternal perception labels and ultrasound labels was evaluated by estimating the probability of the null hypothesis (e.g., equal AUROC performance between models) using non-symmetric bootstrap p-values [32].

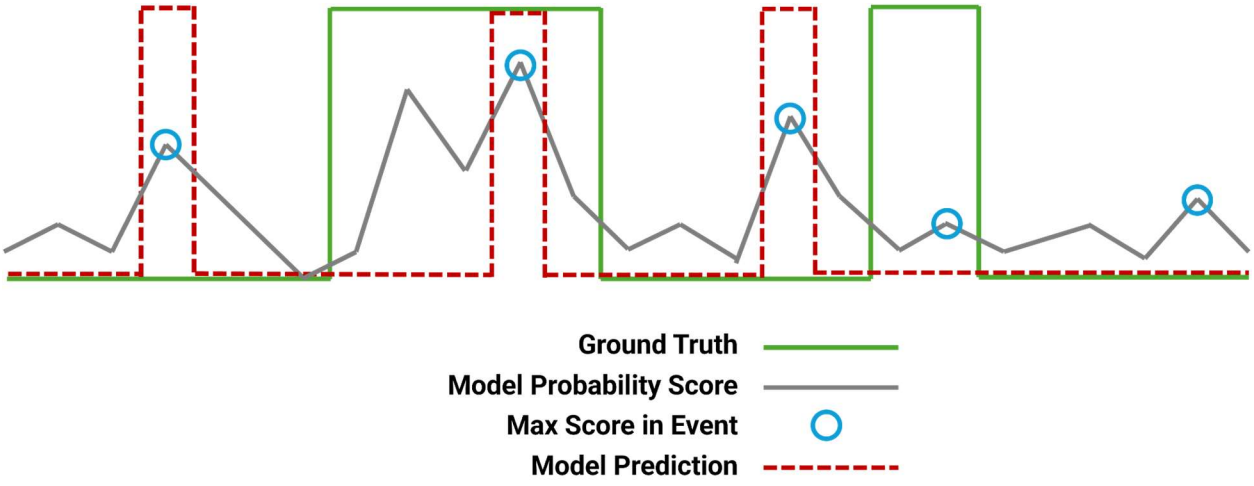


Figure 24: Strategy used to extract probability scores from model predictions at the event-level

The probability scores within each event were then used to classify predictions, based on a chosen threshold, into {0,1} to evaluate the accuracy, specificity, sensitivity, and precision of each model. First, performance metrics were calculated when models were evaluated at a constant specificity of 0.95 (e.g., 5% false positive rate). Then, performance metrics were calculated when models were evaluated at a constant sensitivity of 0.50 (e.g., correctly predicting 50% of the fetal movement events in the held-out test data).

Due to the randomization of selected samples for the testing set resulting from the random 8:2 train/test split, it is not feasible or informative to calculate event-level performance. The performance metrics reported for the models developed with this approach were calculated on a window-by-window basis.

4.5 Results

The results presented in this study demonstrate that maternal perception is a less reliable label source for training machine learning classifiers to detect ultrasound-defined fetal movement events under real-world conditions. By comparing RF classifiers trained with maternal perception and ultrasound annotations to detect fetal movement events, evaluation against the ultrasound clinical standard resulted in higher performance for models trained with ultrasound labels in terms of AUROC and specificity at a constant sensitivity level. Both models achieved similar sensitivity at a false positive rate of 5%. Additionally, the results of this study highlight the overly optimistic performance metrics reported in prior work and the importance of considering real-world deployment environments during model development.

4.5.1 Dataset Contextualization

Between one and three trials were collected for each participant for a total of 34 trials in the dataset. Each trial was between 7.5 and 12 minutes in duration, resulting in a total of 354 minutes of data, or approximately 6 hours. 1077 fetal movement events were identified via ultrasound across all participants. This high density of discrete events provided the necessary statistical granularity to address the study aims of testing label reliability and data splitting approaches. 489 of those events were perceived by participants, resulting in a mean sensitivity of 32% (SD: 18%) across the dataset. Participants indicated perceived fetal movement 97 times that were not observed by ultrasound, resulting in a mean precision (positive predictive value) of 83% (SD: 19%). A participant's ability to accurately perceive fetal movement is influenced by

factors such as gestational age, body mass index (BMI), and placental position [1,14]. A participant-level analysis of the ability to perceive fetal movement compared to the ultrasound clinical standard, along with relevant participant demographics, is detailed in Table 7.

Table 7: Participant-level analysis of perceived fetal movement compared to the ultrasound clinical standard

Subject #	Trials	FMs	Non-FMs	Precision	Recall	GA	BMI	Placenta Location
1	1	26	27	1.00	0.52	31	28.8	Anterior
2	2	38	39	0.80	0.08	26	36.3	Anterior
3	2	63	63	1.00	0.12	26	25.7	Anterior
4	*Excluded due to signal acquisition issues*							
5	1	36	37	0.83	0.26	25	27.7	Anterior
6	2	16	17	0.27	0.18	31	25.0	Anterior
7	2	84	84	1.00	0.11	28	23.5	Posterior
8	1	29	30	0.48	0.22	24	27.1	Posterior
9	3	101	101	0.76	0.44	26	35.0	Anterior
10	2	82	83	0.94	0.15	28	25.0	Anterior
11	*Excluded due to signal acquisition issues*							
12	3	73	74	0.78	0.40	32	39.8	Anterior
13	2	69	70	1.00	0.19	27	30.3	Anterior
14	2	70	71	0.97	0.35	25	27.1	Anterior
15	3	91	92	0.77	0.45	25	30.5	Posterior
16	3	106	107	0.83	0.54	30	35.9	Anterior
17	3	144	144	0.91	0.70	24	24.8	Posterior
18	2	49	50	0.90	0.34	27	25.8	Anterior
TOTAL	**34**	**1077**	**1089**	**MEAN: 0.83 (SD: 0.19)**	**MEAN: 0.32 (SD: 0.18)**			

After data segmentation, 18% of the total windows in the dataset were labeled positively for fetal movement according to the maternal perception annotations. 32% of the windows were labeled positively according to the ultrasound annotations.

4.5.2 Reliability of Maternal Perception Annotations for Model Training

Following Approach A (Fig. 23) for the train/test split, where every participant has at least one trial in the training and testing sets, the model trained with ultrasound labels outperformed the model trained with maternal perception labels. Both models were evaluated

against ultrasound annotated events. The model trained with ultrasound labels achieved an AUROC of 0.65 (95% CI 0.61-0.69), while the model trained with maternal perception labels achieved an AUROC of 0.61 (95% CI 0.57-0.65) (Fig. 25).

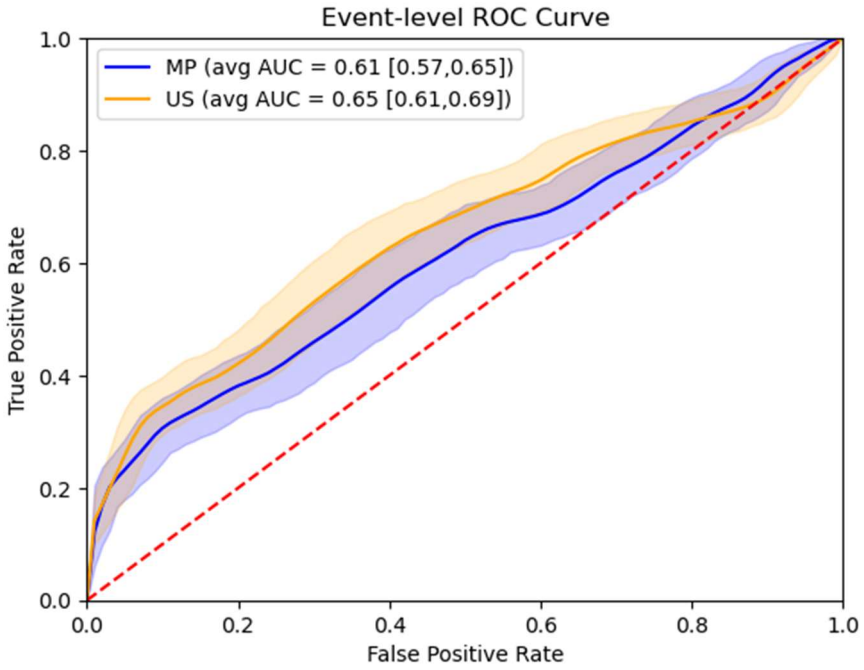


Figure 25: Receiver-operating characteristic curves associated with each model following a train/test split where every participant has at least one trial in the training and testing sets. The blue curve represents predictions resulting from the model trained with maternal perception (MP) labels, while the orange curve represents predictions resulting from the model trained with ultrasound (US) labels. Bootstrapped 95% confidence intervals are illustrated to show performance separation.

At a false positive rate of 5%, the model trained with ultrasound labels achieved a sensitivity of 25% compared to 24% for the maternal perception trained model. Finally, when capturing 50% of the fetal movement events in the test set, the ultrasound trained model achieved a specificity of 72% compared to 65% for the maternal perception trained model. Ultimately, models trained with ultrasound labels are better able to distinguish and correctly identify fetal movement events compared to models trained with maternal perception labels. However, both models have low sensitivity to fetal movement events. Additional performance metrics are summarized in Table 8.

Table 8: Event-level performance, unseen trials

Training Labels	Testing Labels	Threshold	ACC	SPEC	SEN	PPV	F1	AUROC

US	US	0.76	0.60	**0.95**	0.25	0.83	0.39	0.65 (0.61-0.69)
MP	US	0.61	0.59	**0.95**	0.24	0.82	0.37	0.61 (0.57-0.65)
US	US	0.635	0.61	0.72	**0.50**	0.64	0.56	0.65 (0.61-0.69)
MP	US	0.45	0.58	0.65	**0.50**	0.59	0.54	0.61 (0.57-0.65)

Following Approach B (Fig. 23) for the train/test split, where one participant was held-out in the testing set, the model trained with ultrasound labels outperformed the model trained with maternal perception labels, achieving mean AUROC scores of 0.61 (SD = 0.14) and 0.57 (SD = 0.14), respectively. These results were not significant, however, as there was wide variation in performance between subjects.

At a false positive rate of 5%, the model trained with ultrasound labels achieved a mean sensitivity of 21% (SD = 12%) compared to 17% (SD = 12%) for the maternal perception trained model. Finally, when capturing 50% of the fetal movement events in the test set, the ultrasound trained model achieved a mean specificity of 68% (SD = 17%) compared to 59% (SD = 16%) for the maternal perception trained model. Following both real-world approaches to the train/test split, models trained with ultrasound labels outperform those trained with maternal perception labels. Participant-level performance metrics are summarized in Table 9 and ranked by the number of fetal movement events present for each participant.

Table 9: Event-level performance, unseen subjects/LOSO

# of FMs (ranked)	**Subject #**	**AUROC**		**Sensitivity at <5% FPR**		**Specificity when capture >50% of FM events**	
Training Label Source		MP	US	MP	US	MP	US
16	6	0.15	0.18	–	–	–	–
26	1	0.44	0.54	–	0.12	0.44	0.41
29	8	0.49	0.51	0.03	0	0.43	0.47
36	5	0.62	0.70	0.22	0.31	0.51	0.86
38	2	0.56	0.59	0.08	0.05	0.46	0.64
49	18	0.58	0.61	0.31	0.37	0.58	0.78
63	3	0.66	0.62	0.27	0.25	0.71	0.73
69	13	0.64	0.67	0.09	0.20	0.61	0.70
70	14	0.55	0.59	0.16	0.27	0.54	0.56
73	12	0.50	0.52	0.10	0.12	0.53	0.42
82	10	0.60	0.65	0.16	0.22	0.53	0.42
84	7	0.77	0.77	0.30	0.32	0.89	0.90
91	15	0.62	0.65	0.12	0.15	0.63	0.70
101	9	0.52	0.58	0.10	0.10	0.42	0.55
106	16	0.55	0.73	0.13	0.28	0.52	0.82
144	17	0.82	0.84	0.49	0.42	0.94	0.94
–	4	*Excluded due to signal acquisition issues*					
–	11	*Excluded due to signal acquisition issues*					
MEAN		**0.57 (SD=0.14)**	**0.61 (SD=0.14)**	**0.17 (SD=0.12)**	**0.21 (SD=0.12)**	**0.59 (SD=0.16)**	**0.68 (SD=0.17)**

4.5.3 Random Sampling Inflates Performance

Models were developed according to Approach C (Fig. 23), where windows were randomly sampled following an 8:2 train/test split. Because of the random split, model performance was evaluated on a window-by-window basis, rather than at the event-level, in this approach. Due to random sampling, it was not feasible or informative to evaluate the

performance of the model trained with maternal perception labels against ultrasound-labeled windows. In this case, the model trained with maternal perception labels was evaluated against maternal perception-labeled windows as a proxy for the clinical standard. Using the same dataset, feature set, and model architecture, model performance was significantly increased when a random train/test split was employed. The maternal perception model achieved AUROC = 0.80, while the ultrasound model achieved AUROC = 0.75 (Fig. 26).

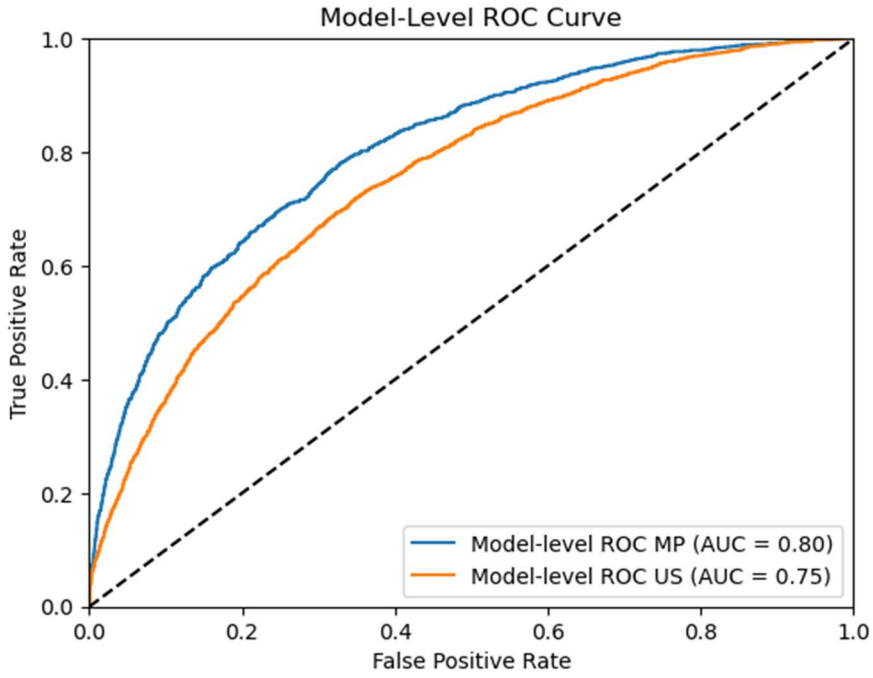


Figure 26: Receiver-operating characteristic curves associated with each model follow a randomly sampled 8:2 train/test split. The blue curve represents predictions resulting from the model trained with maternal perception (MP) labels, while the orange curve represents predictions resulting from the model trained with ultrasound (US) labels.

4.6 Discussion

In this study, the comparison of model performance across different labeling sources and validation strategies revealed a significant gap between experimental results and clinical utility. To understand the tradeoffs of using labels based on maternal perception vs ultrasound during model training, the physiological factors influencing data quality must first be considered. During the data collection process, participants indicated perceived fetal movement 97 times that were not observed through ultrasound imaging. While ultrasound is clinically reassuring, this imaging modality is constrained by a two-dimensional field of view within a three-dimensional space. Consequently, fetal movement occurring outside of the plane of the ultrasound probe may not be observed during event annotation. This discrepancy highlights that maternal perception

likely captures internal, physical perception of fetal movement that, while difficult to verify via 2D imaging, represent real physical events that a wearable IMU system might still capture. When comparing maternal perception to the ultrasound clinical standard, the sensitivity of maternal perception in this study (32%) falls below the lower end of the range (33-88%) reported in literature [14]. This lower sensitivity may be partially explained by the prevalence of anterior placentas among participants (12 of 16), as having an anterior placenta is known to reduce maternal perception of fetal movement [14]. Body mass index (BMI) may be another factor contributing to reduced sensitivity [11]. The mean BMI in this dataset (29.3) was higher than values reported in prior studies [19,33]. Because BMI increases with gestational age, thickness of subcutaneous adipose tissue may be a more effective indicator of fetal movement attenuation. These demographic factors likely influence the signal-to-noise ratio of wearable IMU sensors, and further investigations should prioritize direct measurements of adipose tissue thickness to better characterize the physical limitations of IMU-based systems.

Beyond the impact of physiological factors, the reliability of event labels themselves impacts the ability of a classifier to learn representative features of fetal movement. Models trained with ultrasound labels outperformed those trained with maternal perception labels in terms of AUROC and specificity. Though maternal perception of fetal movement as a clinical standard is less resource intensive and more accessible, variable reaction delays in maternal annotations may introduce error in event labeling. Some studies have tried to account for this reaction delay through label buffering [34]. However, when utilizing small window sizes (e.g., 0.5 seconds), data windows immediately preceding or following an event may be mislabeled despite these buffers. Additionally, maternal perception may not be sensitive to subtle fetal movements visible by ultrasound imaging [14]. These subtle movements may be captured by the wearable sensors but incorrectly labeled as no fetal movement by the mother. Training machine learning classifiers with inconsistent labels complicates the identification of optimal decision boundaries [35-37]. Evaluating models against maternal perception labels can also lead to an inaccurate assessment of performance. Model outputs may correctly identify FMs that a mother did not perceive, which would be incorrectly labeled as a false positive. In contrast, ultrasound annotations provide precise onset and completion timestamps and capture subtle movements, reducing labeling errors and allowing models to more effectively learn signal characteristics associated with true fetal movement events. Ultrasound labels also provide a less ambiguous and

more clinically meaningful estimation of model performance. Consequently, while maternal perception may serve as a scalable source for training labels, wearable systems must be evaluated against objective ultrasound to determine their true sensitivity and clinical utility.

While the results established the technical feasibility of IMU-based FM detection, they highlighted the substantial hurdles remaining before these systems can transition from a proof-of-concept to a clinically reliable tool. While models trained with ultrasound labels were slightly better at distinguishing fetal movement events, the utility of IMU-based systems must be evaluated within the context of clinical safety and the consequences of detection errors. Both models, regardless of training label quality, achieved low sensitivity when controlling for false positive rate. Under the personalized validation approach (Approach A), while models trained with ultrasound and maternal perception labels achieved 25% and 24% sensitivity at a false positive rate of 5%, respectively, their performance remained lower than maternal sensitivity of 39%. This difference highlights a critical safety consideration regarding the risk of false reassurance. When model sensitivity was tuned to 50% to outperform maternal perception, the specificity of the ultrasound-labeled model dropped to 72%. In the clinical context, this high false-positive rate could provide a dangerous sense of security to an at-risk patient, potentially delaying necessary intervention. The ability to detect FM between subjects achieved in this study represents a foundational technical milestone; however, the trade-off between sensitivity and specificity indicates that these models are not yet a viable substitute for clinical observation. Until detection models can reliably exceed maternal perception without compromising specificity, the ethical deployability of these models in unsupervised settings remains limited.

The challenge of achieving clinically acceptable sensitivity is further complicated by methods used to validate these models across existing literature. The significant increase in performance observed under a random 8:2 split (AUROC = 0.80) compared to the LOSO approach (AUROC = 0.61) highlights a widespread methodological issue in the field. Random window-level sampling allows for data leakage, where the model interpolates between highly correlated adjacent windows or learns subject-specific signatures rather than generalizable fetal movement patterns. LOSO, on the other hand, enables participant-level generalization, deployment realism, and prevention of data leakage. Furthermore, random sampling necessitates window-by-window evaluation, rather than event-level interpretation, which is not representative of clinical deployment. Because fetal movement events vary in duration, window-level metrics

may overrepresent long-duration movements and underrepresent brief kicks and startle motion. Clinicians prioritize the detection of the fetal movement event itself. Therefore, the development and evaluation of these models should occur at the event-level using LOSO or unseen trials to accurately reflect real-world deployment conditions.

The disparity in performance between validation approaches suggests that subject-specific physiological and signal characteristics heavily influence model performance. While the LOSO approach provides the most conservative estimate of generalizability, the higher performance observed when using unseen trials from known participants indicates that individualized patterns are significant. This subject-level variability was evidenced by the range of AUROC scores, with the model achieving the highest performance (0.84) on Subject 17 and the lowest performance (below 0.50) on Subject 6. Notably, these extremes in performance aligned with the ability of participants to perceive movement and their respective placental locations. Subject 17 had a posterior placenta and high perception accuracy, while Subject 6 had an anterior placenta and low perception accuracy. These results suggest that the factors complicating maternal perception likewise complicate IMU-based detection. Consequently, a deployed system may require a personalized calibration phase, potentially leveraging a brief, ultrasound-labeled trial, to fine-tune the model to the participant's specific noise profile and physiology. However, the results across both personalized and generalized validation strategies suggest that while wearable fetal monitoring shows promise, current performance levels remain insufficient for unsupervised reassurance applications, underscoring the importance of employing rigorous validation methodologies before clinical deployment.

Because model performance is highly sensitive to methodological choices like label source and evaluation strategy, the results of this study provide a necessary framework for benchmarking against other wearable fetal movement detection systems. Altini et al. [18] used a LOSO split during model training and evaluated model performance at the event-level, providing the closest comparison to benchmark the results of this study. They achieved 75% sensitivity and 65% precision across six participants. However, maternal perception was used as a clinical standard for evaluation, which may limit the reliability of the performance metrics. Sensitivity may be optimistically biased if the true number of fetal movement events was larger than the number of movements perceived by participants. Because this study used ultrasound as the

clinical standard, the results provide a more rigorous account of the current capabilities and limitations of wearable IMU systems.

