Fetal Movement Identification from Multi-Accelerometer Measurements using Recurrent Neural Networks

Janith Bandara Senanayaka
Dept. Electrical & Electronic Eng.
University of Peradeniya
Peradeniya, Sri Lanka
janith.b.senanayaka@eng.pdn.ac.lk

Eranda Somathilake
Dept. of Mechanical Eng.
University of Peradeniya
Peradeniya, Sri Lanka
eranda.somathilake@eng.pdn.ac.lk

Upekha Delay
Dept. Electrical & Electronic Eng.
University of Peradeniya
Peradeniya, Sri Lanka
upekha.delay@eng.pdn.ac.lk

Samitha Gunarathne
Dept. Electrical & Electronic Eng.
University of Peradeniya
Peradeniya, Sri Lanka
samithalg@eng.pdn.ac.lk

Roshan Godaliyadda
Dept. Electrical & Electronic Eng.
University of Peradeniya
Peradeniya, Sri Lanka
roshangodd@ee.pdn.ac.lk

Parakrama Ekanayake
Dept. Electrical & Electronic Eng.
University of Peradeniya
Peradeniya, Sri Lanka
mpb.ekanayake@ee.pdn.ac.lk

Janaka Wijayakulasooriya
Dept. Electrical & Electronic Eng.
University of Peradeniya
Peradeniya, Sri Lanka
jan@ee.pdn.ac.lk

Chathura Rathnayake
Dept. of Obstetrics and Gynaecology
University of Peradeniya
Sri Lanka
chathura67@hotmail.com

***Abstract*—This work presents two approaches, the many-to-many method and the many-to-one method, to identify fetal movements, especially fetal kicks, using recurrent neural networks from multi-accelerometer readings acquired from four accelerometers mounted on the abdomen of a pregnant woman. Additionally, we discuss some issues associated with the imperfect ground-truth labelling, acquired through maternal perception, that deteriorate and misrepresent the network performance.**

***Index Terms*—Fetal movements, accelerometer, inertial measurement unit, multi-sensor, spectrogram, recurrent neural network**

I. Introduction

Being a non-invasive method that pregnant mothers can perform without the necessity of specialized knowledge, counting fetal movement, in other words, "kick count," plays a significant role in monitoring fetal well-being [1]. Despite this appeal, it maybe difficult for a pregnant mother to focus solely on kick counting amidst all other day-to-day work. Medical devices are also available in clinical settings, such as ultrasonic scanners, to identify fetal movements [2]. Those instruments, however, require expert knowledge and can be harmful if exposed for a long time. Therefore, it is desirable to have a non-invasive device that pregnant mothers can use outside the clinical environment.

This work explores the possibility of using the accelerometer readings acquired from four accelerometers mounted on the abdomen of a pregnant mother for fetal movement identification. In particular, two approaches consisting of recurrent neural networks, namely the many-to-many and the many-to-one method, are considered to identify the fetal kicks. Further, the issues associated with imperfect ground-truth labels, acquired through maternal perception, are discussed.

Section II presents the related works. Section III and IV give details about the instrument and the dataset, while section V describes the methodology. Section VI then discuss the results and associated problems. Finally, section VII concludes the paper.

II. Related Work

Drawbacks of the existing fetal movement analysis methods has incentivised the study of the viability of having simpler and cheaper monitoring techniques while having reasonable accuracies [3], [4].

Although this is quite a novel concept, use of simple devices as an replacement for manual maternal kick counting has been researched in several instances using different types of sensors. The use of acoustic sensors for fetal movement detection was analysed in [4] and states that they are too sensitive for this particular application when compared with accelerometers. Other studies have either used a single accelerometer [3], [5]–[7] or multiple ones [8], [9]. Single sensor configurations have reasonable accuracy, but multiple sensors will observe a wider

978-1-6654-2637-4/21/$31.00 ©2021 IEEE

surface over the abdomen providing a better observation of any movement.

