AI-Based Risk Stratification and Explainable Models for Wearable Mental Health Monitoring Systems

Ramji Gupta,[1,2], Shashi Kant Gupta[3]*

[1]Computer Science and Engineering, Lincoln University College, Petaling Jaya, Selangor Darul Ehsan-47301, Malaysia. pdf.ramji@lincoln.edu.my

[2]Department of Electronics & Communication Engineering, Parul Institute of Engineering and Technology, Faculty of Engineering & Technology, Parul University, Vadodara, Gujarat, India, ramji.gupta270086@paruluniversity.ac.in

[3]Computer Science and Engineering Lincoln University College, Petaling Jaya, Selangor-47301, Malaysia, **shashigupta@lincoln.edu.my**

*Corresponding Author E-mail: ramjigupta38@gmail.com

Abstract: Wearable artificial-intelligence systems using explainable-AI methods have been proved effective for neurological disorders like epilepsy, Parkinson's disease, stroke and Alzheimer's disease. Adapting the same concepts to depression, anxiety, and stress is not as easy; meta-analytic findings show that mood disorders identification largely continues to base itself on one public dataset only, prefer past studies with very few studies extending outside of the wrist area, and hardly any tests beyond the wrist have been done. We highlight these issues and propose a multimodal sensing and modeling scheme that incorporates photoplethysmography, electrodermal activity, actigraphy, skin temperature, and sleep stages with the help of a CNN-BiLSTM-Attention model which is the extension of CNN-BiLSTM-Attention and XAI system already tested on CHB-MIT scalp EEG for predicting seizures to biosignals concerned with mental health. Continuous physiological measurements open an opportunity for hacking as well. So, besides the modeling component, we include a secure-transmission five-layer stack adapted from a wireless-sensor-network one tested within a military body-area-network environment. We wrap up this piece with a clinical effectiveness, explainability consistency, and transmission security plan under a simulated breach which aims at closing the specific gaps addressed by our design.

Keywords: Wearable Sensing, Mental Health Monitoring, Explainable-AI, CNN-BiLSTM-Attention, Secure Data Transmission.

1. INTRODUCTION

Devices designed for wearing on the body can monitor the user's state constantly while the user is not at the hospital by providing heart rate, movement and sleep information at any time. There has already been sufficient number of studies turning these data inputs, together with some wearable technologies into AI systems that could predict a neurological condition, risk stratification wise. We have tested a CNN-BiLSTM-Attention combination on CHB-MIT scalp EEG data set. Our findings suggest that a multi-component approach consisting of convolutional feature extraction, bidirectional temporal modeling, and attention-based explanation might enable the detection of pre-ictal states in epilepsy with a time window of several days that would benefit the clinical management. The same architectural concepts have later been applied to other areas like Parkinson's disease, stroke, and Alzheimer's disease. This contribution is the second in an ongoing conference series: the first publication focused on the main features of the diseases, and now we present one of the possible extensions of the method - the use of the same model to detect depression, anxiety or acute stress. Though, in mood and anxiety conditions there is no direct counterpart of the seizures on one hand and, then again, the diagnostics do not involve an EEG device but self-reporting questionnaires or similar instruments.

So, the wearable-mood-detection study body is not that much larger than that of the epilepsy literature which we mainly refer to. Section 2 shows us the points that are lacking in the field of research. Section 3 is an introduction to the new research design that builds on the preprocessing and architectural choices from the epilepsy work rather than coming up with something totally new. Section 4 describes each step along the pipeline in detail also mentioning the layer for secure transmission that guards the physiological data upon which the models are based. Section 5 discusses evaluation of the system at the stage of collection of the data, while section 6 brings a review and analysis of the solutions the design proposes and what it still leaves open or doesn't.