The findings of this study offer a critical framework for the future development of wearable fetal monitoring systems, shifting the focus from high laboratory performance to clinical reliability, through the identification of physiological and methodological considerations. While the cohort size was appropriate for this proof-of-concept investigation of methodological variables, the prevalence of anterior placentas and modest participant count likely limited the model's exposure to a broader diversity of movement signatures. Future investigations should prioritize larger, more diverse cohorts and incorporate objective measurements of maternal physiology, such as subcutaneous adipose tissue thickness, to better characterize the physical limits of IMU-based sensing. Furthermore, the results suggest three essential recommendations for the field: (1) a transition away from random window-based data splitting in favor of LOSO or trial-based protocols to ensure generalizability, (2) ultrasound must be maintained as the evaluation standard even when leveraging maternal perception for scalable training, and (3) performance should be reported at the event-level to align with clinical decision-making.

While ultrasound provides the objective precision necessary for clinical validation, maternal perception labels remain a highly valuable asset for model training due to their scalability and reflection of real-world physical perception of fetal movement. Rather than treating these two annotation sources as mutually exclusive standards, future research should integrate them into a hybrid, tiered machine learning pipeline consisting of pre-training, fine-tuning, and personalized calibration. First, large-scale, weakly-supervised pre-training could be conducted using extensive, easily scalable maternal perception datasets to teach the model generalized IMU signal architectures across a massive, diverse population. Then, the model could undergo clinical fine-tuning using smaller, highly precise ultrasound-labeled datasets to establish objective decision boundaries and precise movement onset and offset timestamps. Finally, to address the profound subject-level variability highlighted by the extremes of Subject 6 and Subject 17, a personalized calibration phase could be deployed. This phase could utilize a brief, ultrasound-validated session at the clinic or a structured at-home maternal annotation period to adapt the model’s sensitivity to an individual participant's unique noise profile, placental location, and adipose tissue thickness. By explicitly structuring the roles of both

modalities, future systems can bridge the gap between scalable, consumer-accessible subjective data and the objective precision required for safe, unsupervised clinical deployment.

4.7 Conclusion

In conclusion, this study investigated the reliability of maternal perception and ultrasound labels for training machine learning classifiers under various train/test split approaches. Models trained with ultrasound labels were better able to distinguish and correctly identify fetal movement events, but the low sensitivity to fetal movement at a controlled 5% false positive rate demonstrated by both models suggests that the deployability of IMU-based systems in unsupervised settings remains limited. Additionally, careful investigation of various train/test split approaches revealed that existing fetal movement detection systems may not generalize to wider populations due to the significant effect of subject-specific signal characteristics. Future work should utilize subject specific calibration trials or a LOSO approach to model training to provide more rigorous evaluations of model performance. Ultimately, the methods presented in this study provide a more rigorous account of the capabilities and limitations of wearable IMU-based systems for fetal movement detection.

4.8 References

[1] Walton, J. R., and Peaceman, A. M., 2012, "Identification, Assessment and Management of Fetal Compromise," Clin. Perinatol., 39(4), pp. 753–768. https://doi.org/10.1016/j.clp.2012.09.001.

[2] Gregory, E. C. W., Valenzuela, C. P., and Hoyert, D. L., 2020, National Vital Statistics Reports Volume 71, Number 8 July 26, 2023. [Online]. Available: https://www.cdc.gov/nchs/products/index.htm.

[3] Raynes-Greenow, C. H., Gordon, A., Li, Q., and Hyett, J. A., 2013, "A Cross-Sectional Study of Maternal Perception of Fetal Movements and Antenatal Advice in a General Pregnant Population, Using a Qualitative Framework," BMC Pregnancy Childbirth, 13. https://doi.org/10.1186/1471-2393-13-32.

[4] Scala, C., Bhide, A., Familiari, A., Pagani, G., Khalil, A., Papageorghiou, A., and Thilaganathan, B., 2015, "Number of Episodes of Reduced Fetal Movement at Term: Association with Adverse Perinatal Outcome," Am. J. Obstet. Gynecol., 213(5), pp. 678.e1-678.e6. https://doi.org/10.1016/j.ajog.2015.07.015.

[5] Ryo, E., Kamata, H., and Yatsuki, K., 2025, "Correlation between Perinatal Abnormalities and Decreased Fetal Movement as Counted by a Fetal Movement Acceleration Measurement Recorder: A Prospective Cohort Study," International Journal of Gynecology and Obstetrics, 169(2), pp. 656–662. https://doi.org/10.1002/ijgo.16101.

[6] Patel, S., Conway, A. E., Adjei, T., Abati, I., Dhawan, S., Yu, Z., Vaidyanathan, R., and Lees, C., 2025, "Is It Possible to Monitor Fetal Movements with a Wearable Device? A Review of Novel Technologies," European Journal of Obstetrics and Gynecology and Reproductive Biology, 305, pp. 329–338. https://doi.org/10.1016/j.ejogrb.2024.12.011.

[7] Olesen, A. G., and Svare, J. A., 2004, "Decreased Fetal Movements: Background, Assessment, and Clinical Management," Acta Obstet. Gynecol. Scand., 83(9), pp. 818–826. https://doi.org/10.1111/j.0001-6349.2004.00603.x.

[8] Heazell, A. E. P., and Frøen, J. F., 2008, "Methods of Fetal Movement Counting and the Detection of Fetal Compromise," J. Obstet. Gynaecol. (Lahore)., 28(2), pp. 147–154. https://doi.org/10.1080/01443610801912618.

[9] Mangesi, L., Hofmeyr, G. J., Smith, V., and Smyth, R. M. D., 2015, "Fetal Movement Counting for Assessment of Fetal Wellbeing," Cochrane Database of Systematic Reviews, 2015(10). https://doi.org/10.1002/14651858.CD004909.pub3.

[10] Barros, J. G., Rosado, R., and Ayres-de-Campos, C., 2021, "Surveillance of Fetal Well-Being," The Continuous Textbook of Women's Medicine Series—Obstetrics Module, Global Library of Women's Medicine.

[11] Brown, R., Johnstone, E. D., and Heazell, A. E. P., 2016, "Professionals' Views of Fetal-Monitoring Support the Development of Devices to Provide Objective Longer-Term Assessment of Fetal Wellbeing," Journal of Maternal-Fetal and Neonatal Medicine, 29(10), pp. 1680–1686. https://doi.org/10.3109/14767058.2015.1059808.

[12] Bradford, B. F., Cronin, R. S., McKinlay, C. J. D., Thompson, J. M. D., Mitchell, E. A., Stone, P. R., and McCowan, L. M. E., 2019, "A Diurnal Fetal Movement Pattern: Findings from a Cross-Sectional Study of Maternally Perceived Fetal Movements in the Third Trimester of Pregnancy," PLoS One, 14(6). https://doi.org/10.1371/journal.pone.0217583.

[13] Ozkaya, E., Baser, E., Cinar, M., Korkmaz, V., and Kucukozkan, T., 2012, "Does Diurnal Rhythm Have an Impact on Fetal Biophysical Profile?," The Journal of

Maternal-Fetal & Neonatal Medicine, 25(4), pp. 335–338. https://doi.org/http://dx.doi.org/10.3109/14767058.2011.576721.

[14] Hijazi, Z. R., Callan, S. E., and East, C. E., 2010, "Maternal Perception of Foetal Movement Compared with Movement Detected by Real-Time Ultrasound: An Exploratory Study," Australian and New Zealand Journal of Obstetrics and Gynaecology, 50(2), pp. 144–147. https://doi.org/10.1111/j.1479-828X.2009.01122.x.

[15] Heazell, A. E. P., Sumathi, G. M., and Bhatti, N. R., 2005, "What Investigation Is Appropriate Following Maternal Perception of Reduced Fetal Movements?," J. Obstet. Gynaecol. (Lahore)., 25(7), pp. 648–650. https://doi.org/10.1080/01443610500278303.

[16] Frøen, J. F., and Frøen, J. F., 2004, A Kick from Within-Fetal Movement Counting and the Cancelled Progress in Antenatal Care Introduction and History.

[17] Rodriguez-Villegas, E., Iranmanesh, S., and Imtiaz, S. A., 2018, "Wearable Medical Devices: High-Level System Design Considerations and Tradeoffs," IEEE Solid-State Circuits Magazine, 10(4), pp. 43–52. https://doi.org/10.1109/MSSC.2018.2867247.

[18] Altini, M., Mullan, P., Rooijakkers, M., Gradl, S., Penders, J., Geusens, N., Grieten, L., and Eskofier, B., 2016, "Detection of Fetal Kicks Using Body-Worn Accelerometers During Pregnancy: Trade-Offs Between Sensors Number and Positioning," IEEE EMBC.

[19] Lai, J., Woodward, R., Alexandrov, Y., ain Munnee, Q., Lees, C. C., Vaidyanathan, R., and Nowlan, N. C., 2018, "Performance of a Wearable Acoustic System for Fetal Movement Discrimination," PLoS One, 13(5). https://doi.org/10.1371/journal.pone.0195728.

[20] Ghosh, A. K., Catelli, D. S., Wilson, S., Nowlan, N. C., and Vaidyanathan, R., 2024, "Multi-Modal Detection of Fetal Movements Using a Wearable Monitor," Information Fusion, 103. https://doi.org/10.1016/j.inffus.2023.102124.

[21] Xu, J., Zhao, C., Ding, B., Gu, X., Zeng, W., Qiu, L., Yu, H., Shen, Y., and Liu, H., 2022, "Fetal Movement Detection by Wearable Accelerometer Duo Based on Machine Learning," IEEE Sens. J., 22(12), pp. 11526–11534. https://doi.org/10.1109/JSEN.2022.3172451.

[22] Delay, U., Nawarathne, T., Dissanayake, S., Gunarathne, S., Withanage, T., Godaliyadda, R., Rathnayake, C., Ekanayake, P., and Wijayakulasooriya, J., 2021, "Novel Non-Invasive in-House Fabricated Wearable System with a Hybrid Algorithm for Fetal

Movement Recognition," PLoS One, 16(7 July), pp. 1–23. https://doi.org/10.1371/journal.pone.0254560.

[23] Mesbah, M., Khlif, M. S., Layeghy, S., East, C. E., Dong, S., Brodtmann, A., Colditz, P. B., and Boashash, B., 2021, "Automatic Fetal Movement Recognition from Multi-Channel Accelerometry Data," Comput. Methods Programs Biomed., 210.

[24] Zhao, X., Zeng, X., Koehl, L., Tartare, G., and Jonckheere, J. De, 2020, "A Wearable System for In-Home and Long-Term Assessment of Fetal Movement," IRBM, (41), pp. 205–211.

[25] Spicher, L., Bell, C., Sienko, K. H., and Huan, X., 2025, "Comparative Analysis of Machine Learning Approaches for Fetal Movement Detection with Linear Acceleration and Angular Rate Signals," Sensors, 25(9). https://doi.org/10.3390/s25092944.

[26] Rattanasak, A., Jumphoo, T., Pathonsuwan, W., Kokkhunthod, K., Orkweha, K., Phapatanaburi, K., Tongdee, P., Nimkuntod, P., Uthansakul, M., and Uthansakul, P., 2025, "An IoT-Enabled Wearable Device for Fetal Movement Detection Using Accelerometer and Gyroscope Sensors," Sensors, 25(5). https://doi.org/10.3390/s25051552.

[27] Friedrich, S., and Friede, T., 2022, "On the Role of Benchmarking Data Sets and Simulations in Method Comparison Studies." [Online]. Available: http://arxiv.org/abs/2208.01457.

[28] Lones, M. A., 2024, "Avoiding Common Machine Learning Pitfalls," Patterns, 5(10). https://doi.org/10.1016/j.patter.2024.101046.

[29] Matoba, H., Kusaka, T., Shimatani, K., Kim, D., Kanazawa, H., Kuniyoshi, Y., and Tanaka, T., 2026, "MRF-Based Mapping of Fetal and Maternal Deformations for Understanding and Supporting Fetal Movement Perception," 2026 IEEE/SICE International Symposium on System Integration, SII 2026, Institute of Electrical and Electronics Engineers Inc., pp. 832–837. https://doi.org/10.1109/SII64115.2026.11404623.

[30] Vitali, R. V., and Perkins, N. C., 2020, "Determining Anatomical Frames via Inertial Motion Capture: A Survey of Methods," J. Biomech., 106. https://doi.org/10.1016/j.jbiomech.2020.109832.

[31] Spicher, L, Bell, C, Sienko, KH, & Huan, X. "Four Data-Driven Design Choices for IMU-Based Fetal Movement Detection: Sensitivity, Thresholding, Windowing, and Filtering." Proceedings of the 2026 Design of Medical Devices Conference. 2026 Design of Medical Devices Conference. Minneapolis, MN, USA. April 20–22, 2026. V001T04A009. ASME. https://doi.org/10.1115/DMD2026-1071

[32] Asparouhov, T., and Muthén, B., 2023, Bootstrap P-Value Computation.

[33] Qin, M., Xu, Y., Liang, Y., and Sun, T., 2023, "A Wearable Fetal Movement Detection System for Pregnant Women," Front. Med. (Lausanne)., 10. https://doi.org/10.3389/fmed.2023.1160373.

[34] Ghosh, A. K., Balasubramanian, S., Devasahayam, S., Vaidyanathan, R., Cherian, A., Prasad, J., and Nowlan, N. C., 2020, "Detection and Analysis of Fetal Movements Using an Acoustic Sensor-Based Wearable Monitor," Proceedings - 2020 7th International Conference on Information Science and Control Engineering, ICISCE 2020, pp. 512–516. https://doi.org/10.1109/ICISCE50968.2020.00113.

[35] Ward, J. A., Lukowicz, P., Trö, G., and Starner, T. E., Activity Recognition of Assembly Tasks Using Body-Worn Microphones and Accelerometers.

[36] Frénay, B., and Verleysen, M., 2014, "Classification in the Presence of Label Noise: A Survey," IEEE Trans. Neural Netw. Learn. Syst., 25(5), pp. 845–869. https://doi.org/10.1109/TNNLS.2013.2292894.

[37] Bekker, A. J., and Goldberger, J., TRAINING DEEP NEURAL-NETWORKS BASED ON UNRELIABLE LABELS.

Chapter 5 Contextual Factors Incorporated by Experienced Design Practitioners: A Study Across Three Design Domains[3]

5.1 Abstract

Engineering design is shaped by a wide range of contextual factors, yet existing frameworks for characterizing these influences are often developed and evaluated within a single domain, limiting their broader applicability. This study examined the validity of an existing contextual factor taxonomy across design domains and its ability to capture a shared structure of contextual reasoning in engineering design. Using qualitative data collected from practitioners across engineering design domains and income settings, an existing taxonomy of contextual factors was applied and critically evaluated. Qualitative coding and summative content analysis showed that many contextual factor categories are generalizable across domains, suggesting the presence of a core, underlying structure in how designers account for context. At the same time, domain-specific patterns were identified that motivated nuanced design behaviors within the taxonomy. These findings contribute (1) empirical evidence supporting the cross-domain validity and potential generalizability of contextual factor frameworks, and (2) a refined taxonomy that integrates both generalizable core elements and domain-specific considerations. Together, this work advances theoretical understanding of how context is incorporated into design processes and informs the development of more adaptable design methods, tools, and educational approaches.

5.2 Introduction

Engineering design is inherently shaped by contextual factors, including social, cultural, institutional, and economic constraints that influence how problems are framed and solutions are developed [1–3]. For example, a designer may consider the frequency of power outages, average temperature, or available modes of transportation in a certain region during development [4,5]. Contextual factors must not only be considered but also be incorporated into design decision-

[3] Submitted to the ASME Journal of Mechanical Design for publication

making by evaluating their implications throughout a design process [6,7]. While prior research has identified a range of these factors within specific domains, studies have yet to investigate whether these contextual considerations reflect domain-specific behavior or point to more generalizable structures in design reasoning. This gap limits the field from building cumulative theory and developing design methods that are robust across application areas.

A failure to incorporate contextual information into engineering design decisions has been linked to design shortcomings and product failures across domains including medical devices, information technology, and consumer products [3]. For example, the da Vinci robotic surgery system gained widespread adoption as a robotic platform intended to improve surgical precision and clinical outcomes. However, many reported adverse events stemmed from insufficient consideration of the operating room context rather than purely technical faults. Communication barriers caused by face masks, reflective surfaces, and alarm noise disrupted coordination among surgical teams, while institutional factors such as off-site instrument storage and inconsistent surgeon training further contributed to workflow disruptions and system malfunctions [8]. Similarly, barcoded medication administration systems were introduced to reduce medication errors by requiring clinicians to scan patient wristbands before administering medication. In practice, contextual constraints such as lost or damaged wristbands, unreadable medication barcodes, and unreliable network connectivity led to workarounds that undermined the system's intended safety benefits and contributed to medication errors [9]. In consumer technology, Google Glass was introduced as an early wearable device resembling a pair of eyeglasses that could display information directly within the user's field of vision. Although the device represented a significant technical innovation, its designers failed to fully anticipate the sociocultural and political implications of wearable cameras in public spaces, leading to concerns about surveillance and privacy that generated widespread backlash [10]. Together, these examples illustrate how overlooking institutional, infrastructural, sociocultural, and political contexts can undermine otherwise sophisticated technologies, highlighting the importance of incorporating contextual knowledge into engineering design processes.

Though there is evidence that incorporating contextual factors into design decisions is necessary for successful product deployment, engineering practice tends to prioritize technical performance over appropriateness for context. Technical rigor is necessary for a solution to work, but it is not sufficient for a solution to be adopted or effective [11]. Even if the designer

and user have the same objective, the factors that drive the engineer's decision-making process may ignore the needs and context of the user, resulting in infrastructure failure or design rejection. In some cases, technologies that are poorly aligned with their context can disrupt existing social structures and exacerbate divisions related to cultural identity, economic class, or social status [11]. Engineering designers are typically trained to prioritize technical feasibility, but contextual factors define a set of non-negotiable constraints that are often undervalued in engineering analysis. To address this gap, approaches such as participatory design, empathic design, and human centered design seek to uncover contextual factors and user needs [12–14]. While these approaches encourage designers to gather rich contextual information, they provide limited guidance on which aspects of context are most relevant and how these factors are incorporated into engineering design processes. Researchers have partially addressed this gap through the development of frameworks to guide the collection [15] and incorporation [5] of contextual factors, but these frameworks have only been developed for isolated contexts (i.e., global health). To extend prior work, this study explored and compared how experienced designers incorporated contextual factors across design domains and income settings.

5.3 Background

Across engineering design literature, researchers have proposed a variety of ways to characterize context, reflecting different perspectives on the factors that influence design decisions. For example, context has been described in terms of "broad" and "close" factors in relationship to the design problem [16]. Other work separates context into customer, market, and usage contexts, emphasizing the different stakeholders and environments that shape product development [17]. Similarly, a context model proposed by Hales and Gooch [18] organizes context into five levels: the external environment, market, company, project and personal levels. More recently, Jagtap [19] identified several contextual aspects relevant to engineering design, including income, rural or urban setting, design sector, national context, and gender.

While these categorizations highlight important dimensions of context, it can still be difficult to organize and interpret contextual information in the complex environments where products will be used. Designers can easily become overwhelmed by the massive amount of information required to understand these settings [15,20]. Taxonomic tools simplify complex information into clear, structured categories [20–22]. These frameworks help teams from different backgrounds establish a shared language, ensuring that everyone has a common

understanding of user needs and project requirements [20,23]. Beyond organizing data, taxonomies provide continuous support throughout design processes, guiding decision-making across all stages of development, from problem definition through implementation [5,24]. More broadly, frameworks, tools, and methods that support design decision-making are critical for identifying what information is missing, helping designers spot potential "blind spots" before they lead to costly project failures [20,21]. Ultimately, using taxonomic frameworks can help organize broad aspects of context into practical insights that can be incorporated into design decisions to align solutions with their intended context [5,22].

To address the need for structured approaches to organizing contextual information, two taxonomical frameworks have been developed to categorize contextual factors based on the experiences of engineering designers working in global health contexts. Aranda-Jan et al. [15] classified contextual factors into eight categories: infrastructure, geographical/environmental, institutional, economic, public health, technology, industrial, and sociocultural. Building on this work, Burleson et al. [5] developed an expanded framework consisting of 32 contextual factors classified into nine categories: infrastructure, environmental, institutional, economic, public health, technological, industrial, sociocultural, and political. In addition to identifying which contextual factors were considered, the researchers synthesized how contextual information was incorporated into engineering design decision-making, describing themes of contextual incorporation across the nine categories. Together, these frameworks provide designers with structured tools for identifying relevant contextual factors and understanding how such factors may influence engineering design decisions. However, both frameworks were developed within a specific design domain, global health technologies in low- and middle-income countries, which may limit their applicability to other engineering design contexts.

While these frameworks were developed in the context of global health technologies, engineering design practices vary across product domains that operate within different regulatory, technological, and market environments. Medical devices are typically physical artifacts designed to address unmet clinical needs, improve safety, and manage risk within healthcare systems [25,26]. Because safety is the primary concern in medical device design, these products are subject to strict regulatory standards and extensive documentation requirements [26,27]. Medical devices are often tested through simulation-based studies prior to clinical deployment, though these simulations may fail to capture the cognitive load,

environmental distractions, and time pressures present in real clinical workflows, making real-world validation essential [27]. Nearly 80% of all medical devices are designed in North America, Europe, Japan, or China, and most are designed for these contexts [28]. In effect, a vast majority of medical equipment in low-resource settings is donated. Because this equipment was not designed for the context, 70-90% of donated equipment is not operational [28,29]. Designing and implementing medical devices in low-resources settings involves navigating systemic barriers that span manufacturing and regulatory requirements, procurement and market competition, environmental and power instability, and long-term maintenance and workforce capacity [4].

