The 15th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH 2025)
October 28-30, 2025, Istanbul, Türkiye

Recognition of Anxiety-Related Activities using 1DCNNBiLSTM on Sensor Data from a Commercial Wearable Device

Ugonna Oleh[a,*], Roman Obermaisser[a]

[a]Embedded System Institute, University of Siegen, 57076, Siegen, Germany

Abstract

Globally, anxiety disorders are one of the most prevalent mental health disorders, yet monitoring often relies on subjective self-reports. Commercial wearable devices offer a promising avenue for objective assessment of behavioural biomarkers, but a significant "lab-to-life" gap exists where models trained on research-grade sensors fail on consumer hardware. This paper presents a proof-of-concept methodology for bridging this gap. We employ a domain adaptation strategy using a 1DCNNBiLSTM model, which is pre-trained on the public ADAMSense dataset and then fine-tuned on a custom dataset from a Garmin Venu 3 smartwatch. Our results first quantify the domain shift, showing a performance drop from 89.11% to 49% accuracy when applying the base model to the new device. Subsequently, a 5-fold cross-validation demonstrates that fine-tuning achieves a robust and stable mean accuracy of 87.0% (± 1.2%). This significant improvement validates our methodology as a feasible and effective pathway for adapting research models to real-world applications. This pilot study provides a practical and rigorously validated foundation for future, larger-scale research into wearable-based mental health monitoring.

Keywords: m-health; Anxiety Detection; Consumer Electronics; Deep Learning; Human Activity Recognition (HAR).

1. Introduction

In 2021, it was estimated that about 4.4% of the world population suffered from anxiety disorder [18], making it one of the most prevalent mental health challenges globally. Conventionally, the diagnosis and monitoring of anxiety disorders rely on self-report questionnaires and clinical interviews, which are susceptible to recall bias. The use of wearable devices with Artificial Intelligence (AI) offers a solution to these shortcomings [1]. In 2021, Khan et al. proposed that Human Activity Recognition (HAR) can be used to recognise activities related to Anxiety Disorders

---

* Corresponding author. Tel.: +49 (0)271 740 3350
  E-mail address: ugonna.oleh@uni-siegen.de

using data from sensors on a wrist-worn IMU (Inertial Measurement Unit) and a smartphone placed in the volunteers' pockets [11].

The rapid consumer adoption of commercial wearable devices, such as smartwatches, has created a technological opportunity for mental health assessment [10] [14]. Utilising these commercial wearable devices will enhance the adoption of AI systems, which can be deployed on these devices. However, models developed using data from research-grade sensors often fail when applied to data from consumer-grade devices [12]. Researchers have identified key challenges that define this gap, such as data accessibility, interoperability between devices and whether the hardware and software are fit-for-purpose in a real-world deployment [12]. Unlike research-grade sensors that provide raw data and full control [15], commercial wearables are 'black-box' systems whose proprietary, on-board processing fundamentally alters the signal characteristics, making direct application of research models challenging [13]. The central problem is the lack of validated methodologies to bridge this domain shift, especially for the complex task of recognising subtle, anxiety-related activities.

To address this challenge, this paper presents a proof-of-concept methodology for adapting a HAR model from a public, research-grade dataset to a limited commercial-grade dataset. The primary contributions of this work are:

- The design and application of a hybrid deep learning architecture, a 1DCNNBiLSTM, specifically tailored to classify subtle, anxiety-related activities from only wrist-worn sensor data.
- Quantifying the domain shift for anxiety-related HAR by demonstrating significant performance degradation when a model trained on the public ADAMSense dataset [11] is applied directly to data from a commercial Garmin Venu 3 smartwatch.
- Validating a practical domain adaptation strategy that uses fine-tuning on a small, custom dataset to successfully bridge this performance gap, establishing a feasible pathway for developing real-world mental health monitoring tools.

The remainder of this paper is organised as follows: Section 2 presents a review of related work. Section 3 covers the method, including the proposed model architecture. Section 4 presents the results of the model performance. Section 5 discussed the implications of the results gotten. Section 6 concludes the paper and outlines future work.