Fetal movement is, to some extent, clearly manifested in the accelerometer readings, which is evident by [8] as they were able to identify fetal movement from the accelerometer readings purely by observing the raw accelerometer signal. Hence, using a thorough analysis procedure will result in a more reproducible and accurate results.

Machine learning techniques were used in [3], [5] where different signal preprocessing methods were analysed prior to using convolutional neural networks for the predictions. The use of Long Short-Term Memory (LSTM) units for fetal movement detection using a single accelerometer was considered in [10]. In this study, we focus on using Gated Recurrent Units (GRU) for the analysis [11].

III. Instrument

The instrument used for data acquisition, in particular, to record accelerations, consists of four inertial measurement units (IMUs), each containing a tri-axial accelerometer and a tri-axial gyroscope, which are mounted on the pregnant mother's abdomen as depicted in Fig. 1. Whenever a fetal movement, particularly a fetal kick, occurs, the pregnant mother was advised to press a push button on the instrument, which records the mother's perception as the ground-truth label.

Furthermore, since the instrument should apply to domestic use and be non-invasive for long-term monitoring, IMUs were used as they are passive and were relatively cheap. Specifically, the device uses the MPU6050 model as the IMUs and a Raspberry Pi 1 to record readings.

IV. Dataset

Due to the prevailing Covid-19 pandemic, the dataset solely consists of recordings of a single subject, and the data acquisition was carried out at home rather than in a clinical setting. Consequently, the dataset does not contain perfect ground-truth labels recorded from a medical device, such as an ultrasonic scanner. Nonetheless, the instrument design and data collection were conducted under the guidance of a consultant obstetrician and gynaecologist.

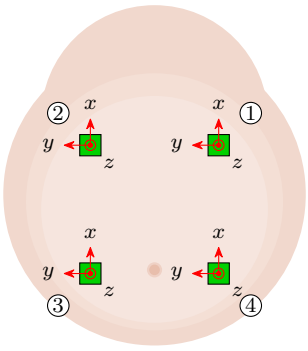

Fig. 1. Illustration of the front view of pregnant mother's abdomen showing the sensor placement. Small green squares represent IMUs with their axes marked on top of that and the accelerometer numbers are shown inside the circles.

The dataset consists of the accelerometer and gyroscope readings acquired from the four IMU units mounted on the abdomen, including ground truth labels obtained through the pregnant mother's perception. In this instance, however, only the accelerometer readings in the axis orthogonal to the abdomen (z-axis) were utilized since they contain the most information about fetal movement – accelerometer readings in this direction have the highest correlation with fetal movement.

V. Methodology

A. Data Preprocessing

Dataset is one of the most important parts when it comes to deep learning. Neural networks, not all, learn from the dataset; it plays a crucial role in training better-performing neural networks. Nevertheless, the dataset cannot be used as it is and should be preprocessed. Here, we outline the data prepossessing procedures used in the proposed work, where we use the term positive (or '1') to indicate the presence of fetal movement while negative (or '0') to indicate absence.

First of all, in both approaches, the accelerometer readings in the z-axis and ground-truth labels were selected from the raw data files as the training data.

1) Many-to-Many Approach: In this approach, training samples were then created by extracting windows from the data around the occurrences of the positive ground-truth labels – with random perturbations to allow data augmentation. Also, the length of these positive ground-truth labels is extended to improve the training. Finally, from those samples, the spectrograms were created, which are used as the input.

2) Many-to-One Approach: The raw accelerometer readings were then converted into spectrograms after removing the segments that contain no positive labels for a long time. Next, training samples were created using the sliding window approach. Afterward, a sample is labeled as a positive example if the latter half of the window contains any positive ground-truth labels or a negative otherwise. Finally, these preprocessed spectrogram samples were used as the input.

Removing the segments that do not contain any fetal movements for a long time was done to reduce the severe class imbalance in the dataset – positive instances were less than 1%. This was done to ensure the effective learning of the neural network. However, even after doing so, the dataset has more negative instances than positive ones, following a similar distribution as the true distribution since it has a class imbalance by nature. Therefore, it can be considered that there is less possibility of missing potential areas with artefacts that may have been falsely classified.