Information and communication technologies (ICTs) encompass both hardware and software systems used to store and transmit data and automate complex processes [30]. These technologies support activities such as business decision-making, process automation, and efficiency improvements, while also expanding access to services such as government resources, education, and healthcare through mobile broadband and internet connectivity [30,31]. Security and data privacy remain central concerns in ICT design [32,33]. In low-resource contexts, ICTs have expanded opportunities for virtual education, improved agricultural supply chain management, and supported patient care and medication management in healthcare settings [34–36]. However, technological advancement alone cannot produce social progress; the effectiveness of ICT solutions depends on alignment with the social and institutional context in which they are deployed [37].

Consumer products encompass physical or abstract artifacts designed for everyday use [38]. Compared with other product domains, consumer product design places greater emphasis on aesthetics, functionality, and cultural symbolism, while market competition often makes product uniqueness and perceived value central determinants of success [39–41]. Consumer product development is typically driven by market demand and user experience, with profitability serving as a primary objective [38,41]. More recently, consumer product design has increasingly emphasized resource efficiency and circular economy principles in response to sustainability concerns [42]. In low-resource settings, irregular income, limited access to infrastructure and basic services, and short-term consumption behaviors influence product adoption and use [43,44]. As a result, frugal innovation approaches have emerged to design products that address the needs of users in resource-constrained environments [45].

While engineering designers often possess the skills needed to recognize and incorporate contextual factors into their design processes, these considerations are frequently examined within specific domains. As a result, it remains unclear whether the contextual factors identified in these frameworks are unique to that domain or whether they reflect broader patterns of reasoning used by designers across engineering practice. Prior work suggests that engineering designers rely on cognitive strategies that include both domain-specific and domain-general knowledge [46], raising the possibility that the contextual categories identified in global health design may apply to complex engineering problems more broadly. This study addresses the possibility of generalization by examining the applicability of an existing taxonomy of contextual factors across multiple design domains. Rather than focusing solely on whether the taxonomy can be applied in new settings, the extent to which it can capture a shared structure of contextual reasoning across diverse design contexts was investigated. Our findings demonstrated that the taxonomy not only generalizes across domains but also benefits from defined considerations that account for domain-specific nuances. The contributions of this work are two-fold: (1) empirical evidence supporting the cross-domain validity and generalizability of contextual factor frameworks in engineering design and (2) an expanded taxonomy that advances theoretical understanding of how context shapes design processes. These contributions encourage more unified design theories and more adaptable design support tools.

5.4 Methods

The goal of this study was to investigate the transferability and applicability of existing contextual factor categories to other design domains and income settings and to compare the incorporation of contextual factors across the design domains of medical devices, information technology, and consumer products. This study was guided by the following research questions: (1) What contextual factors do engineering designers incorporate into their design processes? and (2) How are the contextual factors incorporated similar or different across design domains and income settings?

To address our research questions, an exploratory study of experienced designers' practices was conducted through retrospective semi-structured interviews with participants across various domains and contexts. Using a summative content analysis approach, we examined how participants engaged with and applied specific contextual factors within their

previous design processes. This study was deemed exempt by the University of Michigan Institutional Review Board (HUM00199822).

5.4.1 Participants

Study participants (Table 10) included 58 engineering designers from three design domains: medical devices (MED), information technologies (IT), and consumer products (CP). Approximately half of the designers from each domain implemented their work in low- and middle-income country (LMIC) contexts and half in high-income country (HIC) contexts, classified by the World Bank [47]. Eligible participants were required to (1) have a minimum of three years of professional design experience and (2) have participated actively in the design of at least one product within the last five years from one of three industry sectors: medical devices, information technology, or consumer products.

Table 10: Participant Information

Identifier	Gender	Region based in	Race	Regions worked in/products implemented in	Years of design experience
MED-L-01	Male	North America	White	Sub-Saharan Africa, Southern and Southeast Asia	20
MED-L-02	Female	North America	White	All regions	12
MED-L-03	Female	North America	South Asian	South Asia	10
MED-L-04	Female	Europe	N.p.	Sub-Saharan Africa, South Asia	5
MED-L-05	Male	North America	White	All regions	27
MED-L-06	Male	South Asia	South Asian	South Asia	5
MED-L-07	Female	North America	White	Sub-Saharan Africa, South America, Southern and Southeast Asia	9

MED-L-08	Male	Sub-Saharan Africa	African	Sub-Saharan Africa	3
MED-L-09	Male	Sub-Saharan Africa	African	Sub-Saharan Africa	5
MED-L-10	Male	Europe	White	Sub-Saharan Africa, South Asia	28
MED-L-11	Male	Sub-Saharan Africa	South Asian	Sub-Saharan Africa	5
MED-H-01	Female	North America	South Asian	North America, Europe, Southern and Southeast Asia	4
MED-H-02	Male	North America	Asian	North America	4
MED-H-03	Female	North America	White	North America	5
MED-H-04	Female	North America	Hispanic/Latina	All regions	4
MED-H-05	Female	North America	White	North America	4
MED-H-06	Male	North America	White	All regions	15
MED-H-07	Male	North America	White	North America, Europe, Southeast Asia	13
MED-H-08	Male	North America	White	All regions	38
MED-H-09	Male	North America	Asian	North America, Europe	3
MED-H-10	Male	North America	White	All regions	35

IT-L-01	Male	North America	White	South Asia	11
IT-L-02	Male	North America	South Asian	Sub-Saharan Africa, Southern and Southeast Asia, South America	10
IT-L-03	Female	North America	South Asian	Sub-Saharan Africa	3
IT-L-04	Male	Sub-Saharan Africa	African	Sub-Saharan Africa	8
IT-L-05	Male	Sub-Saharan Africa	African	Sub-Saharan Africa, South Asia	4
IT-L-06	Female	West Asia / Middle East	White	North America, Sub-Saharan Africa, West Asia / Middle East	23
IT-L-07	Male	North America	Pakistani / Asian	North America	8
IT-L-08	Female	South Asia	Asian	Sub-Saharan Africa, Asia	7
IT-L-09	Male	South Asia	Asian	South Asia	7
IT-L-10	Male	South Asia	Asian	South Asia	7
IT-H-01	Female	North America	White	North America, Europe, Central America, Oceania	10
IT-H-02	Female	Europe	White	Europe	20
IT-H-03	Female	North America	White	All regions	7
IT-H-04	Female	North America	Asian	North America, Europe	5
IT-H-05	Female	North America	Mixed	North America	12
IT-H-06	Female	North America	White	North America	4
IT-H-07	Female	North America	White	North America, Central America, Europe, Asia	4
IT-H-08	N.p.	North America	N.p.	North America	15

IT-H-09	Male	North America	White	All regions	23
IT-H-10	Male	North America	Asian	North America, South Asia	4
CP-L-01	Male	Oceania	White	Europe, Southeast Asian, Oceania	9
CP-L-02	Male	Sub-Saharan Africa	African	North America, Sub-Saharan Africa	3
CP-L-03	Male	North America, South America	Hispanic/Latino	South America	10
CP-L-04	Male	Sub-Saharan Africa	African	Sub-Saharan Africa	5
CP-L-05	Male	Europe	N.p.	Sub-Saharan Africa, Southeast Asia	6
CP-L-06	Male	South Asia	Asian	All regions	4
CP-L-07	Male	Sub-Saharan Africa	African	Sub-Saharan Africa	7
CP-L-08	Male	North America	White	North America, Sub-Saharan Africa, Southern and Southeast Asia	18
CP-H-01	Male	Europe	N.p.	Europe, Oceania	30
CP-H-02	Female	North America	White	North America, South America, Europe	3
CP-H-03	Female	North America	White	North America, Europe, East Asia	7
CP-H-04	Female	North America	African	All regions	7
CP-H-05	Male	North America	White	All regions	40
CP-H-06	Male	North America	White	North America	6

CP-H-07	Female	North America	N.p.	North America, Southern and Southeast Asia	15
CP-H-08	Female	North America	Asian	North America, Europe, East Asia	16
CP-H-09	Male	North America	White	North America	6

[a]Note: Some participants chose not to provide data related to gender and/or race (noted by "N.p.")

[b]MED, medical devices; IT, information technologies; CP, consumer products; H, high-income countries; L, low-income countries; M, male; F, female; USA, United States of America; SSA, Sub-Saharan Africa; SEA, Southeast Asia

The research team used a purposeful and convenience sampling strategy [48,49] for participant recruitment. Participants who voluntarily expressed interest were screened for eligibility through an online recruitment questionnaire. For each admitted participant, a 90-minute video interview was scheduled, a survey to collect relevant demographic information was distributed, and participant consent was collected via email.

5.4.2 Data Collection

A semi-structured interview approach was used to collect data about what contextual information participants paid attention to and how they incorporated this contextual information into their engineering design processes. The interviews were conducted in six main stages: (1) introduction, overview of study goals, and warmup, (2) design experience overview, (3) understanding project context and stakeholders, (4) understanding the design context, (5) deep dive into incorporating contextual factors, and (6) wrap-up. First, participants were introduced to the concept of context and asked to focus on only one project they had recently worked on so that they could easily recall details more thoroughly. The interview protocol prompted participants to share further details about the project's users, stakeholders, and context. Then, the "deep dive" was conducted to understand which contextual factors were considered and incorporated into the participants' design processes. During the "deep dive", a series of slides were screen-shared one at a time. Each slide contained one to two contextual factor "cards" that provided a definition of each of the nine contextual factor categories and examples of contextual factors within that category. Finally, participants were asked to reflect on their experiences incorporating contextual factors into their design process and share any final thoughts.

Our research team used the same interview protocol used in a prior study that investigated the incorporation of contextual factors by global health designers [5]. From this previous version, the slide deck shared during the "deep dive" was updated to include all 32 contextual factors organized into 9 contextual factor "cards" (Appendix A). Minimal changes were made to the original protocol to adapt terms so they translated across domains (e.g., regions or institutions used to explain concepts, examples provided to explain each contextual factor classification, etc.). Example questions from the interview protocol are detailed in Appendix B. Two study team members conducted interviews independently. Interviews were conducted via videoconferencing software (n=58), audio recorded, and transcribed by a third-party transcription service.

5.4.3 Data Analysis

To explore the incorporation of specific contextual factors across design practice, we followed a summative content analysis approach, to quantify and analyze the frequency of themes across a sample [50]. All transcripts were read in depth to identify and annotate instances in which a participant provided evidence of a contextual factor incorporated into their design process. An initial analysis was conducted with a subset of 15 participants (MED LMIC: n=11, IT LMIC: n=4) to characterize the incorporation of context factors by global health designers [5]. These interviews had previously been coded by two independent coders using the contextual factor taxonomy until an intercoder reliability (ICR) defined as the number of agreements divided by the sum of agreements and disagreements [51] of 0.88 was achieved.

The taxonomy was then applied to 43 additional interviews to deductively assess the transferability of global health contextual factors to broader domains (e.g., medical device, information technology, and consumer product), an approach consistent with the extension of existing frameworks [52]. Inductive analysis was also conducted to identify emergent themes not previously captured in the codebook [53]. Using both deductive and inductive analysis ensured systematic testing of the existing taxonomy and the discovery of domain-specific nuances within the raw data.

Two research team members independently coded transcripts and calculated ICR until the average ICR for approximately one-quarter of our full dataset (13 transcripts) reached 0.76 at the secondary contextual factor level and 0.81 at the primary classification level. The remaining transcripts (n=30) were coded independently by the two coders.

The primary disagreements were: (1) determining the bounds of the "institution" and (2) differentiating between contextual factors that are dependent on one another, such as technologies that are required to be compatible with technologies available in the context. Therefore, three sets of two closely related contextual factors were combined. These included "distance" combined with "transportation and road quality", "availability of technologies" combined with "compatibility with the technical context", and "existing practices and procedures" combined with "healthcare system and practices". The final codebook contained 29 contextual factors organized into 9 primary classifications (Table 11). Because no new contextual factors were introduced, the original interviews from Burleson et al. [5] were not re-coded but the contextual factors above were merged for analysis.

Table 11: Final codebook containing 29 contextual factors organized into 9 primary classifications

Classifications (primary codebook)	**Contextual Factor (secondary codebook)**
Economic (1)	Individual and household characteristics (1a), Regional and national characteristics (1b), Labor market characteristics (1c)
Environmental (2)	Weather (2a), Natural environment (2b)
Industrial (3)	Supply & manufacturing (3a), Distribution (3b), Maintenance (3c)
Infrastructure (4)	Utilities (4a), Transportation & road quality (4b), Attributes of the built environment (4c)
Institutional (5)	Financial capacity (5a), Resources (5b), Indoor environment (5c), Existing practices and procedures (5d), Capacity and capability of staff (5e)
Political (6)	Regulations and regulatory processes (6a), Stakeholder power dynamics (6b), Global priorities (6c), Political systems & culture (6d)
Public Health (7)	Health demographics (7a)
Socio-cultural (8)	Cultural tradition and practices (8a), Stigmas and taboos (8b), Symbols (8c), Aesthetics (8d), Education and literacy rates (8e), Language (8f)
Technological (9)	Compatibility with the technical context (9a), Availability of consumables (9b)

First, to understand the transferability of the existing codebook and incorporation strategies, the key patterns identified in Burleson et al. [5] were compiled. Then, the excerpts from the five remaining domains (e.g., MED-H, IT-L, IT-H, CP-L, and CP-H) were reviewed and synthesized according to these key patterns. Simultaneously, the excerpts were reviewed to investigate new ways in which contextual factors are incorporated across product domains and patterns were identified. This iterative thematic analysis was conducted to understand the similarities and differences across product domains and income settings.

5.5 Findings

The findings are structured in four main parts. The first part consists of an overview of the participant-level classification of the contextual factors incorporated during design across product domains and income settings using the taxonomy developed by Burleson et al. [5]. Part two describes a detailed comparison of selected themes of incorporation of contextual factors that were observed in each product domain. Part three describes a detailed comparison of selected themes of incorporation that were specific to each income setting. Finally, part four describes instances where participants incorporated multiple contextual factors into a single design decision.

5.5.1 Participant-level classification of contextual factors incorporated during design

In this study, an existing taxonomical framework of contextual factors [5] was applied to evaluate its applicability and transferability across contexts. A breakdown of the observed contextual factors by domain and income setting is detailed in Table 3. Most contextual factors (27 of 29) were identified by at least one additional participant, while nearly one quarter of the contextual factors (7 of 29) were identified by at least one additional participant across all product domains and income settings. The two contextual factors mentioned only by global health designers in the previous study (Burleson et al. 2024) were Availability of consumables and Global priorities.

Across product domains, most participants incorporated sociocultural (44 of 58), political (43 of 58), and institutional (42 of 58) factors. All nine categories of context were mentioned by at least one participant, suggesting that all aspects of context are important and potentially impactful to design regardless of the domain. Medical device designers incorporated the most contextual factors across domains (139 factors across 21 participants). All 29 contextual factors

were incorporated by at least one participant. Information technology designers incorporated the fewest contextual factors (67 factors across 20 participants). Consumer product designers incorporated 74 factors across 17 participants. Participants from all three product domains described the incorporation of 21 of the 29 contextual factors. Six of the 29 contextual factors were used by participants from two of the three domains. Consumer product and medical device participants used three of the seven contextual factors (Natural environment, Distribution, and Maintenance), while information technology and medical device participants used three of the seven contextual factors (Regional and national characteristics, Capacity and capability of institutional staff, and Education and literacy rates). The frequencies by contextual factor and participant domain are summarized in Table 12.

Table 12: Contextual factor usage counts by domain and context (unit of analysis: participants)

Contextual Factor	MED-L (n=11)	MED-H (n=10)	IT-L (n=10)	IT-H (n=10)	CP-L (n=8)	CP-H (n=9)
Economic (22)	**9**	**4**	**4**	**2**	**3**	**0**
Individual and Household Characteristics (10)	5	0	2	0	3	0
Regional and national characteristics (7)	4	2	1	0	0	0
Labor market characteristics (10)	3	2	1	2	2	0
Environmental (26)	**11**	**4**	**1**	**0**	**6**	**4**
Weather (22)	11	3	1	0	5	2
Natural environment (11)	4	4	0	0	1	2
Industrial (24)	**11**	**3**	**1**	**1**	**5**	**3**
Supply & manufacturing (18)	10	3	1	1	2	1
Distribution (10)	5	1	0	0	2	2
Maintenance (13)	8	1	0	0	3	1

Infrastructure (33)	**11**	**4**	**6**	**1**	**6**	**5**
Utilities (20)	8	2	5	0	2	3
Transportation & Road quality (16)	6	2	2	1	5	0
Attributes of the built environment (11)	1	2	1	0	3	4
Institutional (42)	**10**	**10**	**4**	**8**	**7**	**3**
Institutional financial capacity (10)	4	3	1	0	2	0
Institutional resources (11)	4	4	0	1	2	0
Indoor environment (11)	4	3	0	1	2	1
Existing practices and procedures (33)	9	9	5	9	7	4
Capacity and capability of institutional staff (12)	5	4	2	1	0	0
Political (43)	**11**	**8**	**4**	**7**	**7**	**6**
Regulations and regulatory processes (34)	8	7	2	6	6	5
Stakeholder power dynamics (24)	8	2	2	4	5	3
Global priorities (5)	5	0	0	0	0	0
Political systems and culture (10)	6	0	0	1	3	0
Public Health (15)	**5**	**6**	**1**	**2**	**0**	**1**
Health demographics (15)	5	6	1	2	0	1
Sociocultural (44)	**11**	**7**	**8**	**5**	**6**	**7**
Cultural tradition and practices (24)	7	0	5	1	4	7
Stigmas and taboos (20)	5	3	4	3	3	2

Symbols (13)	7	0	4	0	2	0
Aesthetics (14)	4	1	1	1	4	3
Education and literacy rates (9)	6	0	2	1	0	0
Language (21)	8	4	5	2	2	0
Technological (31)	**10**	**4**	**8**	**4**	**3**	**2**
Compatibility with the technical context (30)	9	4	8	4	3	2
Availability of consumables (8)	8	0	0	0	0	0

The incorporation of contextual factors was also compared across low- and middle-income country (LMIC) and high-income country (HIC) contexts. Designers in LMICs incorporated a broader range of contextual factors (169 factors across 29 participants). These participants prioritized factors in the sociocultural (25 of 29), infrastructural (23 of 29), political (22 of 29), institutional (21 of 29), and technological (21 of 29) classifications. Designers in HICs incorporated 111 factors across 29 participants. They prioritized a narrower range of contextual factors across institutional (21 of 29), political (21 of 29), and sociocultural (19 of 29) classifications. Participants in HICs described using 25 of the 29 contextual factors, while 4 of the 29 contextual factors were exclusively used by participants in LMICs (Individual and household characteristics, Global priorities, Symbols, and Availability of consumables). The frequencies by contextual factor and income setting are summarized in Table 12.

The expanded dataset largely agreed with the themes identified by Burleson et al. [5], demonstrating significant overlap in how designers incorporate contextual factors across various domains and contexts. Most previously identified themes were observed by at least one additional participant from the expanded dataset, with a few exceptions that were highly specialized in the global health (MED LMIC) context (e.g., Global priorities and Availability of consumables). Beyond confirmation of the original framework, this analysis yielded a collection of new themes. While many of these new themes were distinct considerations for design decision-making, others provided a more nuanced understanding of existing themes, illustrating how similar contextual factors can drive different design outcomes depending on the specific

engineering domain. The most observed themes across the dataset and detailed considerations are summarized in Appendix C.

5.5.2 Detailed Comparison of contextual factor incorporation across design domains

Across all product domains, most designers incorporated contextual factors from sociocultural, political, and institutional classifications. However, differences in the contextual factors incorporated between product domains suggest a deeper investigation of context is required. This section describes selected themes of incorporation within each product domain.

5.5.2.1 Medical Device Designers

Medical device designers incorporated the most extensive set of contextual factors across product domains. They described the incorporation of all 29 contextual factors identified in the taxonomy, and the greatest number of contextual factors per participant (169 factors across 21 participants). Aligned with the general trend across the full dataset, most medical device designers incorporated institutional (20 of 21), political (19 of 21), and sociocultural (18 of 21) factors. These participants described integration with existing clinical workflows, mitigation of safety risk through regulatory compliance, and satisfaction of cultural norms.

Many medical device designers incorporated existing practices and procedures by designing their products to function within an institution's available storage and cleaning procedures (MED-L-4,7,8,10; MED-H-5:8,10). Commonly used chemicals, protocols, and frequency of cleaning helped participant MED-H-5 define durability specifications:

> "And then I guess the other thing with cleaning really comes down to chemicals that they're going to use. [...] it can be really hard to find a material that works for all the hospitals because some are going to use bleach and some are going to use alcohol and some are going to use hydrogen peroxide and different things like that. And they all have different procedures and they're not really willing to change those procedures. So you have to be able to work within whatever they have." – MED-H-5

Medical device designers also described the incorporation of strict regulatory requirements (MED-L-1:3,4,6,10:11, MED-H-1:4,6,9,10). Participants explained that the United States Food and Drug Administration (USFDA) and the European Union Medical Device Regulation (EUMDR) are the two main bodies that set regulatory standards, though each country has its own process. Most participants who met global regulations were motivated to meet stricter standards than what was required in the use country to ensure transferability to other contexts. Participant MED-H-9 described the navigation of medical device regulation:

> "So if we're going to sell it to the US, that's a set of requirements that we have to meet based off FDA and what they expect to see. [...] if we were to bring it to Europe or China, we'd have to do another review of, okay, who's the regulatory body there and what do they need for them to accept the device and sell it in their country? So for Europe, we have to meet MDR requirements [...] So they require a clinical or evaluation report to sell a product over there. And they have different classifications of devices. […] So they have different levels of scrutiny, and those levels drive different requirements." – MED-H-9

Because health conditions are often stigmatized in society, it was especially important to medical device designers to understand and incorporate local stigmas and taboos by designing for dignity (MED-L-1,3,4,11; MED-H-2,6,8). Participant MED-L-1 described how redesigning an artificial knee to look more natural reduced social stigma in public:

> "Well, first of all, a person who has an amputated leg wears pants all the time, because they don't want people to see their artificial leg [...] And when they sit down the corners of the knee made points on the pants. [...] So we redesigned the knee to be rounded, so when you sit down, it looks like your other knee. That was an interesting stigma that we heard from people. [...] That social stigma is huge. Being able to kind of pass as not having a deformity, made people so excited about having our new design because they knew that they could walk around in public." – MED-L-1

5.5.2.2 Consumer Product Designers

Consumer product designers also incorporated a broad set of contextual factors (24 of the 29 factors from the taxonomy) and identified 74 distinct factors across 17 participants. Compared to the other product domains, consumer product designers did not describe incorporating Regional and national characteristics, Capacity and capability of institutional staff, Global priorities, Education and literacy rates, and Availability of consumables. Most incorporated sociocultural (13 of 17) and political (13 of 17) factors. They described generating value for the user by meeting community needs and aligning with powerful stakeholders in the context. Uniquely, consumer product designers also prioritized infrastructural factors (11 of 17), often in the context of the home environment.