2. Related Work

Human Activity Recognition (HAR) involves transforming multi-modal sensor data into high-level insights of activities performed for various daily applications [19]. Wearable and smartphone sensors are widely used for HAR systems, typically leveraging data from accelerometers, gyroscopes, and magnetometers. Deep Learning architectures have significantly advanced sensor-based HAR. Hybrid models, combining the strengths of Convolutional Neural Networks (CNNs) for feature extraction and LSTMs or Bidirectional LSTMs (BiLSTMs) for sequence learning, have demonstrated high performance in HAR [16]. For instance, a hybrid CNN-BiLSTM model has achieved high accuracies on datasets such as UCI-HAR and PAMAP2 [20]. Our proposed 1DCNNBiLSTM model aligns with these advanced hybrid architectures, leveraging their combined capabilities for classifying anxiety-related activities.

The application of HAR directly extends to mental health monitoring. Studies have explored detecting stress, mood changes, and anxiety using passively sensed data from smartphones and wearables. Khan et al specifically proposed using HAR to recognise anxiety-related activities, conducting experiments with wrist-worn IMU sensors and smartphones, and provided the ADAMSense dataset, which focuses on eleven anxiety-related activities [11]. Reviews by Dobson et al. [7] and Abd-Alrazaq et al. [2] further highlight the role of sensors and wearable AI in detecting physiological signs of anxiety for "in-the-moment" interventions. Despite these advancements, a notable "lab-to-life" gap persists. Models developed and trained using data from research-grade sensors often experience a substantial drop in performance when applied to data collected from commercial-grade devices [3]. This discrepancy arises because of the black-box nature of commercial wearables [2]. The absence of validated methodologies to bridge this domain shift, especially for complex and subtle activities like those related to anxiety, remains a significant problem.

To overcome the "lab-to-life" gap and device heterogeneity, domain adaptation techniques are crucial. This involves leveraging knowledge from a source domain (e.g., research-grade data) to enhance learning in a target domain (e.g., commercial device data) [5]. This is particularly relevant in HAR, where variations can occur across different sensor

devices, on-body positions, and target users [6]. Various studies have explored transfer learning to address data scarcity and device heterogeneity [4]. A common strategy involves pre-training models on large, publicly available research datasets and then fine-tuning them on smaller, custom datasets from the target commercial devices [6]. This allows models to learn general features from extensive public data and then adapt to the specific characteristics of commercial devices. The challenge of cross-user scenarios, where data distribution differences exist between training and real-world usage, is also actively addressed by domain adaptation techniques [4]. The exploration of domain adaptation for IMU-based HAR is a significant research area, focusing on addressing heterogeneities caused by sensor placement, device biases, and environmental diversities. This work, which demonstrates a methodology of pre-training on a public dataset like ADAMSense and fine-tuning on data from a commercial Garmin Venu 3 smartwatch, contributes to bridging this performance gap for anxiety-related activity recognition, providing a practical pathway for leveraging consumer wearables in mental health monitoring.

3. Method

The complete methodology follows a two-stage domain adaptation strategy designed to adapt a model from a public, research-grade dataset to a custom, commercial-grade dataset. The workflow involves pre-training a base 1DCNNBiLSTM model on the public ADAMSense dataset, followed by fine-tuning and evaluation on a custom dataset collected from a Garmin Venu 3 smartwatch using a robust cross-validation protocol.

3.1. Datasets and Data acquisition

The training and evaluation of the proposed model utilises a public dataset for initial training and a custom-collected dataset from a commercial smartwatch for fine-tuning and validation.

3.1.1. ADAMSense Dataset

The public dataset used is the ADAMSense Dataset, which is a dataset of anxiety-related activities. It contains sensor data obtained from wrist-worn and chest-worn IMU sensors (accelerometer, gyroscope and magnetometer) [11]. The data obtained is classified into 11 activities: ear rubbing or scratching, forehead rubbing or scratching, hair pulling, hand rubbing or scratching, hand tapping, knuckle cracking, nail biting, nape rubbing or scratching, smoking, idle sitting, and idle standing. In this study, the ADAMSense dataset serves as the source domain, enabling the proposed model to be trained on a larger dataset before being adapted to the specific target hardware.

3.1.2. Custom Garmin Venu 3 Dataset

The commercially available device used to fine-tune the model for real-world application is the Garmin Venu 3. A custom dataset was collected using the Garmin Venu 3 smartwatch. First, a Garmin watch app was developed to collect sensor data from the IMU sensors. The app was programmed using Monkey C [9]. Unlike research-grade devices, commercial devices have certain constraints, such as limitations in the size of data that can be stored. Although Garmin watch apps can be programmed to store data as FIT (Flexible and Interoperable Training) files [8], message length limitations in the API restricted our custom logger to collect only data from the accelerometer and gyroscope.

