

## Article

# Investigating the Feasibility of Assessing Depression Severity and Valence-Arousal with Wearable Sensors Using Discrete Wavelet Transforms and Machine Learning

Abdullah Ahmed <sup>1,\*</sup> , Jayroop Ramesh <sup>2,\*</sup> , Sandipan Ganguly <sup>3</sup>, Raafat Aburukba <sup>2</sup> , Assim Sagahyroon <sup>2</sup> and Fadi Aloul <sup>2</sup> 

<sup>1</sup> Department of Electrical and Computer Engineering, University of Massachusetts Amherst, Amherst, MA 01003, USA

<sup>2</sup> Department of Computer Science and Engineering, American University of Sharjah, Sharjah 26666, United Arab Emirates

<sup>3</sup> Department of Computer Science, University College London, London WC1TE 6BT, UK

\* Correspondence: amahmed@umass.edu (A.A.); b00057412@aus.edu (J.R.); Tel.: +971-5-59517842 (J.R.)

**Abstract:** Depression is one of the most common mental health disorders, affecting approximately 280 million people worldwide. This condition is defined as emotional dysregulation resulting in persistent feelings of sadness, loss of interest and inability to experience pleasure. Early detection can facilitate timely intervention in the form of psychological therapy and/or medication. With the widespread public adoption of wearable devices such as smartwatches and fitness trackers, it is becoming increasingly possible to gain insights relating the mental states of individuals in an unobtrusive manner within free-living conditions. This work presents a machine learning (ML) approach that utilizes retrospectively collected data-derived consumer-grade wearables for passive detection of depression severity. The experiments conducted in this work reveal that multimodal analysis of physiological signals in terms of their discrete wavelet transform (DWT) features exhibit considerably better performance than unimodal scenarios. Additionally, we conduct experiments to view the impact of severity on emotional valence-arousal detection. We believe that our work has implications towards guiding development in the domain of multimodal wearable-based screening of mental health disorders and necessitates appropriate treatment interventions.

**Keywords:** affective; depression screening; digital phenotype; emotion; machine learning; passive sensing; wavelet transforms; wearable devices



**Citation:** Ahmed, A.; Ramesh, J.; Ganguly, S.; Aburukba, R.; Sagahyroon, A.; Aloul, F. Investigating the Feasibility of Assessing Depression Severity and Valence-Arousal with Wearable Sensors Using Discrete Wavelet Transforms and Machine Learning. *Information* **2022**, *13*, 406. <https://doi.org/10.3390/info13090406>

Academic Editors: Sidong Liu, Cristián Castillo Olea and Shlomo Berkovsky

Received: 19 July 2022

Accepted: 25 August 2022

Published: 27 August 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Depression is one of the prevailing mental health disorders, affecting approximately 280 million people worldwide [1,2]. In the wake of the crippling effects of the pandemic, the global burden of depression continued to worsen, by at least 27.6% on average. This debilitating disorder differs from general mood fluctuations and fleeting emotional reactions along the aspects of intensity, recurrence, and anhedonia [1,2]. Depression often manifests owing to a complex interaction between socioeconomic, psychological and biological factors, and is exacerbated by comorbid physical conditions in an adverse feedback loop. Early screening is critical for timely intervention and the delivery of effective treatment methods such as cognitive behavioral therapy and/or antidepressant medications. However, the social stigma associated with mental health, the scarcity of trained health-care providers, relatively high rates of misdiagnoses and exorbitant service costs discourage individuals from seeking assistance [3].

To improve diagnoses fidelity, it is vital to develop universal screening approaches which reveal useful information without being intrusive or biased, as is the case with the currently adopted surveys [4]. With the proliferation of machine learning (ML) methods

in the healthcare domain, where the automatic identification of relevant patterns and relationships among data without specification of a priori hypotheses holds considerable prognostic utility, there is the potential for detecting the presence of elusive disorders such as depression. In parallel, the ubiquity of smartphones and wearable devices such as smartwatches and fitness trackers offer the much sought-after capabilities of long-term unobtrusive monitoring in free-living conditions. Recent studies have employed ML while leveraging passive, non-intrusive modalities such as smartphone call/text logs [4], social media posts [5], gyroscope readings [6] GPS [7] and heart rate [8] for detecting states of depression and mental health distress.

These worthwhile approaches show the potential for depression screening using passive smartphone data while maintaining a fair degree of user privacy. Out of the different parameters measured, those acquired through wearable devices enable continuous and objective monitoring of patients [9]. Moreover, physiological signals, or a combination of them in tandem with ML, can predict symptoms of depression and anxiety [10,11]. This fact, coupled with the increasing ownership of wearables, estimated at 1 billion in 2022, suggest that personalized screening be introduced without revealing any user-specific information as opposed to partial alternatives such as location data or personal message history.

Noticeably, in the most recent literature, the sample population typically consists of depressed and completely healthy participants, and physiological signals are compounded with smartphone-derived biomarkers such as device usage, step count, sleep measures and location. In this work, we aim to quantify the effects of using only unimodal and multimodal signals from wearable sensors in terms of heart rate (HR), galvanic skin response (GSR) and accelerometry (ACC) among a predominantly depressed population.

Another facet of emotional state assessment which can benefit from wearable monitoring is affective experience quantification. As purported by [12], the interplay between the severity of depression and affective emotional activation in terms of valence and arousal can have implications for individual behavior and response to daily stimuli [13]. This manifests a strong influence on processes such as attention, perception, decision-making, learning and mental well-being [14]. Valence captures the extent to which an emotion is positive/negative, while arousal captures the intensity of the experienced emotion. Thus, we also explore this domain as an additional aspect of our central work.