B. Networks

The network architectures used for both methods are almost the same as shown Fig. 2, with minor differences in the last layers to match the respective output types. In the many-to-many model, the return sequences attribute in the final recurrent layer is set to 'True', and the following layers are

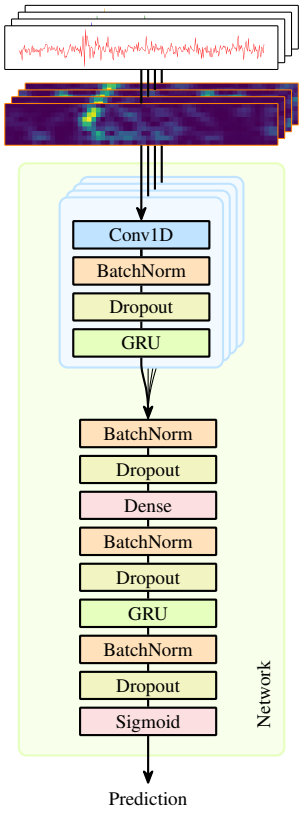

Fig. 2. Base network architecture used for both approaches.

time distributed. In contrast, in the many-to-many model, the return sequences attribute in the final recurrent layer is set to 'False', and the following layers are not time distributed.

For the recurrent layers, most common choices are LSTM and GRU. Both of these units have similar performance [12], with GRU having the added advantages of computational stability [11] and low computational complexity [12]. Therefore, GRUs were used in the recurrent layers of the proposed models.

Accelerometer readings were transformed into corresponding spectrograms for each accelerometer individually before feeding into the network. There is a separate block for each spectrogram dedicated to distilling information from the respective accelerometer separately. The outputs of these blocks are concatenated and then passed to the final network block, which performs the final prediction.

Specific parameter values used in the networks are given in Table I. Layers with the same name have the same parameters as given in the table and the Sigmoid layer is also a dense layer with one unit that have the sigmoid activation. For more specific details, please refer to the implementations[1].

1) Many-to-Many Approach: In this approach, for each windowed sample, the network produces a sequence of predictions similar to the ground-truth labels, as illustrated in Fig. 3. This model can be used to detect fetal movement from streaming accelerometer readings on the fly.

[1]The implementations, including all the codes and Jupyter notebooks, are available at https://github.com/janith-bandara/fetal-kicks-multi-imu

TABLE I
PARAMETERS OF THE LAYERS

Layer	Parameter	Value
Conv1D	Filters	32
	Kernel Size	16
	Stride	1
	Padding	Same
BatchNorm	-	-
Dropout	Rate	0.5
GRU (first)	Units	16
Dense	Units	30
	Activation	ReLU
GRU (second)	Units	10
Sigmoid (dense)	Units	1
	Activation	Sigmoid

2) Many-to-One Approach: In this approach, for each window, the network produces a single output label identifying whether the window contains a fetal movement or not, as illustrated in Fig. 4. This model can be used to detect fetal movement from a data window. Moreover, using the sliding window approach, this model can be extended to produce a sequence of predictions, mimicking the previous model. However, doing so will increase the computational and memory complexity compared to the previous model. Additionally, the long-term dependencies further away from window length will be lost; in this particular problem, however, such long-term dependencies may not be necessary.

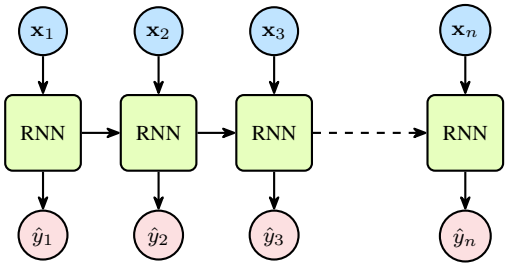

Fig. 3. Illustration of the many-to-many approach. $\mathbf{x}$ and $\hat{y}$ denote the inputs and the outputs of the model respectively.