2. KNOWLEDGE GAPS IN WEARABLE MENTAL HEALTH AI

2.1. Data Validation and Generalization Gaps

A 2026 systematic review and meta-analysis of wearable-based methods for identifying depression, revealed that 96.5% of the included studies used only the Depresjon public dataset and the pooled sensitivity estimates greatly varied between the studies (high I^2 statistics compared to those ones could expect from comparable clinical populations) [1]. About two thirds of the studies under the review were retrospective, i.e. they were based on existing data sets rather than collecting new data prospectively, and the specificity decreased from 0.93 when the task was the detection of same-day depression to about 0.65 when the task was predictive of a future depressive episode. Except a single one, which focused on multi-site sensors (not the only sensor placement, rather sensors from multiple sites), all these works are restricted to measurements from the wrist. Such other sensor placements that might be relevant to mood disorders, like head or chest sensors are barely studied except in seizure detection research which routinely relies on multi-site recordings both of EEG and accelerometry. External validation, by which we mean that the performance of a model developed using data from site A is tested on data from site B or even other sites - is extremely scarce [1, 2]

Table 1: Data and validation gaps in wearable depression detection

Gap	Reported figure	Consequence
Single-dataset dependence [1, 2]	96.5% of studies use Depresjon	Limited demographic and comorbidity diversity behind reported accuracy
Retrospective-design bias [1, 2]	69% retrospective vs. prospective	Prospective studies show measurably higher sensitivity and specificity
Detection-to-prediction drop [1,2]	Specificity 0.93 → 0.65	Models trained for detection generalize poorly to episode prediction
Wrist-worn concentration [1, 2]	Near-universal wrist placement	Chest, head, and multi-site configurations remain unexplored

2.2. Modality, Explainability, and Clinical-Trust Gaps

The models that consider just movement or activity data do not give satisfactory results when the model is a pure activity model versus when it is fused with sleep, heart rate, and location. The accuracy of such sensors has been reported to drop by several percentage points when only one stream of inputs is considered [2]. Besides, the physiological responses of anxiety and depression are different as well: anxiety mostly results in a sudden sympathetic nervous system response which shows up as a heart rate, and electrodermal activity in minutes. Meanwhile the effects of depression last much longer; the sleep and circadian rhythm slowly shift over days. No sensing-and-model design has been published as yet which could detect both patterns with one type of algorithm. As far as the explanation of results goes, there is still no common standard how one could measure the completeness or faithfulness of SHAP or attention-based explanations, that one could use in mental-health wearables, and clinician comprehension of such outputs is rarely assessed directly [3].The issue also comes down to the regulatory standards which compound the situation: FDA rules for software-as-a-medical-device and the EU AI Act both are expected to lead to continually-learning adaptive algorithms, yet the validation of such systems is still ongoing in the field of psychiatry [4-8].

3. PROPOSED FRAMEWORK

The setup is proposed as an extension of CNN-BiLSTM-Attention and XAI that were validated on CHB-MIT scalp EEG, for seizure prediction, but applied not to a mental health wearable biosignal instead of being re-invented from scratch. The complete flow: multimodal sensors supply data to a preprocessing phase that gets the clean, synchronized signals; a feature extraction stage then builds time-domain, frequency-domain, time-frequency and non-linear features; the CNN-BiLSTM-Attention classifies each time window; explainability layer outputs SHAP, attention-weight, and Grad CAM through the internal model of clinician; and a secure-transmission layer transports the data from the wearable to cloud-based model serving through a clinician dashboard, with a feedback loop enabling model updates and personalization.

4. METHODOLOGY

4.1. Multimodal Sensing

Overall, there are five kinds of sensors which have been selected to reflect both types of features (i.e. acute and chronic) that are listed and explained in Section 2.2. Photoplethysmography or PPG is the heart-rate variability or HRV detector which indicates the depressive condition after weeks of decreased HRV and also warns about acute anxiety when HRV plunges. Electrodermal activity or EDA/galvanic skin response (GSR) is the one which is directly related to sympathetic system functioning. That means, it captures even those short-lived stress situations which HRV alone might overlook. Inertial-measurement-unit (IMU) actigraphy is used to track the level of physical and movement activities of a person (e.g. sleep/wakefulness cycles and sleep fragmentation). Decreased daytime activity, and altered nocturnal rest or activity are features linked to depression. These features can be captured or followed through the IMU actigraphy data. Together, skin temperature and sleep-stage monitoring act as indicators of the circadian rhythm. Sleep fragmentation and blunted temperature rhythm are considered to be the earliest indicators of depression.