The primary contributions of this work are the following:

- Validating the potential of implementing ML algorithms with retrospectively collected wearable-derived physiological data for classifying between moderately and severely depressed individuals.
- Assessing the quality of low frequency, general signal features extracted using discrete wavelet transforms (DWT) for developing ML algorithms.
- Examining the relative efficacies of heart rate, galvanic skin response and accelerometry readings in distinguishing between depression severity and emotional states.
- Investigating the role of depression severity in emotional valence and arousal detection.

This paper is organized such that Section 2 introduces the dataset and outlines the methodology and Section 3 presents the results and its discussion, with Section 4 concluding the work.

## 2. Methodology

### 2.1. Dataset

The DAPPER dataset [15] is an aggregation of ambulatory physiological and psychological data reported over a period of five days by 142 participants. It is segregated into sections as displayed in Table 1.

**Table 1.** DAPPER Contents.

| Physiological Data           | Psychological Data                |
|------------------------------|-----------------------------------|
| Heart Rate (HR)              | ESM (Experience Sampling Method)  |
| Galvanic Skin Response (GSR) | DRM (Daily Reconstruction Method) |
| ACC Data                     | -                                 |

The volunteers were asked to engage in a pre-test (a data collection procedure preceding the start of the experiment) where they submitted their BDI-II measures amongst other details. During the main experiment, the psychological details were reported by the participants through answering questionnaires on their smartphones. They received six ESM questionnaires per day over 9 a.m.–11 p.m. with a minimum 3-h gap between each questionnaire. The DRM questionnaires were sent at 11 p.m. each day. Nonetheless, our research aimed to utilize the physiological recordings and focus on heart rate, which was collected using photoplethysmography, galvanic skin response, collected using surface electrodes from the wrist, and triaxial accelerometer data to assess depression severity. All the physiological data was recorded using Psychorus, a customized wristband, during Monday–Friday, 9 a.m.–11 p.m. Out of 142 total patients, 87 patients had valid physiological readings. The data was originally recorded with the following sampling rates (HR:20 Hz, GSR:40 Hz, ACC:20 Hz) but later downsampled to 1 Hz by the original authors to reduce complexity for future use by the research community. DAPPER followed the Helsinki standard and all the participants submitted their written permission.

At first, the raw data comprised 2249 signals. However, any signals of duration of less than 30 min were rejected, since the goal was to follow the 30-min ESM event periods. The concluding number of signals were 2034, each over a span of 1800 s for 87 patients. The distinguishing factor of DAPPER is the environment in which the data was collected in. Instead of the more popular laboratory-based controlled experiments, this dataset aimed to provide data gathered during natural day-to-day activities. Furthermore, this approach replicates accurate everyday situations, thus providing a resource with reliable data that can be used in further applications.

The distribution of the BDI-II scores had a mean with standard deviation of  $29.73 \pm 7.0$ , with a minimum of 21 and a maximum of 60. Discretizing the data for binary classification, the following four cut-offs are applied to generate four depression severity ranges: minimal severity ( $\leq 13$ ), mild severity ( $\geq 14$  and  $\leq 19$ ), moderate severity ( $\geq 20$  and  $\leq 28$ ) and severe severity ( $\geq 29$ ). Out of the 2034 available data instances, 960 belong to the moderate class and 1074 belong to the severe class. This suggests that all 87 patients in the original study suffered from clinical depression to some degree.

With a pure focus on studying depressed populations, we opt to consider the cases exhibiting minimal depression severity as our experimental control group. In accordance with [16], employing a control group of such close proximity to other experimental cases (in terms of depressive mental state) alleviates the control group's inherent selection bias, and potentially enhances the study's validity.

For binary segregation of arousal and valences, the Likert scale-reported scores of 1 and 2 are treated as low, whereas 3, 4 and 5 are considered as high. For valence, this resulted in 1003 and 1031 for low and high categories, respectively. For arousal, this resulted in 1361 and 673 for low and high categories, respectively. It is observable from Figure 1 that the emotional states stratified across the depressed population congregate towards the higher intensity values.

