mTanaaw: A System for Assessment and Analysis of Mental Health with Wearables

Anuja Pinge† Dheryta Jaisinghani‡, Surjya Ghosh†, Aditya Challa†, Sougata Sen†
† BITS Pilani Goa Campus, India, ‡University of North Iowa, USA

2024 16th International Conference on COMmunication Systems & NETworkS (COMSNETS) | 979-8-3503-8311-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/COMSNETS59351.2024.10427432

***Abstract*—Researchers commonly develop application specific data collection systems for capturing user data in mental health monitoring studies. Although useful for their study, these systems typically capture specific data and are often not extendable in other studies – a scalability issue. To address this issue, we, in this paper, propose mTanaaw, a wearable and mobile-based data collection system that is capable of collecting sensor data in a plug-and-play manner. We provide detailed design and implementation of the system. We provide design and implementation details of the system. Using mTanaaw, we demonstrate its capabilities of collecting mental health data in a stress detection study. We evaluated the proposed system with five participant's data, which is collected using the data collection module. We present the entire pipeline of mTanaaw with multiple possible changes in configuration of parameters via the system. Overall, we observe that such a system can reduce the burden on researchers. We will make the system available for mental health researchers.**

***Index Terms*—Wearable Technology, Affective Computing**

I. Introduction

Detecting mental health using wearables has been a long standing research goal in the mobile and wearable healthcare community. Although researchers have made substantial progress in the area over the last several years, yet these advances often appear minuscule as compared to advances in several other research fields, such as Natural Language Processing [1], and Computer Vision [2]. One possible reason for this slow advancement is because of the unavailability of very large datasets; fields like Computer Vision have very large datasets such as ImageNet [3]. This unavailability can be attributed to the lack of agreement in collecting specific homogeneous signals, and challenges involved in performing human subject studies. Indeed, researchers in the mental health domain collect data using various devices, via various sensors, and at various sampling rates. Furthermore, the differences in data collectors often hinder researchers from collecting data from specific type of devices – e.g., a data collector might allow pairing only smartwatches, but not chest worn heart rate monitors.

Fields which have large datasets are currently relying on using deep learning techniques for completing certain types of tasks. Although, mental health monitoring research has recently aimed to use deep learning [4], yet, unfortunately, the performance of deep learning models in mental health research is quite low as compared to performance in other fields. This low performance is often attributed to the small size of the datasets. To take advantage of the advances in deep learning-based data analysis, a uniform data collection tool specifically designed for mental health monitoring research is urgently required. To address this gap, in this paper, we present *mTanaaw*, a data collection tool that is designed specifically for collecting data related to mental health monitoring. Although data collection applications for human activity recognition is currently available, however, these applications do not cater to all requirements of a passive data collector for mental health monitoring.

Although the task of developing a mental health data collection application might sound trivial, however, there are several inherent challenges that must be addressed while building such a system. Firstly, researchers store data in a specific manner, based on their requirements. This prevents having uniformity in datasets. Researchers who venture into using existing datasets often spend sub-

stantial time understanding the data storage pattern. Secondly, the devices used in these studies are heterogeneous. Different studies must collect data at different sampling rates. The task of introducing homogeneity in the data is a humongous task. Thirdly, mental health research has started relying on Just-in-Time Adaptive Interventions (JITAIs). Current activity recognition systems, however, do not often have the possibility of providing JITAIs. Fourthly, questionnaires specific to mental heath monitoring are often diverse leading to non-standard data storage methods. Finally, the data processing and intervention pipeline vary substantially based on the type of application.

mTanaaw provides a platform that addresses these aforementioned challenges. We will provide a tool that can collect data and store them in a uniform manner. We have developed a common data processing pipeline. Our system design is extremely modular – researchers can easily integrate additional modules to the system. Furthermore, we plan to evolve the system over time to introduce additional features.

We have currently tested the system on a small user study for physiological stress monitoring – one use case of the mTanaaw system. Overall, we observed that the system could collect multiple data types from multiple data sources (devices). We could run both deep and shallow learning techniques and obtain desired evaluation metric based outcomes. Key contributions of this work are:

- We present the design and implementation details of *mTanaaw*, a generic framework for collecting, processing, and analyzing data in mental health monitoring studies.
- We use the *mTanaaw* system to collect data from a small stress detection study. We demonstrate that using shallow learning techniques, we could detect stress with a F1-score of 64.75%.