Data Collection Protocol: The study involved six volunteers. Before the start of the data collection process, volunteers were informed of the project details, and they provided informed consent. Each participant wore a Garmin Venu 3 on their dominant wrist with the sensor logger app enabled and was prompted to perform the activities classified from the ADAMSense dataset.

3.2. Data Processing

Following data acquisition, a multi-stage processing pipeline was implemented to prepare the data from both the source (ADAMSense) and target (Garmin Venu 3) domains for input into the 1DCNNBiLSTM model.

ADAMSense Data Processing: The full ADAMSense dataset contains 25 features from various sensors. To align the source data with the capabilities of our target device, we created a specific subset of the dataset. We extracted only the three-axis accelerometer and three-axis gyroscope data from the wrist-worn IMU, discarding the pocket

and magnetometer sensor data. This resulted in a 6-feature time-series dataset that directly mirrors the data streams collected from the Garmin Venu 3.

Custom Data Processing: The raw sensor data collected from the Garmin Venu 3 was saved in the proprietary FIT file format. The first processing step was to convert these FIT files into a standard CSV format for analysis. The raw data was recorded at a frequency of 100Hz. However, to maintain consistency with the ADAMSense dataset and reduce computational load, the data was downsampled to 50Hz. This was achieved by resampling the data to 20-millisecond intervals and taking the mean of the values within each interval. Data streams were synchronised using their timestamps, and accelerometer data was converted from milli-g to m/s². Finally, the processed time-series data was manually labelled against video recordings of the sessions.

Shared Processing (Windowing and Segmentation): After the initial processing, a uniform windowing strategy was applied to both datasets to create fixed-size segments for the model. The continuous data streams were segmented using a window size of 200 time steps (equivalent to 4 seconds at 50Hz) with a 50% overlap between consecutive windows. This approach follows standard practice in HAR; the 4-second window is long enough to capture a complete cycle of the target activities, while the 50% overlap prevents loss of information at window boundaries and serves as a form of data augmentation. The resulting segments were then used as input for the training and evaluation pipeline.

3.3. Proposed Model Architecture

The proposed model is a hybrid 1DCNNBiLSTM model that leverages the feature extraction properties of a CNN and temporal context awareness of the BiLSTM [17]. This architecture was selected because it combines the strengths of both models for time-series classification. The first layer of the model is a 1DCNN layer. The one-dimensional nature of the model enables it to accept time series data as input without requiring data reshaping. This layer excels at automatically extracting hierarchical features from raw sensor signals and identifying salient local patterns within the data. Unlike traditional machine learning approaches that require manual feature engineering, IDCNNs automatically learn features directly from the raw time-series signals. The model also contains three BiLSTM layers. The BiLSTM layers model the temporal dependencies between features, processing the sequence in both forward and backwards directions. This provides a complete contextual understanding of an activity, which is crucial for distinguishing between similar activities (e.g., 'hand rubbing' vs. 'knuckle cracking') and offers an advantage over using an LSTM alone. Batch Normalisation is applied after each layer in the model to stabilise the learning process and improve convergence. Dropout is also used consistently throughout the network as a regularisation technique to prevent overfitting.

3.4. Experimental Validation

To evaluate our methodology, we designed a three-stage experimental protocol.

1. **Baseline Performance:** First, the 1DCNNBiLSTM model was trained and validated on the full ADAMSense dataset to establish a baseline performance on research-grade data.
2. **Quantifying Domain Shift:** Next, to quantify the "lab-to-life" gap, the model was trained only on the 6-feature subset of the ADAMSense data (wrist-worn accelerometer and gyroscope) and then evaluated directly on the entire custom Garmin Venu 3 dataset without any fine-tuning.
3. **Fine-Tuning and Cross-Validation:** Finally, to validate our domain adaptation strategy and ensure the robustness of our findings, we fine-tuned and validated the pre-trained model using a 5-fold cross-validation on our custom dataset. The data was partitioned into five folds. In each iteration, four folds were used for fine-tuning the pre-trained model, and the remaining fold was used for testing. This process was repeated five times, with each fold serving as the test set once. The final performance metrics reported in Section 4 are the mean and standard deviation of the results across all five folds.