4.2. Signal Preprocessing

The processing of signals is a key step as movement artifacts, which is by far the biggest source of noise for on-body measurement devices like wrist-based PPG or EDA sensors are eliminated during this step. Artifact rejection stage takes care of movement contaminations which is one of the principal sources of disturbance during normal activities for wrist-worn PPG and EDA. Then the Discrete Wavelet Transform (DWT, Daubechies-4) is used to break down each signal into a number of frequency sub-bands, the same wavelet family that is applied in the CHB-MIT EEG dataset pipeline, and finally, Welch's power spectral density estimation is utilized to obtain HRV frequency-domain features including the LF/HF ratio, which is one of the most commonly used indices of autonomic balance. Rolling 24-hour statistics (interdaily stability, intradaily variability) are the features used to capture the circadian drift signature, the characteristic distinguishing chronic depression from acute anxiety. The decision of using this pipeline in our work is not an accident: based on the literature, activity-only models only achieve around 69% performance accuracy, and the addition of sleep and heart-rate fusion reportedly leads to a significant jump to around 77% accuracy - this is a direct reply to the methodological aspect of the unimodal-fusion penalty that has been introduced in Section 2.2. [3, 9, 19]

4.3. Candidate Datasets

Three dataset sources are planned to resolve the single-dataset gap highlighted in Section 2.1. WESAD and DREAMER offer multimodal data for stress and affect such as EDA, heart rate/HRV respiration skin temperature to be used in pretraining and validation of the architecture. Depresjon supplies motor-activity recordings at night that identify a distinction of depressive disorders, both unipolar and bipolar, from controls. A future prospective multi-site own cohort dataset annotated with PHQ-9, GAD-7, and MADRS clinical scales will work as the external-validation ground, something that the Section 2.1 found to lack greatly in the existing research [9-11].

4.4. Model Architecture: CNN-BiLSTM-Attention

First, three consecutive convolutional layers (kernel sizes 7, 5, and 3) with 64, 128, and 256 filters respectively, batch normalization, ReLU activation and max pooling follow each convolutional layer, which allows to learn local patterns in a multimodal stack. Next, the bidirectional LSTM is employed to capture temporal context from both the past and the future directions because the depressive state has been found to be a matter of a few days and not a fraction of a second. An Attention layer is a component that determines how to combine different hidden states together to provide the final prediction. This mechanism can be also treated as an interpretability tool because it shows to the researchers the key parts (moments & modalities) influencing a certain decision. A fully connected layer, after which dropout and the softmax activation functions were introduced, leads to the final stage where the probabilities for each class among four possible outcomes: euthymic, at-risk, depressive, and anxious/acute-stress, are determined [19]. Fig. 1 presents the classifier. Multimodal fused features for each time windows make the input tensor.