II. Related Work

Traditionally. mental health studies have relied on Experience Sampling Method, where user response was collected via pen and paper [5]. Obvious issues such as human effort, recall bias, under reporting, etc. exist in such form of data collections. To overcome these challenges, developing a digital, mobile-based data collection system became obvious. This prompted researchers to develop digital data collection systems such as *andWellness* [6]. However, with the fast advancement in mobile and wearable systems, platforms like *andWellness* have gradually become dated. It is challenging to integrate devices such as smartwatches in the system.

Modern wearable-based mental health monitoring systems often require data from a diverse range of sensors. This is in contrast with research in the areas of NLP and computer vision where the data is uniform – either images or speech signal collected at specific intervals [2]. This diversity forces researchers to develop a specific data collection system that suffices their needs in the short time period. However, the lack of uniformity in the manner in which data is collected, and lack of desire to share the data with others often leads to existence of small, study-related datasets. A recent effort in developing a common data collection system and providing a uniform dataset by Xu et al. aims to bridge this issue [7]. In their work, Xu et al. collected data from 618 participant-years of data. The data collected is from a convenience sample – student population– collected from a specific geographic location. It is unclear whether building models using this data will generalize to other geographic regions. Making the data collector system public opens up the scope for others to use this app and collect more data. However, it is unclear whether the authors aim to create a large dataset (such as ImageNet) where researchers globally can contribute directly. This again brings back the age old challenge – "where do we get a large dataset to train our model?" This problem of having small datasets or no uniform data collector does not exist only for mental health monitoring systems, but with overall human activity recognition systems. Challenges of data collection with human participants is one of the major hindrance in this effort. Indeed researchers have aimed to develop synthetic

datasets for Inertial Measurement Unit (IMU) data [8]. However, such datasets might be useful in HAR studies, not for physiological sensing studies, which often rely on changes in physiological markers, behavioral patterns, or social signals [9].

One specific wearable-based mental health monitoring task that has been of interest to researcher for years is physiological stress detection [10]. Yet advances in the field is low due to lack of large datasets. Researchers such as Chalmers et al. used only Electrocardiogram (ECG) features to detect stress [11], while Mishra et al. used ECG data along with GSR data in their studies [12], [13]. Furthermore, while Chalmers et al. collected data using an off-the-shelf FitBit wristband, Mishra et al. used a Polar HRM and a custom made GSR sensor. These devices do not collect similar data and thus it is a challenge to develop a dataset where both the data can be used together.

Several researchers have trained the stress detection model using Supervised learning techniques using classifiers and regressors such as Support Vector Machine (SVM), K-NN, Naive Bayes, Decision Tree, Random Forest, XGBoost [14], [15]. In addition to the classifier and regressor choice, there is also a difference in the validation approach. While some researchers prefer using the train-validate-test split, others perform leave-one-out cross validation [15].

Some have also attempted deep learning approaches such as 1D CNN and LSTM for stress detection. Lee et al. collected EEG signals using laboratory study protocol. They used CNN for stress detection and observed that CNN has capability to classify into stressful and non-stressful situation [16]. Li et al. used LSTM to train the model [4]. Researchers such as Yu & Sano used Autoencoder for performing stress detection [17]. Autoencoder, which is unsupervised deep learning is widely used in other domains such as Computer Vision. Although promising, it is unknown whether Autoencoders can perform well for small datasets.

All these works indicate the diversity in sensor signals, devices, data collection approaches, machine learning techniques, and output metrics. We aim to develop mTanaaw that will handle these diversities and create a system that will seamlessly collect data for any mental health study.

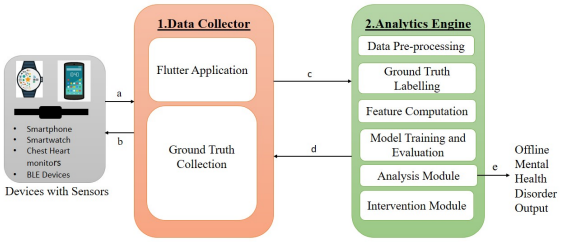


Fig. 1. Architecture of mTanaaw, our proposed data collection tool for Mental Health Monitoring.