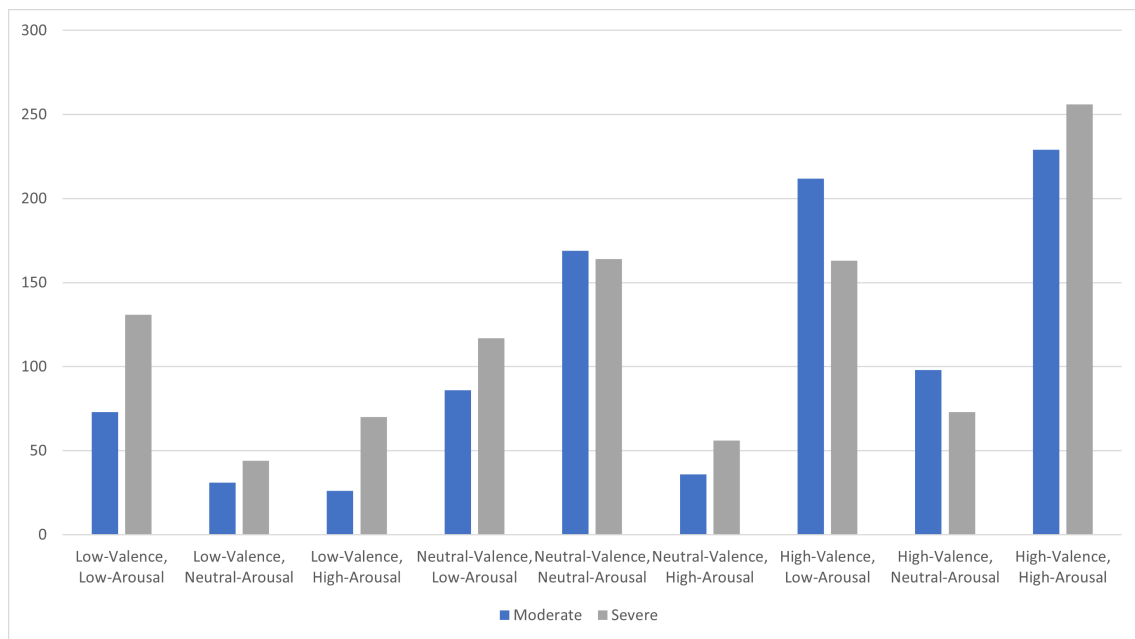


**Figure 1.** Valence and Arousal frequency grouped by depression severity.

## 2.2. Feature Extraction

Our complete approach is outlined in Figure 2, consisting of data processing, feature extraction and ML implementation. Discrete wavelet transform (DWT) can be implemented on-devices directly for real-time signal monitoring with relatively lower battery consumption. We derive generic statistical features for all three modalities after DWT analysis with the goals of minimizing computational overhead and preserving the same processing pipeline and introducing a notion of translation invariance to noise or other motion artifacts. Furthermore, a secondary goal is to view the differences in emotional state activation in terms of valence and arousal across the two depressed populations, and the learnable differences with machine learning.

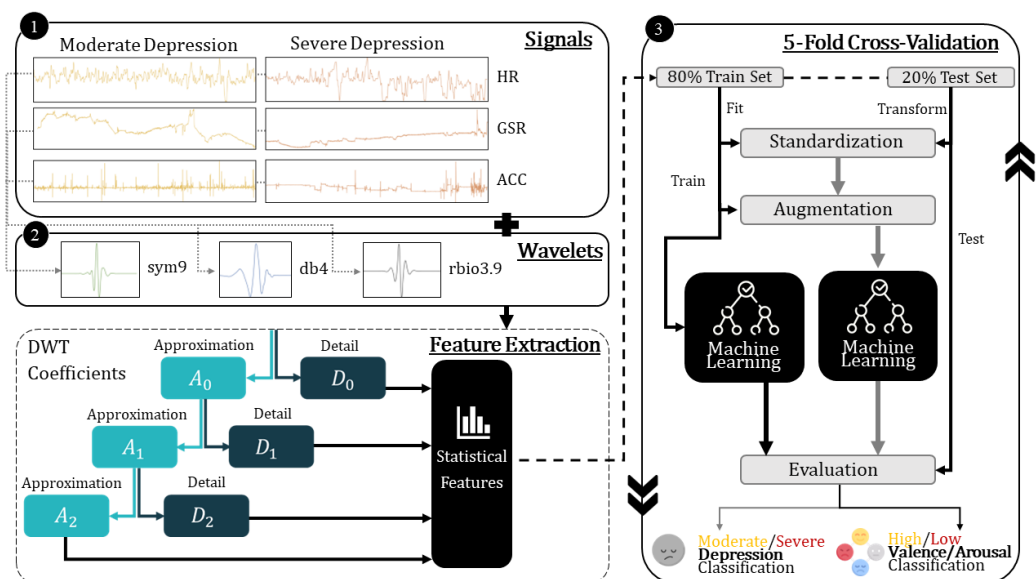


**Figure 2.** End-to-end model pipeline.

DWT can be considered as the projection of a sum vectored and zero mean signal  $S$  into a set of basis functions called wavelets  $\phi_{i,k}(n)$  and  $\psi_{i,k}(n)$ , which localize the time–frequency domain characteristics of the signal. These wavelets  $\Psi_{a,b}(t)$  are generated from a single-

base wavelet  $\Psi$  called the mother wavelet, through a series of dilations and translations of a scaling function as defined in.

$$\begin{aligned}\phi_{i,k}(n) &= 2^{-\frac{i}{2}} \phi(2^{-i}n - k) \\ \psi_{i,k}(n) &= 2^{-\frac{i}{2}} \psi(2^{-i}n - k)\end{aligned}\quad (1)$$

In Equation (1),  $n$  is the length of the signal  $S$ ,  $k$  is the discrete translations and  $2^i$  are the dyadic dilations.

More specifically, each series is a decomposed time series of coefficients describing the evolution the temporal component within a corresponding frequency band. This is performed using a pair of finite impulse response filters, taken as low-pass and high-pass respectively. With this, the DWT on a signal results in the approximation  $A_i(k)$  and detailed coefficients  $D_i(k)$  formulated by:

$$\begin{aligned}A_i(k) &= \sum_n S(n) \phi_{i,k}(n) \\ D_i(k) &= \sum_n S(n) \psi_{i,k}(n)\end{aligned}\quad (2)$$

In Equation (2), the approximation coefficients  $A_i(k)$  are the low-pass elements, and the detailed coefficients  $D_i(k)$  are high-pass elements for a signal  $S$ , at a decomposition level  $i$ . In the case of multi-level decomposition, numerous stages of decomposition occurs, beginning within the original signal. The former is analyzed further in the same manner to a certain decomposition level, while the latter is not. The decomposition level is determined by the mother wavelet, sampling frequency, and signal length and is generally treated as the optimal decomposition level.

After experimenting with candidate discrete mother wavelets across the families of coiflets, biorthogonal, daubechies, reverse biorthogonal, symlet and haar, the following are selected based on their high performances and accordance with prior physiological signal analysis in literature. For HR, we use the symlet-9 (sym9) which has the properties of near symmetry, orthogonality and biorthogonality. For GSR, we use the daubechies-4 (db4) which has the properties of asymmetry, orthogonality and biorthogonality. For ACC, we use the reverse biorthogonal-3.9 (rbio3.9) which has the properties of symmetry, no orthogonality and biorthogonality. Empirically, we find that the optimal levels of decomposition are 1, 3 and 2 for PPG, GSR and ACC signals, respectively.