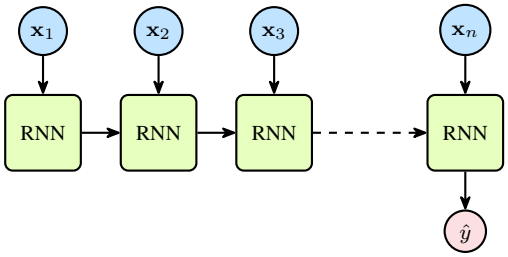

Fig. 4. Illustration of the many-to-one approach. $\mathbf{x}$ and $\hat{y}$ denote the inputs and the output of the model respectively.

C. Training

Initially, 75%, 15%, and 10% of the dataset were split into training-, validation-, and test-set, respectively. Both models were trained using the 'Adam' optimizer and the 'binary cross-entropy loss' for 100 epochs with a batch size of 10.

Training and validation accuracy curves, and training and validation loss curves for former model are shown in Fig. 5a and Fig. 5b, respectively. Training and validation accuracy curves, and training and validation loss curves for latter model are shown in Fig. 6a and Fig. 6b, respectively.

VI. Results and Discussion

One of the main contributing factors to the rise of deep learning was the availability of large datasets. Additionally, the loss function conveys the desired objective to the neural network via a training algorithm. Consequently, a large dataset and a loss function that accurately represents the desired objective are crucial factors to elicit better performance from a neural network in state-of-the-art deep learning.

Fig. 7 shows the many-to-many network's predictions on an example in the training dataset (see Fig. 7b). Further, Fig. 8 shows the many-to-many network's predictions on an example in the validation dataset before (see Fig. 8b) and after (see Fig. 8c) the training. Fig. 7a and Fig. 8a show the corresponding spectrograms of the respective accelerometer readings in the chosen examples.

According to Fig. 7, there is severe overfitting to the training dataset, which could be primarily due to the scarcity of training data. One could likewise argue that the training curves depicted in Fig. 5 are also a clear indication of the overfitting: in general, the answer is yes, but to our knowledge, this argument holds water if and only if the loss function and the accuracy metric precisely represent the intended outcomes. For instance, consider Fig. 8: at the start of the training, the network's predictions on the shown example in the validation set are merely random (see Fig. 8b), whereas at the end of the training, the network's predictions on the same example look pretty good (see Fig. 8c); in contrast to the increased loss value when going from the initial state to the final state. Hence, even though the validation loss has increased during the training, the network's performance has improved. Therefore, the increased validation loss does not necessarily imply that the network's performance has worsened. Also, those results were obtained from the network trained using the given loss function, which means that the used loss function is not entirely unsatisfactory: but it is not the most accurate representation of the desired objective.

The delay between the occurrence of fetal movement and positive ground-truth labelling varies depending on the re-

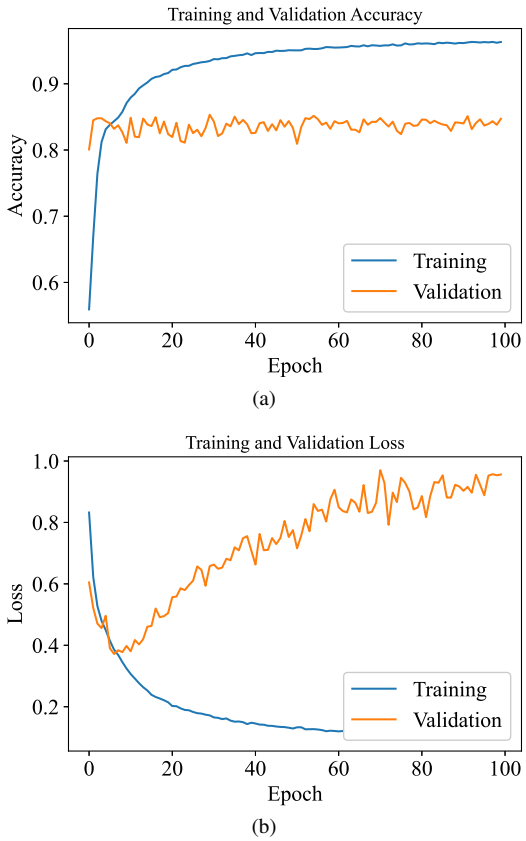

Fig. 5. Training accuracy and loss curves of the many-to-many model. (a) training and validation accuracy; (b) training and validation loss.