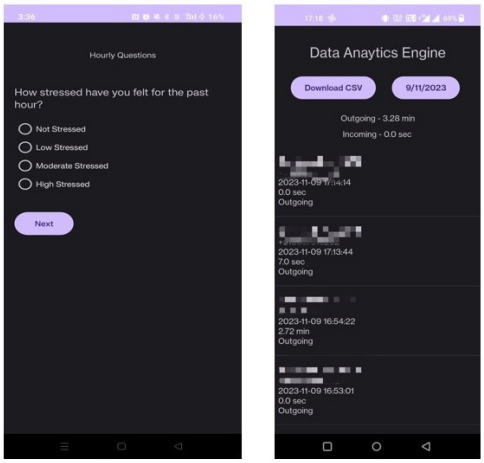
III. System Implementation Details

We next provide implementation details of the mTanaaw system. The mTanaaw system consists of two major modules – a platform independent data collector module that is implemented using Flutter, and a backend module that has both data management and analytics modules. The architecture of the proposed system is shown in Figure 1.

A. System Description

1. Data Collector Data collector consists of a Flutter App for both sensor data collection and ground truth collection.

Sensor Data Collector: The data collector is developed in Flutter. It can be configured to run either on the smartwatch or on the smartphone. Based on the device where it is executing, specific capabilities are either turned on or turned off – e.g., the smartphone application can collect data from the smartphone, as well as aggregate data from other wearable devices. Users can turn on the sensors of interest. Subsequently, users can select the sampling frequency for each sensor. Currently, we have just enabled collection of heart rate, accelerometer, and gyroscope. However, additional modules can be installed on the devices to collect and store data from other sensors. Arrow (a) in Figure 1 represents the data aggregation pipeline. The data collected from the devices is stored locally and transferred to the smartphone in batches. Batch size can be configured based on the



(a) (b)

Fig. 2. Ground Truth Collection Application a) Hourly Questionnaire b) Call Log Application

application that will use it. Data from the smartphone is also sent to the server in batches (Arrow (c)).

Ground Truth Collector: Ground truth is collected using self-reported questionnaires on the same smartphone app. The interval of sending questionnaires can be specified by the person running the study. Currently, we support both hourly and morning-evening(daily) questionnaires.

2. Analytics Engine The data analytics engine performs data pre-processing and analysis tasks.

Data Pre-processing: The pre-processing steps required for physiological signals include outliers removal, interpolating missing values, removing duplicates, filtering out invalid readings, and normalization. Data received from different devices can have missing values for timestamps. Hence there is a need to detect such in-between missing timestamps and assign value. Currently we have the option of using pre-built interpolation techniques or users can deploy their custom interpolation technique. We also have certain outlier removal and normalization techniques pre-built that are optional for users to include.

Feature Computation: We developed a feature computation module that can compute different sensor-specific features with the provided window size and overlap percentage. These features can include time-domain, frequency-domain, or statistical-features. Normalized data is grouped in varying window sizes with or without overlaps for computing features. Users can add more or remove existing features.

Training and Evaluation of the Model: mTanaaw currently can perform train-test split based validation, as well as cross validations techniques such as 10-fold- cross validation or leave-one-person-out cross validation.

Intervention Module: mTanaaw's backend stores various questionnaires. For specific studies, researchers can indicate which questionnaires to send to participants when certain criteria are met. We currently have a FireBase implementation for the same. When specific criteria are met, the questionnaires are sent to the user's phone and wearables (Figure 1(d)). Users can respond to a questionnaire via any connected device.

B. USE CASE

We next demonstrate the usage of mTanaaw in a specific use case – stress detection. To attain this, we collected smartwatch based heart rate, accelerometer, and gyroscope data using mTanaaw.

Dataset: We recruited 5 participants (3 males, and 2 females aged between 20 years to 30 years). Participants were asked to wear a Samsung Watch 4 during the study. The Samsung Watch 4 was configured to collect heart rate at 1 reading per second and accelerometer data at 6 readings per second. The average study duration was around 41 minutes (min 40 minutes, max 43 minutes).

Data Collection Method: The study was performed in a controlled environment by inducing stress. The protocol used was the same as the one used in prior work by Pinge et al. [18]. We used the known sequence of tasks as the ground truth.

Feature Computation: We computed time domain features by varying window sizes between 15 secs and 60 secs with 0% to 50% of overlap. We computed time domain features from heart rate – i.e., Min(HR), Max(HR), Mean(HR), St.Dev(HR). We also computed the magnitude of the accelerometer data.

Training the Model: We used the above features to train a model using a super-

vised approach and a deep learning approach. Specifically, we used Random Forest, 1D-CNN and Auto Encoder in mTanaaw for training stress detection models.