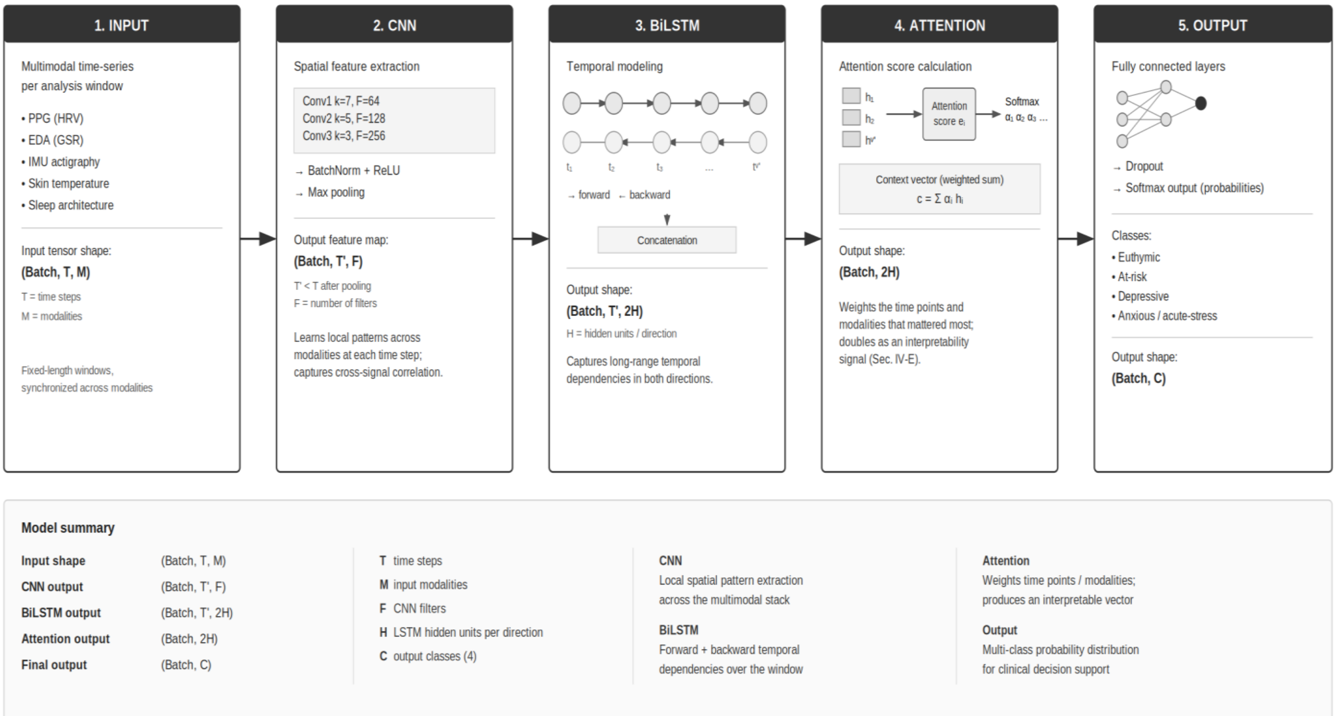

Fig. 1: CNN-BiLSTM-Attention architecture mapping multimodal wearable biosignals to mental-health risk classes.

4.5. Explainability Layer

Three types of explainability methods explain the decisions of the model after it has been trained. The attention-weight visualization is generated by the deep neural network itself and points out segments of the time series that were most attentively processed by the model. Grad-CAM identifies which frequency bands were most contributing to the classification decision, for example, which HRV LF/HF power, on the time-frequency representations; following the same method as for the epileptic seizures-prediction system [7]. SHAP values indicate feature effects both at a global and a local level showing the main source among the modalities (HRV, sleep, or activity) that led to a specific classification [5]. One alone or two together among these three methods do not resolve the difference between metrics of a standardized-fidelity metric, which is the problem outlined in Section 2.2; Section 5's evaluation strategy sees this as an open measurement issue to be addressed rather than the problem which could be solved only by an architecture alone [12-14].

4.6. Secure Data Transmission

Personal mental-health data is also subject to threats identified when it comes to tactical wireless sensor networks: listening in, sending fake packets, impersonation of legitimate devices, and analyzing traffic on the data sent by the device. The proposed platform is based on a five-level secure-routing architecture which was initially checked against the security requirements of sensor networks [10, 15].

Table 2: Secure-Transmission layers adapted from a validated WSN architecture

Layer	Mechanism	Purpose
1. Encryption	AES-128-CCM authenticated encryption	Low-compute-cost protection of physiological packets on wearable-class processors
2. Authentication	ECC-based mutual authentication	Compact keys and lower overhead than RSA between wearable and gateway
3. Trust management	Direct + indirect trust scoring	Combines data-consistency and forwarding-reliability metrics with neighbor-reported trust
4. Adaptive routing	DQN-informed relay selection	Favors the most trustworthy, energy-sufficient path in body-area-network configurations
5. Intrusion detection	Hybrid deep-learning classifier	Flags spoofed or injected physiological streams before they reach the classification pipeline

Using this architecture in a validation exercise of military body area networks, an intrusion detection of 99.46% was attained in different types of attacks, such as sinkhole, hello flood DDoS Sybil, and traffic analysis [10, 11]. The average energy overhead from the security part was about 8%. These numbers are from a different deployment context. No reproduction of them for wearable-class hardware with physiological data has been presented yet; Section 5 presents them as a target envelope rather than an established result for this case [15, 16].

5. EXPERIMENTAL DESIGN AND EVALUATION PLAN

Evaluation is organized around the two things the framework has to get right at once: clinical accuracy and transmission security. Table 3 lists both sets of metrics together with the targets the design is meant to reach.