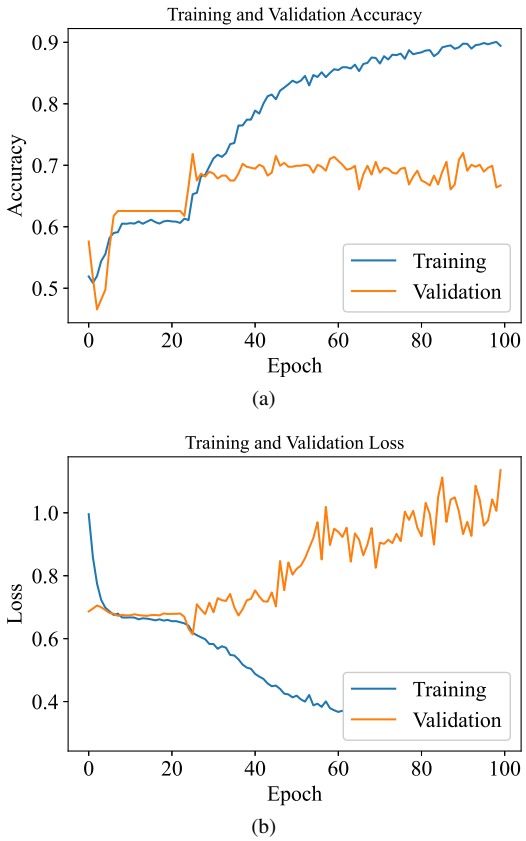

Fig. 6. Training accuracy and loss curves of the many-to-one model. (a) training and validation accuracy; (b) training and validation loss.

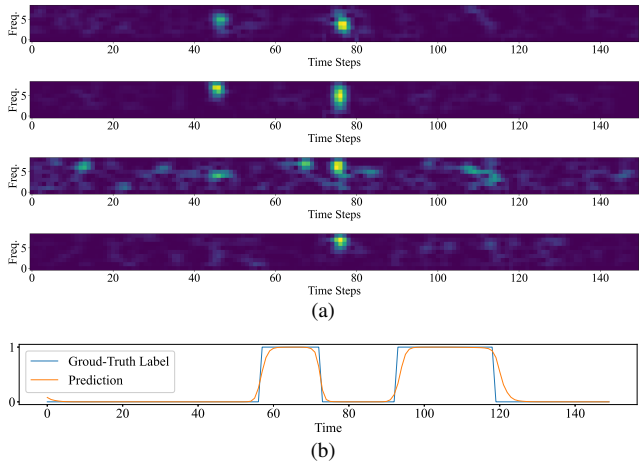

Fig. 7. Predictions of many-to-may network and ground-truth labels of a sample of the training set with its spectrograms: (a) corresponding spectrograms (accelerometer numbers are in the order of top to bottom); and (b) ground-truth and predictions.

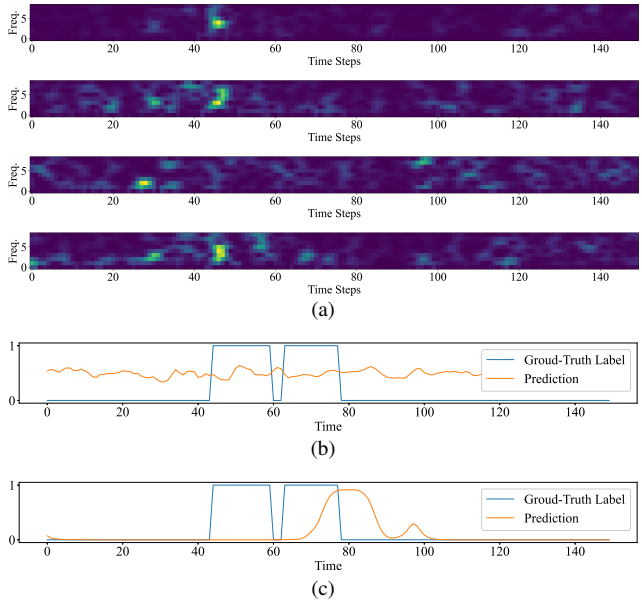

Fig. 8. Predictions of many-to-one network and ground-truth labels of a sample of the training set with its spectrograms: (a) corresponding spectrograms (accelerometer numbers are in the order of top to bottom); (b) initial predictions; and (c) final predictions.

