Combining Multiple tinyML Models for Multimodal Context-Aware Stress Recognition on Constrained Microcontrollers

Dr. Kieran Woodward, Michael Gibbs and Prof. Eiman Kanjo

Smart Sensing Lab, Department of Computer Science, Nottingham Trent University, UK.

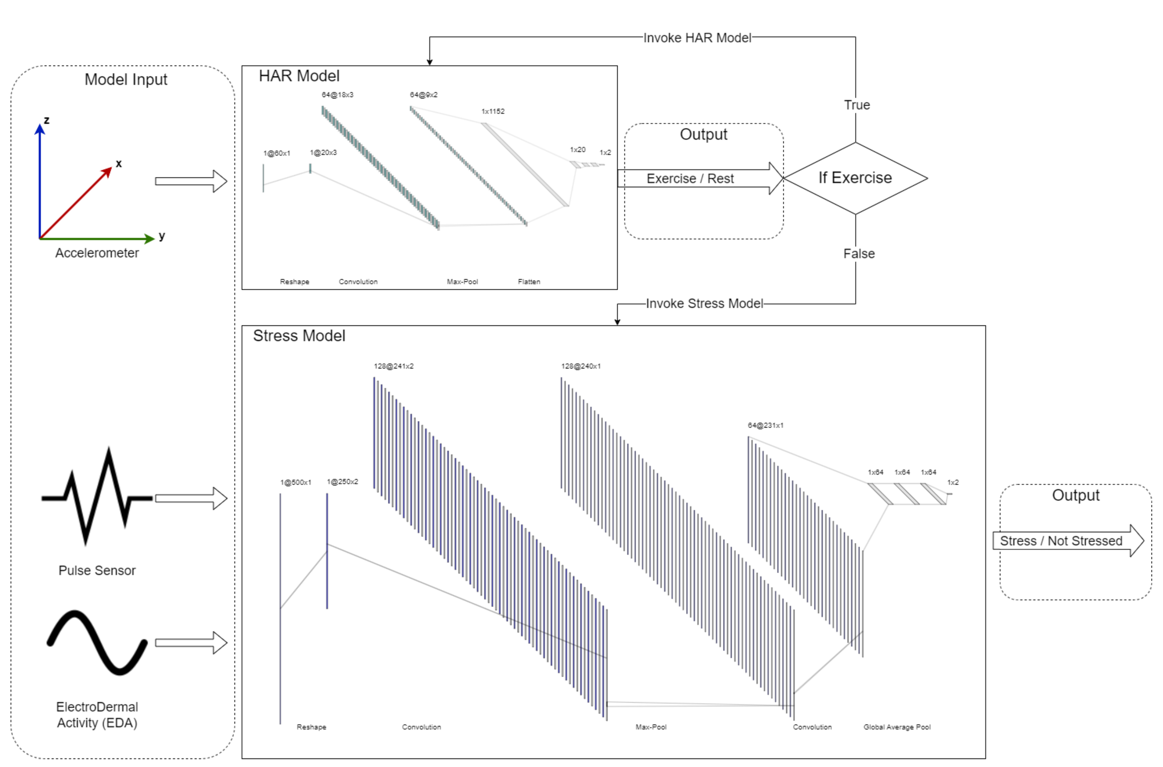
Abstract

As stress continues to be a major health concern, there is growing interest in developing effective stress management systems that can detect and mitigate stress in real-world environments. Deep Neural Networks (DNNs) have shown their effectiveness in accurately classifying stress, but most existing solutions rely on the cloud or large obtrusive devices for inference. The emergence of tinyML provides an opportunity to bridge this gap and enable ubiquitous intelligent systems.

In this work, we developed a context-aware stress detection approach that uses utilises tinyML to perform continuous inference of physical activity to mitigate motion artifacts when inferring stress from heart rate and electrodermal activity. We explore the challenges and trade-offs in deploying DNNs on resource-constrained microcontrollers for real-world stress recognition. Our proposed context-aware approach can improve the accuracy and privacy of stress detection systems while eliminating the need to store or transmit sensitive health data to the cloud.

Introduction

Photoplethysmography (PPG) sensors enable the non-invasive monitoring of Heart Rate (HR). Automatic analysis of PPG has rendered it valuable in clinical and non-clinical settings. However, tracking heart rate with PPG is difficult due to motion artifacts, which are major causes of signal degradation, as they obscure the location of the heart rate peak in the spectra. Therefore, context recognition in the form of Human Activity Recognition (HAR) is essential to improve the performance of stress detection. By integrating exercise recognition, stress detection systems can become more reliable by mitigating any motion artifacts from the signal. Therefore, we propose the continuous inference of physical activity and only once no physical activity is detected will stress inference be completed, in order to reduce motion artifacts.



The proposed system containing the HAR model (top) and stress model (bottom), along with the internal logic.

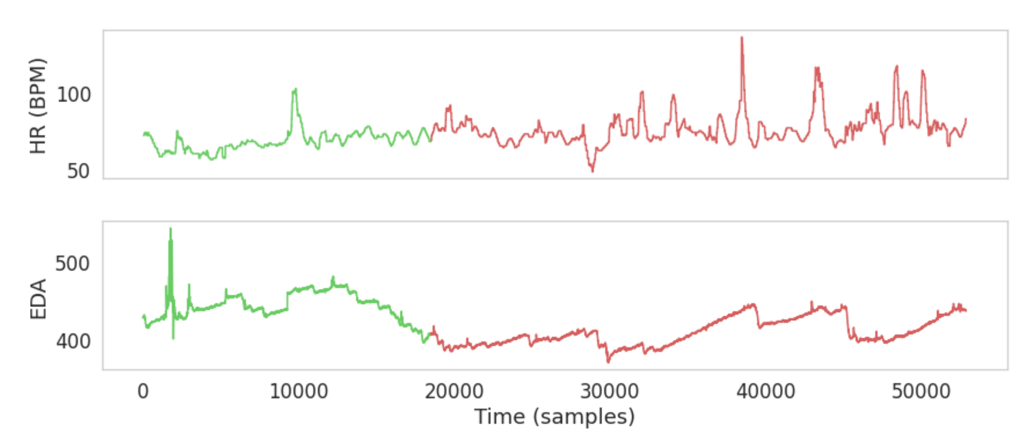
Methods

A non-invasive sensor-based approach presents the most significant opportunity to assess stress, as sensors can be easily connected to microcontrollers and used inconspicuously in the real-world.

- HR sensors are commonly used within wearable computing systems as they can be embedded within a wide range of devices due to their small footprint and provide insights into the autonomous nervous system.
- EDA is often used to train affective models to classify stress as it is directly related to the sympathetic nervous system.
- Motion data measures movement through accelerometers, gyroscopes, and magnetometers which can be used to measure physical activity.

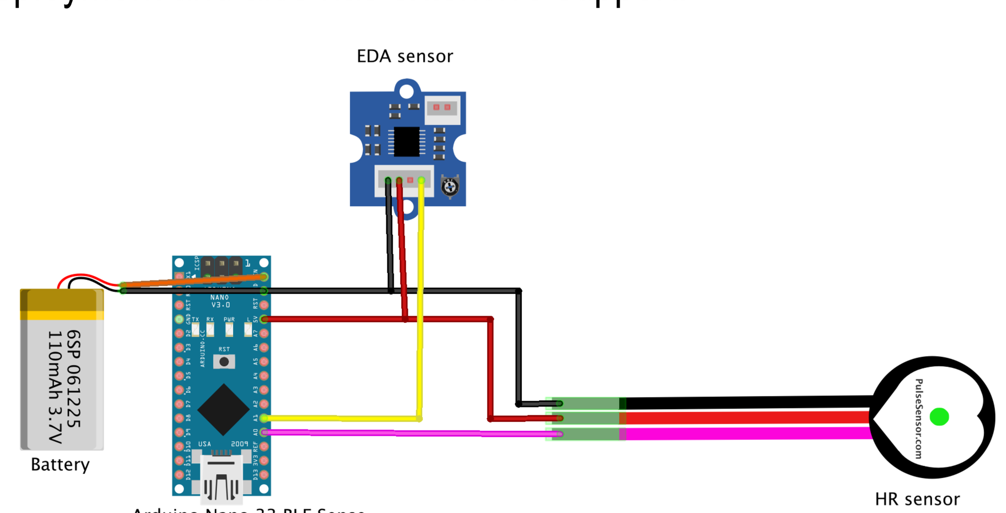
In order to complete the multi-model context-aware approach a dataset for both the HAR and stress recognition were required to train the classification models.

- For the HAR we used the WISDM: Wireless Sensor Data Mining dataset from Kwapisz et al. [1].
- For the stress recognition dataset, a lab-based stressor experiment was conducted using the Montreal stress test [2]. Sensors recorded raw HR amplitude and EDA, each sampled at 30Hz to collect physiological data while experiencing relaxed and stressed states.



HR and EDA dataset sample.

We have used the Arduino Nano 33 Sense as it is a small form factor, low power, and cost-effective microcontroller. It is based on the ARM Cortex-M4 32-bit processor. While there are other microcontrollers that are more powerful, the Arduino Nano 33 Sense is an extremely low-cost microcontroller enabling wide deployment and also has extensive support.



Microcontroller and sensor circuit.

Classification Models

Upon model deployment, a stress model using 2 convolutional layers with int 8 quantisation achieved an accuracy of 88% and the HAR model with one convolutional layer with float 16 quantisation 99% using hold-out validation with a 20% test split. Further breakdown of precision, recall, and F1 score for each model can be found in tables below.

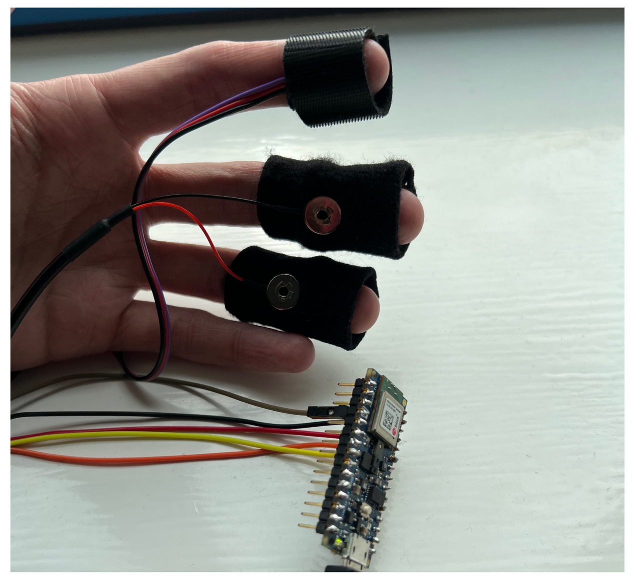
HAR Model			
Class	Precision	Recall	F1
Rest	0.95	0.98	0.96
Exercise	1.00	0.99	1.00
		Accuracy	0.99
		Model Size (.h file)	607KB

Performance metrics for the HAR classification model.

Stress Model			
Class	Precision	Recall	F1
Non-Stress	0.86	0.95	0.90
Stress	0.94	0.83	0.88
		Accuracy	0.88
		Model Size (.h file)	1.1MB

Performance metrics for the stress classification model.

It was necessary to ensure that both models could fit onto the target board, namely the Arduino Nano 33 Sense. The stress model was initially tested using float 16 quantisation but this resulted in 92% of the microcontroller's storage being used. Therefore, to ensure both models were able to fit on the microcontroller, they each required quantisation, with the HAR model requiring float 16 quantisation and the stress model int 8 quantisation.



Individual wearing the HR and EDA sensors connected to an Arduino Nano 33 Sense.

Typically, as a consequence of quantisation, the accuracy of a model is reduced. In this case, the accuracy of the HAR model was only reduced by 0.02% while the latency was reduced from 186ms to 11ms, the flash usage from 108.5KB to 49.8KB and the RAM usage from 21.2KB to 7.8KB. Therefore, the extremely minor accuracy sacrifice was deemed appropriate to demonstrate two models on a single microcontroller while still performing at a high level of significance.

Discussion

Context plays an important role in stress detection, as the activity in which an individual is engaged can significantly impact physiology. In particular, physical activity can drastically alter both HR and EDA, which are common biomarkers for stress detection.

Our results confirm the potential to run two classification models on a resource-constrained, low-cost microcontroller to mitigate the challenges of motion artifacts. Each of the models achieved high accuracy, demonstrating the potential for real-world inference.

However, limitations were encountered for real-world deployment:

- Only raw HR and EDA data used for stress classification. Additional features made model too large for microcontroller with HAR model
- Stress CNN with 4 convolutional layers achieved 95.6% accuracy, a 7.6% increase over 2-layer CNN but used 97% of microcontroller's dynamic memory with int 8 quantisation.
- 2 Layer CNN reduced the model size by 67%.
- Model trained with 1 convolutional layer reduced accuracy to 61.8%
- Float 16 quantisation with 2 convolutional layers used 92% of storage therefore Int 8 quantisation was used

Conclusion

In this work, we have demonstrated how a classification model can be used to provide context to a second model on the edge using a low-cost microcontroller. Here we have examined how activity recognition can be used to help reduce motion artifacts in real-world stress detection but this framework has many potential real-world applications.

References

[1] - Kwapisz, J. R., Weiss, G. M., & Moore, S. A. (2010). Activity Recognition using Cell Phone Accelerometers. *Proceedings of the Fourth International Workshop on Knowledge Discovery from Sensor Data (at KDD-10).*

[2] - Dedovic, K., Renwick, R., Mahani, N. K., & Engert, V. (2005). The Montreal Imaging Stress Task : using functional imaging to investigate the perceiving and processing psychosocial stress in the human brain Psychiatry & Neuroscience.

Connect with me

Read the paper

Connect with Me