technologies

*Article*

# HADD: High-Accuracy Detection of Depressed Mood

**Yu Liu [1], Kyoung-Don Kang [1,*] and Mi Jin Doe [2]**

[1] Department of Computer Science, State University of New York at Binghamton, 4400 Vestal Parkway East, Vestal, NY 13850, USA
[2] Tompkins County Mental Health Services, 201 E Green St., Ithaca, NY 14850, USA
* Correspondence: kang@binghamton.edu

**Abstract:** Depression is a serious mood disorder that is under-recognized and under-treated. Recent advances in mobile/wearable technology and ML (machine learning) have provided opportunities to detect the depressed moods of participants in their daily lives with their consent. To support high-accuracy, ubiquitous detection of depressed mood, we propose HADD, which provides new capabilities. First, HADD supports multimodal data analysis in order to enhance the accuracy of ubiquitous depressed mood detection by analyzing not only objective sensor data, but also subjective EMA (ecological momentary assessment) data collected by using mobile devices. In addition, HADD improves upon the accuracy of state-of-the-art ML algorithms for depressed mood detection via effective feature selection, data augmentation, and two-stage outlier detection. In our evaluation, HADD significantly enhanced the accuracy of a comprehensive set of ML models for depressed mood detection.

**Keywords:** depressed mood detection; multimodal data analysis; ecological momentary assessment; passive sensing; machine learning



**Citation:** Liu, Y.; Kang, K.-D.; Doe, M.J. HADD: High-Accuracy Detection of Depressed Mood. *Technologies* **2022**, *10*, 123. https://doi.org/10.3390/technologies10060123

Academic Editor: Mario Munoz-Organero

Received: 15 October 2022
Accepted: 23 November 2022
Published: 29 November 2022



## 1. Introduction

Depression is a serious mental health problem. According to the World Health Organization (WHO) [1], approximately 5% of adults in the world suffer from depression, and it is a leading cause of disability. In the U.S., over 21 million adults experience one or more major depressive episode each year [2], and the estimated annual economic burden of major depressive disorder is over USD 210 billion [3].

Providing timely recognition and treatment is vital for the mitigation of the development and escalation of depression [4,5]; however, depression is under-recognized and under-treated [6]. Many depressed people do not seek professional mental health services for several reasons, such as stigma, cost, and accessibility [7,8]. Typically, they seek mental health services after their symptoms have become severe [5]. The COVID-19 pandemic has aggravated the problem because of the increased isolation and loneliness and the decreased access to mental health services [9].

Mental health providers (MHPs) use a standard set of screening instruments for depression, such as PHQ-9 (Patient Health Questionnaire-9) [10] and BDI-II (Beck Depression Inventory-II) [11], whose effectiveness has been clinically proven. However, mobile questionnaires that use mobile/wearable devices in a non-clinical setting often suffer from low compliance and sparse responses [12]. In addition, the sensor data available in mobile/wearable devices are not leveraged to detect depressed mood. Thus, questionnaires alone may not be the most effective approach to high-accuracy ubiquitous detection of depressed mood in daily life.

The advances in ML and mobile/wearable technology, such as smartphones, smartwatches, and wristbands, have enabled the ubiquitous detection of depressed mood via passive sensing in daily life. An increasing body of research estimates the presence of depressed mood by analyzing objective sensor data that are passively collected using smart

devices, such as patterns of sleep, activity, location data, vital signs, facial expressions, vocal prosody, phone usage, and social activity [13–23]. With user consent, devices can collect these data automatically without requiring users to manually enter the data. Thus, passive sensing is less obtrusive and burdensome to users, enabling ubiquitous sensing and analysis of users' depressive status. However, it is challenging to detect depressed mood with high accuracy by only analyzing passive sensing data, without associating them with subjective features, such as screening questionnaires or EMA results, which directly represent individual users' mood/symptoms. In addition, some sensors, such as an $\alpha$-amylase sensor that is correlated with depression [20], are not usually available in commodity mobile/wearable devices, such as smartphones or smartwatches. Other sensors in such devices, such as an ECG (electrocardiogram) or EEG (electroencephalogram) [19,24], may not be as accurate as medical equipment.