sponse time of the pregnant mother, leading to imperfect ground-truth labels. Hence, the overlap between the network's predictions and the ground truth positive labels can be less, producing higher loss and low accuracy. Therefore, the binary cross-entropy loss and the accuracy metric used in this instance are not clear indications of the network performance. In other words, even if the network learned to perform better, variation in the response time in the ground-truth labelling could lead to less overlapping with the predictions that ultimately give higher loss and low accuracy. However, with a large amount of data, this effect maybe averaged out.

To mitigate the issues caused by varying response time coupled with inaccurate loss function, the many-to-one method was proposed, because, in this method, the loss and the accuracy metric is more informative than in the previous case. However, the varying response time causes another issue, especially mislabelling, which again deteriorates the network performance. As mentioned in the data preprocessing in section V-A, a windowed sample is labeled as positive if it contains fetal movement and negative otherwise. However, this is done based on the imperfect-ground truth labels. Consequently, there can be windows labeled as positive since they contain positive labels even though they do not contain fetal movement data as illustrated in the third example in Fig. 9, or windows that contain fetal movement data but are labeled as negative because they do not contain positive labels as illustrated in the first example in Fig. 9. These scenarios originated from imperfect ground-truth labelling lead to mislabelling of data, which ultimately affects the network performance as shown in Fig. 6, where the network overfitted to the training data. When faced with randomly shuffled labels, similar to possibly mislabeled examples as in this case, this behavior of making predictions with low accuracy on validation data while making predictions with high accuracy on training data by overfitting is examined in-depth in [13].

The validation loss and accuracy curves exhibit zig-zag behavior because the validation dataset was small.

In order to alleviate overfitting, regularization and data augmentation techniques can be utilized. In this instance, dropout and batch-normalization layers were incorporated as the means of mitigating overfitting. In addition to that, a simple data augmentation technique was also used. However, to our knowledge, enforcing further regularisation can deteriorate the network performance. A sufficient amount of data with a more accurate loss function of the objective will improve the performance.

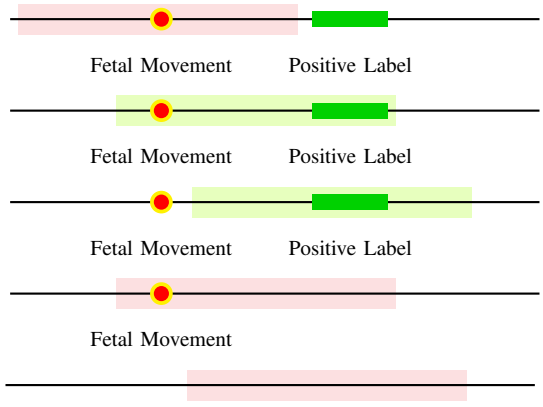

Fig. 9. Illustration of the possible mislabelling scenarios that can arise from the imperfect ground-truth labels. Green windows denote positive samples while red windows denote negative samples. These can be described from top to bottom as follows: false negative, true positive, false positive, false negative, and true negative.

VII. Conclusion

This analysis emphasizes the importance of having clear insight into what the network is doing, rather than superficially rely on the loss and the accuracy values to evaluate the performance. Also, preprocessing the dataset carefully and having a loss function that accurately captures the desired objectives are imperative. Further, since it is costly and time-consuming to acquire a large amount of data with perfect ground-truth labels, which should be done in a clinical setting using a medical device such as an ultrasonic scanner, devising a loss function or a model that could deal with imperfect ground-truth labels is of paramount importance.