The final features pertaining to the coefficients of the signals are Shannon Entropy, Zero Crossing Rate, Mean Crossing Rate, Mean, Standard Deviation, Median, Variance, Root Mean Square, 5th Percentile, 25th Percentile, 75th Percentile, and 95th Percentile values.

### 2.3. Data Augmentation

While the disparity between the majority and minority classes (0 and 1, no and yes, respectively) were not as extreme, augmentation was applied to observe any quantifiable improvement in overall model performance. Five techniques for augmentation were used, and their breakdown by type is as follows: oversampling: Synthetic Minority Oversampling Technique (SMOTE) and Adaptive Synthetic Sampling (ADASYN), oversampling + undersampling: SMOTE + Tomek links (SMOTEK), SMOTE + Edited Nearest Neighbours (SMOTEEN), and undersampling: Random Under sampling and Edited Nearest Neighbours (ENN).

- SMOTE is an oversampling technique that increases the number of minority samples in the dataset, by generating new samples from existing minority class samples. The approach generates new samples that are not duplicates, but convex combinations of two or more randomly chosen neighboring data samples in the feature space.
- ADASYN is an adaptive data generation method which creates synthetic samples to reduce class imbalances in a dataset. The approach uses weighted distribution for

different minority class samples as per relative difficulty in learning, and generates more samples similar to the harder-to-learn samples. This therefore reduces overall bias present in the dataset, and should improve learning performance of models trained on this data as well.

- SMOTEK and SMOTENN are hybrid techniques that consist of both undersampling and oversampling. Initially, SMOTE performs the over-sampling, then the resulting clusters that overlap on nearby points causing overfitting are removed using Tomek Links, or Nearest Neighbors, respectively, in the methods. The idea here is to clean distributions and lead to a distinct class separation.
- Random under sampling involves randomly discarding samples from the majority class until a balanced class distribution is attained.
- Condensed nearest neighbor undersampling involves the selection of prototypes from the training data, in order to essentially reduce the dataset size for instance-based classification. During this process, the the prototypical instances of the majority class are retained, while likely redundant instances are eliminated from the dataset.

#### 2.4. Machine Learning

We utilized a wide range of standard ML algorithms such as Logistic Regression (LR), Support Vector Classifier (SVC), K-Nearest Neighbors (KNN), Light Gradient Boosting Machines (LGB), and Random Forest (RF) and eXtreme Gradient Boosting (XGB), and CatBoost (CB) to cover both traditional and ensemble techniques. These supervised algorithms span a variety of strategies in learning complexity, and their implementations generally elicit good bias-variance trade-off when data is not particularly voluminous (few hundred to a few thousand unique samples) as shown in [17–19]. Stratified sampling was used to divided 80% training set and 20% testing set through a five-fold cross-validation approach. Random search yielded the final hyperparameters for all algorithms.

### 3. Results

This section presents the results for (i) depression severity classification, and (ii) valence-arousal detection for each depression category. The performance measures are quantified with the metrics of accuracy, sensitivity, specificity, F1-score and, where pertinent, the Area Under the ROC (receiver operating characteristic) Curve (AUC).

#### 3.1. Detecting Depression Severity

To evaluate the efficacy of the proposed multi-modal inference model, we test a variety of candidate machine learning models and tabulate their performance in Table 2. Despite most of the models scoring similarly (in terms of accuracy, sensitivity, specificity, F1-score, and AUC), we decided on using CatBoost (CB) Classifier in our trails for performing most reliably with an accuracy, sensitivity, and F1-score values of  $64 \pm 2.0$ ,  $68.7 \pm 3.0$ , and  $66.8 \pm 2.0$ , respectively. It was observed that data augmentation did not yield any statistically significant increase in performance with this scenario.

**Table 2.** Quantitative model performance metrics for the multimodal cases.

| Model | Accuracy       | Sensitivity    | Specificity    | F1-Score       | AUC            |
|-------|----------------|----------------|----------------|----------------|----------------|
| LGBM  | $64.0 \pm 3.0$ | $68.2 \pm 2.0$ | $59.4 \pm 5.0$ | $66.7 \pm 2.0$ | $63.8 \pm 3.0$ |
| RF    | $64.0 \pm 2.0$ | $67.3 \pm 2.0$ | $60.2 \pm 5.0$ | $66.4 \pm 2.0$ | $63.8 \pm 2.0$ |
| XGB   | $60.7 \pm 2.0$ | $64.8 \pm 2.0$ | $56.1 \pm 5.0$ | $63.5 \pm 1.0$ | $60.5 \pm 2.0$ |
| CB    | $64.0 \pm 2.0$ | $68.7 \pm 3.0$ | $58.8 \pm 4.0$ | $66.8 \pm 2.0$ | $63.7 \pm 2.0$ |
| KNN   | $58.2 \pm 2.0$ | $61.0 \pm 3.0$ | $55.0 \pm 4.0$ | $60.6 \pm 2.0$ | $58.0 \pm 2.0$ |
| SVC   | $60.2 \pm 2.0$ | $63.0 \pm 2.0$ | $57.1 \pm 5.0$ | $62.6 \pm 1.0$ | $60.1 \pm 2.0$ |
| LR    | $55.8 \pm 2.0$ | $60.2 \pm 1.0$ | $50.8 \pm 4.0$ | $59.1 \pm 1.0$ | $55.5 \pm 2.0$ |

In Table 3 we evaluate the performance of uni-modal inference and contrast it against bi- and tri- modal inference to evaluate the added value of multiple-modality and quantify its effectiveness in improving inference quality. Derived from the marginal inference quality increase (in comparison to the tri-modal model) based on the highlighted performance criteria, a holistic score of each permutation's performance is given in the last column of Table 3.