Evaluation and Results: We evaluated the model using Leave-One-Person-Out cross-validation (LOPO-CV). We considered heart rate features and also a combination of heart rate with accelerometer features.

Random Forest: We observed that considering heart rate window size of 60 seconds without overlap performed better then window size of 15 seconds. In the best case, we obtained a F1-score of 60.6%. We also noted that the F1-score improved to 64.6% when accelerometer features were used along with heart rate. Table I presents results for the Random Forest classifier with heart rate features. Table II presents results for the Random Forest classifier using heart rate and accelerometer features.

Window, Overlap	Accuracy	F1-Score	Precision	Recall
15s, 50%	61.2%	59.5%	51.2%	36.8%
15s, 0%	61.3%	59.3%	53.5%	35.7%
60s, 50%	59.5%	58.3%	54.0%	44.9%
60s, 0%	**62.0%**	**60.6%**	**61.7%**	**46.1%**

TABLE I
RF CLASSIFIER WITH HR FEATURES

Window, Overlap	Accuracy	F1-Score	Precision	Recall
15s, 50%	62.1%	60.3%	53.2%	36.6%
15s, 0%	63.0%	61.2%	54.8 %	36.8%
60s, 50%	63.5%	62.5 %	57.0%	44.7 %
60s, 0%	**65.4%**	**64.8%**	**57.5%**	**50%**

TABLE II
RF CLASSIFIER WITH HR AND ACCELERATION FEATURES

AutoEncoder: The AutoEncoder model performed better with heart rate and accelerometer features for window size of 60 seconds in terms of accuracy, but the F1-score was lower. We also observed that the accuracy of the AutoEncoder model increased as compared to the Random Forest model from 65.4% to 69.3% but there is a substantial reduction in the F1-score from 64.6% to 47.4%. Table III presents the results for the AutoEncoder model with heart rate features. Table IV presents the results for the AutoEncoder model using heart rate and accelerometer features.

Window, Overlap	Accuracy	F1-Score	Precision	Recall
15s, 50%	67.5%	44.4%	61.0%	38.9%
15s, 0%	68.4%	43.1%	63.1%	35.8%
60s, 50%	66.7%	46.0%	**65.3%**	38.4%
60s, 0%	**68.8%**	**47.4%**	63.5%	**41.1%**

TABLE III
AUTOENCODER WITH HR FEATURES

Window, Overlap	Accuracy	F1-Score	Precision	Recall
15s, 50%	60.7%	26.1%	37.2%	22.8%
15s, 0%,	65.2%	46%	59.3 %	40.8%
60s, 50%	63.2 %	**46.7 %**	**61.6%**	**42%**
60s, 0%	**69.3%**	44.1%	59.8%	38.6%

TABLE IV
AUTOENCODER WITH HR AND ACCELERATION FEATURES

1D CNN: We also performed the classification using 1D-CNN and observed that 1D CNN model gave higher accuracy for heart features alone, as compared to heart rate with accelerometer features. Furthermore, 1D CNN resulted in higher accuracy and F1-score for window size of 15 seconds with overlap of 50% between the windows. Overall, we observed that 1D CNN resulted in the highest recall amongst all models – 96.5%. Tables V and VI present the performance of the 1D CNN classifier.

Window, Overlap	Accuracy	F1-Score	Precision	Recall
15s, 50%	**64.5%**	**53.4%**	**38.8%**	**96.5%**
15s, 0%	67.0%	48.4%	35.5 %	87.2%
60s, 50%	60 %	41.4%	30.1%	73.8%
60s, 0%	47.8%	33.8%	25.6%	56.6%

TABLE V
1D CNN WITH HR FEATURES

Window, Overlap	Accuracy	F1-Score	Precision	Recall
15s, 50%	**63.1%**	**53.6%**	**38.8%**	**96.5%**
15s, 0%	55.8%	52.2%	38.5%	93.3%
60s, 50%	55.1%	49.0 %	36.3%	87.8%
60s, 0%	60%	45.9%	33.4%	80.8%

TABLE VI
1D CNN WITH HR AND ACCELERATION FEATURES

IV. DISCUSSION AND FUTURE WORK

Currently, we have developed mTanaaw, a framework that can collect and analyse mental health data. We have tested the system using diverse sensors and in a stress detection study. This system can be further extended to monitor other mental health disorders. In

future, we intend to add more modules and sensor data extractors such as Electrodermal activity (EDA), and Respiration sensors. As a part of future work, we intend to develop an intervention system for providing timely interventions or guidance that can help individuals to tackle their state of mind. We observed that deep learning techniques performed poorly as compared to the shallow learning models. One reason for this could be the very limited amount of data. In future, we will perform larger user studies, share the dataset as well as the tool so that other researchers can provide additional data collected from diverse population.