In conclusion, given a sufficient amount of data, an accuracy metric that precisely evaluates the network performance, and a proper loss function that accurately captures the desired objective, a recurrent neural network with similar architecture will serve as a promising means to detect fetal movements, enabling long-term non-invasive fetal movement monitoring outside the clinical environment. Moreover, multiple accelerometers are well suited to detect fetal movement in a non-invasive manner.

Acknowledgment

We would like to extend our gratitude to the pregnant mother who volunteered for the data collection. This research was made possible through the contribution from the citizens of Sri Lanka towards the state-funded university system.

References

[1] G. Stephen, E. Martindale, and A. Heazell, "Predicting poor perinatal outcome in women who present with decreased fetal movements-a preliminary study," *Journal of Obstetrics and Gynaecology*, vol. 29, pp. 705–710, 2009.

[2] C. Gribbin and D. James, "Assessing fetal health," *Current Obstetrics & Gynaecology*, vol. 15, no. 4, pp. 221–227, 2005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957584705000478

[3] U. Delay, T. Nawarathne, S. Dissanayake, S. Gunarathne, T. Withanage, R. Godaliyadda, C. Rathnayake, P. Ekanayake, and J. Wijayakulasooriya, "Novel non-invasive in-house fabricated wearable system with a hybrid algorithm for fetal movement recognition," *PLOS ONE*, vol. 16, no. 7, pp. 1–22, 07 2021. [Online]. Available: https://doi.org/10.1371/journal.pone.0254560

[4] J. Lai, R. Woodward, Y. Alexandrov, Q. ain Munnee, C. C. Lees, R. Vaidyanathan, and N. C. Nowlan, "Performance of a wearable acoustic system for fetal movement discrimination," *PloS one*, vol. 13, no. 5, p. e0195728, 2018.

[5] U. H. Delay, B. M. T. M. Nawarathne, D. W. S. V. B. Dissanayake, M. P. B. Ekanayake, G. M. R. I. Godaliyadda, J. V. Wijayakulasooriya, and R. M. C. J. Rathnayake, "Non invasive wearable device for fetal movement detection," in *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)*, 2020, pp. 285–290.

[6] W. A. W. S. Wasalaarachchi, H. A. A. Subhashini, S. A. Y. Abeywardhana, M. S. L. Gunarathne, W. T. Ruwanga, G. M. R. I. Godaliyadda, M. P. B. Ekanayake, J. V. Wijayakulasooriya, and R. M. C. J. Rathnayake, "Fetal movements identification based on non-negative matrix factorization and spectral clustering," in *2019 14th Conference on Industrial and Information Systems (ICIIS)*, 2019, pp. 266–271.

[7] T. Nawarathne, T. Withanage, S. Gunarathne, U. Delay, E. Somathilake, J. Senanayake, R. Godaliyadda, P. Ekanayake, J. Wijayakulasooriya, and C. Rathnayake, "Comprehensive study on denoising of medical images utilizing neural network based auto-encoder," *arXiv preprint arXiv:2102.01903*, 2021.

[8] N. Zakaria, E. Paulson, and M. Balakrishnan, "Fetal movements recording system using accelerometer sensor," *ARPN Journal of Engineering and Applied Sciences*, vol. 13, pp. 1022–1032, 02 2018.

[9] M. A. Alsheikh, A. Selim, D. Niyato, L. Doyle, S. Lin, and H.-P. Tan, "Deep activity recognition models with triaxial accelerometers," *arXiv preprint arXiv:1511.04664*, 2015.

[10] E. Somathilake, J. B. Senanayaka, U. Delay, S. Gunarathne, T. Nawarathne, T. Withanage, R. Godaliyadda, P. Ekanayake, J. Wijayakulasooriya, and C. Rathnayake, "Fetal movement detection using long short-term memory network," in *2021 10th International Conference on Information and Automation for Sustainability (ICIAfS)*, 2021, pp. 464–469.

[11] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[12] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[13] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, 2021.