**Table 3.** Quantitative model performance metrics for exhaustive unimodal and multimodal cases with CB.

| Model           | Accuracy       | Sensitivity    | Specificity    | F1-Score       | AUC            | Delta $\delta_{avg}$ |
|-----------------|----------------|----------------|----------------|----------------|----------------|----------------------|
| PPG             | 54.0 $\pm$ 2.0 | 62.5 $\pm$ 3.0 | 44.5 $\pm$ 2.0 | 58.9 $\pm$ 2.0 | 53.5 $\pm$ 2.0 | +9.7                 |
| GSR             | 56.2 $\pm$ 1.0 | 70.7 $\pm$ 3.0 | 40.1 $\pm$ 1.0 | 63.0 $\pm$ 2.0 | 55.4 $\pm$ 1.0 | +6.8                 |
| ACC             | 59.0 $\pm$ 2.0 | 65.1 $\pm$ 6.0 | 52.3 $\pm$ 6.0 | 62.6 $\pm$ 3.0 | 58.7 $\pm$ 2.0 | +4.9                 |
| PPG + GSR       | 57.5 $\pm$ 2.0 | 67.9 $\pm$ 2.0 | 45.8 $\pm$ 5.0 | 62.8 $\pm$ 1.0 | 56.9 $\pm$ 2.0 | +6.2                 |
| PPG + ACC       | 62.1 $\pm$ 1.0 | 65.8 $\pm$ 4.0 | 57.9 $\pm$ 4.0 | 64.7 $\pm$ 2.0 | 61.9 $\pm$ 1.0 | +1.9                 |
| GSR + ACC       | 62.5 $\pm$ 1.0 | 68.9 $\pm$ 2.0 | 55.4 $\pm$ 1.0 | 66.0 $\pm$ 1.0 | 62.2 $\pm$ 1.0 | +1.4                 |
| PPG + GSR + ACC | 64.0 $\pm$ 2.0 | 68.7 $\pm$ 3.0 | 58.8 $\pm$ 4.0 | 66.8 $\pm$ 2.0 | 63.7 $\pm$ 2.0 | -                    |

For the DAPPER dataset, all possible single-modal models—using either PPG, GSR, or ACC, result in similarly performing models with an average accuracy of 56.4% and average F1-score of 61.5%. When observing combinations of any two modalities, however, we observe an average accuracy and F1-score values of both 60.7% and 64.5%, respectively. Finally, when all three modalities are combined, the resulting model has an accuracy of 64.0% and an F1-score of 66.8%.

Merely looking at the average performance for the learned models of varying number of modalities, we observe a performance increasing trend with each added modality. The move from uni- to bi-modal inference increases accuracy by 7.62%, and F1-score by 4.88%. Analogously, going from bi-modal to tri-modal inference improves the accuracy and F1-score by 5.44% and 3.57%, respectively. This upward trend validates the variability of the proposed multi-modal paradigm for depression severity inference.

### 3.2. Detecting Emotional Arousal and Valence

Much of performance measures across the metrics range from  $\sim 55$  to  $\sim 60$ , which belies that only an average level of separability can be obtained between high and low states. In contrast to the depression severity classification case, Tables 4 and 5 exhibit results where one modality fares better than multi-modality. To detect high- and low-valence/arousal states in a moderately depressed population, GSR and ACC proved feasible, with contribution from SMOTE augmentation. The specificity being higher than sensitivity indicates a better recall of the low-valence/arousal states. To detect high- and low-valence/arousal states in the severely depressed population, HR and GSR proved feasible, with contribution from SMOTE and SMOTEEN. The sensitivity for valence is better, while the specificity for arousal is marginally better, indicating that high-valence and low-arousal states are more discernible in the severely depressed population. However, there appears to be no common modality for each kind of emotional state activation, and the results appear to be characteristic of this dataset population with varying degrees of applicability to a general population.

**Table 4.** Quantitative performance metrics for the moderately depressed population.

| State   | Modality | Approach    | Accuracy | Sensitivity | Specificity | F1-Score |
|---------|----------|-------------|----------|-------------|-------------|----------|
| Valence | GSR      | SVC + SMOTE | 62.9     | 62.8        | 66.7        | 76.6     |
| Arousal | ACC      | LR + SMOTE  | 63.9     | 48.8        | 75.9        | 54.5     |

**Table 5.** Quantitative performance metrics for the severely depressed population.

| State   | Modality | Approach      | Accuracy | Sensitivity | Specificity | F1-Score |
|---------|----------|---------------|----------|-------------|-------------|----------|
| Valence | HR       | KNN + SMOTE   | 61.2     | 64.0        | 53.3        | 71.0     |
| Arousal | GSR      | SVC + SMOTEEN | 56.9     | 56.3        | 57.8        | 61.5     |

#### 4. Discussion

The goal was to separate behaviors of people suffering from depression based on emotional stimuli and possibly identifying affective emotional triggers in response to experienced events. As mentioned previously, there is a prevalence of depression among all surveyed individuals in this dataset who also experience higher-intensity emotional states. This warrants an investigation of the impact of depression in dampening/exacerbating emotional responses to various stimuli or events. As reported by [20], states of depression induce marked dysfunctional regulations of affective experience and affective quality perception. Generally, the neutral arousal/valence state is the most commonly experienced [21], because of the impaired emotional modulation to affective stimuli [22].