V. Conclusion

In this paper, we present the design and development of mTanaaw, a mental health monitoring system. We evaluated the system in a 5-participant stress detection study and obtained performance values for various sensors, classifiers and parameters. In future, we will make this system open source and available for researchers to use and collect uniform mental health data.

Acknowledgment

This research results from a research program supported by BITS Pilani's grants BPGC/RIG/2020-21/03-2021/01 and BPGC/RIG/2021-22/08-2021/01. All findings and recommendations are those of the authors and do not necessarily reflect the views of the sponsors.

References

[1] U. Ahmed, K. Iqbal, and M. Aoun, "Natural language processing for clinical decision support systems: A review of recent advances in healthcare," *JICET*, vol. 8, no. 2, pp. 1–16, 2023.

[2] D. Kitaguchi, N. Takeshita, H. Hasegawa, and M. Ito, "Artificial intelligence-based computer vision in surgery: Recent advances & future perspective," *Annals of gastroenterological surgery*, vol. 6, no. 1, 2022.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[4] B. Li and A. Sano, "Extraction and interpretation of deep autoencoder-based temporal features from wearables for forecasting personalized mood, health, and stress," *Proc. of the ACM on Inter., Mob., Wearable and Ubiq. Technol. (IMWUT)*, vol. 4, no. 2, 2020.

[5] N. Schmitz *et al.*, "Diagnosing mental disorders in primary care: the general health questionnaire (ghq) and the symptom check list (scl-90-r) as screening instruments," *Social psychiatry and psychiatric epidemiology*, vol. 34, 1999.

[6] J. Hicks *et al.*, "Andwellness: An open mobile system for activity and experience sampling," in *Wireless Health 2010*, ser. WH '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 34–43. [Online]. Available: https://doi.org/10.1145/1921081.1921087

[7] X. Xu *et al.*, "Globem: Cross-dataset generalization of longitudinal human behavior modeling," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 6, no. 4, jan 2023. [Online]. Available: https://doi.org/10.1145/3569485

[8] H. Kwon *et al.*, "Imutube: Automatic extraction of virtual on-body accelerometry from video for human activity recognition," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 4, no. 3, sep 2020. [Online]. Available: https://doi.org/10.1145/3411841

[9] S. Abdullah and T. Choudhury, "Sensing technologies for monitoring serious mental illnesses," *IEEE MultiMedia*, vol. 25, no. 1, pp. 61–75, 2018.

[10] J. Healey and R. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 2, pp. 156–166, 2005.

[11] T. Chalmers *et al.*, "Stress watch: The use of heart rate and heart rate variability to detect stress: A pilot study using smart watch wearables," *Sensors*, vol. 22, no. 1.

[12] V. Mishra *et al.*, "Continuous detection of physiological stress with commodity hardware," *ACM transactions on computing for healthcare*, vol. 1, no. 2, 2020.

[13] V. Mishra, S. Sen, G. Chen, T. Hao, J. Rogers, C.-H. Chen, and D. Kotz, "Evaluating the reproducibility of physiological stress detection models," *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, vol. 4, no. 4, pp. 1–29, 2020.

[14] S. Campanella *et al.*, "A method for stress detection using empatica e4 bracelet and machine-learning techniques," *Sensors*, vol. 23, no. 7, p. 3565, 2023.

[15] A. Tazarv *et al.*, "Personalized stress monitoring using wearable sensors in everyday settings," in *International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2021.

[16] J. H. Lee *et al.*, "Stress monitoring using multimodal bio-sensing headset," in *Extended Abstracts of Conference on Human Factors in Computing Systems*, 2020.

[17] H. Yu and A. Sano, "Semi-supervised learning for wearable-based momentary stress detection in the wild," *Proceedings of the ACM on Inter., Mob., Wearable and Ubiq. Technol. (IMWUT)*, vol. 7, no. 2, 2023.

[18] A. Pinge, S. Bandyopadhyay, S. Ghosh, and S. Sen, "A comparative study between ECG-based and PPG-based heart rate monitors for stress detection," in *In International Conference on COMmunication Systems & NETworkS (COMSNETS)*. IEEE, 2022.