Consumer product designers incorporated sociocultural factors in many ways, but uniquely, they prioritized aesthetics in the end-use context. Many participants incorporated local aesthetics by making detailed design changes to convey status and prestige (CP-L-1,5; CP-H-3,7). Participant CP-H-7 described how wealthy, city architecture motivated rural communities to engage in improved sanitation practices:

> "People viewed their toilet as a piece of infrastructure that is available in the cities. And so it was a prestige issue. [...] People said that a thatched structure makes me look poor. People in the cities have concrete. [...]

> We would rather have no toilet and just poop in the open than have a toilet that makes us look poor. [...] If I buy this, I'm showing off to my neighbors that I'm just like the [city] people. [...] And so our motivation is addressing this health issue, but none of the motivators for why people were buying the toilet had anything to do with health." – CP-H-7

Unlike the large regulatory bodies dealt with by medical device designers, consumer product designers incorporated political factors by working with local political officials and navigating stakeholder dynamics. Many designers aligned their solutions with powerful stakeholders' interests by identifying budgets and goals to incorporate into their designs (CP-L-1,2,3,6,8; CP-H-5,6). Participant CP-H-8 described choosing a more expensive and complex technology because it was trusted by the local community:

> "Let's do the cheap one and train people how to do it. That's what we went in there for. That was our plan. And of course, that's not what ended up happening. The ultimate plant that we decided to build was actually something that was designed by the [national] government. [...] And so from a technological standpoint, the engineering design decision really was almost out of our hands. It was really like what's trusted in the local community." – CP-H-8

Uniquely, consumer product designers considered and incorporated infrastructural factors. They often incorporated attributes of the built environment by making detailed design decisions based on common features in the home like construction materials, home size, and compatibility with electrical and HVAC systems. (CP-L-2,4,5; CP-H-2,8,9). For example, participant CP-L-5 described the incorporation of construction materials:

> "Normally what you would do is put a screw in the wall and then you would fix this. Well that's assuming that the wall is going to be flat and the wall is concrete. Yes. Nothing of that happens out there. The wall is not flat and the wall is made of adobe and as soon as you put the screw, you may bring the whole wall down. You need to be very careful." – CP-L-5

Similarly, participant CP-H-2 described how the size of home in different contexts may change the way a consumer product is designed:

> "So a lot of our US appliances are just twice as big sometimes as appliances in other regions, and that's due to the constraints of home sizes mostly. […] Refrigerators and ovens specifically end up being a lot more compact to be able to fit into more of an apartment lifestyle as opposed to a single family home." – CP-H-2

5.5.2.3 Information Technology Designers

Information technology designers incorporated a broad set of contextual factors (24 of the 29 factors from the taxonomy) but identified the fewest contextual factors at the participant level (67 factors across 20 participants). Contextual factors not incorporated by information

technology designers included Natural environment, Distribution, Maintenance, Global priorities, and Availability of consumables. Most designers in this domain incorporated sociocultural (13 of 20) and institutional (12 of 20) factors. They prioritized user experiences by meeting cultural norms and optimized or automated existing workflows. Uniquely, information technology designers incorporated technological factors (12 of 20), likely because many information technologies are digital tools or applications.

When incorporating sociocultural factors, many information technology designers incorporated local education and literacy rates by using context-specific graphics on products and manuals (IT-L-1,3,7,10; IT-H-1). For example, participant IT-L-1 described using icons for users who were not literate:

> "Often we can't assume literacy on the part of users in different parts of [country]. And so, relying instead on numeracy and on heavy use of graphics, for example, that we would have graphical icons that are printed on the envelopes and so on." – IT-L-1

Information technology designers considered institutional factors when identifying user needs. They often incorporated institutional practices and procedures by identifying systemic gaps in existing systems to inform needs finding during early stages of design (IT-L-2; IT-H-4,5,6,8,10). Specifically, participant IT-H-8 described pain points in hospital administration processes:

> "So hospitals always have a backlog of stuff to bill for. They always have a pile of patient encounters that people need to review and go through daily and that's the only way they get paid." – IT-H-8

Uniquely, information technology designers considered and incorporated technological factors. Information technologies are commonly digital tools, so designers often considered available smartphones and internet connectivity to ensure solutions were compatible in the end-use context. Participant IT-L-1 described technology access within a specific rural context that was unexpected:

> "That was an assumption we made basically that most patients could access, like a basic mobile phone on a daily basis, along with connectivity. That turned out not to be true in many locations. […] Even though there are so many hundreds of millions of mobile accounts, they're not evenly distributed. And so, if you look at older adults, especially women, especially in rural areas, the chances of having daily access to a mobile phone really decrease. […] I think that's really influenced the environments that we can deploy." – IT-L-1

5.5.3 Detailed Comparison of contextual factor incorporation across income settings

Across income settings, designers in LMICs incorporated more contextual factors (169 factors across 29 participants) across a broader range of classifications. Designers in LMICs incorporated all 29 contextual factors identified in the taxonomy. They prioritized factors from sociocultural, infrastructural, political, technological, and institutional classifications. Designers in HICs incorporated fewer contextual factors (111 across 29 participants) and only described 21 of the 29 factors from the taxonomy. These designers did not describe incorporating Individual and household characteristics, Global priorities, Symbols, and Availability of consumables. They prioritized a narrower range of classifications including institutional, political, and sociocultural factors. Designers from both income settings mentioned factors from all other classifications, but to a lesser extent, suggesting that a broad understanding of context is important. This section describes selected themes of incorporation within each income setting.

5.5.3.1 Low- and Middle-Income Country Settings

In LMIC settings, designers uniquely incorporated individual and household characteristics when setting cost requirements (MED-L-3,6,9,11; IT-L-1,3,4; CP-L-1,2), frequently citing low affordability among users. Unique strategies to drive down costs in LMIC settings included government sponsorship or subsidy (MED-L-8, IT-L-2) or financing plans (CP-L-1, CP-L-2). However, participants further explained that pricing products too low could generate suspicion of quality (MED-L-3, MED-L-7, CP-L-8). For example, participant MED-L-7 explained how their team navigated pricing:

> "If you ask a doctor how much a device should cost, they won't necessarily know. […] Everybody wants everything as cheaply as possible, but sometimes we'll price our technologies so low that they actually think it's not valuable or clinically accurate. […] We found in talking with hospital administrators, if you're like, this thing is $20. This thing is $50. This thing is $100. They'll actually be like, "Well, I want the $50 one because I don't want the $20 one. That's garbage." – MED-L-7

Designers in LMIC settings also uniquely incorporated local symbolism by selecting colors and shapes with shared meaning in their specific context (MED-L-2,4,5,7:11; IT-L-1,3,4; CP-L-4). Participant IT-L-6 described how a basic circle had different meanings in different contexts:

> "As she was showing them this very minimalist circle that anyone who is familiar with the BBC weather iconography would be like, oh, there's the sun. These Burmese farmers who were very focused on the moon, because that was very important to them in their iconography as Buddhist, they saw a circle and they

said, oh, this must be the moon. And so she ended up designing a visual iconography that was a little bit more photorealistic, even though that wasn't the style at the time." – IT-L-6

The most common theme mentioned by designers in LMIC settings was the incorporation of local weather into their design processes, primarily by selecting materials that withstood local temperature, humidity, and rainfall (MED-L-1:11; IT-L-1,4; CP-L-1:4,6). Many designers described how exposure to harsh environments can be more common in occupational and daily life (MED-L-1,3,7:9; IT-L-1; CP-L-1:3). Specifically, participant MED-L-1 described how material selection was influenced by exposure to rain in the farming industry:

"We knew that people needed to get it wet… It's wet in some parts of India. It's wet in Vietnam. It's wet in Indonesia. And people worked in the farming [industry]. And the other [prosthetic limbs] that had the same mechanism design, or titanium and stainless, and couldn't get wet. And we were intentionally picking plastics that could tolerate getting wet. Because we knew that the patients needed them to be durable and useful in conditions that were consistent with their work environment. And a lot of them worked in farming or worked at other places where they did get their leg wet every day." – MED-L-1

Compared to HIC settings, where large regulatory bodies provide structure, designers in LMICs described a larger emphasis on political systems and culture. Participants described the importance of understanding local political movements (IT-L-4; CP-L-2,8) and learning and participating in local politics (IT-L-2; CP-L-3,8). In one example, however, a participant (CP-L-4) described how the end of an election cycle can halt interest in a product. Participant CP-L-8 described the incorporation of local politics through participation:

"We had to get a transformer installed on a power pole [...] for the plant to run. And that required a more regional-level government official to basically say, yes, go build this thing. We've submitted permits to the local office, but ultimately me [and my team] had to go meet with this government official at their office [...] And so that direct face-to-face connection [...] made a huge difference to get their buy-in on just installing that one small, really critical component." – CP-L-8

Finally, designers in LMIC settings incorporated institutional practices and procedures by embedding training materials within their solutions (MED-L-10; IT-L-1,4). Participant IT-L-1 explained:

"We can't assume that people have been trained to, let's say, use the mobile application. [...] Embedding training materials to be more responsive to their baseline levels of knowledge has been one design implication of this, training anytime, anywhere as opposed to assuming that everyone is awake and alert in these huge, hot, dense, food coma, training sessions that I've witnessed [...] Making that more flexible has certainly been important." – IT-L-1

Medical device designers in LMIC settings were the only ones to incorporate global priorities and the availability of consumables into their solutions. These designers described aligning their solutions with problems identified by large funding organizations like the Gate's Foundation or the United Nations Sustainable Development Goals (MED-L-1,4,5,6). They also described overcoming the cost or availability of consumables by including spare parts or minimizing the need for them (MED-L-1,2,4:8,10). For example, participant MED-L-5 included spare parts that would not be available in the context:

> "We also put a spare parts kit for commonly replaced parts [...] in a Ziploc bag inside of the unit. So if a service tech goes to take the unit apart and try to fix something, they'll have [...] most of the components that they would need to switch out. And that was also something we learned in discussions. There just aren't that many spare parts [...] And by putting those components in there, we can help them do their job." – MED-L-5

5.5.3.2 High-Income Country Settings

In HIC settings, institutional factors were most incorporated into design processes. First, they incorporated institutional practices and procedures by designing different modes of use within the same solution for different end-use contexts (IT-H-4,6,7; CP-H-6). For example, participant IT-H-6 described how different living situations for the elderly would define the mode of use for their solution:

> "So that was one of our first questions was what environment are they currently in? It does define a lot of the experience moving forward. [...] These were the three different categories, are they living with you, living at home, their own home? Is it a living facility? [...] So all of which are different because there's a whole series of daily tasks. [...] So that was something that was in the intake survey. It defined a lot of the experience of how you'd be using it." – IT-H-6

Compared to LMIC settings, where designers incorporated institutional practices and procedures by embedding training materials, designers in HIC settings opted to simplify product features (MED-H-1,5,8). For example, participant MED-H-1 described the simplification of a technology to avoid additional training:

> "But if it was something that had a remote controlled scope at the end, that's where the surgeon would feel more comfortable than the anesthesiologist. And in talking with some anesthesiologists, we showed them, we said, "Would you want a little scope when you're placing the device, on the end?" And they said, "To be honest, no, because I've never used a scope before, I don't know how to place them. So if I were to use a device like this, I'd probably have to go through some type of training to be able to use it, which just isn't super ideal." – MED-H-1

Finally, designers in HIC settings incorporated institutional factors to increase the usability of their solutions. Some designers incorporated institutional practices and procedures (MED-H-4,5,9; IT-H-2,7), while others incorporated factors related to the indoor environment (MED-H-2,3,7; IT-H-5). For example, participant MED-H-7 designed a feature to increase usability in the darkness of a catheter lab:

> "A cath lab is usually a darker room with bright monitors shining directly at the physician. So their ability to look down and see some barely visible mark is very limited. So sometimes from a design perspective, we tend to implement some other forms like, oh, there's instead of a slider that needs to go to a specific location, and so the physician sees it and stops, we put a button there. So when he gets to that slide, it clicks the button and then maybe he hears it or maybe just straight up stops, and if he needs to go further, he needs to click the button and move forward. All of it is to really make sure that the physician looks at the things he wants to look at and not at the things he's forced to look at in your design." – MED-H-7

In addition to institutional factors, in HIC settings, designers frequently designed to the extreme to ensure transferability across contexts. This strategy was common when incorporating aspects of the natural environment like pressure, temperature, and humidity (MED-H-2,5,10) and meeting higher regulatory requirements than what was required in the end-use context (MED-H-1,2,3,4,6; CP-H-3,8). For example, participant MED-H-5 described designing to the highest environmental standard:

> "These devices are used in the pre-market space, and so you have to consider that they could be at different elevations. So the pressure difference is a big thing, the temperature difference, so what is your operating temperature going to be? It's very different if they're down in Florida with super high humidity versus being up in New York where it's cold. [...] Those we tend to pretty much design to essentially the highest standard." – MED-H-5

Designers in both LMIC and HIC settings incorporated health demographics from the end-use context into their designs, but designers in HIC settings had two unique considerations: the size of users in contexts where obesity is more common (MED-H-1,2) and tailoring solutions to elderly users in the workforce (MED-H-3; IT-H-5,6). For example, participant MED-H-2 described how the form factor of their solution changed based on context:

> "So in America especially and then you start going more and more south, the obesity population starts increasing, so it's actually a certain length of the instrument you have to consider depending on what area you're in. For most of the world, especially for Europe and Asia, your instruments don't have to be necessarily as long navigation wise" – MED-H-2

Finally, designers in HIC settings incorporated sociocultural factors in unique ways through their design processes. For example, designers incorporated cultural tradition and

practices by designing features that aligned with community preferences (CP-H-2,3,5,7,8). Participant CP-H-2 described community perceptions of cleanliness when washing clothes:

> "So for example, our South American market, we use the most water per cycle, per average, twice as much water compared to other regions simply because consumers think that their clothes are getting cleaner when they see more water in the washing machine. Even though our studies show that that's not always the case, there's a right amount of water to effectively clean your clothes and adding more water will honestly maybe even make them not clean as well at a certain point." – CP-H-2

5.5.4 Incorporating two or more contextual factors into a single design decision

Across all domains and contexts, participants described the simultaneous incorporation of two or more contextual factors from varying classifications. Following the same trends as single-factor incorporation, this behavior was most common among medical device designers (15 of 21) and designers in LMIC settings (20 of 29). Rather than considering contextual factors in isolation when making decisions, participants described navigating a complex landscape of interdependence where factors from one classification frequently constrained or informed others. In these instances, the design decision resulted from the combination of these factors, with each playing a vital role in the final technical or procedural requirement. Institutional and political factors were most frequently incorporated alongside factors from other classifications, and political factors were incorporated alongside factors from every other classification. A summary of the interplay between contextual factor classifications is illustrated in Figure 27.

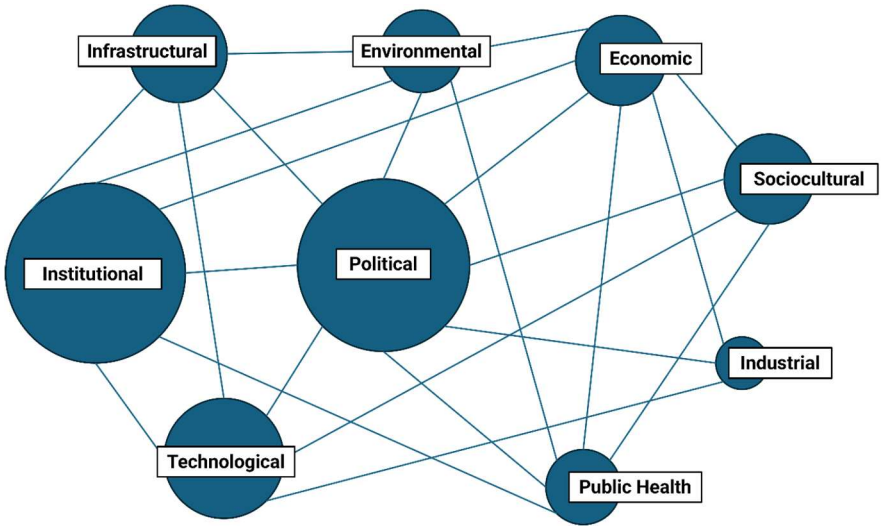


Figure 27: Multi-factor incorporation of contextual factor classifications. Node size represents the frequency with which a classification was incorporated in combination with at least one other classification. Lines connecting classifications indicate co-incorporation, illustrating the breadth of joint incorporation across the dataset.

Participants often described political factors as a filter through which other design decisions must pass through. When incorporated alongside other contextual factors, political factors often constrained solutions, overriding technical utility or user preference. For example, participant IT-H-4 described how stakeholder power dynamics were prioritized over the capacity and capability of staff members:

> "We wanted to develop a 3D view of the machine so users can rotate the machine and to see different parts of the machine. [The users] mentioned this actually can't help their work because they're very familiar with the sensor data and each sensor connected to the machines. So what they really wanted to know is the sensor reading of each sensor of the machine. [...] However, because the executive is very excited about this, we still designed and developed that version even though it was not very helpful to our end users, but this is a trade off because the exec will pay for the contract and we have to satisfy them." – IT-H-4

Similarly, political regulations can conflict with sociocultural traditions or economic realities. Participant CP-L-3 described how political regulations were in conflict with farmers' cultural values and the economic potential of producing their own crop:

> "The [economic government agency] and the [national agricultural organization], they work in tandem to maintain [crop] as a main product of export and a main economic influx for the country. [...] They're asking us to grow the seeds that they have engineered, but we know that those seeds do not benefit the soil. And that is one of our main values. We know that if we are able to [process this crop], we can sell it three x price. So we want to be able to do that instead of having to sell it to the [national agricultural organization]. [...] How can we bypass the [national agricultural organization]? How can we escape the cornering that the prerogatives that they're putting us under?" – CP-L-3

Like political factors, sociocultural factors were often prioritized in multi-factor decisions because cultural behaviors and domestic realities were complex constraints to design around. For example, participant CP-L-5 described how space-saving behaviors in small homes were at conflict with their solution's ability to capture data for government incentives:

> "In [Asian country] people normally, if they have an induction cook stove, they put it on the table, but normally once they finish cooking they will remove it from the table because they need the table for something else. It's not like they have a huge kitchen or a huge table and the [data capture technology] needs some time to communicate. [...] Well, it was bringing big struggles to this manufacturer because one of their key points of their business was collecting this data for claiming [government incentives], but the end users' habits were not allowing it." – CP-L-5

Participants also described navigating the sociocultural complexities at the intersection of public health standards. They incorporated sociocultural context by balancing "prestige" aesthetics to improve sanitation practices (CP-H-7) and designing around sociocultural stigmas

against touch that developed in response to COVID-19 (IT-H-1). Participant IT-H-6 described how disease burden can put immense pressure on family dynamics and how these factors were incorporated during the early stages of their design process:

> "And there's a lot in the news as I was working on this project that caregiving is this crazy new issue […] there's a bit of a sandwich situation like all these people in their thirties and forties who are still getting their life and their career and their family off the ground and building a life are ending up having to take care of parents because of sickness or covid or all these other situations. So they are a bit straddled by both of these factors." – IT-H-6

The interdependence of contextual factors was also commonly described as a "waterfall" effect. The effect of one contextual factor can trickle down to constrain or add complexity to a design decision. For example, participant CP-H-8 described how local weather interacted with the built environment of the home to cause a technical failure:

> "In the winter, it's cold, it's dry, and we get this effect where the [robotic vacuum] actually builds up static electricity, just like if you're scuffing along your wool and carpets, and this robot is not grounded because it is autonomous [...] What happens, it drives over one of these metal grates. It grounds itself, it gets snapped. All of our testing in [Asian country], tends to be much more humid and does not have this type of forced air ventilation system. So this was a failure mode that was never observed in any of the [Asian country] testing, but as soon as they ship it to American homes, Americans are like, yeah, my robot got hit by lightning and it doesn't work. [...] Subsequent additions of the [robot vacuum] were redesigned to have some shielding and some different details about how to dissipate that electrostatic charge." – CP-H-8

In other cases, the intersection of contextual factors can reveal unique needs in complex environments. Participant IT-L-2 incorporated institutional and infrastructural factors into a single, inseparable need:

> "In [large Asian city] traffic was a big problem. [...] You've identified a patient with a [critical cardiac event], you need to get onto care, but the ambulance with traffic… this hub center that you typically go to, the beds are full. It's going to take two hours in traffic to get to. Maybe it's better to go to this center that's a little bit further away distance-wise, but traffic-wise, it's not as congested, and they have beds available. Having this kind of data-enabled network where you're able to understand bed capacity, traffic, that was an example of how we could [optimize the network]. How do you get a patient from here to treatment, but in the most efficient way possible accounting for the local infrastructure?" – IT-L-2

Ultimately, these findings suggest that design decision-making is not a linear response to discrete factors, but a holistic negotiation of a multidimensional contextual landscape. While design decisions can be a site of conflict, where multiple categories of context compete for priority, they can also provide opportunities to discover and solve complex needs.