The achieved performance measures are in line with a recent study conducted with a similar demographic (race, age and education levels) asserting the potentially useful yet limited predictability of depression severity with wearable devices [23]. It appeared that greater severity of depressive symptoms showed associations with larger variation of night-time heart rate between the hours of 4 a.m. and 6 a.m. Additionally, this led to the findings (adjusted for covariates) that severity was also correlated robustly with weekday circadian activity rhythms. Thereby, our work focusing on individuals during conscious ESM activities serves as a complementary study to the general continuous diurnal and nocturnal biomarkers evaluated in [23].

According to [24], low-arousal states being associated with low-valence fine-grained states such as sadness, lethargy or fatigue is correlated with stronger levels of depression. This is because high arousal occurs when the cortical circuits in the brain are engaged and allocates attention in response to a particular stimuli [25]. However, our results do not agree with these findings, leading to our hypothesis that either the self-reported BDI-II scores were not reflective of the true underlying mental state, or that the self-reported valence and arousal is overly positive by choice of omission. A potential psychological link unifying our results in light of the previous studies is the theory of ambivalence over emotional expression. This is a condition wherein individuals have the propensity to avoid expression of emotions [26], owing to the effects of depression. In [27], it is rationalized that inciting high-arousal states in people with heightened depression thorough memory recalling tests increases help-seeking intentions. It could very well be that the stimuli or events experienced by the participants of the DAPPER dataset creation were involved in positive or familiar environments during the course of study.

Although the heterogeneous nature of different consumer-grade wearables occasionally leads to likely noise saturation, inaccurate values and uncalibrated errors [28], it is rationalized that the characteristic patterns associated with certain mental/physical diseases can indeed be reflected [29]. Additionally, it is not known if the individuals were on prescribed anti-depressants, engaged in psychological counselling or under any treatment which may introduce confounding variables that cannot be adjusted for.

While Deep Learning has demonstrated exemplary results in several fields and recent studies, there are still few limitations and challenges in the biomedical domain pertaining to class imbalance and data complexity, which discouraged its use in our work. As put forward in [30], Deep Learning models tend to capture spurious relations in the training data within clinical studies involving biomedical signals. This occurs, in context of this work, due to the implicit nature of some emotional states such as valence/arousal and depression, which do not manifest across all subjects with the same intensity or magnitude [31]. Thus, it appears that higher volumes of data readings from wearables are necessary to combat the relative sparsity of the wearable measurements, i.e., the ratio of the duration of *normal* physiological

behavior to the duration of context-specific instantaneous responses to certain stimuli, and achieve higher performance scores.

Ref. [32] also indicates that with skewed data, the models prioritize the majority group due to higher prior probability. Unlike with summarized measures such as wavelet decomposed features, augmenting continuous raw signals and the rectification of class imbalance requires more complex techniques to account for the profound understanding of the morphology and patterns [33]. We believe the lightweight approaches (DWT + standard ML) are relatively more conducive towards power-efficient deployment on wearable or edge devices, as shown in [34].

We envision this research as an initial baseline for performance benchmarking on the DAPPER dataset, as well as an assessment of the relationship between depression and ephemeral emotional states.

## 5. Conclusions

In this paper we presented multi-modal depression detection machine learning models while juxtaposing its performance to that of single and double modality models. More specifically, we leverage low-frequency HR, GSR and ACC signals belonging to 87 patients from the Daily Ambulatory Psychological and Physiological recording for Emotional Research (DAPPER) dataset to train multiple ML algorithms to classify between moderate and severe states, as defined by the categorized Beck's Depression Inventory (BDI-II) score. After listing the top performing depression detection models on our dataset's mobile/wearable device-based biomedical signals, we make a case for the introduction of additional modality, with respect to the DAPPER dataset.

In comparison to attempts present in the literature, using a variety of datasets for the detection and prediction of depression trends of individuals using mobile and power-constrained devices, there is a case to be made for the use of multiple-modality devices. One contemporary work [35], describes a smartphone activity monitoring model for the collection of medically relevant metrics and tracking of user activities for the detection of PHQ-9 levels, achieving detection performance metrics of 59.1–60% accuracy, 62.3–72% sensitivity and 47.3–60.8% specificity. In contrast, the metrics of our proposed approach are  $64 \pm 2.0\%$  accuracy,  $68.7 \pm 3.0\%$  sensitivity and  $58.8 \pm 4.0\%$  specificity, which consistently surpass the results in [35]. Therefore, we ultimately conclude that additional modality incorporation could potentially improve the detection accuracy of depression by accounting for different aspects of latent or conscious physiological responses to psychological events.

**Author Contributions:** Conceptualization, R.A., A.S. and F.A.; data curation, A.A. and S.G.; investigation, J.R., A.A. and S.G.; methodology, J.R.; project administration, R.A. and A.S.; resources, F.A. and A.S.; software, J.R.; supervision, R.A., A.S. and F.A.; validation, J.R., A.A. and S.G.; writing—original draft, J.R., A.A. and S.G.; writing—review and editing, R.A., A.S. and F.A. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** A consent has been obtained from the patient(s) to use this dataset for research purposes and publishing this paper.

**Data Availability Statement:** The dataset adopted in this research is openly available in [Synapse] at <https://doi.org/10.7303/syn22418021> (accessed on 20 April 2022).