Table 3: Evaluation metrics and targets

Domain	Metric	Target
Clinical	Sensitivity / specificity	Detection and episode-prediction accuracy
Clinical	External-validation AUC	≥ 0.90
Clinical	DOR / F1	Composite diagnostic performance
Clinical	XAI fidelity score	SHAP / attention faithfulness (metric under development, see VI)
Security	Packet delivery ratio under attack	> 97%
Security	Intrusion-detection accuracy	Target informed by [10], not yet reproduced here

Domain	Metric	Target
Security	Energy overhead from security stack	< 10%
Security	Key-rotation policy	Adaptive refresh tied to device trust and battery state

The scheme plans to carry out activities in three separate steps. At the beginning, they plan to validate the architecture using WESAD, DREAMER and Depresjon datasets. It is expected to produce the baseline performance and reproduce the gap of the performance between the unimodal and the multimodal approach (from approximately 69 to 77%) when the data is properly handled. In the second phase, they will use the prospective multi-site cohort that has been PHQ-9, GAD-7, and MADRS annotated. It is the external-validation test that we pointed out in Section 2.1 as missing in previous studies. And, it will be the chance to see if the clinicians could get the idea or if it was just something SHAP and attention and Grad-CAM techniques were generating. The third phase will involve assessing the secure transmission component for various attack strategies. The secure layer of the network for the wireless sensor network is expected to remain the same as the original one; yet, rather than using the body-area-network testbed that provided the reference figures, the attacks will target the wearable-class hardware (sinkhole, hello-flood DDoS Sybil, and traffic analysis) [3, 7].

6. DISCUSSION

All four design decisions in Section 4 address a particular deficiency from Section 2, not something general to a particular kind of approach. The multi-site prospective cohort directly addresses the single-dataset and retrospective-design deficiencies; the five-sensor fusion directly addresses the unimodal-fusion penalty; the dual acute/chronic feature set (HRV and EDA against sleep and circadian drift) addresses the modality gap between anxiety and depression; and the secure-transmission layer addresses a deficiency the earlier neurological model shared, but failed to correct unprotected physiological data leaving the device.

Two of these constraints warrant further mention, to be as explicit as possible about what the system can and cannot do. The intrusion-detection and energy-overhead data points quoted in the secure-transmission layer are derived from a military-wireless-sensor-network validation, so the question of whether the same intrusion detection precision is obtainable on these physiologically--imaged sensors, on a fraction of a compute budget, and within a different traffic pattern, is something for the evaluation plan rather than a provided-in-advance assumption.

The XAI fidelity metric in Table 3 is said to have a target value because one does not yet exist; sat building or using one for this particular application is part of the remaining work, rather than a previously defined parameter. Continuous mental-state monitoring has some issues not as applicable to a seizure alert. False-positive seizure alert is a slight annoyance in a clinical setting; a false-positive depressive-episode alert sent dozens of times, to a patient who already has enough problems, are of a different sort of nature, because of this the threshold for an affect-alert, and its presentation, will need clinical input in advance of implementation (not only a statistical calibration). Issues of interests, consent, and data-sovereignty in the continuous effect labelling project are same thing of a different nature, and also need to be addressed by the multi-site cohort protocol during IRB submission [3, 7].

7. CONCLUSION AND FUTURE WORK

In addition to validating wearable-AI-plus-XAI design pattern for detecting seizures, Parkinson's disease, stroke, and Alzheimer's, the paper highlighted where the design pattern fails to deliver when depression, anxiety, and stress detection are considered. The work has also proposed a system which leverages the CNN-BiLSTM-attention backbone and its DWT/Welch-PSD preprocessing pipeline that has already been validated by the CHB-MIT database, instead of having to develop a new one. The subsequent steps are clear, that is, to obtain IRB approval and the multi-site subject pool required for the prospective PHQ-9/GAD-7/MADRS-labelled data collection, to run the fusion ablation to replicate the 69% to 77% accuracy gain, as documented in similar work, in this particular database, and lastly, to verify that the secure-transmission stack is effective in real wearables-class hardware for different attack scenarios, instead of taking the military WSN figures as is, it was adapted from.

Acknowledgements

The authors gratefully acknowledge the financial support provided by the **Indian Council of Medical Research (ICMR)** under the project titled “Seizure-Sentry: Wearable Epilepsy Prediction Device for Enhanced Patient Safety through Timely Alerts and Personalized Seizure Management” (Project ID: **IIRPSG-2024-01-01197**).