5.6 Discussion

The findings from this study provided evidence that incorporation of contextual factors in engineering design is not solely domain-specific but instead reflects a partially shared structure of reasoning that extends across diverse design settings. By demonstrating the generalizability of an existing taxonomy [5] and defining domain-specific nuances, these findings not only address whether contextual frameworks can be applied in new domains but demonstrate how they can support more generalizable design theory. Theoretically, this work contributes to ongoing efforts to build cumulative knowledge in engineering design [46] by identifying a set of core contextual factors that apply across domains. This agreement suggests that designers, regardless of application area, may draw on a common set of considerations when navigating context, even as these are expressed differently across domains. The introduction of domain-specific considerations further highlights that generalizability in design does not imply uniformity, but rather a layered structure consisting of shared core elements and adaptable peripheries (Fig. 28).

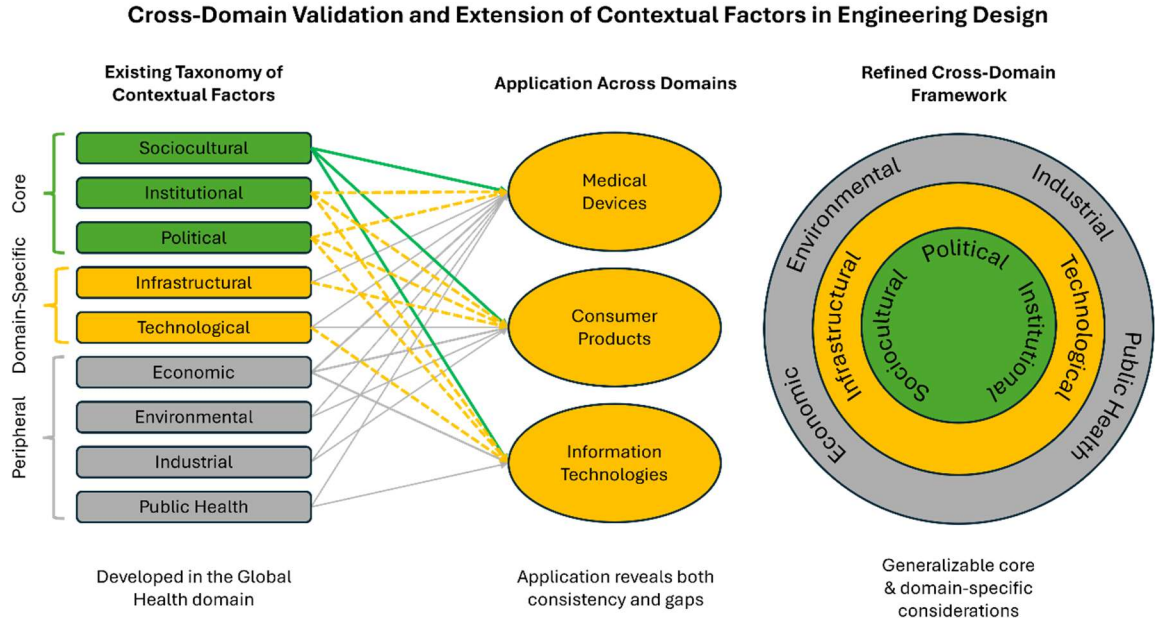


Figure 28: Cross-domain validation and extension of contextual factors in engineering design. Green factors are representative of core contextual factors that are generalizable across domains. Yellow factors are domain-specific contextual factors that are adaptable considerations within domains. Grey factors are contextual factors that may or may not be important considerations, depending on the domain, but are key peripheral explorations.

The core contextual factors incorporated by designers, regardless of domain, were sociocultural, political, and institutional classifications. This general trend aligns with the findings from Burleson et al. [5]. First, sociocultural factors were critical to product adoption

where cultural expectations and social practices within a context can shape how problems are framed and which technologies are accepted within a community [11,37,54]. Language was incorporated often, reflecting the adaptation of products to the needs and culture of local markets driven by globalization and widespread internet connectivity [55]. Participants in every domain also emphasized the importance of inclusive design. Design practices to include traditionally excluded individuals and groups by tech companies like Microsoft and Apple have enabled the innovation of accessible technologies. For example, Apple's Voice Control feature was developed for users with limited physical mobility but is now a hands-free solution for the general population [56]. Sociocultural factors shape the adoption of new technologies, regardless of the product domain, and are a core consideration in engineering design.

Though designers incorporated a common set of core factors, nuances in how these factors were adapted demonstrated key domain-specific design behaviors. Medical device designers incorporated the broadest and greatest number of contextual factors, likely explained by the complexity of clinical workflows, regulatory requirements, and cultural perceptions of health [26,27]. Medical device regulatory processes are complex, and regulatory bureaucracy can be expensive, but they are intended to ensure the safety and efficacy of devices that have the potential to sustain or end human life [4,57]. Understanding and incorporating the regulatory expectations in the end-use context is critical to the acceptance and adoption of new technologies. Clinical workflows also presented a unique complexity, commonly related to storage and cleaning. Most institutions have their own standard operating procedures for cleaning, and no universal standard exists. As a result, major differences between institutions exist, particularly across income-contexts [58]. Clinical workflows also require complex logistical coordination between stakeholders and resources, including people, equipment, materials, and waste streams [59]. Storage and location of equipment therefore vary across institutions. Even within a single institution, procedures occur within distinct indoor environments like operating rooms, catheter labs, and neonatal intensive care units; participants in this study emphasized the importance of designing features to operate within these spatial constraints. Ultimately, medical device designers are uniquely constrained by strict regulatory requirements and unique institutional workflows that must be incorporated in addition to the core context.

Consumer product participants uniquely prioritized infrastructural factors alongside core political and sociocultural factors. Unlike medical devices and information technologies that are constrained within clinical and industrial institutions, consumer products are designed for the everyday person [38]. Participants in this study incorporated users' built environments rather than the norms of an organization or institution, as domestic spaces are a common consideration for household appliances like refrigerators, vacuum cleaners, and washing machines [60]. Though consumer products are less formally regulated than medical devices or information technologies, political factors are still a core consideration because public policy and government subsidies often drive the development of new innovations [61]. Participants often aligned product goals with powerful stakeholders' interests. Sociocultural factors were distinctly incorporated by consumer product participants, like using local aesthetics to convey status and prestige. This behavior follows the close relationship between self-image and consumer behavior; how we think of ourselves and how we want others to see us shape our product choices and preferences [62]. Cultural tradition and practices were also incorporated to educate encourage behavior change in end-users in this study. Because consumer products are deeply connected to identity and culture, consumer products have the potential for drastic impact on everyday life [62]. Consumer product designers uniquely incorporated core and domain-specific context that was deeply dependent on understanding the values, beliefs, and behaviors that govern a local context.

Information technology participants prioritized technological compatibility alongside core factors related to data privacy regulations, institutional workflows, and user experiences. Information technologies are digital tools, enabled by the internet, to connect, store, and communicate information within an integrated platform [63]. Therefore, technological constraints were a more significant consideration in this domain, particularly smartphone availability and internet connectivity. Political factors related to data privacy and security were unique because the digitization of information requires information technologies to generate, process, and exchange vast amounts of data [63]. Because users expect their data to be protected, information technologies are subject to strict data privacy regulations in some contexts. Information technologies are often implemented to improve existing institutional workflows to automate tasks and support decision-making [63] or provide access to hard-to-reach resources [30,31,64] Participants in this study uniquely incorporated institutional factors into their design

processes by identifying systemic gaps in existing workflows during early stages of design and designing multiple versions of a similar solution when they identified similar needs in different contexts. Information technology design processes face distinct constraints related to the available technology alongside core data regulations and existing institutional practices and procedures within the end-use context.

Between income settings, participants mentioned a broader incorporation of contextual factors when designing for LMIC settings compared to HIC settings. LMIC settings are rich with place-based tradition, local identity, and diverse knowledge, beliefs and values [11]. Yet, these contexts are also home to the world's poorest populations. Irregular income, limited access to basic resources, and unmaintained infrastructure are common [11,44,65]. Consequently, technology is often donated to LMIC settings, but most solutions are rejected or fail in the context because they do not meet the contextual needs of the community [4,11]. Participants designing for LMIC settings considered a broader number of contextual factors to capture the unique values, behaviors, and standards in these settings. Individual and household characteristics associated with limited and irregular income are far more significant in these contexts. As is the tendency for symbolism to be incorporated into design decisions; capturing shared meanings within a culture is more critical to adoption. When considering who the designer is in relationship to the user, it is also particularly important, then, for those who are designing for contexts they are unfamiliar with to capture, organize, and synthesize the breadth of contextual factors in the end-use environment [11]. Designers in LMIC settings incorporated core, domain-specific, and peripheral factors to better explore, understand, and synthesize contextual information in these unique environments.

At the intersection of product domains and income settings, medical device participants in LMIC settings were met with the most complex and unique settings and therefore incorporated the broadest number of contextual factors. These designers navigated regulatory bureaucracy, procurement systems, global aid, and government-controlled healthcare spending [4]. Within each institution, healthcare workers are often unskilled and devices cannot be regularly maintained [4]. Creative solutions in these contexts can help overcome the complexity of the setting, requiring medical device designers to collect a broader understanding of context and incorporate many factors into their design decisions. It follows that medical device participants in LMIC settings in this study shared a broader and deeper understanding of contextual factors.

Interestingly, the existing taxonomy used in this study was developed for these complex settings, explaining its generalizability across diverse contexts [5,15].

In addition to a nuanced understanding of contextual factors, participants in this study also demonstrated nuanced decision-making behavior. While most contextual factors were mapped to a single design decision, participants described incorporating multiple contextual factors into the same design decision. This nuanced, real-world behavior was defined through the navigation of complex contextual landscapes, where factors were described as co-dependent and interrelated. Existing literature has described similar behavior, where contextual factors act as both the source of and resolution to trade-offs in design decision-making [66]. In some cases, complex design contexts paired with technical requirements and constraints can force designers to accept trade-offs that compromise design performance and, in some cases, result in termination of a project [5]. However, the complexity of these contexts can also reveal knowledge and information that can increase the degrees of freedom available to navigate these trade-offs, providing designers with opportunities to bypass trade-offs through expansion of the problem space, problem reframing, feature development, or decisions related to use and implementation [5,67]. Navigating these tradeoffs is only possible when a designer is equipped with a broad understanding of context. Designers who truly understand the contextual constraints of their design problem can be more creative, agile, and effective during decision-making [66].

The refined taxonomy and evaluation of design behavior presented in this study make key contributions to design research, practice, and education. For design research, the findings from this study present an opportunity to develop more unified frameworks for studying context in engineering design, rather than parallel taxonomies for different domains, enabling more direct comparison across studies and supporting the accumulation of contextual knowledge. Additionally, unified frameworks can support the integration of contextual considerations with broader models of design cognition and decision-making. For design practice, more generalizable contextual knowledge can inform the development of design support tools and methods that are adaptable across domains. For example, checklists, prompts, or computational tools derived from a unified framework may help designers more systematically account for context without requiring entirely new taxonomies for each application area. A broader understanding of context also supports effective decision-making, especially when navigating the tradeoffs between conflicting requirements [66]. For design education, a unified framework

suggests the value of teaching contextual awareness as a transferable skill, grounded in a shared set of core principles rather than strategies siloed by domain. Prior work showed that students tend to prioritize institutional and technological factors, but rarely consider core sociocultural or political factors [7]. Students also struggle to know when or how to incorporate contextual information, especially when that information conflicts [7]. A unified framework can help students recognize which aspects of context to consider and patterns in how context shapes design decisions, better preparing them to work across diverse problem spaces.

While this study draws on designers' experiences from specific domains and settings, the generalizability of the framework warrants further testing, as universal applicability can only be framed as a future goal. Future research should examine this refined taxonomy in additional domains and with broader populations to assess its capabilities and limitations. Additionally, this study implied patterned design behavior for incorporating contextual factors into design decisions. Future research should explore, in more detail, how contextual information is incorporated into decision-making and when throughout a design process contextual information can be incorporated into specific design decisions. Overall, this work advances engineering designers' ability to consider and incorporate context into their design processes by moving beyond descriptions siloed by domain toward a more generalizable and extensible framework. By identifying both shared considerations and domain-specific nuances, the findings from this study encourage the accumulation of contextual knowledge and the development of design methods and educational approaches that are both robust and adaptable across domains.

5.7 Conclusion

Engineering design is inherently shaped by the end-use context of a potential solution. However, frameworks used to characterize these contextual factors are often developed within a single domain, limiting their broader applicability. This study evaluated the validity of existing contextual frameworks [5,15] across diverse domains, resulting in a refined taxonomy that integrates generalizable core factors with domain-specific nuances. This refined taxonomy and the design behaviors observed in this study enable researchers to accumulate generalizable design knowledge, empowers practitioners to systematically account for context using evidence-based tools, and provides educators with a foundation for teaching contextual awareness as a transferable skill. Ultimately, these insights advance the theoretical understanding of contextual

reasoning, bridging the gap between theory and practice through the development of adaptable methods, tools, and educational frameworks that better account for context.

5.8 Appendix A: Contextual Factor Cards

Technological *Related to existing technology that directly affects performance, adequate operability, and functionality* Ex: Compatibility with the technical context, available technologies in the context, availability of consumables	Industrial *Broader technical enterprises that are required for the artifact's performance or operability* Ex: Supply & manufacturing, maintenance, distribution	Institutional *Constraints of norms within the organization or institution where the solution must operate* Ex: Existing practices/procedures, capacity and capability of staff, institutional financial capacity, institutional resources, indoor environment
Economic *Context-specific economic capacity, structures, and individual and household attributes* Ex: individual and household characteristics, regional and national characteristics, labor market characteristics	Environmental *External factors for functionality and operability within the natural environment* Ex: Weather / climate, Natural environment (e.g., water quality, air quality, dust)	Infrastructure *External factors for functionality and operability within the built environment* Ex: Utilities available, Transportation and road quality, Distance between resources, Attributes of the built environment (e.g., commonly used building materials)
Public Health *Local health status and health outcomes of specific groups of people, as well as organized local efforts to prolong life and prevent disease*	Socio-cultural *Social & cultural characteristics of individuals and groups (behaviors, norms, knowledge, beliefs, arts, customs, etc.)* Ex: Symbols, language, education/literacy, aesthetics,	Political *Systems, institutions, and culture of decision-making and forms of power* Ex: regulations and regulatory processes (e.g., policies, regulations, initiatives), stakeholder power dynamics (e.g., influential stakeholders), political systems and culture (formal and informal mechanisms of decision-making), global priorities

Ex: Healthcare system and practices, Health demographics	cultural traditions/practices, stigmas and taboos	(e.g., Gates funding priorities, WHO recommendations)

5.9 Appendix B: Example Questions From Interview Protocol

- To ground some of our questions today, I want to talk about one of your previous projects where you learned a lot about your solution's end users and setting and used that information to make design decisions. Can you think of a specific instance or project that could be a good example of your work that we could use throughout the interview?
- What motivated the project? What are the project objectives?
- Can you tell me who your end user is and what you know about them? How did you go about learning about your end user? How did you apply that information about your end user's context to your design process?
- Who were the key stakeholders and partners? What role did they play in making decisions? What perspective, expertise, or resources did they provide? How did you incorporate that information into your design? How did your stakeholders influence your design or process?
- Can you tell me more about the setting or environment that the final product is used in? How did you incorporate those contextual factors into the final design or design process? Were you familiar with this context? How do different contexts differ between different end use cases?
- Did other existing technology in the context influence your design? Did your design interface with existing technology in the context?
- Did you consider the availability or quality of certain manufacturing methods? Did the types of available maintenance influence your design?
- Did you change your design to accommodate the staff who would be using it? Did the typical procedures at the [institution] influence your design in any way?
- Was the setting resource-constrained or abundant?
- Did the weather or natural conditions affect your design?
- Did utilities available in that area or the quality of transportation affect the way your product was designed?
- Did you make any design decisions based on public health?
- Did you run into any stigmas or taboos that changed your design? Culturally, were there any symbols, aesthetics, or languages used specific to that context?
- Were there any regulations you had to adhere to that influenced your design process? Were there stakeholders that influenced the decisions you made or exercised power over the design? Were peoples' interests influenced by global initiatives or government desires?

5.10 Appendix C: Selected Themes of Incorporation of Contextual Factors Common Across Domain and Income Settings

Category	Theme	Frequency	Detailed Considerations

Environmental	Material selection to withstand weather conditions (n=21)	Less common among IT designers (n=2) More common among designers in LMICs (n=18)	Environmental exposure is common during distribution, even if the end use context is controlled Exposure to harsh environments in everyday life is more common in LMICs
Technological	Incorporate available technology (n=17)	Most common among IT designers (n=10)	Consider availability of smartphones and internet connectivity Consider technology that is NOT available
Sociocultural	Use context-specific languages on products and manuals (n=14)	Common across domains and contexts	Critical to test translated text, especially for longer languages and unfamiliar characters Multi-language support can help reach more users
Infrastructure	Include features so fluctuating power does not damage product components (n=12)	Common across domains and contexts	Consider using power supplies tolerant to fluctuations, reducing electricity reliance, using rechargeable batteries, or design for a backup power source (e.g., solar)
Sociocultural	Design for dignity by respecting cultural stigmas and taboos (n=11)	Common across domains and contexts	Make design decisions to help users blend into the end-use context Incorporate features to protect user privacy
Political	Prioritize global regulations (n=14)	More common among MED (n=9) and CP (n=4) designers	Meet stricter standards to ensure transferability to other contexts Be aware that global regulations may not exist in the context Avoid regulatory hurdles by showing equivalence to existing products or using off-the-shelf components
Political	Prioritize local regulations (n=13)	More common among IT designers (n=9)	Consider data privacy regulations

Economic	Set cost requirements based on individual and household income (n=9) or national and regional characteristics (n=8)	Individual and household income considered only by designers in LMICs (n=9)	Strategies to meet cost requirements included setting different prices in different contexts for the same design, designing multiple versions of the same product at different price points, or utilizing government sponsorship/subsidy or financing plans Pricing products too low can generate suspicion of quality Making products reusable can drive down cost, but may increase regulatory burden
Institutional	Set cost requirements based on institutional financial capacity (n=9)	Common across domains and contexts	

5.11 References

[1] Jagtap, S., 2019, "Design and Poverty: A Review of Contexts, Roles of Poor People, and Methods," Res. Eng. Des., 30(1), pp. 41–62. https://doi.org/10.1007/s00163-018-0294-7.

[2] Mattson, C. A., and Wood, A. E., 2014, "Nine Principles for Design for the Developing World as Derived from the Engineering Literature," Journal of Mechanical Design, 136(12). https://doi.org/10.1115/1.4027984.

[3] Wood, A. E., and Mattson, C. A., 2016, "Design for the Developing World: Common Pitfalls and How to Avoid Them," Journal of Mechanical Design, 138(3). https://doi.org/10.1115/1.4032195.

[4] Overhoff, M., Lumpe, T. S., and Shea, K., 2023, A Product Life Cycle Approach to Medical Device Development for Low-Resource Settings - A Systematic Literature Review.

[5] Burleson, G., Wojciechowski, L., Toyama, K., and Sienko, K. H., 2024, "Characterizing the Use of Contextual Factors in Engineering Design: An Exploration of Global Health Designer Practice," Design Science, 10. https://doi.org/10.1017/dsj.2024.28.

[6] Mosyjowski, E., Espinoza von Bischhoffshausen, J., Lattuca, L., and Daly, S., 2020, "Student and Practitioner Approaches to Systems Thinking: Integrating Technical and Contextual Considerations," ASEE Annual Conference, American Society of Engineering Education.

[7] Burleson, G., Herrera, S. V. S., Toyama, K., and Sienko, K. H., 2023, "Incorporating Contextual Factors Into Engineering Design Processes: An Analysis of Novice Practice," Journal of Mechanical Design, 145(2). https://doi.org/10.1115/1.4055780.

[8] Catchpole, K. R., Hallett, E., Curtis, S., Mirchi, T., Souders, C. P., and Anger, J. T., 2018, "Diagnosing Barriers to Safety and Efficiency in Robotic Surgery," Ergonomics, 61(1), pp. 26–39. https://doi.org/10.1080/00140139.2017.1298845.

[9] Koppel, R., Wetterneck, T., Telles, J. L., and Karsh, B. T., 2008, "Workarounds to Barcode Medication Administration Systems: Their Occurrences, Causes, and Threats to Patient Safety," Journal of the American Medical Informatics Association, 15(4), pp. 408–423. https://doi.org/10.1197/jamia.M2616.

[10] Yu, J., Ferniany, W., Guthrie, B., Parekh, S. G., and Ponce, B., 2016, "Lessons Learned from Google Glass: Telemedical Spark or Unfulfilled Promise?," Surg. Innov., 23(2), pp. 156–165. https://doi.org/10.1177/1553350615597085.

[11] A.P. Witmer, 2022, Contextual Engineering: Translating User Voice into Design, Springer.

[12] Heylighen, A., and Dong, A., 2019, "To Empathise or Not to Empathise? Empathy and Its Limits in Design," Des. Stud., 65, pp. 107–124. https://doi.org/10.1016/j.destud.2019.10.007.

[13] Holeman, I., and Kane, D., 2020, "Human-Centered Design for Global Health Equity," Inf. Technol. Dev., 26(3), pp. 477–505. https://doi.org/10.1080/02681102.2019.1667289.

[14] Drain, A., Shekar, A., and Grigg, N., 2021, "Insights, Solutions and Empowerment: A Framework for Evaluating Participatory Design," CoDesign, 17(1), pp. 1–21. https://doi.org/10.1080/15710882.2018.1540641.

[15] Aranda-Jan, C., Jagtap, S., and Moultrie, J., 2016, "Towards A Framework for Holistic Contextual Design for Low-Resource Settings," International Journal of Design, 10(3), pp. 43–63. [Online]. Available: www.ijdesign.org.

[16] Atman, C. J., 2019, "Design Timelines: Concrete and Sticky Representations of Design Process Expertise," Des. Stud., 65, pp. 125–151. https://doi.org/10.1016/j.destud.2019.10.004.

[17] Green, M. G., Rajan, P., and Wood, K. L., 2004, "PRODUCT USAGE CONTEXT: IMPROVING CUSTOMER NEEDS GATHERING AND DESIGN TARGET SETTING," ASME Design Engineering Technical Conferences and Computers and Information in Engineering Conference, ASME, Salt Lake City. [Online]. Available:

http://asmedigitalcollection.asme.org/IDETC-CIE/proceedings-pdf/IDETC-CIE2004/46962a/393/2616871/393_1.pdf.

[18] Hales, C., and Gooch, S., 2004, "Engineering Design Process: Review and Analysis," Managing Engineering Design, Springer, London, pp. 233–240.

[19] Jagtap, S., 2022, "Co-Design with Marginalised People: Designers' Perceptions of Barriers and Enablers," CoDesign, 18(3), pp. 279–302. https://doi.org/10.1080/15710882.2021.1883065.

[20] Disselkamp, J. P., Seidenberg, T., Westphal, S., Lick, J., Ptock, L., Wyrwich, F., Hovemann, A., and Dumitrescu, R., 2025, "Integrative and Integrated Product and Production System Development: A Taxonomy for Managing Dependencies and Processes," Proceedings of the Design Society, Cambridge University Press, pp. 2121–2130. https://doi.org/10.1017/pds.2025.10226.

[21] Zani, C. M., Brandalise, F. M. P., Pedó, B., Formoso, C. T., Tommelein, I. D., Ortiz-Marcos, I., Viana, D. D., Bataglin, F. S., and Koskela, L., 2025, "Devising a Taxonomy for Visual Management Requirements by Considering Different Purposes and Contexts," Construction Management and Economics, 43(7), pp. 518–536. https://doi.org/10.1080/01446193.2025.2476396.

[22] Duan, J., Evans, M., Hurn, K., Storer, I., and Bai, Z., 2024, "A Creative Industrial Design Framework of the Taxonomy for Chinese Indigenous Materials and Relevant Crafts," Humanit. Soc. Sci. Commun., 11(1). https://doi.org/10.1057/s41599-024-02768-1.

[23] Duarte Poblete, S. S., Romani, A., and Rognoli, V., 2024, "Emerging Materials for Transition: A Taxonomy Proposal from a Design Perspective," Sustainable Futures, 7. https://doi.org/10.1016/j.sftr.2024.100155.

[24] Sierra-Fontalvo, L., Ruiz-Pastor, L., Gonzalez-Quiroga, A., and Mesa, J. A., 2024, "Decoding Product Obsolescence: A Taxonomic Approach from Product Design Attributes," J. Clean. Prod., 475. https://doi.org/10.1016/j.jclepro.2024.143635.

[25] Martin, J. L., Norris, B. J., Murphy, E., and Crowe, J. A., 2008, "Medical Device Development: The Challenge for Ergonomics," Appl. Ergon., 39(3), pp. 271–283. https://doi.org/10.1016/j.apergo.2007.10.002.

[26] Aimer, O., and Baldridge, C., 2025, “Navigating Medical Device Safety: Current Status, Challenges, and Future Regulatory Directions,” Drug Saf. https://doi.org/10.1007/s40264-025-01599-6.

[27] Beitia Kraemer Moraes, A. C., Yusef Bakri, S. F., Karlinski Scherer, C., Daneluz Vaz, L. O., Yonara da Silva Galhardo, J., Granemann Souza, E., Nascimento, C. das D. do, and Lund, R. G., 2026, “Systematic Review of Usability Evaluations in Medical Devices: Methodological Choices, Heuristic Application, and Confounding Factors,” Appl. Ergon., 134. https://doi.org/10.1016/j.apergo.2025.104715.

[28] World Health Organization, 2010, Medical Devices: Managing the Mismatch: An Outcome of the Priority Medical Devices Project, World Health Organization.

[29] Lugossy AM, Malkin R, and Teninty B, 2025, “Donations of Medical Imaging Equipment in Global Health,” Radiology in Global Health: Strategies, Implementation, and Applications, Springer.

[30] Thakur K., Pathan AS.K., and Ismat S., 2023, “An Overview of ICT Technology Advancement,” Emerging ICT Technologies and Cybersecurity, Springer.

[31] Umidjon, T., and Ugli, S., 2026, CENTRAL ASIAN JOURNAL OF INNOVATIONS ON TOURISM MANAGEMENT AND FINANCE Smart Tourism Concept: Innovative Significance and Future Prospects in Modern Tourism Development. [Online]. Available: https://cajitmf.centralasianstudies.org/index.php/CAJITMF154.cajitmf.centralasianstudies.org/index.php/CAJITMF.

[32] Büscher, M., Perng, S. Y., and Liegl, M., 2019, “Privacy, Security, and Liberty: ICT in Crises,” Cyber Law, Privacy, and Security: Concepts, Methodologies, Tools, and Applications, IGI Global, pp. 248–266. https://doi.org/10.4018/978-1-5225-8897-9.ch013.

[33] Canedo, E. D., Do Vale, A. P. M., Patrão, R. L., de Souza, L. C., Gravina, R. M., Dos Reis, V. E., Mendonça, F. L. L., and de Sousa, R. T., 2020, “Information and Communication Technology (ICT) Governance Processes: A Case Study,” Information (Switzerland), 11(10), pp. 1–28. https://doi.org/10.3390/info11100462.

[34] Deichmann, U., Goyal, A., and Mishra, D., 2016, “Will Digital Technologies Transform Agriculture in Developing Countries?,” Agricultural Economics (United Kingdom), 47, pp. 21–33. https://doi.org/10.1111/agec.12300.

[35] Aftab, M., and Ismail, I., 2015, Defeating Poverty through Education: The Role of ICT.

[36] Chib, A., Van Velthoven, M. H., and Car, J., 2015, "MHealth Adoption in Low-Resource Environments: A Review of the Use of Mobile Healthcare in Developing Countries," J. Health Commun., 20(1), pp. 4–34. https://doi.org/10.1080/10810730.2013.864735.

[37] Toyama, K., 2015, Geek Heresy: Rescuing Social Change from the Cult of Technology, PublicAffairs.

[38] Norman, D., 2013, The Design of Everyday Things: Revised and Expanded Edition, Basic books.

[39] Homburg, C., Schwemmle, M., and Kuehnl, C., 2015, "New Product Design: Concept, Measurement, and Consequences," J. Mark., 79(3), pp. 41–56. https://doi.org/10.1509/jm.14.0199.

[40] Gilal, N. G., Zhang, J., and Gilal, F. G., 2018, "Linking Product Design to Consumer Behavior: The Moderating Role of Consumption Experience," Psychol. Res. Behav. Manag., 11, pp. 169–185. https://doi.org/10.2147/PRBM.S161384.

[41] Cooper, R. G., 2011, Winning at New Products: Creating Value through Innovation, Basic books.

[42] Singh, J., Cooper, T., Cole, C., Gnanapragasam, A., and Shapley, M., 2019, "Evaluating Approaches to Resource Management in Consumer Product Sectors - An Overview of Global Practices," J. Clean. Prod., 224, pp. 218–237. https://doi.org/10.1016/j.jclepro.2019.03.203.

[43] Jagtap, S., Larsson, A., and Kandachar, P., 2013, "Design and Development of Products and Services at the Base of the Pyramid: A Review of Issues and Solutions," Int. J. Sustainable Society, 5(3), pp. 207–231.

[44] World Bank, 2017, World Development Indicators (English), Washington, D.C.

[45] Jagtap, S., and Larsson, A., 2013, "Design of Product Service Systems at the Base of The Pyramid," pp. 581–592. https://doi.org/10.1007/978-81-322-1050-4_46.

[46] Cross, N., 2025, Designerly Ways of Thinking and Knowing, Springer.

[47] World Bank Group, 2024, "World Bank Country Classifications by Income Level for 2024-2025."

[48] Etikan, I., 2016, “Comparison of Convenience Sampling and Purposive Sampling,” American Journal of Theoretical and Applied Statistics, 5(1), p. 1. https://doi.org/10.11648/j.ajtas.20160501.11.

[49] Patton, M. Q., 2018, “Expert Sampling,” The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation, SAGE Publications, Inc., pp. 648–649.

[50] Hsieh, H. F., and Shannon, S. E., 2005, “Three Approaches to Qualitative Content Analysis,” Qual. Health Res., 15(9), pp. 1277–1288. https://doi.org/10.1177/1049732305276687.

[51] O’Connor, C., and Joffe, H., 2020, “Intercoder Reliability in Qualitative Research: Debates and Practical Guidelines,” Int. J. Qual. Methods, 19. https://doi.org/10.1177/1609406919899220.

[52] Moss, P. A., and Haertel, E. H., 2016, “Engaging Methodological Pluralism,” Handbook of Research on Teaching, D. Gitomer, and C. Bell, eds., AERA, Washington D.C., pp. 127–247.

[53] Thomas, D. R., 2006, “A General Inductive Approach for Analyzing Qualitative Evaluation Data,” American Journal of Evaluation, 27(2), pp. 237–246. https://doi.org/10.1177/1098214005283748.

[54] Sabet Sarvestani, A., and Sienko, K. H., 2013, User-Based Design Approach to Develop a Traditional Adult Male Circumcision Device.

[55] Sudirjo, F., 2023, “Marketing Strategy in Improving Product Competitiveness in the Global Market,” Journal of Contemporary Administration and Management (ADMAN), 1(2), pp. 63–69. https://doi.org/10.61100/adman.v1i2.24.

[56] Gilbert, R. M., 2025, “Building Inclusive Systems: From Planning to Implementation,” Inclusive Design for a Digital World, Apress, Berkeley, CA, pp. 305–338.

[57] Bianchini, E., and Mayer, C. C., 2022, “Medical Device Regulation: Should We Care About It?,” Artery Res., 28(2), pp. 55–60. https://doi.org/10.1007/s44200-022-00014-0.

[58] Dancer, S. J., 2023, “Hospital Cleaning: Past, Present, and Future,” Antimicrob. Resist. Infect. Control, 12(1). https://doi.org/10.1186/s13756-023-01275-3.

[59] Ransolin, N., Saurin, T. A., Clay-Williams, R., Formoso, C. T., Rapport, F., and Cartmill, J., 2025, “Beyond the Operating Room: Built Environment Design Knowledge

Supportive of Resilient Surgical Services," Engineering, Construction and Architectural Management, 32(6), pp. 3804–3831. https://doi.org/10.1108/ECAM-10-2023-1063.

[60] Spiliotopoulos, C., Stamminger, R., and Siderius, H. P., 2019, "Bringing the Home in the Lab: Consumer-Relevant Testing for Household Electrical Products," Energy Effic., 12(1), pp. 281–298. https://doi.org/10.1007/s12053-018-9718-5.

[61] Tian, J., Wang, P., and Zhu, D., 2024, "Overview of Chinese New Energy Vehicle Industry and Policy Development," Green Energy and Resources, 2(2). https://doi.org/10.1016/j.gerr.2024.100075.

[62] Kimmel, A. J., 2015, People and Products: Consumer Behavior and Product Design., Routledge.

[63] Aceto, G., Persico, V., and Pescapé, A., 2019, "A Survey on Information and Communication Technologies for Industry 4.0: State-of-the-Art, Taxonomies, Perspectives, and Challenges," IEEE Communications Surveys and Tutorials, 21(4), pp. 3467–3501. https://doi.org/10.1109/COMST.2019.2938259.

[64] Beardsley, S. C., Enriquez, L., Bonini, S., Sandoval, S., and Brun, N., 2010, Fostering the Economic and Social Benefits of ICT.

[65] Radhakrishnan, P., and Arunachalam, R., 2017, Study on Factors Responsible for Shifting of Rural Youth from Agriculture to Other Occupation.

[66] Nickel, J., Hurst, A., and Duimering, P. R., 2024, "Contextual Influences on Trade-Offs in Engineering Design: A Qualitative Study," Design Science, 10. https://doi.org/10.1017/dsj.2024.34.

[67] Nickel, J., Duimering, P. R., and Hurst, A., 2022, "Manipulating the Design Space to Resolve Trade-Offs: Theory and Evidence," Des. Stud., 79. https://doi.org/10.1016/j.destud.2022.101095.

Chapter 6 Discussion, Implications, and Future Work

This chapter presents a summary of the findings and contributions of this dissertation. First, the main findings from Chapters 2-5 are summarized and discussed based on their common themes. Then, the implications of this dissertation are organized by their domains of impact. Finally, this chapter concludes with a discussion of limitations, potential future work, and overall conclusions.

6.1 Chapter Summaries

In this dissertation, I investigated the use of wearable inertial measurement units (IMUs) for fetal movement (FM) detection to support the development of a continuous, objective fetal monitoring system. While wearable sensors enabled by machine learning (ML) show promise, progress toward consumer or clinical deployment remains limited by inconsistent methodological benchmarking and a failure to account for model generalizability. To address these gaps, this work (1) established evidence-based benchmarks for signal processing and algorithm evaluation and (2) defined a contextual framework for design-decision making for future systems.

Chapters 2 through 4 of this dissertation focused on the application of wearable IMUs and ML methods for rigorous characterization, classification, and evaluation of FM events. Chapter 2 aimed to quantify the impact of augmenting accelerometer-based systems with angular rate data across three data representations and ML architectures for detecting FM events with wearable IMUs. The study population included 23 participants carrying healthy, singleton pregnancies. Three datasets were utilized for the comparison of accelerometer and sensor fusion approaches: acceleration data, angular rate data, and a combination of both data types. These datasets were used to benchmark ML models (random forest RF, bi-directional long short-term memory BiLSTM, and convolutional neural network CNN models) for classification of FM events. All models learned discriminative features of FM as evidenced by their test performance on held-out trial data (mean AUROC > 0.73). RF models enabled by the combined dataset achieved the highest performance (AUROC = 0.81). However, FM events in this study were

defined by maternal perception, which inherently is subjective and may not demonstrate a model's true ability to detect FM events. The algorithms benchmarked in this study supported further work in Chapter 4 to evaluate the reliability of maternal perception as a clinical standard for model training.

Chapter 3 aimed to use clinician-validated ultrasound data to precisely define FM events, enabling the characterization of IMU signals for data-driven signal processing and hardware design parameters to support future algorithm development. Recordings of ultrasound imaging from 16 study participants were annotated with time-series labels to define four event types: weak FMs, strong FMs, maternal movement artifacts, and stationary periods with no FMs. These annotations were used to segment and group IMU signals for statistical analysis with linear mixed-effects models. Statistically significant differences between the estimated marginal means of each group were used to characterize FM events and define data-driven parameters for future algorithm development: minimum hardware sensitivity to reliability capture weak FMs, artifact threshold and frequency cutoffs to inform filter design for artifact rejection, and window size based on duration of FM events. The data-driven parameters derived in this study informed the algorithms used in Chapter 4 for benchmarking model training label sources and evaluation approaches.

Chapter 4 aimed to evaluate the reliability of maternal perception as a clinical standard for model training and benchmark generalizable model evaluation approaches. A subset of the study population from Chapter 2 (16 of 23 participants) received ultrasound imaging during additional data collection trials for comparison with maternal perception-labeled events. RF models were trained with data labeled by maternal perception or ultrasound and evaluated on held-out test data labeled with ultrasound. Two data splitting approaches were compared: (1) one trial from each participant contained in the training and testing sets and (2) holding out all trials from a single participant in the testing set (leave-one-subject-out LOSO). Random sampling of windows was tested for comparison to existing studies. Models trained with ultrasound labels outperformed those trained with maternal perception labels under both data splitting approaches. Still, both models had low specificity (< 72%) due to a high rate of false positive predictions, limiting their deployment to unsupervised settings. Models had a limited ability to generalize due to subject-specific physiological factors, evidenced by a drop in performance when using the LOSO approach. The random sampling approach showed significant performance improvements

(AUROC = 0.75), demonstrating the over-optimism of algorithm capability by existing studies in the field. Ultimately, the findings from Chapter 4 provided a rigorous evaluation of the capabilities and limitations of wearable IMU-based systems for deployment to clinical or consumer settings.

Chapter 5 of this dissertation focused on the contextual considerations that may impact the design of fetal monitoring systems. Though technical rigor is necessary for a system to work, it is not sufficient for a solution to be adopted or effective. Incorporating context into technical design decisions supports the likelihood of success in clinical or consumer settings, which is a current limitation of existing fetal monitoring systems.

In Chapter 5, I adapted an existing, domain-specific framework of contextual factors to capture a shared structure of contextual reasoning across diverse design contexts. The existing framework was applied across three product domains (e.g., medical devices, information technology, and consumer products) and two income settings (e.g., low- and middle-income countries LMIC and high-income countries HIC). Summative content analysis supported the identification of patterned designer behavior within each domain and context. The findings from this study demonstrated that sociocultural, political, and institutional factors were core considerations that generalized across domains. However, nuances in designer behavior showed that medical device designers incorporated strict regulatory constraints and integrated their solutions into existing clinical workflows. Information technology designers ensured their solutions were compatible with existing technology and worked to automate or organize existing institutional processes. Consumer product designers adapted their solutions to integrate with domestic spaces and everyday behaviors. Designers in LMIC settings were often constrained by cost requirements and sociocultural perceptions, while designers in HIC settings more often incorporated existing institutional procedures and accommodated powerful stakeholders. The findings from this study suggest that designer behavior is generalizable but not uniform, and the adaptable framework supports future research efforts to systematically incorporate contextual factors into the design of fetal monitoring tools, increasing the likelihood of adoption into clinical and consumer settings.

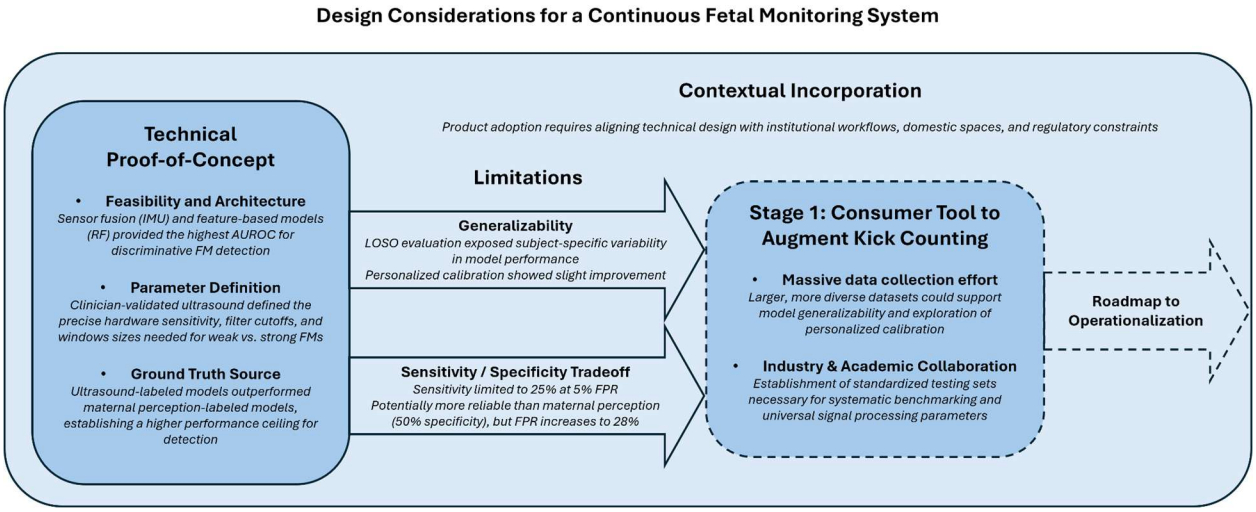


Figure 29: Summary of findings and their implications for future work. Concepts outlined with solid lines illustrate contributions of this dissertation. Concepts outlined with dashed lines are potential next steps informed by the findings from this work.

6.2 Discussion

The four studies presented in this dissertation demonstrated the technical capabilities and limitations of wearable IMU-based systems to support continuous fetal monitoring, as well as contextual considerations for deployment in clinical and consumer settings (Fig. 29). The following sections describe overarching themes that draw from the findings across the presented studies.

6.2.1 Controlled Benchmarking for Technical Development of FM Detection Algorithms

This dissertation establishes a technical framework for FM detection by addressing the critical lack of methodological standardization in sensor selection, signal processing, and model evaluation. Despite over ten years of research, Ghosh et al. [1] noted the "lack of commercial solutions and inconsistencies in sensor types, algorithmic settings, and validation protocols" have severely complicated the interpretation and standardization of wearable sensor-detected FM events. In ML, benchmarking on standardized datasets is best practice for making methodological choices among existing literature, but establishing these benchmark datasets can take many years, as was the case in the more mature field of continuous ECG monitoring [2]. In the absence of a benchmark dataset, a common testing set within a study can be used to evaluate methodological tradeoffs [3,4]. Because there is no benchmark dataset for fetal movement detection, common testing sets were carefully derived within the comparative studies of this dissertation.

First, the data collection system and model architecture were benchmarked and refined in Chapters 2 and 3 to establish system feasibility before investigating the tradeoffs between algorithmic approaches in Chapter 4. Combining acceleration and angular rate signals improved the ability to distinguish fetal movement events (Ch. 2), suggesting that sensor fusion approaches may improve model performance over existing single sensor approaches [1,5-15]. A recent study by Ghosh et al. [1] echoed these results, demonstrating that combining accelerometer, acoustic, and piezoelectric sensors improved model sensitivity to FM. Additionally, feature-based models outperformed deep learning alternatives (Ch. 2), suggesting that features are more robust than raw data representations, which may overfit to background noise in the presence of physiological noise and limited dataset size. However, larger datasets may benefit from deep learning architectures [11], encouraging future benchmarking of model architecture when more data is available for training. After benchmarking, using combined datasets to train feature-based models could support further evaluation of algorithmic approaches. To further refine the methodological rigor of algorithm development, signal processing parameters were derived from clinician-validated FM events (Ch. 3). Existing studies cite varying filter cutoffs [6,10,13,16] and window sizes [6, 11-13] without formal justification or contextualization of their dataset. The statistical methods used in Chapter 3 could be repeated in future studies to define and justify these parameters, contextualizing unique datasets and increasing the interpretability of model decisions.

With a benchmarked data collection system and model architecture and refined signal processing parameters, the tradeoffs between algorithmic approaches were able to be evaluated. Models trained with ultrasound labels had an increased ability to distinguish FM events and correctly classify them compared to models trained with maternal perception labels when both models were evaluated against the same ultrasound-labeled testing set (Ch. 4). While studies have acknowledged the subjectivity of maternal perception [6,17], none have directly compared maternal perception with ultrasound for model training to understand and evaluate the tradeoffs in classification performance. Additionally, models that were able to learn from subject-specific signal characteristics during training (AUROC = 0.65) outperformed those that were trained to learn generalizable patterns (AUROC = 0.61) (Ch. 4). Most existing studies have randomly sampled windows, including from the same time-series trials, to split that dataset into training and testing sets. This practice does not reflect realistic deployment scenarios and can inflate

performance estimates by allowing highly correlated data from the same participant or trial to appear in both sets, thereby limiting generalizability. When compared to the training approaches evaluated in Chapter 4, random sampling resulted in significantly better performance (AUROC = 0.75) (Ch. 4), which can lead to misinterpretation of how an algorithm might perform in real-world settings. By replacing one-off designs with intentional benchmarking, this work provides the rigorous foundation necessary to move wearable fetal monitoring from proof-of-concept to a reproducible clinical standard.

6.2.2 Role of Ultrasound to Promote Algorithm Explainability through Reduced Uncertainty in Event Labels

Beyond technical performance, the integration of ultrasound data resolves the inherent ambiguity of maternal perception, providing a high-fidelity clinical standard for both model training and evaluation. FM events defined by ultrasound annotations had precise onset and completion time points (≥ 0.5s in duration) (Ch. 4). In contrast, FM events defined by maternal perception required wide data buffers (3.5s prior to and 1.5s after annotations) to account for maternal reaction delays and event duration, which introduced uncertainty in the precision of the labels (Ch. 2). Label uncertainty can dilute the FM event signal with noise and the boundary values between features for classification tasks [18-20]. Using ultrasound-labeled events for model training may reduce uncertainty and promote models to make more confident and accurate decisions. Not only do uncertain event boundaries reduce confidence and accuracy, but maternal perception-labels may also miss some FM events entirely. Maternal perception sensitivity to FM events in this study (mean 32%, Ch. 3) aligned with the lower end of reported values in existing literature, 33-88% [17]. When models are trained on events defined only by what a mother feels, effectively they are trained to ignore clinically relevant, low-intensity movements. Ultrasound annotations enable models with the potential to learn the full physiological truth, including movements that do not make contact with the abdominal wall and are more difficult to perceive [21]. Using ultrasound-labeled events for model training promotes models to learn discriminative patterns for all FM events, not just those that are perceived.

Ultrasound-labeled events also reduce uncertainty when interpreting model decisions. If model performance is evaluated against FM events defined by maternal perception, false positive predictions can be ambiguous [22]; there may be uncertainty as to whether the model generated an incorrect prediction or the mother was unable to perceive the movement [17]. Though

maternal perception may serve as a necessary but flawed proxy (Ch. 2), evaluating model performance against ultrasound-labeled FM events (Ch. 4) reduces bias and is essential for proving that the algorithm outputs are clinically meaningful. Similarly, event duration defined by label source influenced the interpretation of model decisions. FM events labeled by maternal perception were primarily the same duration (with the exception of FM events in series, where participants repeatedly indicated perceived FM), while FM events labeled by ultrasound demonstrated widely variable durations (Ch. 3). Variable-duration events necessitated model performance to be evaluated at the event-level (Ch. 4) rather than on a window-by-window basis (Ch. 2). Window-by-window evaluation can lead to overrepresentation of long-duration FM-events, while event-level evaluation fairly represents individual FMs. In the clinic, the presence of FM, not necessarily its duration, carries diagnostic weight [23] making event-level evaluation of model performance a more clinically meaningful and explainable metric [24,25]. Performance metrics that are based on clinically meaningful FMs can be directly communicated with clinicians and consumers to build trust and increase the likelihood of adoption. Ultimately, the ability to explain how model decisions were made and what data they were derived from is critical for the adoption of algorithms contributing to potentially life-impacting healthcare decisions [26,27]. Transitioning from subjective perception to physiological truth is a necessary step for gaining the clinician and consumer trust required for widespread adoption.

6.2.3 Impact of Subject-Specific Physiology on Model Generalizability

The primary hurdle for real-world deployment is the generalizability gap: individual physiology obscures the universal signal patterns required to detect FM in unseen subjects. Benchmarking data splitting approaches revealed a major discrepancy between popular methods in the field (e.g., random sampling) and generalizable methods, like LOSO (Ch. 4). While random sampling is common in machine learning applications to ensure a representative distribution, it is less appropriate for this task because it contributes to data leakage. In fetal monitoring, data windows from the same trial share subject-specific IMU signatures. By randomly sampling these windows into both training and testing sets, models learn to recognize an individual's unique physiological "noise" rather than generalizable fetal movement. The drop in performance for the model trained with the LOSO approach (AUROC = 0.61) suggests that subject-specific factors currently obscure universal signal characteristics of FM. Interestingly, model performance improved slightly when the model was "calibrated" with just one trial from

each participant (AUROC = 0.65). The increase in performance suggests that small amounts of subject-specific data incorporated into training may help models generalize more effectively to subject-specific signal characteristics. Only one prior study used a LOSO approach [13], achieving 75% sensitivity and 65% precision across six participants. However, only maternal perception data was available for labeling FM events, so direct performance comparisons cannot be made.

Using a comparison of maternal perception- and ultrasound-labeled events, participants in this study were less sensitive to FM (mean = 32%, range = 8-70%, Chap. 3 and 4) compared to previously studied populations (33-88% [17]). Physiological factors like body mass index (BMI), gestational age, and placental site may have had an impact on the ability to perceive FM [28] as these factors increase the distance between the fetus, sensory receptors, and the abdominal wall. Not only do these factors affect perception, but they may also impact the utility of wearable motion sensors, which depend on the propagation of FM through the abdominal wall to cause mechanical deformation of the abdomen. Effectively, factors like adipose tissue depth (e.g., BMI) and placental position act as physical damping components [29,30], reducing the amount of abdominal deformation and lowering the signal-to-noise ratio of FM events [21]. Because of the damping effect of these physiological factors, significant variations in FM detection performance between subjects (Ch. 4) may be partially explained by subject-specific physiology. These findings suggest that future commercial systems will not be "plug-and-play" but must require personalized calibration to account for the individual physiology of each user.

6.2.4 Incorporating Context into the Design and Development of FM Detection Algorithms

Despite research efforts to advance the technical capabilities of FM detection systems, technical rigor alone does not support clinical adoption or patient adherence. Limited work has discussed how to integrate these systems into clinical or consumer settings. However, frameworks that support the systematic incorporation of context into design decisions (Ch. 5) may be used to inform algorithmic choices (Ch. 4) and identify barriers to adoption. Sociocultural, political, and institutional context are core considerations for product adoption but have domain-specific nuances that may guide decision-making (Ch. 5).

Because fetal monitoring systems exist at the intersection of medical devices and consumer technology, regulatory context must directly dictate the safety and efficacy standard of algorithm performance. Regulatory bodies such as the United States Food and Drug

Administration (FDA) and the European Union Medical Device Regulation (MDR) dictate the efficacy standards for devices making health-related claims [31,32]. Regulated systems will likely need to perform reliably across general populations, and the current inability of FM detection algorithms to generalize to individual subjects (Ch. 4) may be the primary hurdle to adoption. Even with personalized calibration (Ch. 4), the low sensitivity observed (24% at a 5% FPR) implies a significant tradeoff: developers may need to accept lower detection sensitivity to avoid potentially dangerous false positive detections that could provide false reassurance of fetal health. Ultimately, the risk of false reassurance suggests that developers must prioritize low false-positive rates over high sensitivity to ensure these systems meet the safety expectations of both regulatory agencies and mothers.

Widespread adoption depends on close integration of these systems into existing clinical workflows, as either home-based supplementary tools or ambulatory monitors. Currently, existing methods of kick counting suffer from low adherence (60%) and ineffective response to reduced FM [33]. However, when kick counting is done properly, about 15% of women seek additional clinical testing [34]. FM detection systems enable two strategic entry points into the existing workflow: (1) systems to augment home-based kick counting to improve adherence and data reliability and (2) systems to provide data during the daily or weekly follow-up evaluations that typically follow a report of reduced FM. Systems that act as ambulatory monitors (option 2) will likely require higher safety standards compared to systems acting as supplementary support tools (option 1), creating a critical institutional trade-off for model training. While personalized calibration improves detection performance (Ch. 4), collecting and labeling these data is resource intensive. Adoption of ambulatory systems may be limited to high-resource clinics, where clinical staff can support data collection. Supplementary support tools, on the other hand, could be used in more diverse settings to enable fetal monitoring where access to clinicians or ultrasound imaging may be limited. Interestingly, generalizable, ultrasound-trained models performed similarly to personalized, maternal perception-trained models (Ch. 4) suggesting that developers can leverage maternal-perception based calibration to bridge the performance gap in settings where high-fidelity ultrasound or clinical staffing is unavailable, making the system a viable tool for global health equity.

Beyond institutional trust, patient adherence may be a critical barrier to widespread adoption, as current signal-to-noise ratio limitations necessitate that mothers remain stationary

during data collection (Ch. 2-4). This technical limitation imposes a sociocultural constraint: fetal monitoring systems are not yet capable of continuous data collection but could be used during periods of rest. Incorporating the daily activities and domestic environments of mothers will be essential for identifying data collection opportunities in future home-based longitudinal studies. Because existing systems were built for laboratory testing (Ch. 2-4), future work to support the wearability of continuous fetal monitoring systems during daily activities and everyday environments is critical for adherence. To move beyond the laboratory, future systems must be designed to fit a mother's daily routine, treating data collection as a structured, low-friction habit rather than a disruptive clinical event.

Additional considerations outside of the core framework may be technological, infrastructural, or economic depending on the end-use context. Future work to support digital platforms that communicate insights about fetal health must be compatible with local smartphone and internet availability. These fetal health insights may support telemedicine in rural or remote contexts [35], so distance to care may affect how these insights are communicated or the timeliness of care-seeking guidance provided. Finally, the hardware used for data collection must support pricing constraints, particularly for global health applications. Ultimately, the path to adoption is not a purely technical challenge, but a balancing act where technical limitations, like sensitivity and generalizability, are managed within complex clinical and consumer settings.

6.3 Implications

This dissertation explored methodological considerations for the development of FM detection systems for progress toward future deployment in clinical and consumer settings. The findings from these studies provide insights that would benefit communities interested in fetal monitoring within research and clinical settings and entrepreneurs interested in operationalizing these systems.

6.3.1 Implications for Technical Development: Toward Standardized and Rigorous Benchmarking

The findings from Chapters 2 through 5 provided several useful insights into the technical and contextual considerations for continuous fetal monitoring systems that would interest the research community, especially when designing future studies. Broadly, these results

imply a necessary shift in how researchers approach model validation, data labeling, and methodological decision-making to move the field closer to clinical and commercial viability.

Guidance on model evaluation serves as the most significant implication from this work. Specifically, the exploration of evaluation approaches revealed a major gap between generalizable approaches, such as LOSO, and the popular random sampling methods currently prevalent in the field (Ch. 4). Past reliance on random sampling has produced performance metrics that falsely imply FM detection systems could be safely deployed. Consequently, using LOSO approaches in future studies may provide more realistic estimations of model performance. However, the wide variation in performance observed between subjects (Ch. 4) also suggests that future work must explore the incorporation of personalized calibration data into the training process. Such personalization is likely necessary to achieve FM detection that meets the safety standards of regulatory bodies, clinicians, and consumers (Ch. 5).

Building on this evaluation shift, the transition toward ultrasound-informed modeling represents a critical advancement in the physiological accuracy of model training. While event-level evaluation can technically be applied to maternal perception data, the resulting metrics are less clinically meaningful due to the inherent subjectivity and temporal imprecision of human reporting. In contrast, ultrasound labels enable model evaluation based on the true physiological duration and onset of fetal movement (Ch. 4). Ultrasound evaluation provides a more precise estimation of performance, implying that future studies should prioritize this clinical standard over arbitrary windows or subjective maternal perception.

However, acknowledging the inherent limitations of ultrasound as a clinical standard is critical for contextualizing the performance metrics in this dissertation. Because standard clinical ultrasound provides only a two-dimensional planar view of a three-dimensional uterine space, fetal movements occurring outside the immediate field of view can be easily missed. Furthermore, ultrasound interpretation is subject to operator dependency; systematic reviews demonstrate that fetal biometry assessments can be inaccurate, exhibiting high false-positive rates that risk prompting unnecessary obstetric interventions [38]. This subjectivity is compounded by maternal physiological confounding factors, such as maternal gas, which can visually mimic fetal movement. A notable limitation of the dataset in this dissertation is the reliance on a single OB/GYN for both data collection and subsequent review. Reviewing the ultrasound annotations was particularly resource-intensive; establishing the ground-truth

standard required manual time-synchronization of discontinuous clips, compiling full-length video files, and executing a rigorous double-review process of every ultrasound recording. To mitigate these workflow bottlenecks in future large-scale studies, developing automated synchronization scripts or utilizing AI agents to streamline video compilation and initial annotation would be beneficial. To increase inter-rater reliability, future studies could also separate clinical roles by employing a trained sonographer for data acquisition alongside two independent OB/GYNs for image interpretation. While wearable ultrasound has emerged as a potential solution to generate continuous ground-truth labels for wearable inertial sensors, viability is limited due to safety concerns associated with ultrasonic tissue heating [38]. Consequently, ultrasound must be treated not as a flawless clinical standard, but as a bounded reference that requires rigorous multi-observer validation.

Furthermore, while benchmarking on a common test can support systematic comparisons (Ch. 2 and 4), commercial viability will remain out of reach until a standardized, public benchmark dataset is established. Until a gold standard exists, the ability to make meaningful comparisons between studies will remain limited. Though benchmarking provides technical guidance, accelerating system advancement requires collaborative benchmark datasets from the larger research community. While the dataset compiled in Chapters 2 through 4 represents a valuable benchmarking resource, immediate public release warrants careful consideration. Making benchmark datasets openly available introduces a well-documented risk that they may be scraped and absorbed into large-scale training datasets for foundation models [39]. If this occurs, the dataset can no longer serve as a truly independent test set, since models may have already been exposed to it during training. This type of "benchmark contamination" can artificially inflate performance and make it difficult to fairly compare new methods in the future [39]. Preserving the dataset as a clean, unseen evaluation resource is therefore important for maintaining scientific integrity in model benchmarking. Although the dataset could be withheld entirely or distributed on request, both options have practical limitations, including administrative burden and reduced accessibility for other researchers. An alternative is a controlled-access approach, where the dataset itself is not directly released. Instead, researchers submit their model predictions to a secure evaluation system, which computes performance on the held-out data and returns only summary metrics under formal data-use agreements [40]. This approach is consistent with emerging "private benchmarking" frameworks and allows

researchers to compare methods fairly without exposing the underlying dataset. In this way, the dataset remains a protected reference standard for evaluation while still enabling transparent and reproducible reporting of results across research groups.

Parallel to evaluation strategies, researchers must also consider how data-driven parameters are derived. The statistical methods used to derive data-driven signal processing parameters (Ch. 3) imply that these parameters are likely unique to individual datasets. Therefore, these parameters should be derived directly from study populations to justify design choices and promote increased interpretability of FM detection algorithms through the incorporation of domain knowledge (Ch. 4). While the field still lacks a universal set of signal processing parameters, these study-specific parameters remain a necessary step until larger, more diverse datasets are available to establish a universal standard.

Beyond signal processing, this work highlights the importance of methodological benchmarking in selecting model architecture. Benchmarking found that sensor fusion systems increased FM detection performance and feature-based models outperformed deep learning architectures (Ch. 2). These findings imply that researchers should continue to prioritize sensor fusion approaches to improve model sensitivity and utilize feature-based models to explore edge computing in resource-constrained contexts, while continuing to explore deep learning architectures as larger datasets become available. Additionally, comparisons between ultrasound and maternal perception labels for model training revealed that models trained with maternal perception have a performance ceiling (Ch. 4). These results imply that future models should utilize ultrasound to promote detection of FMs that mothers cannot perceive to bypass the inherent limitations of subjective labeling.

Ultimately, even the most robust technical advancements must be viewed through a contextual lens. Despite significant progress through benchmarking, current FM detection performance (24% sensitivity at 5% FPR) is not yet viable for commercialization (Ch. 4). Even if future studies provide evidence for robust technical solutions, they may still fail in the market if they do not incorporate factors within the end-use context. As a final implication, applying design frameworks to systematically account for context (Ch. 5) can guide the technical decisions that will eventually promote trust and widespread adoption in clinical and consumer settings.

6.3.2 Implications for Clinical Practice: Balancing Accessibility with Clinical Integrity for Decision Support Tools

While the previous section outlines the requirements for future research, the findings from Chapters 2 through 5 are also of direct relevance to obstetricians. Specifically, these insights help clarify the current capabilities and limitations of wearable sensor-based continuous fetal monitoring systems and how such technology might eventually integrate into clinical workflows.

The primary clinical implication regarding current system capabilities involves the balance between sensitivity and diagnostic reality. Chapters 2 and 4 proposed algorithms for continuous FM detection that would allow clinicians further insight into the status of fetal health throughout late-stage pregnancy. Findings from these studies indicated that wearable sensors may be more sensitive to FM than kick counting, based on the potential for high sensitivity when evaluated against ultrasound. However, these performance metrics must be carefully evaluated due to the inherent tradeoffs between sensitivity and specificity. Because increased false positive detections risk dangerous reassurance of fetal health, the current system performance (50% of FM events at a FPR of 28%) remains a concern (Ch. 4). Nevertheless, if false positive predictions can be controlled in future studies, clinicians could use these insights to identify reductions in FM before they become critical.

In addition to system performance, the methodology behind FM detection systems has significant implications for clinical trust. Chapters 3 and 4 relied on ultrasound imaging to define FM events for model evaluation. By deriving parameters from visual confirmation of FM rather than algorithmic determination, these methods promote transparency between clinicians and engineers. This objective approach to defining events is essential for fostering trust in the accuracy of the system. Ultimately, access to such continuous data may help clinicians determine meaningful reductions in FM, filling a critical gap where no current standard exists [33,34,36].

Beyond technical accuracy, the practical integration of these systems into existing workflows presents both opportunities and challenges. Chapters 4 and 5 discuss the considerations for deployment of wearable fetal monitoring systems. For instance, LOSO approaches for model training would allow clinicians to hand off fetal monitoring systems to new patients without additional data collection. However, deployment is currently limited by performance issues related to generalizability between subjects (Ch. 4). While models trained

with personalized calibration data showed improved performance, clinicians may not adopt a tool if it requires resource-intensive calibration trials that conflict with existing institutional procedures.

Despite these technical challenges, the long-term potential for telemedicine and life-saving intervention remains significant, particularly due to expanding rural maternity care deserts. Between 1989 and 2019, over 20% of United States obstetric units closed, degrading geographic access to prenatal care, crowding remaining facilities, and exacerbating rural infant mortality rates relative to urban centers [41]. While some patients bypass these gaps by traveling further for care, others are left entirely unserved. To optimize fetal monitoring in these resource-limited regions, deployment frameworks must leverage "task-shifting" to deliberately delegate monitoring tasks from high-resource hospitals to home-based settings or community health workers. Moving these responsibilities to the domestic environment can expand healthcare access, improve clinical outcomes, and empower patients to self-manage their condition through objective data [42]. Successful implementation could reduce required in-person appointments [37] and provide data-driven insights to help clinicians determine when intervention is necessary to prevent stillbirth or avoid unnecessary delivery.

A critical requirement for deploying task-shifting devices is defining how clinicians interact with massive volumes of continuous, patient-generated health data. This challenge closely mirrors the clinical adoption of the Apple Watch for atrial fibrillation screening [43]. While large-scale trials proved that consumer wearables can successfully identify cardiac anomalies, they initially created a significant administrative burden and alert fatigue for cardiologists who were flooded with raw, unverified data streams [44]. In response, modern digital health frameworks have shifted away from sending continuous raw signals to clinicians, opting instead for integrated telemonitoring workflows that provide high-level, validated summaries [43]. For fetal surveillance, a similar process must be established. Obstetricians should not be responsible for evaluating raw, continuous inertial sensor signals. Instead, automated algorithms must compress days or weeks of monitoring data into clinical-grade summaries or automated alerts triggered only when deviations fall below validated, population-level safety thresholds. Emulating these established cardiovascular care workflows will ensure that at-home fetal monitoring systems alleviate, rather than exacerbate, clinical workflow constraints.

However, as a final cautionary note, the findings from Chapter 4 imply that the risk of false positive detection is still too high for the ethical deployment of these systems to unsupervised settings.

6.3.3 Operationalizing Fetal Monitoring Systems

While the technical and clinical implications highlight that current performance, specifically high FPR, is not yet suitable for unsupervised medical intervention, these limitations provide a foundation for a staged path toward operationalization. The results from Chapters 2 and 4 imply that achieving the classification performance necessary for clinical adoption, specifically generalizing across new subjects with high specificity, will require significantly larger and more diverse datasets. By following the trajectory of other similar consumer-to-clinical wearables, FM monitoring systems may be able to successfully scale through three distinct stages, guided by the contextual factor taxonomy derived in Chapter 5 to uncover "unknown unknowns" and promote long-term product adoption. Because a fetal monitoring system lies at the intersection of all three product domains tested in the taxonomy (e.g., medical devices, information technologies, and consumer products), both core and peripheral contextual factors may be important within each stage of operationalization.

The first stage of operationalization focuses on leveraging consumer adoption to overcome current data limitations, primarily prioritizing cultural and domestic contextual factors associated with the consumer product domain. While current performance metrics (e.g., high FPR) preclude these systems from unsupervised clinical use, they remain viable as consumer-facing tools designed to augment adherence to existing kick-counting protocols. In this initial stage, the sociocultural and infrastructural classifications of context serve as core considerations, directly addressing maternal motivation, anxiety, and ease of at-home use. Deploying FM detection systems initially as consumer products to aid in objectivity and adherence to kick-counting protocols could generate the massive amounts of data required to train more robust deep learning models. Through partnerships between industry and academic institutions, these datasets could be used to establish the standardized testing datasets necessary for systematic benchmarking and the derivation of universal, data-driven methods. In this way, early consumer adoption could provide the data necessary to move from augmenting maternal perception to automating it.

Once a data foundation is established, the second stage requires integrating ultrasound data to provide a more transparent clinical standard, shifting the system toward prioritization of clinical workflow and data privacy factors associated with the information technology domain. As discussed in Chapter 4, the collection and interpretation of ultrasound recordings is resource-intensive and may disrupt existing clinical workflows. The contextual factor taxonomy may be used to discover key institutional and technological factors to set engineering requirements. For example, as FM detection systems transition toward clinical settings, institutional requirements may be creatively met by AI-enabled tools that automatically extract labels from existing clinical ultrasound recordings. This stage mirrors the current progress of novel, commercial solutions like the Bloomlife MFM-Pro, which recently received FDA clearance for fetal heart rate monitoring while remaining an "investigational device" for more complex indications [45]. By clearing baseline biometrics first, continuous fetal monitoring systems build the clinical trust necessary for widespread adoption.

Finally, the third stage involves securing regulatory and reimbursement approval to formalize these systems within the medical device domain, prioritizing both regulatory and systemic insurance factors within the taxonomy. Once efficacy is demonstrated at a clinically acceptable level against ultrasound-labeled events, FM detection systems can be integrated into clinical workflows. This transition requires navigating the political and economic factors discussed in Chapter 5, including regulatory approval and the establishment of insurance reimbursement codes. Ultimately, these approvals enable clinicians to transition from tracking simple binary "kicks" to studying population-level trends in high-risk groups, supporting targeted interventions to reduce the risk of stillbirth.

Ultimately, this dissertation moves the field of FM monitoring from proof-of-concept toward a realistic roadmap for deployment. By establishing evidence-based benchmarks, defining the performance ceiling of maternal perception, and identifying the contextual factors that govern clinical adoption, this work provides a foundation for new approaches to fetal surveillance. The path forward requires persistent commitment to methodological rigor and a deep understanding of the diverse contexts in which mothers and their clinicians interact with technology.

6.4 Recommendations and Future Work

To fully address the research gaps identified within wearable sensor-based FM detection, further work must build upon the foundational contributions of this dissertation. Future research should prioritize the following high-impact trajectories to move these systems from laboratory proof-of-concepts to viable clinical tools.

The first priority for future work is to transition from static, population-level models to generalizable systems that utilize subject-specific calibration. While Chapters 2 through 4 included participants with diverse gestational ages, body sizes, and placental sites, these populations were limited to healthy, singleton pregnancies. To ensure these systems are equitable between groups, the next phase of research must investigate how personalized FM detection algorithms can bridge the generalization gap. Future studies should expand the populations evaluated by recruiting from diverse cohorts to determine the exact impact of subject-specific physiological factors on model efficacy. Furthermore, intentional recruitment accounting for distinct maternal and fetal characteristics could provide an opportunity to minimize signal damping and train physiology-specific models. By utilizing techniques such as transfer learning or personalized model training, researchers can create systems that are universally applicable by tailoring predictions to the individual to ensure consistent performance regardless of maternal or fetal variation.

To advance the baseline capabilities of wearable fetal monitoring, future research might also focus on general architectural and algorithmic optimization. While the statistical features developed in Chapters 2 and 4 captured discriminative characteristics of FM, the baseline feature space could be expanded by calculating kinematic metrics across multiple, overlapping window sizes to capture both transient and prolonged movement dynamics. Furthermore, future modeling could transition from traditional static classifiers to advanced sequential architectures, such as transformer models, to better leverage the temporal dependencies inherent in continuous physiological time-series data. Beyond raw performance metrics, selecting a model architecture should prioritize system transparency. Utilizing heuristic features allows researchers to directly incorporate domain knowledge into the pipeline, providing known information about FM that supports signal discrimination. While deep learning or transformer architectures can autonomously discover complex sequential patterns, heuristic approaches offer greater explainability. In clinical and regulatory environments, this transparency is critical for

demystifying model decision-making, thereby building the user trust necessary to drive long-term product adoption.

Parallel to general algorithmic improvements, the testing of continuous fetal monitoring systems must transition from controlled laboratory environments to real-world activity. The algorithms presented in this work were developed using stationary data collection (Ch. 2–4). While these models learned generalizable patterns (Ch. 2 and 4) and identified statistically significant differences between fetal and maternal movement (Ch. 3), maintaining an acceptable false-positive rate during a mother's daily activities remains a significant challenge. Consequently, future research must build algorithms that are resilient to high-magnitude noise captured during maternal motion. One pathway involves hardware expansion, such as the inclusion of additional sensor types (e.g., acoustic or piezoelectric sensors) to improve baseline sensitivity [1]. From a signal-processing perspective, adaptive sensor selection might be utilized within the IMU array to dynamically isolate the specific node exhibiting the highest signal-to-noise ratio based on proximity to the fetus. Additionally, multi-sensor comparative features (e.g., evaluating the differential angular rates across the array nodes) could help isolate localized FM from maternal motion. Beyond traditional kinematics, investigating time-series entropy might offer novel metrics to assess signal stability during daily life. Additional data in Chapters 2 through 4 was collected while participants performed controlled and repetitive motion. This dataset could serve as an immediate testbed to evaluate these sensor-fusion and artifact-rejection strategies.

Models that are personalized and robust can then be used to shift the focus from short-term detection to long-term fetal health and behavior studies. Because no commercial or clinical technologies currently exist for continuous detection, very little is known about longitudinal fetal activity patterns. Future work should therefore shift from single-session sampling to repeated sampling throughout gestation to develop models capable of detecting changes in patterns over time. Building on the technical design choices in Chapters 2 through 4, researchers can investigate fundamental science regarding fetal activity metrics, such as movement types or rest-period durations. Characterizing these trends in both healthy and at-risk populations could finally enable the clinical definition of "meaningful reductions" in FM, providing the objective evidence needed to prompt clinical intervention.

Beyond FM monitoring, a high-impact progression of this work involves the integration of multiple markers of fetal health. Fetal heart rate is known to accelerate in the presence of FM and acts as a reassuring clinical signal [23]. Fundal measurements are also used to monitor growth and detect fetal growth restriction [46], and the presence of fetal hiccups is associated with neurological function and a decreased risk of stillbirth [47,48]. Consequently, FM detection systems could be tuned to identify predictive fetal movement features like hiccups or strength of movement. Future research could also explore systems that combine FM detection with fetal heart rate and/or frequent fundal measurements. Continuous monitoring of these integrated signals may reveal key insights into targeted intervention windows, ultimately aiding in the prevention of stillbirth and unnecessary early delivery.

Finally, the technical advancements discussed above are only feasible for real-world adoption if contextual factors from the end-use context are simultaneously incorporated. The design framework presented in Chapter 5 provides a foundation for the development of tools that ensure technology is adaptable across diverse design contexts. Future work should focus on the creation of actionable checklists, prompts, or computational tools derived from this framework. Such tools would allow developers to systematically account for context without developing frameworks from scratch. Follow-up studies are necessary to explore how identified contextual factors are incorporated into specific design decisions. Once tested, these tools can be applied to the deployment of FM systems to ensure that technical decisions promote adoption in the complex settings of at-home and clinical care.

6.5 References

[1] Ghosh, A. K., Catelli, D. S., Wilson, S., Nowlan, N. C., and Vaidyanathan, R., 2024, "Multi-Modal Detection of Fetal Movements Using a Wearable Monitor," Information Fusion, 103. https://doi.org/10.1016/j.inffus.2023.102124.

[2] Strodthoff, N., Wagner, P., Schaeffter, T., and Samek, W., 2021, "Deep Learning for ECG Analysis: Benchmarks and Insights from PTB-XL," IEEE J. Biomed. Health Inform., 25(5), pp. 1519–1528. https://doi.org/10.1109/JBHI.2020.3022989.

[3] Friedrich, S., and Friede, T., 2022, "On the Role of Benchmarking Data Sets and Simulations in Method Comparison Studies." [Online]. Available: http://arxiv.org/abs/2208.01457.

[4] Lones, M. A., 2024, “Avoiding Common Machine Learning Pitfalls,” Patterns, 5(10). https://doi.org/10.1016/j.patter.2024.101046.

[5] Boashash, B., Khlif, M. S., Ben-Jabeur, T., East, C. E., and Colditz, P. B., 2014, “Passive Detection of Accelerometer-Recorded Fetal Movements Using a Time-Frequency Signal Processing Approach,” Digital Signal Processing: A Review Journal, 25(1), pp. 134–155. https://doi.org/10.1016/j.dsp.2013.10.002.

[6] Zhao, X., Zeng, X., Koehl, L., Tartare, G., and Jonckheere, J. De, 2020, “A Wearable System for In-Home and Long-Term Assessment of Fetal Movement,” IRBM, (41), pp. 205–211.

[7] Mesbah, M., Khlif, M. S., Layeghy, S., East, C. E., Dong, S., Brodtmann, A., Colditz, P. B., and Boashash, B., 2021, “Automatic Fetal Movement Recognition from Multi-Channel Accelerometry Data,” Comput. Methods Programs Biomed., 210.

[8] Layeghy, S., Azemi, G., Colditz, P., and Boashash, B., 2014, “Non-Invasive Monitoring of Fetal Movements Using Time-Frequency Features of Accelerometry,” IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP).

[9] Liang, S., Peng, J., and Xu, Y., 2021, “Passive Fetal Movement Signal Detection System Based on Intelligent Sensing Technology,” J. Healthc. Eng., 2021. https://doi.org/10.1155/2021/1745292.

[10] Khlif, M. S., Boashash, B., Layeghy, S., Ben-Jabeur, T., Mesbah, M., East, C., and Colditz, P., 2011, “Time-Frequency Characterization of Tri-Axial Accelerometer Data for Fetal Movement Detection,” IEEE International Symposium on Signal Processing and Information Technology, ISSPIT 2011, pp. 466–471. https://doi.org/10.1109/ISSPIT.2011.6151607.

[11] Delay, U. H., Nawarathne, B. M. T. M., Dissanayake, D. W. S. V. B., Ekanayake, M. P. B., Godaliyadda, G. M. R. I., Wijayakulasooriya, J. V., and Rathnayake, R. M. C. J., 2020, “Non Invasive Wearable Device for Fetal Movement Detection,” 2020 IEEE 15th International Conference on Industrial and Information Systems, ICIIS 2020 - Proceedings, (978), pp. 285–290. https://doi.org/10.1109/ICIIS51140.2020.9342662.

[12] Xu, J., Zhao, C., Ding, B., Gu, X., Zeng, W., Qiu, L., Yu, H., Shen, Y., and Liu, H., 2022, “Fetal Movement Detection by Wearable Accelerometer Duo Based on Machine

Learning," IEEE Sens. J., 22(12), pp. 11526–11534. https://doi.org/10.1109/JSEN.2022.3172451.

[13] Altini, M., Mullan, P., Rooijakkers, M., Gradl, S., Penders, J., Geusens, N., Grieten, L., and Eskofier, B., 2016, "Detection of Fetal Kicks Using Body-Worn Accelerometers During Pregnancy: Trade-Offs Between Sensors Number and Positioning," IEEE EMBC.

[14] Abeywardena, C. L., Vanheusden, F. J., Walker, K. F., Arm, R., and Zhang, Q., 2020, "Fetal Movement Counting Using Optical Fibre Sensors," Sensors, 21(48).

[15] Lai, J., Woodward, R., Alexandrov, Y., ain Munnee, Q., Lees, C. C., Vaidyanathan, R., and Nowlan, N. C., 2018, "Performance of a Wearable Acoustic System for Fetal Movement Discrimination," PLoS One, 13(5). https://doi.org/10.1371/journal.pone.0195728.

[16] Bobrova, Y. O., Kapranova, O. N., and Filipenko, K. V., 2021, "Algorithm for Detecting Alarming Fetal Conditions," Proceedings of 2021 24th International Conference on Soft Computing and Measurements, SCM 2021, Institute of Electrical and Electronics Engineers Inc., pp. 243–246. https://doi.org/10.1109/SCM52931.2021.9507177.

[17] Hijazi, Z. R., Callan, S. E., and East, C. E., 2010, "Maternal Perception of Foetal Movement Compared with Movement Detected by Real-Time Ultrasound: An Exploratory Study," Australian and New Zealand Journal of Obstetrics and Gynaecology, 50(2), pp. 144–147. https://doi.org/10.1111/j.1479-828X.2009.01122.x.

[18] Ward, J. A., Lukowicz, P., Trö, G., and Starner, T. E., Activity Recognition of Assembly Tasks Using Body-Worn Microphones and Accelerometers.

[19] Frénay, B., and Verleysen, M., 2014, "Classification in the Presence of Label Noise: A Survey," IEEE Trans. Neural Netw. Learn. Syst., 25(5), pp. 845–869. https://doi.org/10.1109/TNNLS.2013.2292894.

[20] Bekker, A. J., and Goldberger, J., TRAINING DEEP NEURAL-NETWORKS BASED ON UNRELIABLE LABELS.

[21] Matoba, H., Kusaka, T., Shimatani, K., Kim, D., Kanazawa, H., Kuniyoshi, Y., and Tanaka, T., 2026, "MRF-Based Mapping of Fetal and Maternal Deformations for Understanding and Supporting Fetal Movement Perception," 2026 IEEE/SICE International Symposium on System Integration, SII 2026, Institute of Electrical and

Electronics Engineers Inc., pp. 832–837. https://doi.org/10.1109/SII64115.2026.11404623.

[22] Darling, M. C., and Stracuzzi, D. J., Toward Uncertainty Quantification for Supervised Classification. [Online]. Available: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online.

[23] Walton, J. R., and Peaceman, A. M., 2012, "Identification, Assessment and Management of Fetal Compromise," Clin. Perinatol., 39(4), pp. 753–768. https://doi.org/10.1016/j.clp.2012.09.001.

[24] Magnini, M., Ciatto, G., and Omicini, A., 2022, "On the Design of PSyKI: A Platform for Symbolic Knowledge Injection into Sub-Symbolic Predictors," Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Science and Business Media Deutschland GmbH, pp. 90–108. https://doi.org/10.1007/978-3-031-15565-9_6.

[25] Towell, G. G., and Shavlik, J. W., Knowledge-Based Artificial Neural Networks.

[26] Garouani, M., Ahmad, A., Bouneffa, M., Hamlich, M., Bourguin, G., and Lewandowski, A., 2022, "Towards Big Industrial Data Mining through Explainable Automated Machine Learning," International Journal of Advanced Manufacturing Technology, 120(1–2), pp. 1169–1188. https://doi.org/10.1007/s00170-022-08761-9.

[27] Garouani, M., Mothe, J., Barhrhouj, A., and Aligon, J., 2024, "Investigating the Duality of Interpretability and Explainability in Machine Learning," Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI, IEEE Computer Society, pp. 861–867. https://doi.org/10.1109/ICTAI62512.2024.00125.

[28] Gardosi, J., Madurasinghe, V., Williams, M., Malik, A., and Francis, A., 2013, "Maternal and Fetal Risk Factors for Stillbirth: Population Based Study," BMJ (Online), 346(7893), pp. 1–14. https://doi.org/10.1136/bmj.f108.

[29] Paladini, D., 2009, "Sonography in Obese and Overweight Pregnant Women: Clinical, Medicolegal and Technical Issues," Ultrasound in Obstetrics and Gynecology, 33(6), pp. 720–729. https://doi.org/10.1002/uog.6393.

[30] Brown, R., Johnstone, E. D., and Heazell, A. E. P., 2016, "Professionals' Views of Fetal-Monitoring Support the Development of Devices to Provide Objective Longer-Term

Assessment of Fetal Wellbeing," Journal of Maternal-Fetal and Neonatal Medicine, 29(10), pp. 1680–1686. https://doi.org/10.3109/14767058.2015.1059808.

[31] U.S. Food & Drug Administration, 2026, "Overview of Device Regulation."

[32] The European Union Medical Device Regulation, 2026, "Regulation (EU) 2017/745 (EU MDR)."

[33] Patel, S., Conway, A. E., Adjei, T., Abati, I., Dhawan, S., Yu, Z., Vaidyanathan, R., and Lees, C., 2025, "Is It Possible to Monitor Fetal Movements with a Wearable Device? A Review of Novel Technologies," European Journal of Obstetrics and Gynecology and Reproductive Biology, 305, pp. 329–338. https://doi.org/10.1016/j.ejogrb.2024.12.011.

[34] Heazell, A. E. P., and Frøen, J. F., 2008, "Methods of Fetal Movement Counting and the Detection of Fetal Compromise," J. Obstet. Gynaecol. (Lahore)., 28(2), pp. 147–154. https://doi.org/10.1080/01443610801912618.

[35] Stoneburner, A., Lucas, R., Fontenot, J., Brigance, C., Jones, E., and DeMaria, A., 2024, Nowhere to Go: Maternity Care Deserts Across the US. [Online]. Available: https://www.marchofdimes.org/maternity-care-deserts-report. [Accessed: 16-Sep-2025].

[36] Olesen, A. G., and Svare, J. A., 2004, "Decreased Fetal Movements: Background, Assessment, and Clinical Management," Acta Obstet. Gynecol. Scand., 83(9), pp. 818–826. https://doi.org/10.1111/j.0001-6349.2004.00603.x.

[37] Peahl, A. F., Gourevitch, R. A., Luo, E. M., Fryer, K. E., Moniz, M. H., Dalton, V. K., Fendrick, A. M., and Shah, N., 2020, "Right-Sizing Prenatal Care to Meet Patients' Needs and Improve Maternity Care Value," Obstetrics and Gynecology, 135(5), pp. 1027–1037. https://doi.org/10.1097/AOG.0000000000003820.

[38] Brown, R., Wijekoon, J.HB., Fernando, A., Johnstone, E.D., and Heazell, A.EP., 2014, "Continuous objective recording of fetal heart rate and fetal movements could reliably identify fetal compromise, which could reduce stillbirth rates by facilitating timely management." Medical hypotheses, 83(3), pp. 410-417.

[39] Balloccu, S., Schmidtová, P., Lango, M., and Dušek, O., 2024, "Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs," In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 67-93.

[40] Rajore, T., Chandran, N., Sitaram, S., Gupta, D., Sharma, R., Mittal, K., and Swaminathan, M., 2024, "Truce: Private benchmarking to prevent contamination and improve comparative evaluation of llms," arXiv preprint.

[41] Fischer, S., Royer, H., and White, C., 2024, "Health Care Centralization: The Health Impacts of Obstetric Unit Closures in the United States," American Economic Journal: Applied Economics, 16(3), pp. 113-141. https://doi.org/10.1257/app.20220341

[42] Gray, J., 2025, "On Mobile Health, Empowerment, and the Limits of Task Shifting in Healthcare," The American Journal of Bioethics. https://doi.org/10.1080/15265161.2025.2554771

[43] van Steijn, N.J., Blommestijn, I.S., Blok, S., Pepplinkhuizen, S., Somsen, G.A., Knops, R.E., Breukel, L., Tijssen, J.G., Tulevski, I.I., Croon, P.M. and Winter, M.M., 2026, "Enhanced detection and prompt diagnosis of atrial fibrillation using Apple Watch: a randomized controlled trial," JACC.

[44] Senoo, K., 2025, "Atrial Fibrillation Screening: Burden or Benefit for Patients and the Healthcare System," Journal of Cardiovascular Electrophysiology.

[45] Bloomlife, 2026, "Bloomlife."

[46] Lai, J., Nowlan, N. C., Vaidyanathan, R., Shaw, C. J., and Lees, C. C., 2016, "Fetal Movements as a Predictor of Health," Acta Obstet. Gynecol. Scand., 95(9), pp. 968–975. https://doi.org/10.1111/aogs.12944.

[47] Thompson, J.M.D., Wilson, J., Bradford, B.F., Li, M., Cronin, R.S., Gordon, A., Raynes-Greenow, C.H., Stacey, T., Cullling, V.M., Askie, L.M., O'Brien, L.M., Mitchell, E.A., McCowan, L.M.E., Heazell, A.E.P., 2021, "A better understanding of the association between maternal perception of foetal movements and late stillbirth-findings from an individual participant data meta-analysis," BMC Med, 19(1):267. doi: 10.1186/s12916-021-02140-z.

[48] Heazell, A.E.P., Budd, J., Li, M., Cronin, R., Bradford, B., McCowan, L.M.E., Mitchell, E.A., Stacey, T., Martin, B., Roberts, D., Thompson, J.M.D., 2018, "Alterations in maternally perceived fetal movement and their association with late stillbirth: findings from the Midland and North of England stillbirth case-control study," BMJ Open, 8(7). doi: 10.1136/bmjopen-2017-020031.