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

|        |                                                                                   |
|--------|-----------------------------------------------------------------------------------|
| ACC    | Accelerometry                                                                     |
| BDI-II | Beck's Depression Inventory                                                       |
| CB     | CatBoost                                                                          |
| DAPPER | Daily Ambulatory Psychological and Physiological recording for Emotional Research |
| DWT    | Discrete Wavelet Transforms                                                       |
| GSR    | Galvanic Skin Response                                                            |
| HR     | Heart Rate                                                                        |
| KNN    | K-Nearest Neighbors                                                               |
| LGB    | Light Gradient Boosting                                                           |
| ML     | Machine Learning                                                                  |
| RF     | Random Forest                                                                     |
| SVC    | Support Vector Classifier                                                         |
| XGB    | eXtreme Gradient Boosting                                                         |

## References

- Harris, M.G.; Kazdin, A.E.; Chiu, W.T.; Sampson, N.A.; Aguilar-Gaxiola, S.; Al-Hamzawi, A.; Alonso, J.; Altwaijri, Y.; Andrade, L.H.; Cardoso, G.; et al. Findings from world mental health surveys of the perceived helpfulness of treatment for patients with major depressive disorder. *JAMA Psychiatry* **2020**, *77*, 830–841. [\[CrossRef\]](#) [\[PubMed\]](#)
- Daly, M.; Robinson, E. Depression and anxiety during COVID-19. *Lancet* **2022**, *399*, 518. [\[CrossRef\]](#)
- Zhang, F.; Gou, J. Machine Learning Assessment of Risk Factors for Depression in Later Adulthood. *Lancet Reg. Health Eur.* **2022**, *18*, 100399. [\[CrossRef\]](#)
- Tlachac, M.; Melican, V.; Reisch, M.; Rundensteiner, E. Mobile Depression Screening with Time Series of Text Logs and Call Logs. In Proceedings of the 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), Athens, Greece, 27–30 July 2021; pp. 1–4. [\[CrossRef\]](#)
- Zhang, T.; Schoene, A.M.; Ji, S.; Ananiadou, S. Natural Language Processing Applied to Mental Illness Detection: A Narrative Review. *NPJ Digit. Med.* **2022**, *5*, 46. [\[CrossRef\]](#) [\[PubMed\]](#)
- Choudhary, S.; Thomas, N.; Ellenberger, J.; Srinivasan, G.; Cohen, R. A Machine Learning Approach for Detecting Digital Behavioral Patterns of Depression Using Nonintrusive Smartphone Data (Complementary Path to Patient Health Questionnaire-9 Assessment): Prospective Observational Study. *JMIR Form. Res.* **2022**, *6*, e37736. [\[CrossRef\]](#) [\[PubMed\]](#)
- Chikersal, P.; Doryab, A.; Tumminia, M.; Villalba, D.K.; Dutcher, J.M.; Liu, X.; Cohen, S.; Creswell, K.G.; Mankoff, J.; Creswell, J.D.; et al. Detecting Depression and Predicting Its Onset Using Longitudinal Symptoms Captured by Passive Sensing: A Machine Learning Approach With Robust Feature Selection. *ACM Trans.-Comput.-Hum. Interact. (TOCHI)* **2021**, *28*, 1–41. [\[CrossRef\]](#)
- Jacobson, N.C.; Chung, Y.J. Passive Sensing of Prediction of Moment-To-Moment Depressed Mood among Undergraduates with Clinical Levels of Depression Sample Using Smartphones. *Sensors* **2020**, *20*, 3572. [\[CrossRef\]](#)
- Lee, S.; Kim, H.; Park, M.J.; Jeon, H.J. Current Advances in Wearable Devices and Their Sensors in Patients with Depression. *Front. Psychiatry* **2021**, *12*, 672347. [\[CrossRef\]](#)
- Long, Y.; Lin, Y.; Zhang, Z.; Jiang, R.; Wang, Z. Objective Assessment of Depression Using Multiple Physiological Signals. In Proceedings of the 2021 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Online, 23–25 October 2021; pp. 1–6. [\[CrossRef\]](#)
- Moshe, I.; Terhorst, Y.; Opoku Asare, K.; Sander, L.B.; Ferreira, D.; Baumeister, H.; Mohr, D.C.; Pulkki-Råback, L. Predicting Symptoms of Depression and Anxiety Using Smartphone and Wearable Data. *Front. Psychiatry* **2021**, *12*, 625247. [\[CrossRef\]](#)
- Xu, M.L.; De Boeck, P.; Strunk, D. An Affective Space View on Depression and Anxiety. *Int. J. Methods Psychiatr. Res.* **2018**, *27*, e1747. [\[CrossRef\]](#)
- Russell, J.A. Core Affect and the Psychological Construction of Emotion. *Psychol. Rev.* **2003**, *110*, 145–172. [\[CrossRef\]](#) [\[PubMed\]](#)
- Zitouni, M.S.; Park, C.Y.; Lee, U.; Hadjileontiadis, L.; Khandoker, A. Arousal-Valence Classification from Peripheral Physiological Signals Using Long Short-Term Memory Networks. In Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Mexico City, Mexico, 1–5 November 2021; pp. 686–689. [\[CrossRef\]](#)
- Xinyu, S.; Zhang, M.; Li, Z.; Hu, X.; Wang, F.; Zhang, D. A dataset of daily ambulatory psychological and physiological recording for emotion research. *Sci. Data* **2021**, *8*, 161. [\[CrossRef\]](#)
- Malay, S.; Chung, K.C. The Choice of Controls for Providing Validity and Evidence in Clinical Research. *Plast. Reconstr. Surg.* **2012**, *130*, 959–965. [\[CrossRef\]](#)
- Pradhan, A.; Prabhu, S.; Chadaga, K.; Sengupta, S.; Nath, G. Supervised Learning Models for the Preliminary Detection of COVID-19 in Patients Using Demographic and Epidemiological Parameters. *Information* **2022**, *13*, 330. [\[CrossRef\]](#)
- Ramesh, J.; Keeran, N.; Sagahyroon, A.; Aloul, F. Towards Validating the Effectiveness of Obstructive Sleep Apnea Classification from Electronic Health Records Using Machine Learning. *Healthcare* **2021**, *9*, 1450. [\[CrossRef\]](#)

19. Chen, D.; Liu, S.; Kingsbury, P.; Sohn, S.; Storlie, C.B.; Habermann, E.B.; Naessens, J.M.; Larson, D.W.; Liu, H. Deep Learning and Alternative Learning Strategies for Retrospective Real-World Clinical Data. *NPJ Digit. Med.* **2019**, *2*, 43. [[CrossRef](#)] [[PubMed](#)]
20. Laeger, I.; Dobel, C.; Dannlowski, U.; Kugel, H.; Grotegerd, D.; Kissler, J.; Keuper, K.; Eden, A.; Zwitterlood, P.; Zwanzger, P. Amygdala Responsiveness to Emotional Words Is Modulated by Subclinical Anxiety and Depression. *Behav. Brain Res.* **2012**, *233*, 508–516. [[CrossRef](#)]
21. Alskafi, F.A.; Khandoker, A.H.; Jelinek, H.F. A Comparative Study of Arousal and Valence Dimensional Variations for Emotion Recognition Using Peripheral Physiological Signals Acquired from Wearable Sensors. In Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Mexico City, Mexico, 1–5 November 2021; pp. 1104–1107. [[CrossRef](#)]
22. Teismann, H.; Kissler, J.; Berger, K. Investigating the Roles of Age, Sex, Depression, and Anxiety for Valence and Arousal Ratings of Words: A Population-Based Study. *BMC Psychol.* **2020**, *8*, 118. [[CrossRef](#)]
23. Rykov, Y.; Thach, T.Q.; Bojic, I.; Christopoulos, G.; Car, J. Digital Biomarkers for Depression Screening with Wearable Devices: Cross-sectional Study with Machine Learning Modeling. *JMIR mHealth uHealth* **2021**, *9*, e24872. [[CrossRef](#)] [[PubMed](#)]
24. Zhang, L.; Fan, H.; Wang, S.; Li, H. The Effect of Emotional Arousal on Inhibition of Return Among Youth with Depressive Tendency. *Front. Psychol.* **2019**, *10*, 1487. [[CrossRef](#)]
25. Moratti, S. Low Emotional Arousal in Depression as Explained by the Motivated Attention Approach. *Escritos-Psicol.-Psychol. Writ.* **2012**, *5*, 20–26. [[CrossRef](#)]
26. Brockmeyer, T.; Grosse Holtforth, M.; Krieger, T.; Altenstein, D.; Doerig, N.; Friederich, H.C.; Bents, H. Ambivalence over Emotional Expression in Major Depression. *Personal. Individ. Differ.* **2013**, *54*, 862–864. [[CrossRef](#)]
27. Straszewski, T.; Siegel, J.T. Differential Effects of High- and Low-Arousal Positive Emotions on Help-Seeking for Depression. *Appl. Psychol. Health Well-Being* **2020**, *12*, 887–906. [[CrossRef](#)] [[PubMed](#)]
28. Hickey, B.A.; Chalmers, T.; Newton, P.; Lin, C.T.; Sibbritt, D.; McLachlan, C.S.; Clifton-Bligh, R.; Morley, J.; Lal, S. Smart Devices and Wearable Technologies to Detect and Monitor Mental Health Conditions and Stress: A Systematic Review. *Sensors* **2021**, *21*, 3461. [[CrossRef](#)] [[PubMed](#)]
29. De Angel, V.; Lewis, S.; White, K.; Oetzmann, C.; Leightley, D.; Oprea, E.; Lavelle, G.; Matcham, F.; Pace, A.; Mohr, D.C.; et al. Digital Health Tools for the Passive Monitoring of Depression: A Systematic Review of Methods. *NPJ Digit. Med.* **2022**, *5*, 1–14. [[CrossRef](#)] [[PubMed](#)]
30. Dinsdale, N.K.; Bluemke, E.; Sundaresan, V.; Jenkinson, M.; Smith, S.; Namburete, A.I. Challenges for machine learning in clinical translation of big data imaging studies. *arXiv* **2021**, arXiv:2107.05630. [[CrossRef](#)]
31. Christensen, M.C.; Wong, C.M.J.; Baune, B.T. Symptoms of Major Depressive Disorder and Their Impact on Psychosocial Functioning in the Different Phases of the Disease: Do the Perspectives of Patients and Healthcare Providers Differ? *Front. Psychiatry* **2020**, *11*, 280. [[CrossRef](#)]
32. Johnson, J.M.; Khoshgoftaar, T.M. Survey on deep learning with class imbalance. *J. Big Data* **2019**, *6*, 27. [[CrossRef](#)]
33. Zhang, K.; Xu, G.; Han, Z.; Ma, K.; Zheng, X.; Chen, L.; Duan, N.; Zhang, S. Data Augmentation for Motor Imagery Signal Classification Based on a Hybrid Neural Network. *Sensors* **2020**, *20*, 4485. [[CrossRef](#)]
34. Pope, G.C.; Halter, R.J. Design and Implementation of an Ultra-Low Resource Electrodermal Activity Sensor for Wearable Applications. *Sensors* **2019**, *19*, 2450. [[CrossRef](#)]
35. Wahle, F.; Kowatsch, T.; Fleisch, E.; Rufer, M.; Weidt, S. Mobile Sensing and Support for People with Depression: A Pilot Trial in the Wild. *JMIR mHealth uHealth* **2016**, *4*, e111. [[CrossRef](#)] [[PubMed](#)]