4. Results

4.1. Model Performance on Entire ADAMSense Dataset

To demonstrate the accuracy of the 1DCNNBiLSTM model, it was first trained and validated on the entire ADAMSense dataset. The model was trained for 28 epochs using early stopping, which stops the training process if the validation accuracy does not increase further after five epochs. Upon completion of training, the base model achieved an overall accuracy of 98.8% and a loss of 0.037, which is an improvement from the 92% accuracy of the best model in the ADAMSense paper [11]. While this high accuracy on the source dataset could raise concerns of overfitting, the model's architecture includes dropout and batch normalisation layers as regularizers. The model demonstrated excellent performance across all classes, with an average F1-score of 0.99. This strong result confirms that the 1DCNNBiLSTM architecture is highly effective at learning to distinguish between these subtle, anxiety-related activities.

4.2. Base Model Performance on Source Data

The initial stage of our methodology involved training the base model on a subset of the ADAMSense dataset. To ensure the source data mirrored the data available from our target commercial device, we used only the wrist-worn accelerometer and gyroscope data from the ADAMSense dataset, discarding all other sensor streams. Using early stopping, the model was trained for 35 epochs. The training and validation accuracy and loss are shown in Figure 1. After training, the model achieved an average F1-score of 0.90 and accuracy of 89.11%, demonstrating its ability to learn the general features of the target activities from the source domain. The confusion matrix in Figure 2 also shows the high performance of the model on a class-by-class level.
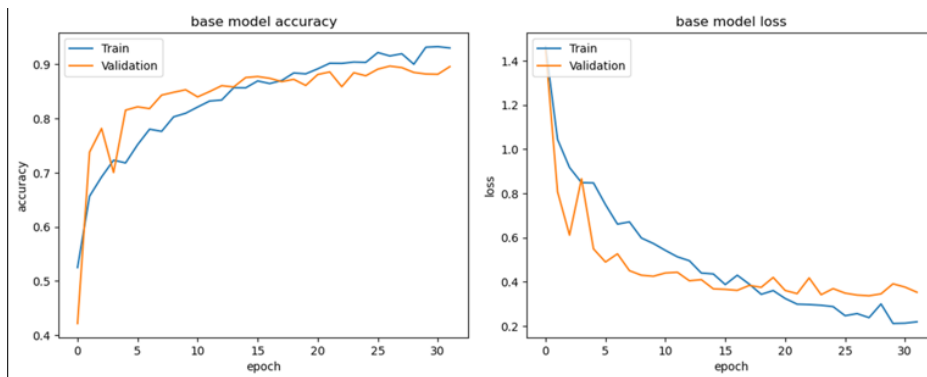

Fig. 1: Accuracy and Loss of the 1DCNNBiLSTM on the Wrist Accelerometer and Gyroscope Data

4.3. Fine-Tuning Performance with Cross-Validation

To quantify the performance gap between the research-grade source domain and the commercial-grade target domain, the base model was first evaluated on the custom dataset collected from the Garmin Venu 3. Without any fine-tuning, the model performed poorly, achieving an accuracy of only 48.7%.

The model then underwent a fine-tuning process evaluated with a 5-fold cross-validation protocol. The fine-tuning resulted in a substantial and stable performance improvement. The model achieved a mean accuracy of 87.0% ($\pm$ 1.2%) across the five folds. The average classification report, presented in Table 1b, shows a substantial performance improvement. This significant increase from 48.7% to 87.0% clearly demonstrates the critical impact and success of the fine-tuning strategy in adapting the model to the commercial wearable sensor data.