Based on these observations, in this paper, we propose a new multimodal data analysis methodology called HADD in order to support high-accuracy, ubiquitous detection of depressed mood by leveraging ML and mobile/wearable data. HADD analyzes mobile EMA data and passive sensing data to generate synergy and enhance the accuracy of depressed mood detection; passive sensing supports unobtrusive, continuous sensing during daily life, and the EMA can capture the mood of a user more directly in a potentially more engaging manner than that of standard questionnaires, such as the PHQ-9. For example, when prompted by the device, the user can simply select an image that best expresses their current mood from a set of provided images or answer a short multiple-choice question about mood. In addition, different images and questions can be provided a few times per day to better engage users and analyze their moods over time.

Furthermore, HADD proposes effective approaches to the support of high-accuracy detection of depressed mood, even when the available dataset is small and imbalanced. This is important because one of the main challenges for digital mHealth (mobile health) is the small data problem, where datasets could be too small to train ML models [25]. A summary of our key contributions is as follows:

- HADD is new in that it supports multimodal analysis of mobile data consisting of subjective EMA data and objective sensor data to enhance the accuracy of the detection of depressed mood. Despite the potential synergy that can be created by analyzing multimodal data, research on detecting depressed mood by analyzing EMA and passive sensing data is in an early stage, and related work is relatively scarce [26–30].
- To support high accuracy in depressed mood detection even when the given/available dataset is small and imbalanced, HADD enhances state-of-the-art ML models as follows: (1) HADD supports effective data augmentation equipped with a self-validity check; (2) it selects key features from a comprehensive set of multimodal mobile/wearable data to support high-accuracy detection of depressed mood with minimal model complexity; (3) HADD supports a new two-stage detection method that detects depressed mood in two stages in order to detect anomalies (depressed mood) with a high accuracy even when the available dataset is small and imbalanced. In the first stage, a state-of-the-art supervised ML model is used to detect/classify depressed mood in users on a daily basis. In the second stage, we analyze the statistical distribution of the daily detection results collected for an extended period of time, such as a month, to detect outliers—users with a depressed mood. By doing this, HADD aims to avoid misclassifying transient mood changes in healthy users as indicating a depressed mood and to further enhance the accuracy of depressed mood detection. Notably, our proposed techniques are not tied to a specific ML model, but are generally applicable to enhance the accuracy of ML models for depressed mood detection.
- HADD is a new general machine learning framework for depressed mood detection and is not tied to a specific machine learning model, unlike the examples in the related work discussed in Section 2. It can support a wide variety of machine learning models and, therefore, enables a mental healthcare provider (MHP) to choose a model that

is most appropriate for their job. In this paper, we use 12 machine learning models and compare their performance by evaluating each model in HADD, one by one. An MHP can choose any of the models based on, for example, the detection accuracy or explainability. Furthermore, HADD is extensible: An MHP or a data scientist can easily add new models if necessary.

- In this paper, we thoroughly evaluate the performance of HADD by using the StudentLife dataset [12]. It is a relatively small dataset that consists of many subjective and objective features that were collected using smartphones. HADD enhances the accuracy of depressed mood detection by 17% on average in comparison with the accuracy of 12 state-of-the-art ML models that all used subjective/objective features without feature selection and did not exploit the data augmentation or two-stage outlier detection supported by HADD.

The rest of this paper is organized as follows. Related work is discussed in Section 2. In Section 3, our approaches to boosted detection/classification of depressed mood are described. In Section 4, the detection performance of HADD is evaluated in comparison with the state-of-the-art baselines. Finally, we conclude the paper and discuss future work in Section 5.

## 2. Related Work

Previous studies have investigated potential correlations between depression and multimodal data, such as subjective EMA/questionnaire and objective sensor data. Beyond merely analyzing their statistical correlations, researchers have also applied various ML algorithms in order to detect depressed mood. In particular, we review digital health techniques that aim to analyze correlations between depression and mobile data or detect depressed mood by using mobile data. In Table 1, we summarize the depressed mood detection techniques that are closely related to HADD, and we discuss them in the following.

**Table 1.** Recent works on ML-based detection of depressed mood that are most closely related to HADD.

| Work | Dataset | Performance | Sensing | EMA | Questionnaire |
|---|---|---|---|---|---|
| [31] | iYouVU | accuracy: 0.76 | ✗ | ✓ | ✗ |
| [32] | iYouVU [31] | RMSE: 0.53 | ✗ | ✓ | ✗ |
| [33] | OSPP [34] | accuracy: 0.96 | ✗ | ✗ | ✓ |
| [35] | proprietary | AUC: 0.886 | ✗ | ✓ | ✓ |
| [17] | proprietary | accuracy: 0.857 | ✓ | ✗ | ✗ |
| [19] | proprietary | accuracy: 0.7698 | ✓ | ✗ | ✗ |
| [22] | proprietary | accuracy: 0.825, F1-score: 0.855 | ✓ | ✗ | ✗ |
| [36] | proprietary | accuracy: 0.601 | ✓ | ✗ | ✗ |
| [37] | proprietary | recall: 0.89, F1-score: 0.86 | ✓ | ✗ | ✗ |
| [38] | StudentLife [12] | recall: 0.815, precision: 0.691, AUC: 0.809 | ✓ | ✗ | ✗ |
| [39] | Depresjon [40] | accuracy: 0.893, AUC: 0.893 | ✓ | ✗ | ✗ |
| [41] | proprietary | RMSE: 4.88 | ✓ | ✗ | ✗ |
| [26] | NESDA [42], MOOVD [20] | AUC: 0.993 for the development set [42] and 0.892 for the validation set [20] | ✓ | ✓ | ✗ |
| [27] | proprietary | accuracy: 0.91, AUC: 0.96 | ✓ | ✓ | ✗ |
| [28] | proprietary | recall: 0.96 | ✓ | ✓ | ✗ |
| [29] | iYouVU [31] | MSE: 0.41 | ✓ | ✓ | ✗ |
| [30] | proprietary | RMSE: 4.5 | ✓ | ✓ | ✗ |
| HADD (this paper) | StudentLife [12] | accuracy: 0.98, precision: 0.92, recall: 1, F1-score: 0.95 | ✓ | ✓ | ✗ |

### *2.1. EMA/Questionnaires*

In [43–47], the authors developed apps for collecting EMA data and studied the associations between symptom features and severity scales. Recent research, such as that of [31–33,35], which are shown in Table 1, took a step further to not only analyze correlations between EMA features and depression, but also to detect depressed mood by applying ML algorithms.

### *2.2. Passive Sensing*

In [13–23], unobtrusive passive sensing was undertaken by using smart devices to analyze patterns of sleep, activity, location data, vital signs, facial expressions, vocal prosody, phone usage, social activity, etc. More specifically, in [13,14,18,20,48–50], correlations between sensor data that were collected using mobile devices and depressed mood were analyzed. Via passive sensing and ML, the authors of [17,19,22,36–39] (shown in Table 1) detected the presence of depressed mood via classification, whereas the authors of [41] estimated the severity of depressive symptoms via regression.

### *2.3. EMA and Passive Sensing*

In [5,51], statistical correlations between depression and EMA + sensor data were analyzed. In [52], an effective technique was proposed to capture both passive sensing and EMA data. Predicting depression by analyzing both mobile EMA and passive sensing data with ML models is a new research direction with relatively little work, such as that of [26–30], which is summarized in Table 1. Among them, the authors of [26–28] detected depressed mood via classification in a way similar to that of HADD. On the other hand, the authors of [29,30] predicted the severity of depressive symptoms via regression. Thus, HADD is most closely related to the work of [26–28] and is complementary to the work of [29,30], since they also performed classification to detect depressed mood. The performance of HADD in terms of depressed mood detection (Section 4) is considerably higher than that reported in [26–28], as summarized in Table 1. HADD achieves a higher performance by supporting an effective multimodal data analysis that leverages effective feature selection, data augmentation, and two-stage outlier detection. (It is nearly impossible for us to reproduce the results of [26–28] in Section 4, since their source code/datasets are not publicly available.) Moreover, the datasets used in [26] were collected in clinical settings, which is impractical for the ubiquitous detection of depressed mood, and they were much more balanced than the StudentLife dataset [12] used in this paper (discussed in more detail in Section 4). Similarly, the dataset used in [27] is much more balanced than [12].

In summary, an increasing amount of work has recently been done to detect depressed mood by using various types of mobile data and ML; however, related work on multimodal data analysis for depressed mood detection in daily life is relatively scarce. HADD is new in that it supports multimodal (subjective and objective) data analysis, effective feature selection, data augmentation, and two-stage prediction to significantly enhance the accuracy of depressed mood detection, even when the available dataset is small and imbalanced. Furthermore, HADD considers a much more comprehensive set of ML models (i.e., 12 ML models) than most works on depressed mood detection did (including the ones discussed in this section).

## 3. HADD

In this section, an overview of HADD is given, followed by a detailed description of the approaches used by HADD to detect depressed mood with high accuracy.

### *3.1. An Overview of the Method*

Figure 1 depicts the high-level structure of HADD, which consists of the following components.

- Data preparation: In this paper, we divided the input data into 24 h slices to enable daily detection of depressed mood, as illustrated in Figure 1. In this paper, we consid-

ered that a user had a depressed mood if their PHQ-9 score was at least 10 (moderately depressed [10]); however, our approach can be used with different depression screening instruments, such as BDI-II. A more detailed discussion of the data preprocessing is given in Section 4.

- Feature selection: As shown in Figure 1, we selected a subset of the features that were closely correlated with depression to support high accuracy with minimal model complexity and computational overhead.
- Data augmentation: To support high accuracy even when the available dataset is small, we augmented the training set, if necessary, to enhance the accuracy by mitigating the imbalance between the samples in depressed and non-depressed moods, as depicted in Figure 1.
- Two-stage detection of depressed mood: To further enhance the prediction performance, we also propose a new method for the two-stage detection of depressed mood. In the first stage, HADD runs a trained ML model against each participant's daily data to estimate their depressed mood over an extended period of time, such as several weeks or months. In the second phase, HADD analyzes the statistical distribution to detect outliers, i.e., users with depressed mood, as illustrated in Figure 1.

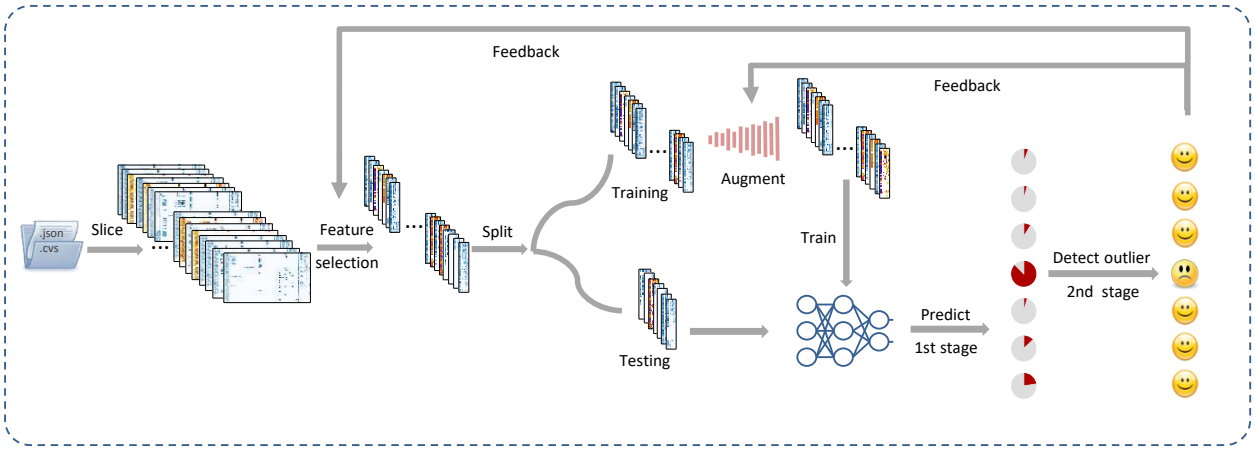


**Figure 1.** An overview of HADD.

A more detailed discussion of HADD, which aims to detect depressed mood with high accuracy through a multimodal data analysis, follows.

### *3.2. Feature Selection*

The goal of feature selection is to extract a smaller yet more effective or useful set of features [53]. By reducing the dimensionality, it is possible to reduce the complexity of ML models, enhance their interpretability, and possibly improve their accuracy by removing hardly correlated features that could be noisy or even misleading. Moreover, it is worth noting that feature selection can reduce the risk of overfitting by using fewer features while decreasing the overhead of collecting/analyzing data. For example, feature selection can pick a compact and effective set of features, e.g., mood, activity, and sleep patterns, which are collected using smartphones and smartwatches. In this paper, we select important subjective and objective features for effective screening of depressive symptoms. First, we calculate the Pearson correlation coefficient for each feature, which is a measure of the linear correlation between a feature and the target variable (e.g., depressed or non-depressed mood):

$$r = \frac{\sum_{i=1}^{n}(\nu_i - \hat{\nu})(\xi_i - \hat{\xi})}{\sqrt{\sum_{i=1}^{n}(\nu_i - \hat{\nu})^2}\sqrt{\sum_{i=1}^{n}(\xi_i - \hat{\xi})^2}} \quad (1)$$

where $n$ is the sample size, $(\nu_i, \xi_i)$ is a pair of an individual feature and the corresponding label, and $\hat{\nu}$ and $\hat{\xi}$ are the sample means of $\nu$ and $\xi$, respectively. Equation (1) returns the correlation coefficient, $r$, which ranges between $-1$ and 1, which indicate strong negative and positive correlations, respectively. For the $r$ value, we also calculate the $p$-value, which represents the confidence of $r$ [54]. In this paper, we use the significance level $\alpha = 0.05$.

In addition, we use an efficient heuristic search to extract key features, since feature selection is an NP-hard problem [53]. If there are $m$ features in total, we first select $m' \leq m$ features with the highest $r$ values, subject to the condition that their $p$ values do not exceed $\alpha$.

From the $m'$ features, we select $m'' \leq m'$ features using the wrapper method. In particular, we apply the Plus-L-Minus-R (PLMR) search method, which adds L features and eliminates R features in each round, instead of using the exhaustive wrapper method, which evaluates every subset of the $m'$ features with the $O(2^{m'})$ complexity. By doing this, we select effective EMA and passive sensing features for high-accuracy detection of depressive mood.

### *3.3. Augmenting Training Data*

A key challenge for mHealth is the small amount of data [25], as discussed before. In addition, depression datasets are typically imbalanced, since a small fraction of the population (e.g., 5% of the world population [1]) is depressed, and depressed symptoms are generally under-recognized [6]. Due to the insufficient information about depressed samples and the lower probability of them being presented to ML models, it is hard to learn patterns of depressed samples.

To increase the accuracy of depressed mood detection and reduce the possibility of overfitting, HADD performs data augmentation. Notably, we only augment the training set; we do not augment the test set. To properly evaluate the performance in depressed mood detection, in this paper, we use an unmodified test set that is unseen during the model training.

Sampling/data augmentation strategies can be categorized as oversampling [55,56] and undersampling [57] methods, which increase the minority class samples and reduce the majority class samples, respectively. In this paper, since both positively and negatively labeled samples are precious, we apply the oversampling strategy to increase the minority (depressed) samples in the training set while avoiding the loss of majority samples.

For data augmentation, we perform *random oversampling with replacement*; we randomly select a depressed user in the training set and duplicate their samples—daily data slices in this paper—and add them to the training set, if the addition of the duplicates does not largely distort the statistical distribution (discussed shortly). In this paper, the number of times that data augmentation is performed is called the *data augmentation factor*.

To simultaneously mitigate data imbalance and avoid excessive oversampling, we carefully perform oversampling step by step as follows:

1. Initialize the augmentation factor $\omega = 1$.
2. Perform data augmentation using $\omega$.
3. Compute the statistical distribution and check the validity.
4. If the validity check is successful, save the augmented dataset and increment $\omega$ by 1. Go to Step 2.
5. Otherwise, abort the augmentation and revert to the dataset generated in the previous iteration. Revert to the original dataset if the validity check fails when $\omega = 1$. Stop data augmentation and exit.

To analyze the validity of each augmentation, let us define a random variable $x$ as the ratio of the days on which certain participants are classified as being in a depressed mood to the total number of days for a depressed mood study. In addition, $\mu$ and $\sigma$ represent the

mean and the standard deviation of $x$, respectively. Given these, we estimate the depressed mood density as follows:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \quad (2)$$

Generally, the distribution of the probability density of depressed mood is likely to be skewed with a long right tail (as illustrated in Section 4) because, globally, about 95% of adults do not have a depressed mood disorder. However, the remaining 5% of the population is in a depressed mood with widely varying severity [1]. Using Equation (2), we compute and plot the density distribution to visually check if the heavy-tailed curve becomes bell-shaped after the augmentation due to excessive oversampling. In such a case, the specific augmentation is aborted and oversampling terminates, as described in Step 5.

### *3.4. Two-Stage Detection of Depressed Mood*

#### 3.4.1. Stage 1

In Stage 1, a trained ML model classifies each user as feeling depressed or not every day by analyzing passive sensing and EMA data. If it estimates that the client $i$ is in a depressed mood for $n_{i,d}$ days out of $n_i$ days for which the user participated in the study, the observed likelihood of this client being in a depressed mood is:

$$x_i = \frac{n_{i,d}}{n_i} \quad (3)$$

Let $n_{min}$ denote the minimum number of the days (e.g., 30 days) for which a user needs to participate in the study. In this paper, we require $n_i \geq n_{min}$ for reliable detection of depressed mood instead of transitory mood changes. If $n_i < n_{min}$, we do not compute Equation (3) for the user $i$ until sufficient samples are collected for them.

In this paper, we consider 12 popular supervised models for Stage 1, which are summarized as follows.

- Logistic Regression (LR) is one of the most popular classification algorithms. It uses the sigmoid function as the cost function:

$$h_\theta(\nu) = \frac{1}{1 + e^{-\theta^T \nu}} \quad (4)$$

  where $\nu$ is the set of the features and $\theta$ represents the model parameters. In the training phase, it minimizes the cost function and learns $\theta$ by using the gradient descent method. It then uses the trained model to classify new data unseen during the training.
- A Support Vector Machine (SVM) finds a hyperplane that has the maximum margin against the support vectors in order to divide training samples into the classes. The SVM supports effective classification if data are linearly separable. Otherwise, its performance may not be significantly different from that provided by LR.
- k-Nearest Neighbors (kNN) computes the distance from a new test data to the stored training data. The $k$-nearest data then determine the class of the new data via voting.
- A Decision Tree (DT) mimics the human decision process by constructing a tree in which each non-leaf node splits based on a feature. Thus, it is human-interpretable if it is short. A deeper tree tends to be harder to interpret and subject to overfitting.
- The Gradient-Boosting Decision Tree (GBDT), Ada Boost (AB), and Random Forest (RF) trade the interpretability of DT for performance by leveraging ensemble learning. In the GBDT and AB, each DT model learns from the error of the previous model. In contrast, the RF applies the bagging method. The DT models in an RF use randomly selected features and learn in parallel. The trained models in the RF classify new data via the majority vote.

- Gaussian Naive Bayes (GNB) is a variant of Naive Bayes that uses the Bayes theorem under the naive assumption that features are independent. GNB assumes that the density of each class is normally distributed.
- Linear Discriminant Analysis (LDA) is simple yet powerful. It uses a linear discriminant function based on the Bayesian and maximum likelihood rules to determine the region (class) to which data belong under the assumption that all classes share the same covariance matrix.
- Quadratic Discriminant Analysis (QDA) uses a quadratic discriminant function. In addition, it does not make the assumption of a shared covariance matrix across classes, unlike LDA. In general, it is more flexible than LDA, but is more susceptible to overfitting than LDA is. For more details on LDA and QDA, interested readers are referred to [58].
- A Deep Neural Network (DNN) is effective for ML due to its capability of nonlinear approximation. In this paper, we designed a DNN that consisted of four fully connected layers with 64, 32, 16, and 2 neurons to classify participants as depressed or non-depressed.
- A Convolutional Neural Network (CNN) is powerful in many ML tasks, e.g., computer vision. In this paper, we designed a CNN that consisted of a convolution layer and three fully connected layers with 64, 32, 16, and 2 neurons, respectively. For convolution, the kernel size was 2 and stride was 1. Following convolution, the CNN performed max pooling.

#### 3.4.2. Stage 2

This stage supports *unsupervised, enhanced classification* in order to further enhance the detection performance by analyzing the output of the first stage. In general, the distribution of the probability density of depressed mood could be skewed with a long right tail, since about 5% of adults in the world suffer from depression and the severity varies widely among different individuals [1]. Thus, we aim to detect outliers, i.e., users in a depressed mood, with characteristics distinguishing them from non-depressed ones.

In this paper, we use the t-test to detect outliers (i.e., users in a depressed mood) because it is a well-established statistical method that is effective for anomaly detection, even when the sample size is small [59]. It also does not require separate training, unlike supervised ML methods. Statistically, an anomaly occurs in the low-probability region of a stochastic model. In our case, it occurs when a participant $i$ holds an abnormally higher likelihood of being in a depressed mood than others. Specifically, HADD performs a t-test to detect users in a depressed mood among $n$ total users by taking the following steps:

1. HADD builds the vector $\boldsymbol{x} = (x_1, \ldots, x_n)^T$, where the random variable $x_i$ is defined in Equation 3 and collected for each user in Stage 1.
2. It computes the mean, $\mu$, and standard deviation, $\sigma$, of the vector $\boldsymbol{x}$.
3. Finally, HADD classifies the client $i$ as class 1 (being in a depressed mood) if the following condition holds:

$$x_i > \mu + t_{\gamma,\xi} \frac{\sigma}{\sqrt{n}} \tag{5}$$

where $t_{\gamma,\xi}$ is looked up in the t-table using the confidence level $\gamma$ and the degrees of freedom $\xi (= n - 1)$ [60].

## 4. Evaluation

In this section, we evaluate the performance of HADD in comparison with several state-of-the-art baselines.

### *4.1. Data Preparation*

#### 4.1.1. Dataset

In this paper, we used the StudentLife dataset [12]. The dataset consists of mobile data collected from 48 students who participated in a 10-week study at Dartmouth College by

using smartphones. The dataset includes passive sensing data, EMA data, and survey responses (self-reports to the PHQ-9 survey). In addition, other features, e.g., the grade-point average and class deadlines, were collected to analyze their possible correlations with the students' mental health and behavioral trends.

We used the StudentLife dataset [12] because: (1) It is a popular feature-rich dataset that was collected using smartphones; (2) the dataset is relatively *small* and *imbalanced*—it has only eight students who had at least a moderately depressed mood. Thus, we could evaluate whether HADD could support high accuracy even when the mobile/wearable dataset was small and imbalanced while enhancing the accuracy in comparison with that of ML models that were not extended by HADD.

#### 4.1.2. Data Preprocessing

Before performing any evaluations, we preprocessed the dataset for effective detection. We sliced the data on a daily basis, as discussed in Section 3.1. For data cleaning, we removed highly sparse EMA/self-report features. For example, we deleted the feature called "Dartmouth now empathic", which only attained 28 responses from all 48 participants during the entire 10-week period. Furthermore, we scaled the remaining features using the min–max normalization method to analyze the features with different ranges of values on a common ground.

In the StudentLife dataset [12], PHQ-9 screening was administered at the beginning and end of the study. In this paper, a student was labeled as being in a depressed mood if they had a PHQ-9 score $\geq 10$ (moderately depressed [10]) both at the beginning and end of the 10-week study. On the contrary, the students whose pre-PHQ-9 and post-PHQ-9 score were <10 were labeled as non-depressed. If a student's pre-PHQ-9 score was $\geq$10 but their post-PHQ-9 score was <10 (or vice versa), the student was removed because it was highly uncertain whether they were in a non-transitory depressed mood. In practice, such users should be monitored for a longer period of time for reliable detection of depressed mood.

After all preprocessing, we obtained a dataset of 2405 records of 34 users with 78 features, where seven participants were labeled as being in a depressed mood.

#### 4.1.3. Data Augmentation

In this paper, we empirically chose the data augmentation factor, $\omega$ (defined in Section 3), to mitigate the data imbalance. Specifically, we chose $\omega = 1$ because it increased the peak density at $x = 0.12$ in comparison with the peak of $\omega = 0$, as well as that at $x = 0.12$, without considerably shifting the entire curve to the right, as shown in Figure 2. In contrast, for $\omega > 1$, the peak became lower and moved to the right, as shown in Figure 2. As a result, the entire density curves in the figure tend to become more bell-shaped for $\omega > 1$.

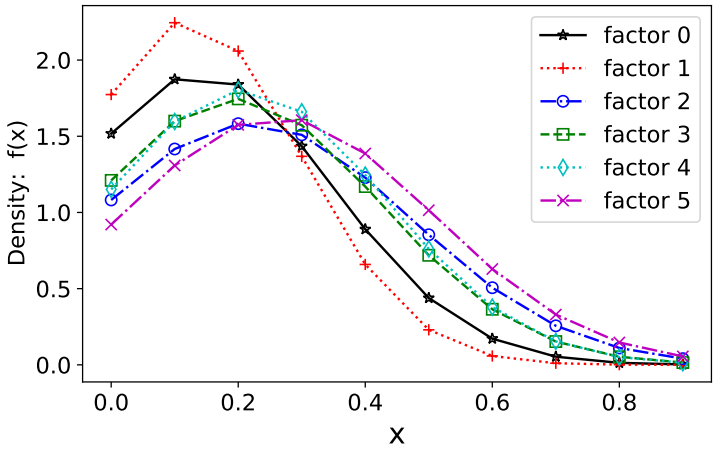


**Figure 2.** Density distribution computed using Equation (2).

#### 4.1.4. Selecting Objective and Subjective Features

As discussed in Section 3.2, we used the Pearson correlation coefficient (PCC) and wrapper method to determine the key features that could cost-efficiently detect depression with high accuracy and low overhead via dimensionality reduction. First, we selected the 24 features (features 1–24 in Table 2) with relatively large absolute $r$ values and $p \leq 0.05$. As shown in Table 2, the $r$ and $p$ values of the features were generally small. Although there were correlations between these features and depression, their correlations were rather weak. Thus, we performed two-stage prediction to enhance the prediction performance. Furthermore, we applied the PLMR method, as described in Section 3. Specifically, we set L = 2 and R = 1. By doing this, we selected the 10 boldface features in Table 2, which consisted of four passive sensing features and six EMA features, as described in the following.

**Table 2.** Pearson correlation coefficients.

| Item | Feature | $r$ | $p$-Value |
|---|---|---|---|
| 1 | Dark_duration | 0.117010485 | $7.33954 \times 10^{-12}$ |
| 2 | Conversation_duration | −0.10050095 | $4.12932 \times 10^{-9}$ |
| 3 | PAM_picture_idx | 0.086854646 | $3.82635 \times 10^{-7}$ |
| 4 | Audio_inference | −0.082883614 | $1.26971 \times 10^{-6}$ |
| 5 | Mood_happyornot | 0.081726937 | $1.78281 \times 10^{-6}$ |
| 6 | Sleep_social | 0.080808105 | $2.32707 \times 10^{-6}$ |
| 7 | Mood_sad | 0.075941272 | $9.10363 \times 10^{-6}$ |
| 8 | Phonecharge_duration | 0.070690366 | $3.63055 \times 10^{-5}$ |
| 9 | Phonelock_duration | 0.063472895 | 0.000209513 |
| 10 | Class_experience | 0.055680536 | 0.001148481 |
| 11 | Sleep_rate | 0.050727791 | 0.003058695 |
| 12 | Administration response_apathetic | 0.049628777 | 0.003761137 |
| 13 | Activity_inference | −0.047794521 | 0.005265698 |
| 14 | Boston Bombing_boston | 0.047554704 | 0.005498195 |
| 15 | Dimensions protestors_appreciative | 0.045303212 | 0.008176002 |
| 16 | Dimensions protestors_proud | 0.044667622 | 0.009118687 |
| 17 | Mood_tomorrow | 0.04460978 | 0.009209115 |
| 18 | Stress_level | 0.044346094 | 0.00963156 |
| 19 | Exercise_exercise | −0.04327199 | 0.01153625 |
| 20 | Dimensions protestors_empathic | 0.042928071 | 0.01221311 |
| 21 | Behavior_sympathetic | 0.042189111 | 0.013787576 |
| 22 | Dining Halls_dinner | -0.03940632 | 0.021438194 |
| 23 | