Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Multi-modal detection of fetal movements using a wearable monitor

Abhishek K. Ghosh [a,e,*], Danilo S. Catelli [c], Samuel Wilson [a,f], Niamh C. Nowlan [b,d,1,*], Ravi Vaidyanathan [a,1,*]

[a] Department of Mechanical Engineering, Imperial College London, London SW7 2AZ, United Kingdom
[b] Department of Bioengineering, Imperial College London, London SW7 2AZ, United Kingdom
[c] Department of Movement Sciences, KU Leuven, Leuven 3000, Belgium
[d] School of Mechanical and Materials Engineering, University College Dublin, Dublin 4, Ireland
[e] Department of Robotics and Mechatronics Engineering, University of Dhaka, Dhaka 1000, Bangladesh
[f] Serg Technologies, London SW7 2LQ, United Kingdom

ARTICLE INFO

Keywords:
Fetal movement monitor
Wearable sensor system
Heterogeneous sensor fusion
Data-dependent thresholding
Machine learning

ABSTRACT

The importance of Fetal Movement (FM) patterns as a biomarker for fetal health has been extensively argued in obstetrics. However, the inability of current FM monitoring methods, such as ultrasonography, to be used outside clinical environments has made it challenging to understand the nature and evolution of FM. A small body of work has introduced wearable sensor-based FM monitors to address this gap. Despite promises in controlled environments, reliable instrumentation to monitor FM out-of-clinic remains unresolved, particularly due to the challenges of separating FMs from interfering artifacts arising from maternal activities. To date, efforts have been focused almost exclusively on homogenous (single) sensing and information fusion modalities, such as decoupled acoustic or accelerometer sensors. However, FM and related signal artifacts have varying power and frequency bandwidths that homogeneous sensor arrays may not capture or separate efficiently. In this investigation, we introduce a novel wearable FM monitor with an embedded heterogeneous sensor suite combining accelerometers, acoustic sensors, and piezoelectric diaphragms designed to capture a broad range of FM and interfering artifact signal features enabling more efficient isolation of both. We further outline a novel data fusion architecture combining data-dependent thresholding and machine learning to automatically detect FM and separate it from signal artifacts in real-world (home) environments. The performance of the device and the data fusion architecture are validated using 33 h of at-home use through concurrent recording of maternal perception of FM. The FM monitor detected an impressive 82 % of maternally sensed FMs with an overall accuracy of 90 % in recognizing FM and non-FM events. Consistency of detection was strongest from 32 gestational weeks onwards, which overlaps with the critical FM monitoring window for stillbirth prevention. We believe the multi-modal sensor fusion approach presented in this research will be a major milestone in the development of low-cost wearable FM monitors enabling pervasive monitoring of FM in unsupervised environments.

1. Introduction

Changes in fetal movement (FM) patterns have long been proposed as a potential biomarker of prenatal health, particularly during the third trimester of pregnancy [1]. Reduced FM has been associated with a range of fetal health issues, including fetal distress, growth restriction, hypoxia, and placental dysfunction [1–7]. Reduced FM has also been correlated with the risk of early induction, emergency cesarean delivery, and small for gestational age babies [8,9]. The relationship between reductions in FM and increased risk of stillbirth is presently under dispute. While some studies reported reductions in maternally sensed FM before stillbirth [10–12], others have not found any significant correlation between them [8,13,14]. Indeed, the relationship between FM and fetal health or birth outcome remains unresolved and has been a topic of contentious debate. A key issue affecting many studies is the subjective nature of maternal perception of FM, which depends on several factors, including the position of the placenta, fetal position, and maternal body mass index [15]. Clinical methods of quantifying FM,

[*] Corresponding authors.
E-mail addresses: abhishek.rme@du.ac.bd (A.K. Ghosh), niamh.nowlan@ucd.ie (N.C. Nowlan), r.vaidyanathan@imperial.ac.uk (R. Vaidyanathan).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.inffus.2023.102124
Received 20 December 2022; Received in revised form 16 September 2023; Accepted 2 November 2023
Available online 4 November 2023

such as ultrasonography, MRI, and cardiotocography, are unsuitable for regular out-of-clinic use and only give short (usually less than 30 min) windows of observation. A wearable device capable of monitoring FM outside of the clinical setting would, therefore, be an invaluable tool in obstetrics and fetal medicine. Ideally, such a device should be able to monitor FM over long durations (hours, rather than minutes), distinguish fetal movements accurately from false positives occurring due to maternal activities, be self-operated by the user, and be non-transmitting.

A small body of work has explored the design and development of passive FM monitors. Most of these devices are based on accelerometers that detect the abdominal vibrations associated with FM. Nishihara et al. [16] used two custom-made capacitive accelerometers to detect FM and expressed the performance of the sensors relative to maternal sensation in terms of prevalence-adjusted bias-adjusted kappa (PABAK) [17]. Using a thresholding-based method with an epoch size of 10 s for data analysis, they obtained an agreement between sensor detections and maternal sensation detections with a mean value of PABAK of 0.75. Ryo et al. [18] extended this study to compare the performance of the sensor system relative to concurrent ultrasound recording. While the sensor system showed promising performance for observed gross fetal movements (PABAK = 0.79), the performance for isolated limb movements was not adequate (PABAK = 0.36). Using four accelerometers and a time-frequency signal processing approach, Boashash et al. [19] obtained a sensitivity and precision of 0.78 and 0.83, respectively, against concurrent ultrasonography over a relatively small data set of around 2.5 h. Mesbah et al. [20] further improved the performance of the same system by using wavelet transform and machine learning-based techniques. Considering a curated data set consisting of 50 % FM epochs (as detected by ultrasound) and 50 % non-FM epochs (background noise and signal artifacts), they achieved binary classification accuracies between 0.87 and 0.95 for different levels of artifact concentration in their non-FM data set. However, the performance of the classifier was not reported for the overall original dataset, which consisted of 798 FM epochs and 8834 non-FM epochs. Using a non-wearable system consisting of four piezoelectric crystals, Valentin et al. [21] detected 78 % of maternally sensed FMs (sensitivity = 0.78) with a large number of false positive detections (precision = 0.40). However, the overall performance of the system was improved when validated against detections by ultrasonography (sensitivity = 0.64, precision = 0.59) [21]. Despite the advances made, none of the aforementioned studies [16,18–21] embedded the sensors in a wearable garment, and therefore, such technologies are not immediately translatable to regular at-home use during normal activities.

Delay et al. [22] reported the performance of an accelerometer embedded in a partially wearable garment against concurrent ultrasound recording. The data acquisition (DAQ) system for the device was relatively large and was not embedded in the garment. For a small test data set consisting of 53 fetal limb movement epochs, 14 maternal laughing epochs, and 103 maternal respiration (background noise) epochs, as determined by concurrent ultrasound, they obtained a sensitivity of 0.81 and a precision of 0.77, in detecting fetal limb movements. However, the performance of the device for detecting gross fetal movements (one of the most prominent types of movement used in other studies [18–20,23]) was not reported. In a previous study from our research group, Lai et al. [23] introduced the application of acoustic sensors to detect FM. Using a non-wearable version of the sensor system, a sensitivity of 0.78 was achieved in detecting startle FM (vigorous, whole-body movements) relative to concurrent ultrasound detections. A wearable version of the system achieved a strong performance (sensitivity = 0.83, precision = 0.54) against maternal sensation detection for a small data set (30 min) [24].

The majority of prior studies on passive FM monitors have quantified the performance of homogeneous sensor arrays in a controlled experimental environment. However, the translation of such a system can only be realized through a self-operated wearable device that can be used during normal activities of everyday life. Wearable devices introduce additional challenges in terms of the inconsistency of signal quality due to the variable sensor attachment quality and increased signal artifacts due to a self-operated data collection procedure. One way to handle such problems is to use a combination of different types of sensors with both redundant and complementary properties to reduce the variance of performance, improve the stability of detection, and expand the sensing ability of the system [25,26]. The fusion of heterogeneous sensors to design wearables for human activity recognition has been explored by numerous researchers in recent years [27]. For example, Talitckii et al. [28] fused data from different inertial sensors (accelerometers, gyroscopes, and magnetometers) to detect symptoms of Parkinson's disease and obtained superior performance compared to vision-based and handwriting-based approaches. Celik et al. [29] proposed a sensor fusion framework using data from inertial (accelerometers, and gyroscopes) and electromyography sensors to perform gait analysis in healthy and stroke-affected participants. In a previous study [30], we tested different sensors in a fetal kick simulator and found both redundant and complementary properties from accelerometers, acoustic sensors, and piezoelectric diaphragms in terms of the intensity (signal-to-noise ratio) and duration of response, frequency response, time-frequency domain representations, and ability to recognize changes in fetal kick characteristics (e.g. intensity, duration, kick distance, etc.). Based on those results, we proposed that a combination of these sensor types will be better equipped to detect and characterize FM and reduce false positive detections compared to a single sensor type. In the current work, we present the design, development, and validation of a multi-modal wearable FM monitor consisting of a combination of accelerometers, acoustic sensors, piezoelectric diaphragms, a force sensor, and a custom-made miniaturized DAQ system embedded in a wearable garment. We also outline the architecture of a novel data analysis algorithm combining data-dependent thresholding and machine learning to automatically detect FM signals. Finally, we evaluate the performance of this novel device through at-home use by the participants to replicate the real-world application environment of a wearable FM monitor.

2. Hardware design

A complete package of hardware, software, and embedded systems was developed for the wearable FM monitor. The hardware system consisted of a heterogeneous sensor network and a bespoke miniaturized DAQ system embedded in an elastic wearable belt as shown in Fig. 1. A detailed description of the design of the hardware system for the current multi-modal FM monitor is provided below.