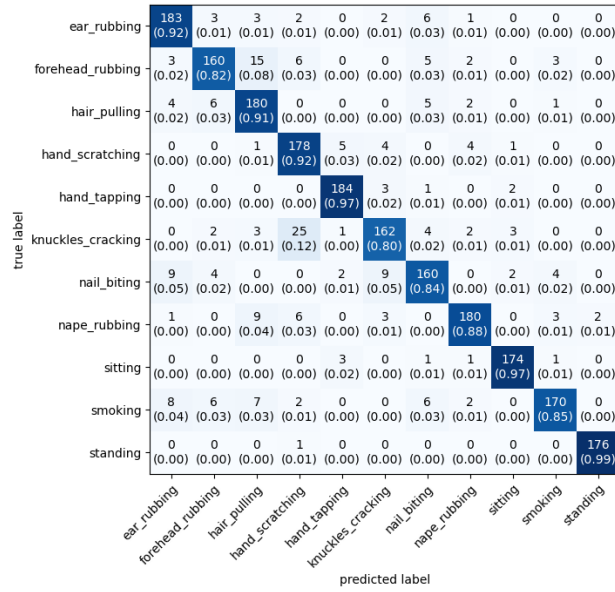| true label \ predicted	ear_rubbing	forehead_rubbing	hair_pulling	hand_scratching	hand_tapping	knuckles_cracking	nail_biting	nape_rubbing	sitting	smoking	standing
ear_rubbing	183 (0.92)	3 (0.01)	3 (0.01)	2 (0.01)	0 (0.00)	2 (0.01)	6 (0.03)	1 (0.01)	0 (0.00)	0 (0.00)	0 (0.00)
forehead_rubbing	3 (0.02)	160 (0.82)	15 (0.08)	6 (0.03)	0 (0.00)	0 (0.00)	5 (0.03)	2 (0.01)	0 (0.00)	3 (0.02)	0 (0.00)
hair_pulling	4 (0.02)	6 (0.03)	180 (0.91)	0 (0.00)	0 (0.00)	0 (0.00)	5 (0.03)	2 (0.01)	0 (0.00)	1 (0.01)	0 (0.00)
hand_scratching	0 (0.00)	0 (0.00)	1 (0.01)	178 (0.92)	5 (0.03)	4 (0.02)	0 (0.00)	4 (0.02)	1 (0.01)	0 (0.00)	0 (0.00)
hand_tapping	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	184 (0.97)	3 (0.02)	1 (0.01)	0 (0.00)	2 (0.01)	0 (0.00)	0 (0.00)
knuckles_cracking	0 (0.00)	2 (0.01)	3 (0.01)	25 (0.12)	1 (0.00)	162 (0.80)	4 (0.02)	2 (0.01)	3 (0.01)	0 (0.00)	0 (0.00)
nail_biting	9 (0.05)	4 (0.02)	0 (0.00)	0 (0.00)	2 (0.01)	9 (0.05)	160 (0.84)	0 (0.00)	2 (0.01)	4 (0.02)	0 (0.00)
nape_rubbing	1 (0.00)	0 (0.00)	9 (0.04)	6 (0.03)	0 (0.00)	3 (0.01)	0 (0.00)	180 (0.88)	0 (0.00)	3 (0.01)	2 (0.01)
sitting	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	3 (0.02)	0 (0.00)	1 (0.01)	1 (0.01)	174 (0.97)	1 (0.01)	0 (0.00)
smoking	8 (0.04)	6 (0.03)	7 (0.03)	2 (0.01)	0 (0.00)	0 (0.00)	6 (0.03)	2 (0.01)	0 (0.00)	170 (0.85)	0 (0.00)
standing	0 (0.00)	0 (0.00)	0 (0.00)	1 (0.01)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	176 (0.99)

Fig. 2: Confusion Matrix of the 1DCNNBiLSTM on the Wrist Accelerometer and Gyroscope Data

Table 1: Comparison of classification performance on the custom Garmin Venu 3 dataset before and after fine-tuning.

(a) Before Fine-tuning

Activity	Precision	Recall	F1-Score
ear_rubbing	0.25	0.26	0.26
forehead_rubbing	0.59	0.63	0.61
hair_pulling	0.63	0.84	0.72
hand_scratching	0.29	0.41	0.34
hand_tapping	0.88	0.60	0.71
knuckles_cracking	0.06	0.07	0.07
nail_biting	0.23	0.24	0.23
nape_rubbing	0.00	0.00	0.00
sitting	0.96	0.96	0.96
smoking	0.67	0.49	0.57
standing	1.00	1.00	1.00
accuracy			0.49
macro avg	0.51	0.50	0.50
weighted avg	0.50	0.49	0.49

(b) After Fine-tuning (5-Fold Avg.)

Activity	Precision	Recall	F1-Score
ear_rubbing	0.77 (±0.06)	0.95 (±0.03)	0.85 (±0.04)
forehead_rubbing	0.90 (±0.05)	0.89 (±0.03)	0.89 (±0.04)
hair_pulling	0.92 (±0.03)	0.74 (±0.09)	0.82 (±0.06)
hand_scratching	0.82 (±0.03)	0.75 (±0.04)	0.78 (±0.03)
hand_tapping	0.77 (±0.03)	0.99 (±0.01)	0.86 (±0.02)
knuckles_cracking	0.82 (±0.05)	0.82 (±0.02)	0.82 (±0.03)
nail_biting	0.91 (±0.03)	0.87 (±0.05)	0.89 (±0.02)
nape_rubbing	0.87 (±0.04)	0.93 (±0.03)	0.90 (±0.02)
sitting	0.93 (±0.05)	0.93 (±0.04)	0.93 (±0.02)
smoking	0.86 (±0.06)	0.82 (±0.06)	0.84 (±0.03)
standing	0.94 (±0.04)	0.95 (±0.04)	0.94 (±0.02)
accuracy			**0.870 (±0.012)**
macro avg	0.86 (±0.02)	0.87 (±0.01)	0.86 (±0.01)
weighted avg	0.86 (±0.02)	0.87 (±0.01)	0.86 (±0.01)

4.4. Ablation Study

To validate the design of our proposed hybrid 1DCNNBiLSTM architecture, we conducted an ablation study comparing its performance with that of several simpler variants. The study was performed on the ADAMSense dataset subset (using only wrist-worn accelerometer and gyroscope data) to isolate the contribution of each architectural component. The models were trained under identical conditions, and their performance was compared using the average F1-score, as shown in Table 2

The results clearly demonstrate the superiority of the proposed hybrid model. The full 1DCNNBiLSTM architecture achieved the highest F1-score of 0.871. The "CNN-Only" model performed the worst (0.697), indicating that while convolutional layers are effective at feature extraction, they are insufficient on their own to model the temporal dynamics of the activities. Both the "LSTM-Only" (0.828) and "BiLSTM-Only" (0.813) models performed better,

Table 2: Ablation study results comparing the F1-score of the proposed model against alternative architectures.

Model Architecture	F1-Score
1DCNNBiLSTM (Proposed)	**0.871**
LSTM-Only	0.828
1DCNNLSTM	0.817
BiLSTM-Only	0.813
1DCNN-Only	0.697

highlighting the importance of temporal sequence modelling. However, their performance was still notably lower than the full hybrid model. This confirms that the CNN layers provide crucial, high-quality features to the recurrent layers, and the combination of both components is synergistic. The bidirectional nature of the LSTM layers also provided a clear advantage over the standard LSTM layers. This analysis justifies our selection of the 1DCNNBiLSTM architecture as the most effective for this task.

5. Discussion

5.1. Summary of Findings

The results of this study validate the proposed methodology for recognising anxiety-related activities on commercial wearable devices through domain adaptation. The initial 1DCNNBiLSTM model, when trained and evaluated solely on the wrist sensor data from the ADAMSense dataset, demonstrated high efficacy, achieving an accuracy of 89.11%.

However, the critical challenge of the "lab-to-life" gap was clearly evidenced when this pre-trained base model was directly applied to the custom dataset collected from the Garmin Venu 3. In this scenario, its performance dropped drastically to an accuracy of approximately 49%. This result confirms that models trained on research-grade data do not generalise well to data from consumer-grade hardware without specific adaptation.

Following the fine-tuning process using a 5-fold cross-validation strategy, the model demonstrated significant improvement, with a mean accuracy of 87.0%. This successful performance increase underscores the effectiveness of the fine-tuning strategy in adapting the model to the unique signal characteristics of the commercial device, thereby bridging the performance gap.

5.2. Significance and Implications

The findings of this study carry significant implications for the field of applied machine learning in wearable health technology. The research makes two primary contributions to the field: first, it provides a clear, quantitative measure of the performance degradation when moving from a research to a consumer-grade device, and second, it validates fine-tuning as a practical and effective solution to this problem.

Quantifying the "Lab-to-Life" Gap: This study quantifies the severe impact of the widely acknowledged domain shift between research and commercial sensors, evidenced by the accuracy dropping from 89.11% to 49% when the base model was applied to the Garmin Venu 3 data. This finding underscores that models developed in controlled lab settings, even with relevant data, are not directly transferable to real-world consumer hardware. It serves as a crucial data point for researchers, emphasising the necessity of planning for domain adaptation from the outset of any project intended for practical deployment.

Validating Fine-Tuning as a Practical Methodology: The most significant implication of this work is the validation of our fine-tuning strategy. By significantly improving the model's accuracy to 87.0% on the custom Garmin data, we demonstrate that a targeted adaptation with a relatively small, custom-collected dataset is a viable and efficient solution. This is highly encouraging for the research community, as it provides a practical pathway for leveraging large, publicly available datasets while still creating effective, specialised models for the noisy and constrained envi-

ronments of consumer devices. This methodology lowers the barrier to entry, as it reduces the need to create massive, new datasets from scratch for every new device or application.

Enabling Future Digital Mental Health Applications: By demonstrating reliable detection of subtle, anxiety-related micro-activities on a commercial smartwatch, this work enables the next generation of digital mental health tools. Moving beyond simple step counting to monitoring meaningful behavioural biomarkers, our approach creates a foundation for future systems that can provide personalised interventions and supply clinicians with objective, longitudinal data.

By successfully fine-tuning a model pre-trained on a public dataset, we demonstrate a viable pathway to overcome issues of data accessibility and data quality inherent in commercial devices [7], creating a specialised and effective tool without needing to build a massive new dataset from scratch.

5.3. Limitations and Future Work

While the results of this proof-of-concept study are promising, it is important to acknowledge limitations that should be addressed in future work.

Dataset Size: The most significant limitation is the small custom dataset, with data collected from only six volunteers. This pilot study was not intended to produce a generalizable model, but rather to test the feasibility of the adaptation methodology. Despite the small dataset, a significant increase in accuracy was observed. Future work must involve collecting data from a larger, more diverse cohort to validate these findings and train a more robust model.

Performance Variability Across Activities: Although the fine-tuned model's overall performance improved significantly, there was still some variability in its ability to recognise specific classes. As shown in the final classification report (Table 1b), activities with less consistent motion signatures like "hand scratching" (average F1-score of 0.78) were recognised with lower accuracy than more distinct activities like "standing" (average F1-score of 0.94). This suggests that certain classes may require more training examples or that their motion signatures are inherently more difficult to distinguish using a wrist-worn sensor.

Building upon the successful validation of our methodology, future work should focus on:

- Expanding the size of the custom dataset for even better accuracy,
- On-device implementation and optimisation using frameworks like LiteRT (formerly TensorFlow Lite).
- integrating the HAR model's output into a higher-level longitudinal monitoring system that can be used for a more holistic assessment of a user's state, moving from simple event detection to meaningful behavioural insight.

5.4. Ethical Considerations

The continuous monitoring of behaviour for mental health assessment carries significant ethical responsibilities. Key considerations include user privacy, data security, and the potential for misinterpretation of results. Any real-world system based on this work must be built with a privacy-by-design approach, utilising on-device processing to minimise data transmission, end-to-end encryption, and a transparent user consent policy. It is imperative that such tools are framed not as diagnostic devices but as aids for self-awareness, providing objective data to supplement clinical care and thereby mitigating the risk of causing user anxiety or over-reliance on the technology.

6. Conclusion

This paper presents a deep learning methodology for recognising subtle, anxiety-related activities using data from a commercial smartwatch. We successfully demonstrated that a 1DCNNBiLSTM model, pre-trained on a public research dataset, can be effectively adapted to a commercial-grade device through a targeted fine-tuning process. Our results empirically validated this approach, demonstrating a significant performance increase from a baseline accuracy of 49% to a robust mean accuracy of 87.0% across a 5-fold cross-validation. This work presents a practical and validated pathway for bridging the "lab-to-life" gap, demonstrating that it is feasible to utilise consumer wearables for the

reliable detection of complex behavioural biomarkers. This methodology serves as a critical foundation for the future development of more sophisticated, accessible, and objective systems for longitudinal mental health monitoring.

References

[1] Abd-alrazaq, A., Alajlani, M., Ahmad, R., AlSaad, R., Aziz, S., Ahmed, A., Alsahli, M., Damseh, R., Sheikh, J., 2024. The performance of wearable ai in detecting stress among students: Systematic review and meta-analysis. J Med Internet Res 26, e52622. URL: https://www.jmir.org/2024/1/e52622, doi:10.2196/52622.

[2] Abd-Alrazaq, A., AlSaad, R., Harfouche, M., Aziz, S., Ahmed, A., Damseh, R., Sheikh, J., 2023. Wearable artificial intelligence for detecting anxiety: Systematic review and meta-analysis. Journal of Medical Internet Research 25, e48754. doi:10.2196/48754.

[3] Akbari, A., Jafari, R., 2019. Transferring activity recognition models for new wearable sensors with deep generative domain adaptation, in: Proceedings of the 18th International Conference on Information Processing in Sensor Networks, pp. 85–96.

[4] Chakma, A., Faridee, A.Z.M., Ghosh, I., Roy, N., 2023. Domain adaptation for inertial measurement unit-based human activity recognition: A survey. URL: https://arxiv.org/abs/2304.06489, arXiv:2304.06489.

[5] Chato, L., Regentova, E., 2023. Survey of transfer learning approaches in the machine learning of digital health sensing data. Journal of Personalized Medicine 13, 1703. doi:10.3390/jpm13121703.

[6] Dhekane, S.G., Ploetz, T., 2024. Transfer learning in human activity recognition: A survey. arXiv preprint arXiv:2401.10185 URL: https://arxiv.org/html/2401.10185v1, arXiv:2401.10185.

[7] Dobson, R., Stowell, M., Warren, J., Tane, T., Ni, L., Gu, Y., McCool, J., Whittaker, R., 2023. Use of consumer wearables in health research: Issues and considerations. J Med Internet Res 25, e52444. URL: https://www.jmir.org/2023/1/e52444, doi:10.2196/52444.

[8] Garmin, 2025. Fit protocol. URL: https://developer.garmin.com/fit/protocol/.

[9] Garmin, 2025. Monkey c. URL: https://developer.garmin.com/connect-iq/monkey-c/.

[10] Hassan, L., Milton, A., Sawyer, C., Casson, A.J., Torous, J., Davies, A., Ruiz-Yu, B., Firth, J., 2025. Utility of consumer-grade wearable devices for inferring physical and mental health outcomes in severe mental illness: Systematic review. JMIR Ment Health 12, e65143. URL: https://mental.jmir.org/2025/1/e65143, doi:10.2196/65143.

[11] Khan, N.S., Ghani, M.S., Anjum, G., 2021. Adam-sense: Anxiety-displaying activities recognition by motion sensors. Pervasive and Mobile Computing 78, 101485. URL: https://www.sciencedirect.com/science/article/pii/S1574119221001140, doi:https://doi.org/10.1016/j.pmcj.2021.101485.

[12] König, J.L., Penaredondo, J., McCullagh, E., Bowen, J., Hinze, A., 2024. Let's make it accessible: The challenges of working with low-cost commercially available wearable devices, in: Proceedings of the 35th Australian Computer-Human Interaction Conference, Association for Computing Machinery, New York, NY, USA. p. 493–503. URL: https://doi.org/10.1145/3638380.3638415, doi:10.1145/3638380.3638415.

[13] Liu, J., Niu, S.Y., Lin, J.H., Liu, J.X., Wang, J.J., He, Q.Q., 2023. Comparing the accuracy of wrist-worn activity trackers for estimating sleep in children and adolescents: A systematic review and meta-analysis. PLOS ONE 18, e0286898. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0286898, doi:10.1371/journal.pone.0286898.

[14] Masoumian Hosseini, M., Masoumian Hosseini, S.T., Qayumi, K., Hosseinzadeh, S., Sajadi Tabar, S.S., 2023. Smartwatches in healthcare medicine: assistance and monitoring; a scoping review. BMC Medical Informatics and Decision Making 23, 248. doi:10.1186/s12911-023-02350-w.

[15] Research, S., 2024. What is raw sensor data and why is it so important for clinical research? URL: https://verisense.net/blog/raw-sensor-data.

[16] Shiranthika, C., Premakumara, N., Chiu, H.L., Samani, H., Shyalika, C., Yang, C.Y., 2020. Human activity recognition using cnn & lstm, in: 2020 5th International Conference on Information Technology Research (ICITR), IEEE. pp. 1–6.

[17] Wang, J., Chen, Y., Hao, S., Peng, X., Hu, L., 2019. Deep learning for sensor-based activity recognition: A survey. Pattern Recognition Letters 119, 3–11. URL: https://www.sciencedirect.com/science/article/pii/S016786551830045X, doi:https://doi.org/10.1016/j.patrec.2018.02.010. deep Learning for Pattern Recognition.

[18] World Health Organization, 2023. Anxiety disorders. URL: https://www.who.int/news-room/fact-sheets/detail/anxiety-disorders.

[19] Ye, X., Wang, K.I.K., 2024. Deep generative domain adaptation with temporal relation knowledge for cross-user activity recognition. arXiv preprint arXiv:2403.14682 .

[20] Zhang, S., Li, Y., Zhang, S., Shahabi, F., Xia, S., Deng, Y., Alshurafa, N., 2022. Deep learning in human activity recognition with wearable sensors: A review on advances. Sensors (Basel) 22, 1476. URL: https://www.mdpi.com/1424-8220/22/4/1476, doi:10.3390/s22041476.