REFERENCES

1. Jiawen Liu, Junhui Wang, Zhaobin Wu. The Performance of Wearable Device–Based Artificial Intelligence in Detecting Depression: Systematic Review and Meta-Analysis. Academic Registry. 2026; Available from: https://doi.org/10.2196/85319

2. Ran Bai, Le Xiao, Yu Guo. Tracking and Monitoring Mood Stability of Patients With Major Depressive Disorder by Machine Learning Models Using Passive Digital Data: Prospective Naturalistic Multicenter Study. JMIR mHealth and uHealth. 2021; Available from: https://doi.org/10.2196/24365

3. Huy Viet Hoang, Huong Ha, Thi-Hiep Nguyen. Advancing mental health diagnostics: a review on the role of smartphones, wearable devices, and artificial intelligence in depression and anxiety detection. Biomedical Engineering Letters. 2025; Available from: https://doi.org/10.1007/s13534-025-00505-4

4. Yekta Said Can, Niaz Chalabianloo, Deniz Ekiz. Continuous Stress Detection Using Wearable Sensors in Real Life: Algorithmic Programming Contest Case Study. Sensors. 2019; Available from: https://doi.org/10.3390/s19081849

5. Xingjian Bai, Christian Coester. Sorting with Predictions. Advances in Neural Information Processing Systems 36. 2023; Available from: https://doi.org/10.52202/075280-1155

6. Marco Túlio Ribeiro, Sameer Singh, Carlos Guestrin. "Why Should I Trust You?". Academic Registry. 2016; Available from: https://doi.org/10.1145/2939672.2939778

7. Authors. Food and Drug Administration (FDA). SpringerReference. 2011; Available from: https://doi.org/10.1007/springerreference_32222

8. Marisa Araújo. THE EUROPEAN UNION'S APPROACH TO ARTIFICIAL INTELLIGENCE REGULATION. The membership of Bulgaria in the European Union Sixteen Years Later. 2024; Available from: https://doi.org/10.37075/iscwe.2023.01

9. Philip Schmidt, Attila Reiss, Robert Duerichen. Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection. Proceedings of the 20th ACM International Conference on Multimodal Interaction. 2018; Available from: https://doi.org/10.1145/3242969.3242985

10. S. Koelstra, C. Muhl, M. Soleymani. DEAP: A Database for Emotion Analysis ;Using Physiological Signals. IEEE Transactions on Affective Computing. 2012; Available from: https://doi.org/10.1109/T-AFFC.2011.15

11. Enrique Garcia-Ceja, Michael A. Riegler, Tine Nordgreen. Mental health monitoring with multimodal sensing and machine learning: A survey. Pervasive and Mobile Computing. 2018; Available from: https://doi.org/10.1016/j.pmcj.2018.09.003

12. Sarthak Garg, Stephan Peitz, Udhyakumar Nallasamy. Jointly Learning to Align and Translate with Transformer Models. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019; Available from: https://doi.org/10.18653/v1/d19-1453

13. Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. 2017 IEEE International Conference on Computer Vision (ICCV). 2017; Available from: https://doi.org/10.1109/ICCV.2017.74

14. Ashish Vaswani, Noam Shazeer, Niki Parmar. Attention Is All You Need. Academic Publication. 2025; Available from: https://doi.org/10.65215/ysbyhc05

15. Ramji Gupta, Shashi Kant Gupta, Saravjeet Singh. Adaptive deep reinforcement learning–based secure routing for wireless sensor networks. Academic Registry. 2026; Available from: https://doi.org/10.1038/s41598-026-62519-x

16. M. Prabha, S. Nandhini, M. Dayanidhy. Edge-AI integrated secure wireless IoT architecture for real time healthcare monitoring and federated anomaly detection. Scientific Reports. 2025; Available from: https://doi.org/10.1038/s41598-025-30150-x

17. M. Sree Vani, Rayapati Venkata Sudhakar, A. Mahendar. Personalized health monitoring using explainable AI: bridging trust in predictive healthcare. Scientific Reports. 2025; Available from: https://doi.org/10.1038/s41598-025-15867-z

18. Naser Hossein Motlagh, Mahsa Mohammadrezaei, Julian David Hunt. Internet of Things (IoT) and the Energy Sector. Academic Registry. 2020; Available from: https://doi.org/10.3390/en13020494

19. Francisco Ordóñez, Daniel Roggen. Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. Sensors. 2016; Available from: https://doi.org/10.3390/s16010115