2.1. Design of the heterogeneous sensor network

A combination of three types of vibration sensors with both redundant and complementary response characteristics to fetal kicks [30], namely the acoustic sensor, the accelerometer, and the piezoelectric diaphragm, was used in the FM monitor (Fig. 1(a)). The acoustic sensor used in the FM monitor is a custom-made proprietary sensor from the Biomechatronics Lab at Imperial College London, UK [31]. It consists of a thin membrane covering a sealed chamber containing a MEMS microphone (Knowles Corp., Itasca, Illinois, USA, MPN-SPU1410LR5H-QB) at the opposite end (Fig. 1(c)). The membrane translates the surface vibration into an intra-chamber pressure fluctuation, which is recorded by the microphone. The dimension of the sealed chamber is adjusted to efficiently capture low-frequency vibrations (1 – 50 Hz) [32]. A breakout board (16 mm × 18.5 mm) of ADXL335, a MEMS accelerometer from Analog Devices Inc. (Massachusetts, USA) (Fig. 1(a)), was selected as the accelerometer sensor for the FM monitor. ADXL335 has a measurement range of ± 3 g and a sensitivity of 300 mV/g. The breakout board has anti-aliasing low-pass filters with a cut-off frequency of 50 Hz connected to each output pin of the breakout board [33], which is compatible with the target frequency range of 1 –

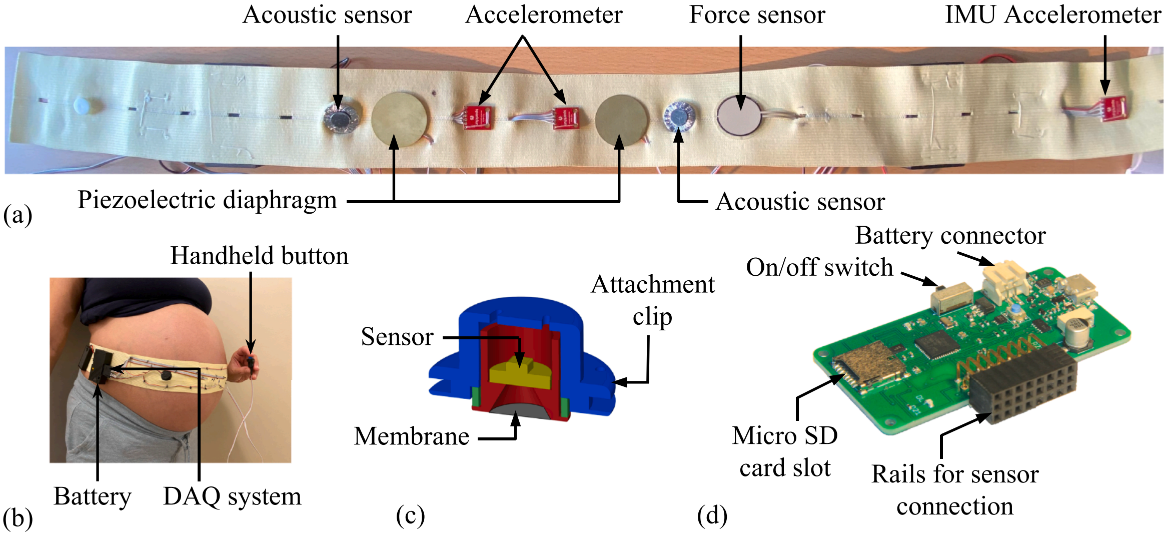

Fig. 1. Hardware system for the wearable FM monitor. (a) Sensors embedded in an elastic belt, (b) belt worn by a pregnant participant, (c) CAD design of the custom-made acoustic sensor, and (d) miniaturized (62 mm × 31 mm) DAQ system designed for the FM monitor.

30 Hz for FM signals [18,19,34]. The piezoelectric diaphragm from Murata Electronics (Kyoto, Japan, MPN- 7BB-35-3L0) was the third type of vibration sensor used in the FM monitor (Fig. 1(a)). It consists of a 25 mm diameter piezoelectric ceramic plate glued to a 35 mm diameter brass plate [35]. Finally, a piezoresistive force sensor, FlexiForce A401 (Tekscan Inc., South Boston, MA, USA) [36], was used to record how tightly the belt was attached to the abdomen (Fig. 1(a)) during the data collection. This sensor was chosen due to its suitability of dimension (31.8 mm diameter) for integration into a wearable belt and excellent repeatability of measurement (±2.5 %) [36].

Two sensors of each selected vibration sensor type, namely, acoustic sensor, accelerometer, and piezoelectric diaphragm, were used as the FM detecting sensors. These sensors were positioned symmetrically with respect to a transverse axis of the belt as shown in Fig. 1(a). Sufficient distances were maintained between consecutive sensors to ensure non-interference with each other's attachment quality. The distance between the end sensors was around 28 cm, which further increased to around 30 cm during use due to the fabric's elongation. Considering the ability of these sensors to detect vibrations due to simulated fetal kicks from a relatively large distance (20 cm) [30], the overall span of the sensors in the belt was considered sufficient to detect vibrations from any regions of the maternal abdomen. An additional accelerometer was placed outside of the abdominal region (Fig. 1(a)) as an inertial measurement unit (IMU) for maternal body movement detection. A latex-free elastic belt from Koninklijke Philips N.V. (Amsterdam, Netherlands, MPN- M2208A) was used as the base belt for the FM monitor (Fig. 1(a)). Due to its stretchability, the belt ensures good contact between the embedded sensors and the skin.

2.2. *Design of the DAQ system*

The main design requirements for the DAQ system were a sufficiently small size to be embedded in a wearable belt, the ability to record data with a sufficiently high sampling rate for capturing FM signal (usually 100 Hz [18–20,34]), the availability of an onboard data storage facility, and to be powered by a small portable battery. It also needed to have sufficient source and ground connections to supply power to all the sensors. Considering all these requirements, a bespoke miniaturized DAQ system (62 mm × 31 mm) was designed based on a low-power 32-bit microcontroller unit (ATSAMD21G18A-48, ARM Cortex-M0+ microcontroller unit) (Fig. 1(d)). The system can acquire data from 8 input channels (7 analogs, 1 digital) simultaneously at a sampling rate of up to 1024 Hz with an ADC resolution of 12-bit. While a sampling rate of 100 Hz is sufficient for capturing FMs [18–20,34], a high sampling rate of 1024 Hz was used for data collection to minimize the aliasing of high-frequency noises, which was particularly important for acoustic sensors and piezoelectric diaphragms as they do not have anti-aliasing filters attached to their outputs. The DAQ system has an onboard micro-SD card slot to store the sampled data (Fig. 1(d)). It runs at an operating voltage of 3.3 V, which can be supplied by a Li-Po single-cell battery. A 3 × 8 angle socket was used in the board to facilitate horizontal connections to the sensors, including the connections for providing power supply (3.3 V, and ground connections) to each sensor (Fig. 1(d)). The DAQ system also includes the onboard signal conditioning circuits necessary for the force sensor and the piezoelectric diaphragms [30].

Two DAQ systems, one on either side of the abdomen, were used to record data from all eight sensors used in the FM monitor, which required a total of 14 analog channels (three analog channels for each accelerometer, and one analog channel for each of the remaining sensors). A handheld button was connected to the digital input pin of both DAQ systems to record maternal sensation detections and to ensure the synchronicity of data collection from both systems (Fig. 1(b)). Each DAQ system was powered by an 850 mAh Li-Po battery. Custom-made 3D-printed boxes were used to embed the DAQ systems and the associated batteries in the belt (Fig. 1(b)).

3. Participants and protocols

Women with a singleton pregnancy at a gestational age between 24 – 40 weeks were invited for data collection to validate the performance of our FM monitor. Ethical permission for the study was approved by the Research Governance and Integrity Team at Imperial College London (ICREC reference- 20IC6329). Five participants were recruited for the data collection, where each participant took part in multiple sessions throughout their pregnancy. All the data collection sessions were self-operated by the participants from their homes to create a real application environment for such a device and were monitored virtually by one of the investigators (A.K. Ghosh).

During the data collection sessions, participants sat comfortably wearing the device and recorded their perceptions of FM by pressing the handheld button attached to the monitor. While participants were instructed to remain relatively stationary to minimize artifacts due to maternal body movements, they could change their sitting positions during the data collection sessions to ensure their comfort. Although the current device can handle signal artifacts due to maternal body movements, data collection during a comfortable sitting position ensured the best possible condition for the pregnant participants to perceive sensations of FMs, which require uninterrupted attention. 33 h of data were collected from all the participants at different stages of their pregnancy (Table 1). It is noteworthy that the overall size of this data set is significantly larger than most of the studies on the performance validation of passive FM monitors [18–20,23,34].

4. The architecture of the data analysis algorithm

Most of the available algorithms for FM monitor data analysis are based on thresholding-based techniques [16,18,19,23,37,38]. However, due to the complex overlapping nature of FM and the associated signal artifacts [19], thresholding-based approaches generally perform poorly in terms of removing signal artifacts. Considering this problem, some recent works [20,34,39] adopted machine learning-based techniques to enhance the performance of the algorithm. Previously described machine learning-based algorithms for FM detection are based mainly on a fixed-length data segmentation approach, which creates a skewed training data set due to a significantly higher number of non-FM segments than FM segments. Additionally, fixed segmentation length prevents the determination of the duration of individual FM activities, as a fixed-length segment may contain a partial FM activity, a single FM activity, or multiple FM activities.

We have designed a new data analysis architecture combining data-dependent thresholding, sensor fusion, and machine learning-based techniques. Our algorithm uses a thresholding process to eliminate regions with an extremely low probability of being FM and hence significantly reduces the skewness of the data set. Additionally, the thresholding process creates data segments with variable lengths based on the continuation of signal amplitude above the threshold value. This feature is critical as the duration of FMs can vary widely, which must be addressed in the signal fusion process. The machine learning classifier is then applied to the segmented data sets to detect FMs and remove signal artifacts. The overall algorithm consisted of seven major steps: pre-processing, segmentation, sensor fusion, feature extraction, classification, detection matching, and post-processing as illustrated in Fig. 2. We quantified the advantages of the machine learning classifier by maintaining algorithm attributes designed to predict FM based on thresholding alone, in which case feature extraction and classification steps were omitted (Fig. 2). Implementation details in each step include:

i. Pre-processing: In this step, raw signals from all sensors were filtered to remove noise components outside of the frequency range of the targeted measurements. In the case of accelerometers, the magnitude of acceleration was considered the raw signal. Signals from the FM-detecting sensors (acoustic sensors, accelerometers, and piezoelectric diaphragms) (S_i) were passed through a 1 – 30 Hz band-pass filter. The lower limit of this passband (= 1 Hz) was chosen to remove the noise components due to maternal breathing [19], and the higher limit (= 30 Hz) was based on the expected spectrum of FM signals [19,23,30]. However, this bandpass filter is not expected to remove signal artifacts due to maternal heartbeats, the frequency of which generally lies between 1.1 and 1.7 Hz [40] and overlaps with the frequency spectrum of FM signals.

Table 1
Summary of participant's information.

Participant no.	Gestational age range (weeks)	Data collection period (hours)	No. of maternally detected FMs
1	33–40	3.42	524
2	24–38	4.55	295
3	27–38	8.41	1229
4	26–40	4.90	306
5	24–38	11.77	1282
Overall	24–40	33.05	3636

Signals from the IMU accelerometer (S_{IMU}) were passed through a 1 – 10 Hz band-pass filter, which was experimentally determined as an optimum passband for detecting maternal body movements (Appendix A). The signal from the force sensor was passed through a low-pass filter with a cutoff frequency of 10 Hz.

ii. Segmentation: Each pre-processed data set (except the force sensor data) was assumed to be a random signal $\hat{S}$. For i th FM-detecting sensor data $\hat{S}_i$, a noise estimate e_i was obtained by taking the median of the set of values in $|\hat{S}_i|$ not exceeding the lower quantile q (= 0.25) of $|\hat{S}_i|$:

$$e_i = median\left(\left\{x \in |\widehat{S}_i| \middle| x \le LQ(|\widehat{S}_i|)_q\right\}\right). \tag{1}$$

Then, a threshold level (h_i) for the i th sensor data was defined as

$$h_i = e_i l, \tag{2}$$

where l is a multiplier representing the minimum signal-to-noise ratio required for the sensor data to be a candidate FM. Due to the estimation of e_i based on the sensor data of the recording session being analyzed, the obtained threshold value (h_i) is data-dependent, which helps the algorithm to better handle the fluctuations in background noise across data collection sessions. After calculating h_i, a binary segmentation (detection) map D_{ij} was created by thresholding the signals above h_i:

$$D_{ij} = \begin{cases} 1 \ if \ |\widehat{S}_i| \ge h_i \\ 0 \ otherwise \end{cases}, \tag{3}$$

where j is the data sample number in the i th sensor data. Based on the assessment of the current data set, a value of $l = 30$ was selected, which captured 97 % of maternal sensation detections in the thresholded data (Section 5.1). $l > 30$ resulted in the capture of a much fewer number of maternally sensed FM signals in D_{ij} while $l < 30$ substantially increased non-FM segments in D_{ij} without any noticeable increase in the number of FM segments. This thresholding process was manually optimized to remove background noise, including signal artifacts due to maternal heartbeats. However, signal artifacts due to maternal heartbeats can also be stronger than FM signals [20], which are expected to be identified and removed by the machine learning classifier described in the classification step of this section.

A candidate FM map C_{ij} was then created by dilating the non-zero values in D_{ij} by 3.0 s (1.5 s forward and 1.5 s backward) (Fig. 2). This allowed the joining of the segmented signals to provide enough signal length for representing an FM event. This length of dilation was chosen based on the 3 s mean duration of fetal movements found in the literature [20,41]. In the case of the IMU accelerometer, the thresholding was done based on a fixed value of h (= 0.002):

$$(D_{IMU})_j = \begin{cases} 1 \ if \ \left|(S_{IMU})_j\right| \ge 0.002 \\ 0 \ otherwise \end{cases}. \tag{4}$$

A maternal body movement map (B_j) was then created by dilating the non-zero values in the $(D_{IMU})_j$ by 4 s (2 s forward and 2 s backward) (Fig. 2). These values of h and the dilation length for the IMU accelerometer were experimentally optimized by recording the accelerometer responses due to maternal body movements (Appendix A).

iii. Sensor fusion: At first, the maternal body movement map (B_j) was checked for temporal overlaps with the candidate FM map (C_{ij}) to remove signal artifacts due to maternal body movements:

$$\widehat{C}_{ij} = C_{ij}.*(1 - B_j), \tag{5}$$

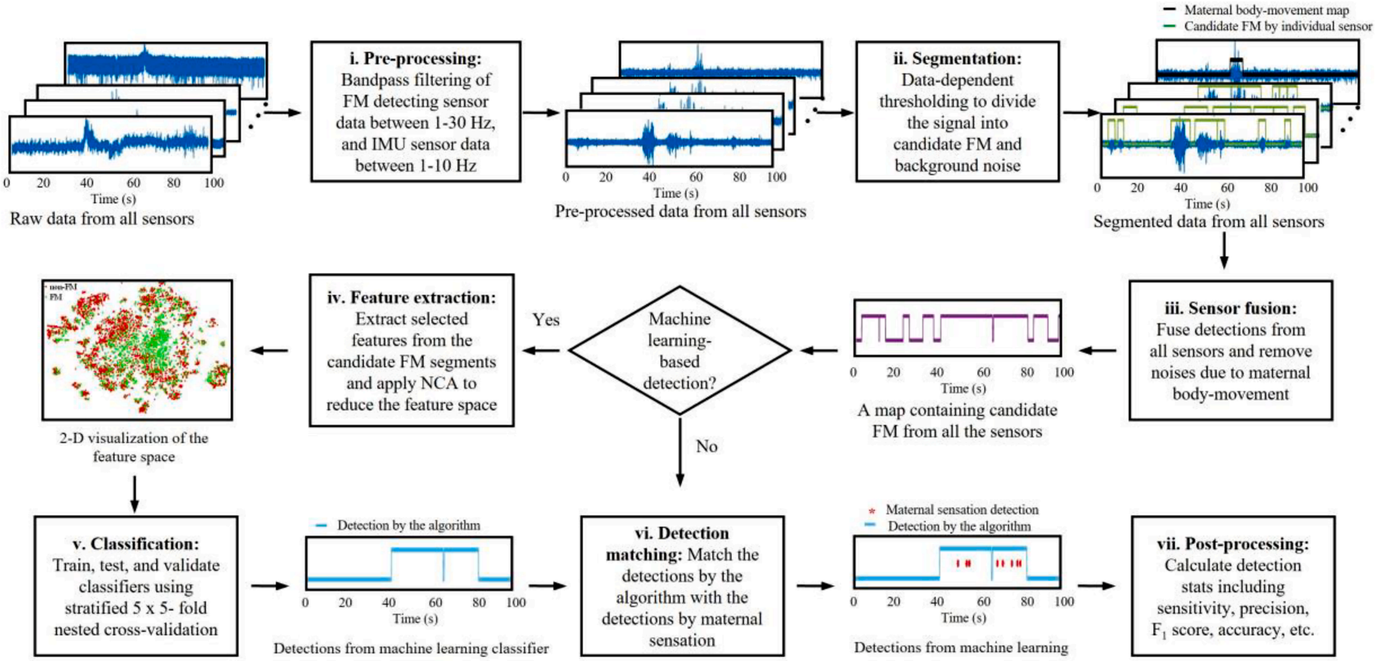
Maternal body-movement map
Candidate FM by individual sensor
0 20 40 60 80 100
Time (s)
Raw data from all sensors
i. Pre-processing: Bandpass filtering of FM detecting sensor data between 1-30 Hz, and IMU sensor data between 1-10 Hz
Pre-processed data from all sensors
ii. Segmentation: Data-dependent thresholding to divide the signal into candidate FM and background noise
Segmented data from all sensors
iii. Sensor fusion: Fuse detections from all sensors and remove noises due to maternal body-movement
A map containing candidate FM from all the sensors
Machine learning-based detection?
Yes
No
iv. Feature extraction: Extract selected features from the candidate FM segments and apply NCA to reduce the feature space
non-FM
FM
2-D visualization of the feature space
v. Classification: Train, test, and validate classifiers using stratified 5 x 5- fold nested cross-validation
Detection by the algorithm
Detections from machine learning classifier
vi. Detection matching: Match the detections by the algorithm with the detections by maternal sensation
Maternal sensation detection
Detection by the algorithm
Detections from machine learning
vii. Post-processing: Calculate detection stats including sensitivity, precision, F₁ score, accuracy, etc.

Fig. 2. Flow diagram of the data analysis algorithm. The algorithm consists of seven major steps and allows detections by two approaches: detection based on thresholding, and detection based on a combination of thresholding and a machine learning classifier. In the case of detection by thresholding only, steps iv (feature extraction) and v (classification) are omitted.

where $\widehat{C}_{ij}$ is the candidate FM map for i th sensor data after the removal of signal artifacts due to maternal body movement. After that, two different cases were considered to determine the final output from this step:

a. Detection based on thresholding only: In this case, candidate FM maps from the same type of sensors on the left and the right sides were combined with a logical OR operator to get the candidate FM map for each sensor type ($\widehat{C}_k$):

$$\widehat{C}_k = \widehat{C}_{kLeft} \| \widehat{C}_{kRight}, \tag{6}$$

where k represents the type of sensor. Three different sensor fusion schemes were considered to determine the final detection by the algorithm: 1) scheme one- detections common to at least one type of sensor are FMs, 2) scheme two- detections common to at least two types of sensors are FMs, and 3) scheme three- detections common to at least three types of sensors are FMs.

b. Detections based on machine learning: In this case, candidate FM maps from all the sensors were fused using the logical OR operator:

$$\widehat{C}_f = \widehat{C}_1 \| \, \widehat{C}_2 \| \ldots \| \widehat{C}_6, \tag{7}$$

where $\widehat{C}_f$ is the fused candidate FM map that holds candidate detections from all the sensors (Fig. 2).

iv. Feature extraction: At first, all non-zero segments in the fused candidate FM map ($\widehat{C}_f$) were labeled as FM or non-FM segments based on their intersection with the sensation map (SN_{map}), which was created by extending each maternal sensation detection by 5 s to the past and 2 s to the future. A larger extension to the past than the future was used to create the sensation map to compensate for maternal reaction time to FM. The reason for choosing this particular length of the extension is explained in Appendix B. If a non-zero segment in $\widehat{C}_f$ overlapped with a non-zero segment in SN_{map}, it was labeled as an FM segment and otherwise a non-FM segment. After that, pre-processed sensor data ($\hat{S}_i$) corresponding to each FM and non-FM segment were collected from all the FM-detecting sensors for feature extraction.

A total of 16 distinctive features, including statistical, time domain, and frequency domain features, were extracted from each sensor data based on a rigorous literature review [20,23,34,39] and observation of the data set (Table 2). The overall duration of each data segment (which is the same for all the sensors for a particular segment) was also considered a feature (11th feature on the statistical and time domain feature type in Table 2). Hence, a total of 97 features from six FM-detecting sensors were extracted for each data segment.

To compensate for the variation of the background noise across the data collection sessions, statistical and time domain features (Table 2) were extracted from the thresholded sensor data ($(\widehat{S}_{ijk})_{thd}$):

Table 2
Features extracted from the sensor data to train machine learning classifiers.

Feature type	Extracted features
Statistical and time domain	(i) max amplitude, (ii) mean amplitude, (iii) standard deviation, (iv) interquartile range, (v) skewness, (vi) kurtosis, (vii) signal energy, (viii) duration of the data points above the threshold, (ix) the mean amplitude of the data points above the threshold, (x) signal energy of the data points above the threshold, (xi) duration of each data segment
Frequency domain	(i) dominant frequency mode, (ii – vi) spectrum energy for five frequency windows: 1 – 2 Hz, 2 – 5 Hz, 5 – 10 Hz, 10 – 20 Hz, and 20 – 30 Hz.

$$\left(\widehat{S}_{ijk}\right)_{thd} = \left|\widehat{S}_{ijk}\right| - h_{ik} \tag{8}$$

where $\widehat{S}_{ijk}$ is the j-th data segment from i th sensor data in the k-th data collection session, and h_{ik} is the threshold value for i th sensor in the k-th data collection session. Frequency domain features (Table 2) were extracted directly from $\widehat{S}_{ijk}$.

To compensate for the variation in the ranges of different features, the normalization of each feature was performed as follows:

$$\left(x_j\right)_{normalized} = \frac{x_j - \mu_j}{\sigma_j}, \tag{9}$$

where μ_j and σ_j are the mean and the standard deviation of the j-th feature (x_j), respectively. Finally, to improve the computational efficiency of the algorithm, feature space was reduced by feature selection through regularized neighborhood component analysis (NCA) [42]. The *fscnca()* function from MATLAB (MathWorks, Inc.) was used to perform the NCA. The optimum regularization parameter for NCA was selected as 1.34×10^{-4} based on a 5-fold cross-validation. The features with weights higher than 0.05 % of the maximum feature weight based on the NCA-based feature ranking were finally selected for training and testing the algorithm. This resulted in the selection of the top 30 out of 97 features as shown in Appendix C.

v. Classification: In this step, machine learning classifiers were trained and tested to distinguish between FM signal and non-FM signal artifacts. We stress that the architecture of the overall algorithm was designed to allow the integration of a range of machine learning classifiers in this phase of processing. Normalized values of features from labeled FM and non-FM data segments were used to train the classifiers. Only the features selected through the NCA-based feature ranking process described in the previous step (iv. Feature extraction) were used in the training process. 5 × 5-fold nested cross-validation was used to select the model parameters and obtain the generalized performance of the trained classifiers. This process involves an inner and an outer loop. In the outer loop, the overall data set was divided into five stratified sets. One by one, each set was selected as the test data set and the rest of the sets were combined to create the training data set (outer). Each training data set was further subdivided into five stratified sets in the inner loop, where one by one each set was selected as the validation data set and the rest of the data sets were combined to create the training data set (inner). The inner loop was used to select the model parameters and the outer loop was used to find the generalized prediction of the algorithm. Predictions on the test data set in each iteration of the outer loop were combined to get the generalized prediction for the whole data set. While the data segmentation process helps to reduce class imbalance in the data set for training a machine learning model, the overall data set still had a higher number of non-FM events than FM events. Therefore, to provide higher importance to detecting true events and to compensate for any remaining skewness (class imbalance) in the data set, the cost of misclassifying a true event was set to be twice the cost of misclassifying a false event.

Four different classifiers, namely the neural network, the random forest, the support vector machine, and the logistic regression, were tested in this research to show the performance of the current algorithm for a wide range of machine learning classifiers. These four classifiers were specifically chosen based on their superior performances in FM detection relative to other machine learning classifiers [20]. A brief description of the machine learning classifiers used in the current research is given below-

a Neural network: A neural network classifier with ReLU (rectified linear unit) activation function for hidden layers, sigmoid activation function for the output layer, and binary cross-entropy loss function was developed using Tensorflow 2.0, an open-source software library for machine learning from Google Brain. The classifier uses the backpropagation algorithm to train the model parameters. The architecture of the neural network (number and size of layers) was optimized using stratified 5-fold cross-validation where the number of hidden layers was varied between 1 and 5 layers and the size of each hidden layer was varied between 10 and 250 neurons. The finally selected neural network had a single hidden layer with 190 neurons. Adding more hidden layers to the classifier, which makes the classifier a deep neural network, did not improve the performance of the algorithm likely due to the relatively small size (3636 FM events in total) and dimension (30 features after NCA-based feature selection) of the current data set. In general, deep learning-based classifiers are more suitable for applications with large and high-dimensional data sets, such as speech recognition, image classification, etc. [43, 44]. For low-dimensional data sets with relatively small sizes, like the one used in the current study, machine-learning classifiers can produce superior performance with better interpretable results than deep neural networks [43].

b Random forest: A random forest classifier was developed by considering an ensemble of 100 classification trees using the *fitcensemble()* function from MATLAB (MathWorks, Inc.). The algorithm randomly selects a portion of the overall feature vector at each split of the tree nodes to create a random forest. Gini's diversity index was used as the splitting criterion, and a minimum leaf node size (= 50) was used as the split-stopping criterion [45]. The number of features randomly selected at each split was optimized to be 17 by stratified 5-fold cross-validation. The cross-validation process includes the case where all the available features (all of the 30 features selected based on NCA) were used for training the classifier to automatically include the bagging classifier into the cross-validation model, which provided the best performance for FM detection in a previous research work [20]. The bagging classifier is basically an extension of the random forest classifier in which instead of choosing a portion of the overall feature vector at each split of the classification trees, the whole feature vector is used.

c Support vector machine (SVM): An SVM classifier with the Gaussian kernel was developed using the *fitcsvm()* function from MATLAB. Hyperparameters of the SVM model, namely the kernel scale and the soft margin constant, were optimized using stratified 5-fold cross-validation. The optimum values of kernel scale and soft margin constants were 0.64 and 3.79, respectively.

d Logistic regression: A regularized logistic regression classifier was developed using the *fitclinear()* function from MATLAB. Stratified 5-fold cross-validation was used to optimize the regularization parameter. The optimum value of the regularization parameter was 5.91×10^{-06}.

At the end of the classification step, the candidate FM map ($\widehat{C}_f$) was reconstructed by removing the segments that were classified as non-FM segments by the classifier.

vi. Detection matching: To match between the detections from the algorithm and the maternal sensation, time windows were created by extending each maternal sensation detection 5 s into the past and 2 s into the future. Time windows that overlapped with maternal body movements (represented by the body movement map B_j) were not considered for further analysis. Candidate FM that overlapped with maternally sensed FM time windows were considered true positive detections (TPD). If multiple candidate FM overlapped with the same time window, only one TPD was considered. The remaining maternally sensed time windows were considered false negative detections (FND). The remaining detections by the algorithm were used to count the false positive detections (FPD) based on a 7 s time window similar to counting the TPD, i.e. if multiple detections fall within the same 7 s window, only a single FPD was counted. The regions in the signal that were not identified as TPD, FPD, or FND were considered true negative detections (TND). Again, the number of TND was calculated based on a time window of 7 s. It should be mentioned here that this 7 s window size was only used to calculate detection statistics for evaluating the performance of the algorithm for the validation study. The original detection by the algorithm as obtained at the end of the classification step for the combined thresholding and machine learning-based algorithm, or obtained at the end of the sensor fusion step for the thresholding-based algorithm does not depend on any predetermined fixed time window length.

vii. Post-processing: In this step, four performance metrics, namely, sensitivity, precision, F_1 score, and accuracy were calculated as follows-

$$Sensitivity = TPD/(TPD + FND), \tag{10}$$

$$Precision = TPD/(TPD + FPD), \tag{11}$$

$$F_1\ score = 2 \times (PPV \times SEN)/(PPV + SEN) \tag{12}$$

$$Accuracy = (TPD + TND)/(TPD + TND + FPD + FND). \tag{13}$$

Sensitivity expresses what proportion of the maternal sensation detection was also detected by the algorithm, and precision expresses what proportion of the total detections by the algorithm were true positives. The F_1 score combines sensitivity and precision as a harmonic mean (reciprocal of the arithmetic mean of reciprocals) to give a single performance metric. Accuracy expresses the overall true detections as a proportion of all the detections by the system. It should be mentioned here that all these performance metrics express the ability of the current device to detect maternally sensed FMs. In addition to the above four parameters, precision vs. recall (PR) curves were plotted to understand the trade-offs between the precision and the recall for

Table 3
Performance of the thresholding-based algorithm for different sensor fusion schemes. AUPRC = area under the precision vs. recall curve.

Sensor fusion scheme	Sensitivity	Precision	F_1 score	Accuracy	AUPRC
Scheme one: detections common to at least one type of sensor	0.97	0.34	0.50	0.58	0.63
Scheme two: detections common to at least two types of sensors	0.88	0.51	0.65	0.79	0.74
Scheme three: detections common to at least three types of sensors	0.71	0.64	0.67	0.85	0.72

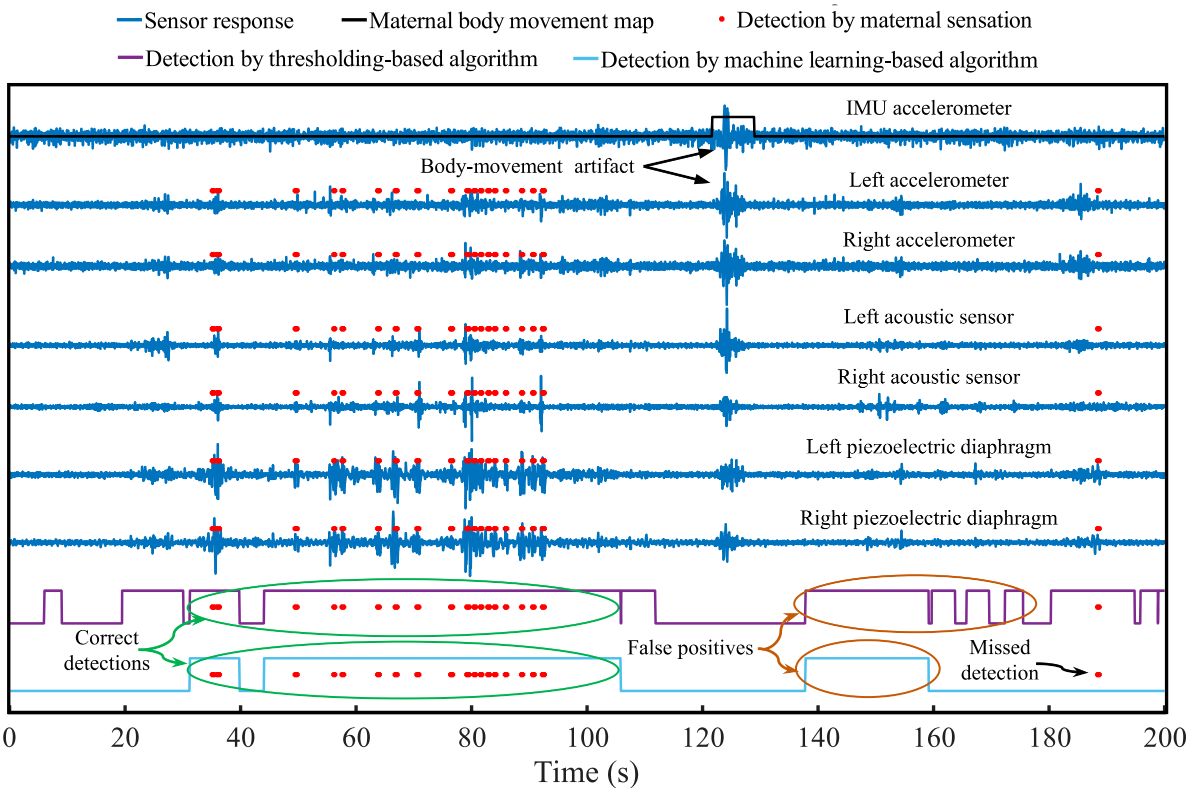

Fig. 3. A typical example of responses from different sensors in the FM monitor and the outputs from the data analysis algorithms. Correspondence between maternal sensation detections and sudden jumps in the signal amplitude represents the ability of the sensors to effectively respond to FMs. The algorithms successfully located and removed the time intervals containing signal artifacts due to maternal body movements using the maternal body movement map, which was created by thresholding the signals from the IMU accelerometer. The machine learning-based algorithm was able to remove the majority of false positive detections observed in the thresholding-based algorithm. Detections common to at least one type of sensor (sensor fusion scheme one) were used to generate the output for the thresholding-based algorithm, and the neural network-based classifier was used in the case of the machine learning-based algorithm.

different values of threshold. Here, recall is the same as sensitivity. The area under the PR curve (AUPRC) was also determined and used as a performance metric to compare different algorithms. The PR curve was selected over the receiver operating characteristic curve (sensitivity vs. false positive rate) here due to the inherent imbalance between FM events and non-FM events in the current data sets [46].

5. Results

The FM monitor was deployed on the participant cohort for unsupervised use in their home. Observation of the direct sensor response indicates individual sensors effectively responded to FMs as indicated by the correspondence between maternal sensation detections and sudden spikes in all sensor outputs as shown in Fig. 3. Focusing next on the thresholding and machine learning-based algorithms, both successfully detected and removed time intervals with maternal body movements using the signals from the IMU accelerometer (Fig. 3). The machine learning-based algorithm removed the majority of false positive detections captured by the thresholding-based algorithm (Fig. 3). A detailed analysis of performances from both the algorithms is provided in the following sections.

5.1. Performance of the thresholding-based algorithm

The performance of the thresholding-based algorithm for three different sensor fusion schemes is shown in Table 3. When detections common to at least one type of sensor (which essentially means every detection from all three types of sensors) were considered (sensor fusion scheme one), the algorithm detected 97 % of maternal sensation detections (sensitivity = 0.97). However, 66 % of all detections by the algorithm were false positives (precision = 0.34), which diminishes overall performance (F_1 score = 0.50, accuracy = 0.58). When detections common to at least 2 types of sensors (sensor fusion scheme two) were considered, the amount of false positive detections substantially dropped (precision = 0.51) at the expense of a moderate reduction in sensitivity (0.88), which indicates the strength of the current heterogeneous suite. As a result, the overall performance of the algorithm improved substantially (F_1 score = 0.65, accuracy = 0.79). Finally, the best performance of the thresholding-based algorithm was obtained when the detections common to all three types of sensors were considered (sensor fusion scheme three) leading to an F_1 score of 0.67 and an accuracy of 0.85. The confusion matrix for the sensor fusion scheme three is shown in Fig. 4(a) demonstrating that the algorithm identified the absence of FM events better than the presence of FM events. Despite substantial improvement in the performance compared to schemes one and two, the sensor fusion scheme three still provided a relatively weak correlation (R^2 = 0.67) with maternal detections for individual data recording sessions as shown in Fig. 4(b).

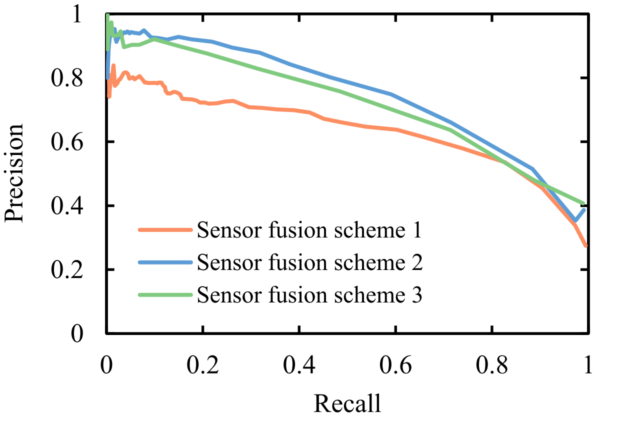

Fig. 5. Precision vs. recall (PR) curve for the thresholding-based data analysis algorithm for different sensor fusion schemes. These curves represent the trade-offs between sensitivity (recall) and precision for different sensor fusion schemes and were generated by varying the threshold multiplier (l) between 20 and 2000 (Eq. (2)) to obtain different combinations of precision and recall.

To demonstrate the trade-offs between the sensitivity and the precision for different sensor fusion schemes, precision vs. recall (PR) curves were plotted as shown in Fig. 5. To obtain different combinations of precision and recall (same as sensitivity) to plot these curves, the threshold value was varied by varying the threshold multiplier (l) between 20 and 2000 (Eq. (2)). Fig. 5 shows that the PR curve for sensor

Table 4
Performance of the machine learning-based algorithm for different classifiers. AUPRC = area under precision vs. recall curve.

Classifier	Sensitivity	Precision	F_1 score	Accuracy	AUPRC
Neural network	0.82	0.76	0.79	0.90	0.86
Random forest	0.84	0.71	0.77	0.89	0.85
Support vector machine	0.81	0.76	0.78	0.90	0.82
Logistic regression	0.78	0.69	0.73	0.88	0.78

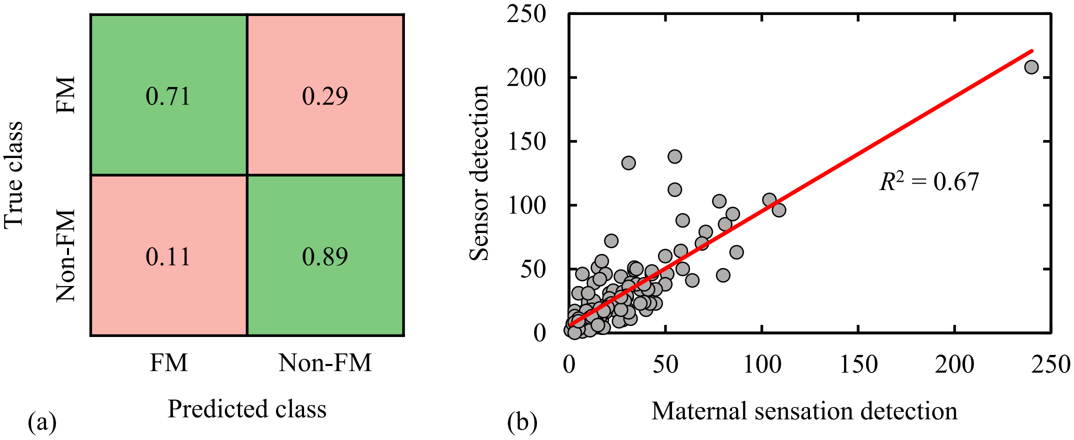

Fig. 4. Output from the thresholding-based algorithm with sensor fusion scheme three in terms of (a) normalized confusion matrix and (b) correlation between the sensor detection and maternal sensation detection for individual data recording sessions. In sensor fusion scheme three, only the detections common to all three types of sensors were considered as detected FMs by the algorithm. Here, the true class represents detection by maternal sensation and R represents Pearson's correlation coefficient.

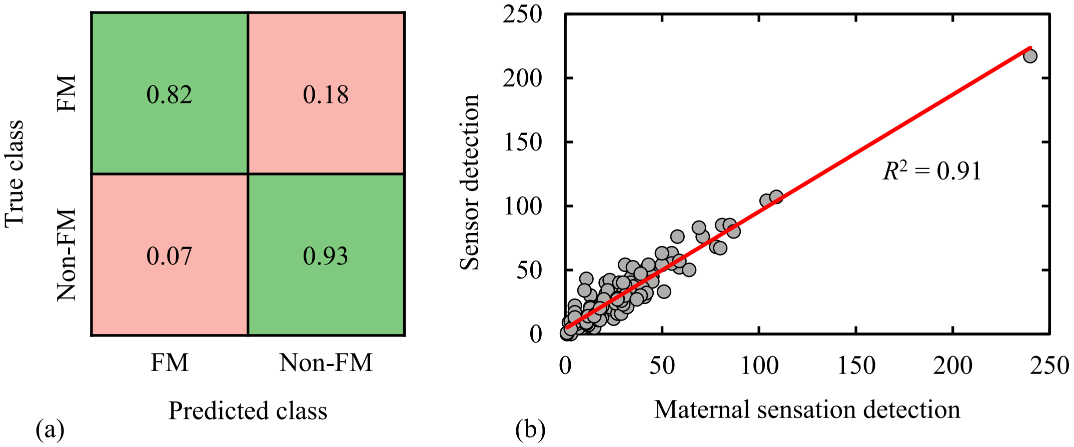

Fig. 6. Output from the machine learning-based algorithm with the neural network-based classifier in terms of (a) normalized confusion matrix and (b) correlation between the sensor detection and maternal sensation detection for individual data recording sessions. Here, the true class represents detection by maternal sensation and R represents Pearson's correlation coefficient.

fusion scheme 2 covers the PR curves for the other two sensor fusion schemes for the majority of the plot area. This is also reflected by a higher area under the PR curve (AUPRC) for the sensor fusion scheme two (AUPRC = 0.74) compared to sensor fusion scheme one and sensor fusion scheme three (AUPRC = 0.63, and 0.72, respectively) (Table 3). This indicates that while sensor fusion scheme three slightly outperformed sensor fusion scheme two for the combination of sensitivity and precision obtained with a specific value of threshold multiplier (l = 30) as presented in Table 3, for the majority of other values of threshold, sensor fusion scheme two will perform better than sensor fusion scheme three.

5.2. Performance of the machine learning-based algorithm

Adding a machine learning classifier led to a significant improvement in the performance of the algorithm relative to the thresholding-based algorithm alone, irrespective of the type of classifier, as shown in Table 4. The best overall performance was achieved by the neural network-based classifier (F_1 score = 0.79, accuracy = 0.90, and AUPRC = 0.86), which was very closely followed by the support vector machine (F_1 score = 0.78) and the random forest (F_1 score = 0.77) classifiers. The performance of the logistic regression classifier was the weakest (F_1 score = 0.73) among different classifiers.

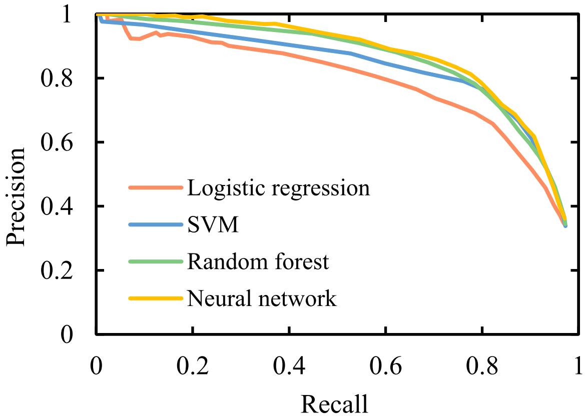

Fig. 7. Precision vs. recall (PR) curve for the machine learning-based data analysis algorithm for different classifiers. These curves represent the trade-offs between sensitivity (recall) and precision for different threshold values used in the machine learning classifiers to convert class probabilities into predictions of specific classes. The threshold multiplier (l) used in the data segmentation step (Eq. (2)) was kept fixed at 30 while generating these curves.

The confusion matrix obtained from the neural network-based classifier (Fig. 6(a)) shows that while the machine learning-based algorithm identified the absence of FM better than the presence of FM, the differential performance between these two classes was substantially reduced compared to the thresholding-based algorithm (Fig. 4(a)). Additionally, incorporating a machine learning classifier led to a large improvement relative to the thresholding-based algorithm in terms of correlation with maternal sensation detections (R^2 = 0.91 for the neural network-based classifier) as shown in Fig. 6(b). Finally, the neural network classifier produced a PR curve that covers the PR curves for all the other classifiers for almost the whole of the plot area (Fig. 7) indicating a superior performance by this classifier for any threshold value or combination of precision and recall.

5.3. Performance across individual participants

We next compared the performance of the thresholding-based and machine learning-based algorithms for individual participants to determine if the same algorithm was consistently performing the best (Fig. 8). We preferentially use the F_1 score over accuracy to represent the overall performance of the algorithm because of the potential risk of bias in the accuracy value due to a large number of TND compared to other types of detection (TPD, FPD, FND) (Eq. (13)). Here, sensor fusion scheme three was used in the case of thresholding-based algorithm and neural network-based classifier was used in the case of machine learning-based algorithm. The performance of the machine learning-based algorithm was consistently better than the thresholding-based algorithm across all the participants (Fig. 8(a) & (b)). Additionally, the machine learning-based algorithm performed more consistently across the participants (standard deviation of F_1 score = 0.05) than the thresholding-based algorithm (standard deviation of F_1 score = 0.07). However, both algorithms performed poorly for participants 2 and 4 compared to the other participants (Fig. 8(a) & (b)). Interestingly, the average force sensor values for participants 2 and 4 were also lower than the other participants as shown in Fig. 8(c). As the force sensor measures the contact pressure between the sensor belt and the participant's abdomen, a decrease in its amplitude indicates a weaker sensor attachment condition. Hence, a considerable drop in the force sensor output could be a reason for the inferior performance observed in the cases of participants 2 and 4. However, it should be mentioned here that no consistent correlation was found between the force sensor values and the device performance in the case of the individual data recording sessions, which can be affected by other factors, such as gestational age

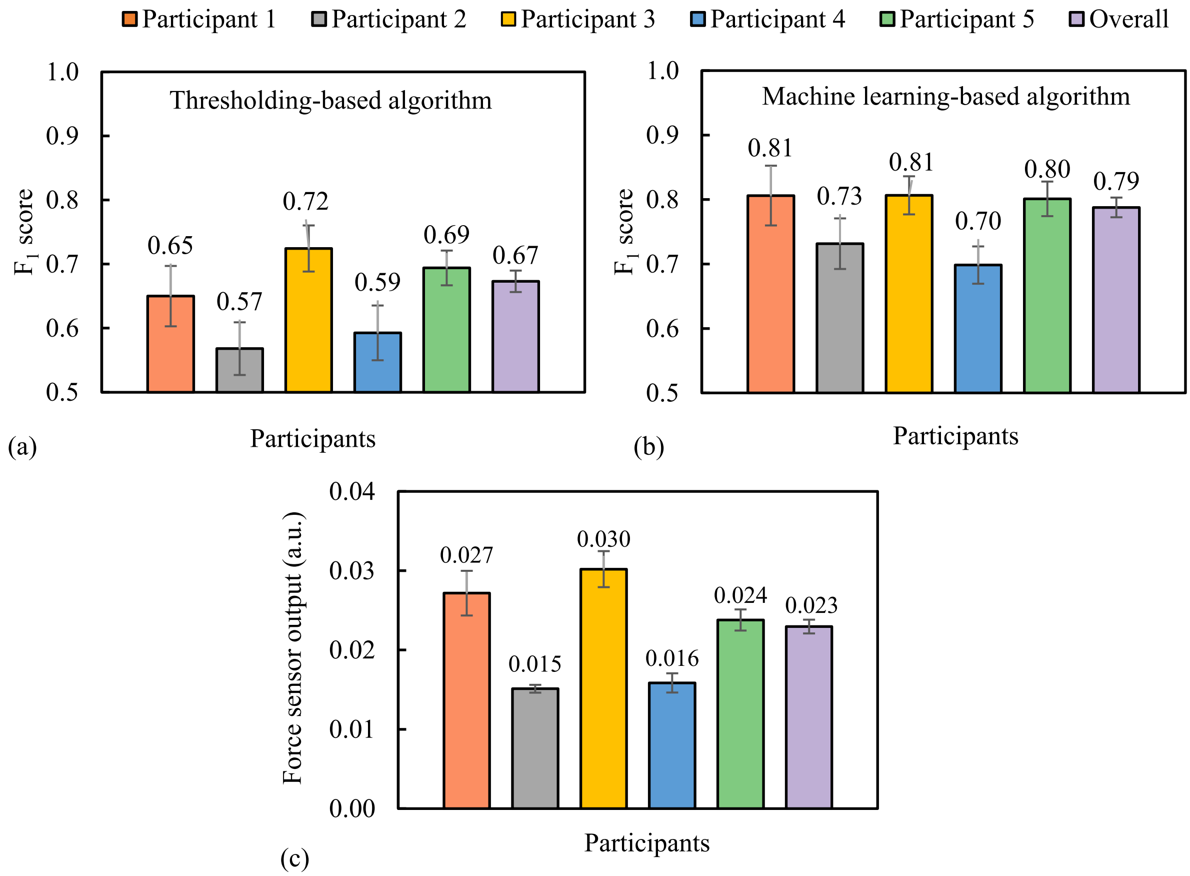

Fig. 8. Performance of algorithms (in terms of F_1 score) across individual participants and its relationship with the belt tightness: (a) performance of the thresholding-based algorithm, (b) performance of the machine learning-based algorithm, and (c) mean force sensor output. Detections common to all three types of sensors (sensor fusion scheme three) were considered for the thresholding-based algorithm and the neural network-based classifier was considered for the machine learning-based algorithm. The error bars for all the plots in this figure are based on standard error ($= \textit{standard deviation}/\sqrt{\textit{no. of samples}}$).

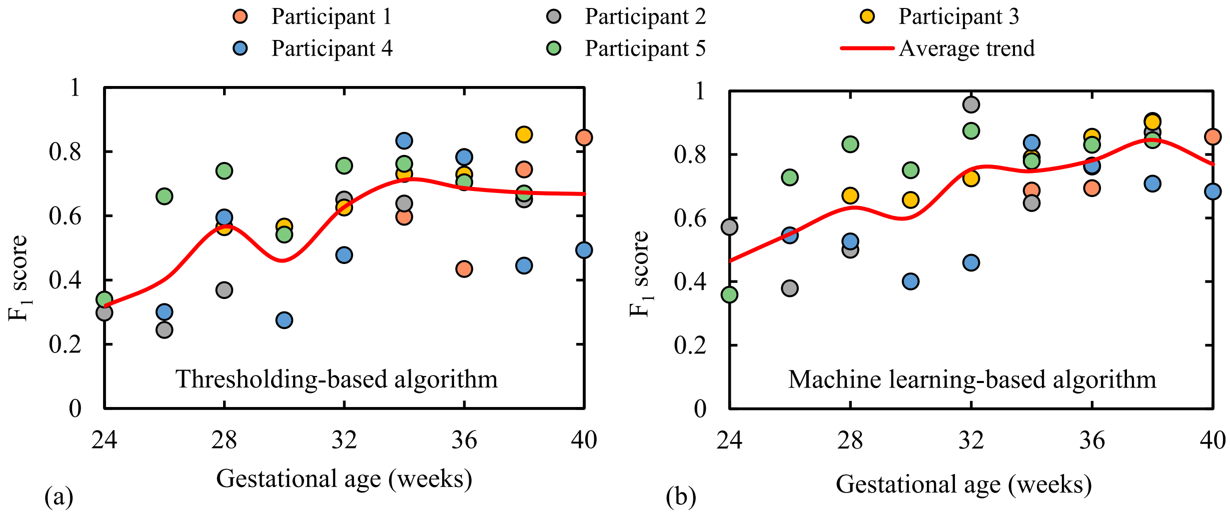

Fig. 9. Variation of performance with gestational age for (a) the thresholding-based algorithm (sensor fusion scheme three), and (b) the machine learning-based algorithm (neural network). Each data point in this plot represents the bi-weekly average F_1 score for an individual participant.

or activity level of the fetus during a particular session for example.

5.4. Variation of performance across gestational age

Our final set of analyses was focused on understanding how the performance of the algorithms varied with the age of the fetus. To analyze that bi-weekly F_1 scores for individual participants were plotted against the gestational age as shown in Fig. 9. The average performance of the machine learning-based (neural network) and thresholding-based (sensor fusion scheme three) algorithms substantially improved with the growth of fetuses between the gestational age of 24 – 32 weeks. Further increases in the gestational age beyond 32 weeks did not have a significant impact on the performance of the algorithms. Both algorithms produced significant variation in performance across the participants for any given gestational age between 24 and 32 weeks, as shown in Fig. 9. However, beyond 32 weeks, the machine learning-based (neural network) algorithm produced a consistent performance with relatively small variation across the participants (Fig 9).

6. Discussion

We have successfully demonstrated the ability of a heterogeneous sensor suite comprised of accelerometers, acoustic sensors, and piezoelectric diaphragms to detect FM when embedded in an elastic wearable garment (Fig. 3). These sensors were combined with an additional IMU accelerometer to detect maternal body movements, making the current FM monitor extremely suitable for usage during regular daily activities.

The performance of the thresholding-based algorithm combining detections from all the sensors (sensor fusion scheme one) was relatively poor (F_1 score = 0.50, accuracy = 0.58). However, a substantial improvement (34 % increase in F_1 score and 47 % increase in accuracy) was achieved by considering a different sensor fusion scheme where only the detections common to all three types of sensors (sensor fusion scheme three) were considered FM events. This demonstrates the power of the current heterogeneous sensor network to improve performance by reducing signal artifacts captured by any individual type of sensor. As all the predicted FM events were simultaneously detected by multiple sensing modalities, this sensor fusion scheme also improved the reliability of detection by the system.

The combination of the machine learning classifier with the thresholding-based data segmentation approach further boosted the ability of the system to distinguish between FM and non-FM signals and delivered a substantially improved performance of the system (accuracy = 0.90, F_1 score = 0.79 for neural network-based classifier). The performance of the machine learning-based algorithm was relatively consistent across the participants (standard deviation of F_1 score = 0.05) and positively correlated with the average tightness between the wearable belt and the abdomen for each participant (Fig. 8), indicating the importance of the sensor attachment quality to obtain optimal performance. The impact of sensor attachment quality on the detection performance also indicates that the performances obtained through non-wearable FM monitors [16,18–20,34], where the sensors were mainly attached using adhesive tapes to ensure a strong and consistent attachment, may not be directly translatable to non-wearable devices.

To understand the performance of the current device in light of prior studies, a comparison was made with the results obtained from previous FM monitors [16,21,24,34] that used maternal sensation as the ground truth similar to the present study (Table 5). Our multi-modal FM monitor significantly outperformed the previous FM monitors, most of which were non-wearable uni-modal devices (Table 5). Considering that the data collection for the current FM monitor was performed through self-operated at-home sessions, the performance of the current device is extremely promising. We believe the use of a multi-modal approach in the current study instead of a uni-modal approach used in the previous studies [16,21,24,34] is a major reason for the improved performance of the current FM monitor. Additionally, the novel data analysis algorithm presented in the current research combined thresholding and machine learning-based schemes to improve upon the algorithms used in [16,21,24,34], which were based on either thresholding-based schemes or machine learning-based schemes. Finally, in line with the findings from previous researchers [16,18,21], we expect further improvements in the performance of our device if evaluated relative to ultrasound detection instead of maternal sensation detection. This is because some of the false positive detections made by the sensor system considering maternal sensation detection as the ground truth may actually be true detections missed by the pregnant participants, which will have a much lower probability of being missed by the ultrasound scanning.

A further key novelty of the current study is the longitudinal analysis of the performance of the FM monitor. The results showed a gradual improvement in the performance of the device with the progression of gestational age (Fig. 9), especially between 24 and 32 weeks, after which the performance was mostly consistent across gestation. Variation of performance across gestation could be due to improvements in maternal awareness of FM with the growth of the fetus [18], stronger vibrations of the maternal abdomen due to fetal movements as the fetus grows, or a combination of both factors. Nevertheless, consistency of performance from 32 weeks onwards indicates the ability of the device to reliably track the patterns of FM during the latter stages of pregnancy.

7. Conclusion

The wearable device proposed in this paper offers several key developments in the design of a passive wearable FM monitor, including the introduction of a heterogeneous sensor fusion-based multi-modal approach to FM detection and a novel data analysis architecture fusing data-dependent thresholding, sensor fusion, and machine learning. The system can monitor FMs outside of a clinical environment with a detection accuracy significantly superior to existing passive FM monitors, most of which are uni-modal devices validated in a controlled experimental environment. Some of the key features of the combined data-dependent thresholding and machine learning-based algorithm presented in this paper include 1) automatic reduction of class imbalance in the data set through thresholding-based data segmentation process, 2) variable-length data segmentation enabling the algorithm to accurately determine the duration of each detected FM activity, 3) data-dependent estimation of threshold level to handle fluctuations of background noises across data collection sessions, and 4) elimination of time periods involving signal artifacts due to maternal body movements enabling the device to be used by pregnant women during regular activities.

The obtained results have shown that the data-dependent thresholding alone does not provide sufficient performance for the FM monitor despite substantial improvements through different sensor fusion schemes. The augmentation of a machine learning classifier with the

Table 5
Comparison of the current FM monitor with the previous FM monitors that used maternal sensation to evaluate the performance. The performance of the current FM monitor presented here is based on the machine learning-based algorithm with a neural network-based classifier. GA = gestational age.

Reference	GA range (weeks)	Sensor system	Wearable/non-wearable	Data collection environment	Performance
Valentin et al. [21]	28 – 39	Four Piezoelectric crystals	Non-wearable	Controlled experimental	F_1 score = 0.53*
Nishihara et al. [16]	19 – 39	Two accelerometers	Non-wearable	Controlled experimental	PABAK = 0.75
Altini et al. [34]	30 – 39	Five accelerometers	Non-wearable	Controlled experimental	F_1 score = 0.70
Our previous work [24]	32 – 39	Five acoustic sensors	Wearable	Controlled experimental	F_1 score = 0.65, PABAK = 0.74
Current work	24 – 40	Two acoustic sensors, two accelerometers, and two piezoelectric diaphragms	Wearable	Self-operated at-home	F_1 score = 0.79, PABAK = 0.81

[*] Calculated from the data provided in the paper.

thresholding-based initial screening of probable FM signals provided the additional artifact removal capability that boosted the performance to a level necessary for the practical application of such a device. Among the different machine learning classifiers tested in this paper (namely, logistic regressing, support vector machine, random forest, and neural network), the neural network performed the best with an overall accuracy of 0.90 and an F_1 score of 0.79. The algorithm also showed a strong correlation ($R^2 = 0.91$) with maternal sensation detections for individual recording sessions. Sufficient tightness of the belt with the abdomen was found to be important for the optimal performance of the device. Finally, the longitudinal study has shown that the device can reliably track FM patterns during the latter stages (32 weeks onwards) of pregnancy.

Based on the obtained results, it can be concluded that the new heterogeneous sensor suite represents a major milestone in the transition of FM detection technology from the non-wearable to the wearable domain. We believe that future iterations of this device will lead to clinical and community translation of a cost-efficient, wearable FM monitor.

Author agreement statement

We the undersigned declare that this manuscript is original, has not been published before and is not currently being considered for publication elsewhere. We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us. We understand that the Corresponding Author is the sole contact for the Editorial process. He/she is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs Signed by all authors as follows:

1. Abhishek K. Ghosh
2. Danilo S. Catelli
3. Samuel Wilson
4. Niamh C. Nowlan
5. Ravi Vaidyanathan

CRediT authorship contribution statement

Abhishek K. Ghosh: Writing – original draft, Methodology, Investigation, Data curation, Software, Formal analysis, Validation, Visualization. **Danilo S. Catelli:** Writing – review & editing, Data curation, Investigation. **Samuel Wilson:** Writing – review & editing, Methodology. **Niamh C. Nowlan:** Writing – review & editing, Conceptualization, Methodology, Resources, Supervision, Funding acquisition, Project administration. **Ravi Vaidyanathan:** Writing – review & editing, Conceptualization, Methodology, Resources, Supervision, Funding acquisition, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the Commonwealth Scholarship Commission, UK (Reference No.: BDCS-2018-46), UK-India Education and Research Initiative (UKIERI) (Grant No.: DST-UKIERI-2016-17-0103), UK Engineering and Physical Sciences Research Council (EPSRC) (Reference: EP/R511547/1), and Imperial College London. The authors would like to express their sincere gratitude to the pregnant women who volunteered for this study.

Appendix

Appendix A. Determination of the threshold value for maternal body movement detection

The maternal body movement map used in the data analysis algorithm (Section 4) was created by thresholding the pre-processed IMU data above a fixed value of 0.002 and then dilating the non-zero values in the thresholded data by 4 s (2 s forward and 2 s backward). These values of threshold and the dilation length for the IMU accelerometer data were experimentally optimized by recording the accelerometer responses due to maternal body movements as shown in Fig. A.1.

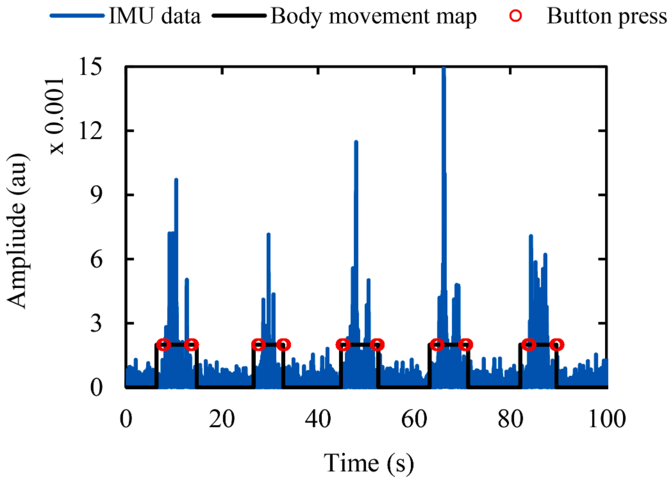

Fig. A.1. Determination of the threshold value for the IMU accelerometer to detect maternal body movements. The start and the end of the body movements are indicated by button presses. The preprocessed IMU data was thresholded above 0.002 and dilated by 4 s (2 s forward and 2 s backward) to create the body movement map.

Appendix B. Effect of detection matching time window size on the performance of algorithms

The performance of any FM detecting algorithm is greatly influenced by the size of the time window (w) used to match the detections between sensors and maternal sensation (or ultrasound detection). However, due to the absence of a universally accepted standard, the value of w has greatly varied (5 – 15 s) across the studies [18,19,23]. To understand the effect of changes in w on the performance of the present algorithms and to find an optimum value of w, the F_1 score and its derivative with respect to w were plotted against w in Fig. B.1. It shows that the performance of both algorithms consistently improved with the increase of w. However, the rate of improvement ($d(F_1 \text{ score})/dw$) gradually decreased and reached a value of around 1 % increase in F_1 score for every 1 s increase in w at around a window size of 7 s for both algorithms. While further increases in w continued to improve the performance of the algorithms, it can also increase the bias of the algorithm toward sensor detections. Hence, 7 s was considered the optimum size of the detection matching time window and was used to generate all results presented in this paper.

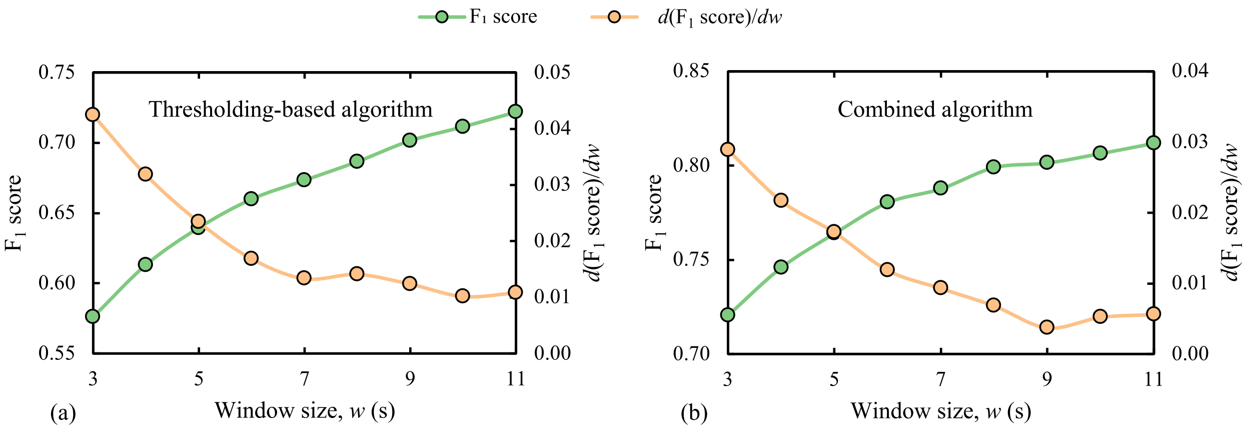

Fig. B.1. Effect of detection matching time window size (w) on the performance of (a) thresholding-based algorithm and (b) combined thresholding and machine learning-based algorithm. Detection matching time windows were created by extending each maternal sensation detection to the past and the future as described in Section 4 (the architecture of the data analysis algorithm).

Appendix C. Feature ranking based on neighborhood component analysis

To improve the computational efficiency of the combined thresholding and machine learning-based algorithm, feature space was reduced through regularized neighborhood component analysis (NCA)-based feature ranking. The features with weights higher than 0.05 % of the maximum feature weight based on the NCA-based feature raking were finally selected (resulting in the selection of the top 30 features) for training and testing the algorithm. Table C.1. shows the selected features and their corresponding NCA-based ranking weights. It can be seen from Table C.1 that the selected top-ranked features consist of an almost equal number of features from each type of sensor (10 features from accelerometer, 10 features from acoustic sensor, and 9 features from piezoelectric diaphragm), which indicates the importance of each type of sensor in the overall sensor combination.

Table C.1
Selected features and their ranking weights based on the neighborhood component analysis (NCA).

Ranking	Feature	Sensor	Ranking weight
1	Mean amplitude above the threshold	Right piezoelectric	7.037
2	Mean amplitude	Left piezoelectric	6.153
3	Interquartile range	Left piezoelectric	6.131
4	Mean amplitude	Right accelerometer	5.093
5	Mean amplitude	Right piezoelectric	4.988
6	Interquartile range	Right acoustic	4.944
7	Skewness	Right acoustic	4.937
8	Mean amplitude	Left acoustic	4.824
9	Mean amplitude	Right acoustic	4.703
10	Skewness	Left piezoelectric	4.583
11	Interquartile range	Right accelerometer	4.322
12	Skewness	Left accelerometer	3.525
13	Duration of segment	Same for all sensors	3.205
14	Skewness	Left acoustic	3.103
15	Dominant frequency mode	Right acoustic	2.662
16	Mean amplitude	Left accelerometer	2.645
17	Dominant Frequency mode	Left acoustic	2.575
18	Dominant Frequency mode	Right accelerometer	2.519
19	Standard deviation	Right accelerometer	1.657
20	Skewness	Right piezoelectric	1.563
21	Interquartile range	Left accelerometer	1.435
22	Duration above threshold	Right acoustic	0.045
23	Mean amplitude above the threshold	Left piezoelectric	0.033
24	Skewness	Right accelerometer	0.026
25	Interquartile range	Left acoustic	0.007
26	Interquartile range	Right piezoelectric	0.006
27	Standard deviation	Left piezoelectric	0.004

(*continued on next page*)

Table C.1 (*continued*)

Ranking	Feature	Sensor	Ranking weight
28	Max amplitude	Right accelerometer	0.004
29	Mean above threshold	Right acoustic	0.004
30	Signal energy	Right accelerometer	0.004

References

[1] J. Lai, N.C. Nowlan, R. Vaidyanathan, C.J. Shaw, C.C. Lees, Fetal movements as a predictor of health, Acta Obstet. Gynecol. Scand. 95 (9) (Sep 2016) 968–975, https://doi.org/10.1111/aogs.12944.

[2] D.J. Bekedam, G.H.A. Visser, J.J. Devries, H.F.R. Prechtl, Motor behavior in the growth retarded fetus, Early Hum. Dev. 12 (2) (1985) 155–165, https://doi.org/10.1016/0378-3782(85)90178-1.

[3] B.S. Richardson, J.E. Patrick, H. Abduljabbar, Cerebral oxidative metabolism in the fetal lamb: relationship to electrocortical state, Am. J. Obstet. Gynecol. 153 (4) (1985) 426–431. Oct 15.

[4] B.S. Richardson, L. Carmichael, J. Homan, J.E. Patrick, Electrocortical activity, electroocular activity, and breathing movements in fetal sheep with prolonged and graded hypoxemia, Am. J. Obstet. Gynecol. 167 (2) (1992) 553–558. Aug.

[5] D.A. Sival, G.H.A. Visser, H.F.R. Prechtl, The effect of intrauterine growth-retardation on the quality of general movements in the human fetus, Early Hum. Dev. 28 (2) (1992) 119–132, https://doi.org/10.1016/0378-3782(92)90107-R. Feb.

[6] M.D. Velazquez, W.F. Rayburn, Antenatal evaluation of the fetus using fetal movement monitoring, Clin. Obstet. Gynecol. 45 (4) (2002) 993–1004, https://doi.org/10.1097/00003081-200212000-00006. Dec.

[7] A.G. Olesen, J.A. Svare, Decreased fetal movements: background, assessment, and clinical management, Acta Obstet. Gynecol. Scand. 83 (9) (2004) 818–826, https://doi.org/10.1111/j.0001-6349.2004.00603.x.

[8] J.M. Turner, V. Flenady, D. Ellwood, M. Coory, S. Kumar, Evaluation of pregnancy outcomes among women with decreased fetal movements, Obstet. Gynecol. Surv. 76 (10) (2021) 583–585, https://doi.org/10.1097/01.ogx.0000798448.64835.4d.

[9] P.J. Dutton, et al., Predictors of poor perinatal outcome following maternal perception of reduced fetal movements - a prospective cohort study, PLoS ONE 7 (7) (2012), https://doi.org/10.1371/journal.pone.0039784. Jul 11.

[10] S. Efkarpidis, E. Alexopoulos, L. Kean, D. Liu, T. Fay, Case-control study of factors associated with intrauterine fetal deaths, MedGenMed 6 (2) (2004,) 53. May 27.

[11] A. Linde, K. Pettersson, I. Rådestad, Women's experiences of fetal movements before the confirmation of fetal death—contractions misinterpreted as fetal movement, Birth 42 (2) (2015) 189–194.

[12] A. Bekiou, K. Gourounti, Reduced fetal movements and perinatal mortality, Mater. Sociomed. 32 (3) (2020) 227–234, https://doi.org/10.5455/msm.2020.32.227-234. Sep.

[13] E. Valencia-Rincon, E. Reyna-Villasmil, D. Torres-Cepeda, J. Mejia-Montilla, N. Reyna-Villasmil, A. Fernandez-Ramirez, M. Rondon-Tapia, Decreased fetal movements and perinatal outcome in term pregnancies, Avances en Biomed. 6 (2) (2017) 98–104.

[14] J.E. Norman, et al., Awareness of fetal movements and care package to reduce fetal mortality (AFFIRM): a stepped wedge, cluster-randomised trial, Lancet 392 (10158) (2018) 1629–1638, https://doi.org/10.1016/s0140-6736(18)31543-5.

[15] Z.R. Hijazi, C.E. East, Factors affecting maternal perception of fetal movement, Obstet. Gynecol. Surv. 64 (7) (2009) 489–497, https://doi.org/10.1097/OGX.0b013e3181a8237a.

[16] K. Nishihara, S. Horiuchi, H. Eto, M.J.E.h.d. Honda, A long-term monitoring of fetal movement at home using a newly developed sensor: an introduction of maternal micro-arousals evoked by fetal movement during maternal sleep, Early Hum. Dev. 84 (9) (2008) 595–603.

[17] T. Byrt, J. Bishop, J.B. Carlin, Bias, prevalence and kappa, J. Clin. Epidemiol. 46 (5) (1993) 423–429, https://doi.org/10.1016/0895-4356(93)90018-v.

[18] E. Ryo, K. Nishihara, S. Matsumoto, H. J. M. e. Kamata, and physics, A new method for long-term home monitoring of fetal movement by pregnant women themselves, Med. Eng. Phys. 34 (5) (2012) 566–572.

[19] B. Boashash, M.S. Khlif, T. Ben-Jabeur, C.E. East, P.B. Colditz, Passive detection of accelerometer-recorded fetal movements using a time-frequency signal processing approach," (in English), Digit Signal Process. 25 (2014) 134–155, https://doi.org/10.1016/j.dsp.2013.10.002. Feb.

[20] M. Mesbah, et al., Automatic fetal movement recognition from multi-channel accelerometry data, Comput. Methods Programs Biomed. 210 (2021), 106377, https://doi.org/10.1016/j.cmpb.2021.106377. Oct.

[21] L. Valentin, K. Marššál, K. Lindström, Recording of foetal movements: a comparison of three methods, J. Med. Eng. Technol. 10 (5) (1986) 239–247, https://doi.org/10.3109/03091908609022914, 1986/01/01.

[22] U. Delay, et al., Novel non-invasive in-house fabricated wearable system with a hybrid algorithm for fetal movement recognition, PLoS ONE 16 (7) (2021), https://doi.org/10.1371/journal.pone.0254560 e0254560.

[23] J. Lai, R. Woodward, Y. Alexandrov, Q.A. Munnee, C.C. Lees, R. Vaidyanathan, N. C. Nowlan, Performance of a wearable acoustic system for fetal movement discrimination, PLoS ONE 13 (5) (2018,), https://doi.org/10.1371/journal.pone.0195728. May 7.

[24] A.K. Ghosh, S. Balasubramanian, S. Devasahayam, R. Vaidyanathan, A. Cherian, J. Prasad, N.C. Nowlan, Detection and analysis of fetal movements using an acoustic sensor-based wearable monitor, presented at the, in: 2020 7th International Conference on Information Science and Control Engineering (ICISCE), 2020.

[25] J. Li, Q. Wang, Multi-modal bioelectrical signal fusion analysis based on different acquisition devices and scene settings: overview, challenges, and novel orientation, Inf. Fusion 79 (2022) 229–247, https://doi.org/10.1016/j.inffus.2021.10.018, 2022/03/01/.

[26] S. Mehrdad, Y. Wang, S.F. Atashzar, Perspective: wearable internet of medical things for remote tracking of symptoms, prediction of health anomalies, implementation of preventative measures, and control of virus spread during the era of COVID-19, Front. Robot. AI 8 (2021), 610653, https://doi.org/10.3389/frobt.2021.610653.

[27] S. Qiu, et al., Multi-sensor information fusion based on machine learning for real applications in human activity recognition: state-of-the-art and research challenges, Inf. Fusion 80 (2022) 241–265, https://doi.org/10.1016/j.inffus.2021.11.006.

[28] A. Talitckii, et al., Comparative study of wearable sensors, video, and handwriting to detect Parkinson's disease, IEEE Trans. Instrum. Meas. 71 (2022) 1–10, https://doi.org/10.1109/tim.2022.3176898.

[29] Y. Celik, S. Stuart, W.L. Woo, E. Sejdic, A. Godfrey, Multi-modal gait: a wearable, algorithm and data fusion approach for clinical and free-living assessment, Inf. Fusion 78 (2022) 57–70, https://doi.org/10.1016/j.inffus.2021.09.016.

[30] A.K. Ghosh, et al., A novel fetal movement simulator for the performance evaluation of vibration sensors for wearable fetal movement monitors, Sensors 20 (21) (2020), https://doi.org/10.3390/s20216020. Oct 23.

[31] R. Vaidyanathan, N. Nowlan, R. Woodward, S. Shefelbine, Biomechanical Activity Monitoring, Google Patents, 2019.

[32] R.B. Woodward, S.J. Shefelbine, R. Vaidyanathan, Pervasive monitoring of motion and muscle activation: inertial and mechanomyography fusion (in English), Ieee-Asme Trans. Mech. 22 (5) (2017) 2022–2033, https://doi.org/10.1109/Tmech.2017.2715163. Oct.

[33] SparkFun Electronics. "Sparkfun Triple Axis Accelerometer Breakout - ADXL335." SparkFun Electronics. https://www.sparkfun.com/products/9269 (accessed March 10, 2022).

[34] M. Altini, et al., Detection of fetal kicks using body-worn accelerometers during pregnancy: trade-offs between sensors number and positioning, in: 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2016, pp. 5319–5322.

[35] Piezoelectric sound components, P37E–23. [Online]. Available: https://www.sparkfun.com/datasheets/Sensors/Flex/p37e.pdf.

[36] Tekscan. "FlexiForce™ Standard Model A401." Tekscan, Inc. https://www.tekscan.com/products-solutions/force-sensors/a401 (accessed March 10, 2022).

[37] E. Ryo, H. Kamata, Fetal movement counting at home with a fetal movement acceleration measurement recorder: a preliminary report, J. Matern.-Fetal Neonatal Med. 25 (12) (2012) 2629–2632, https://doi.org/10.3109/14767058.2012.704449. Dec.

[38] G. Thomas, et al., Detecting fetal movements using non-invasive accelerometers: a preliminary analysis, in: 10th International Conference on Information Science, Signal Processing and their Applications (ISSPA2010), IEEE, 2010, pp. 508–511.

[39] S. Layeghy, G. Azemi, P. Colditz, B. Boashash, Non-invasivemonitoring of fetal movements using time-frequency features of accelerometry, in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2014, pp. 4379–4383.

[40] A. Dennis, L. Hardy, "Defining a reference range for vital signs in healthy term pregnant women undergoing caesarean section," (in eng), Anaesth. Intensive Care 44 (6) (2016) 752–757, https://doi.org/10.1177/0310057×1604400619. Nov.

[41] S.W. Verbruggen, et al., Stresses and strains on the human fetal skeleton during development, J. R. Soc. Interface 15 (138) (2018), https://doi.org/10.1098/rsif.2017.0593. Jan.

[42] N.S. Malan, S. Sharma, Feature selection using regularized neighbourhood component analysis to enhance the classification performance of motor imagery signals, Comput. Biol. Med. 107 (2019) 118–126, https://doi.org/10.1016/j.compbiomed.2019.02.009, 2019/04/01/.

[43] C. Janiesch, P. Zschech, K. Heinrich, Machine learning and deep learning, Electron. Mark. 31 (3) (2021) 685–695, https://doi.org/10.1007/s12525-021-00475-2.

[44] P.P. Shinde, S. Shah, A review of machine learning and deep learning applications, in: 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018, pp. 1–6, https://doi.org/10.1109/ICCUBEA.2018.8697857, 16-18 Aug. 2018.

[45] L. Rokach, Decision forest: twenty years of research, Inf. Fusion 27 (2016) 111–125, https://doi.org/10.1016/j.inffus.2015.06.005, 2016/01/01/.

[46] J. Miao, W. Zhu, Precision–recall curve (PRC) classification trees, Evol. Intell. (2021), https://doi.org/10.1007/s12065-021-00565-2.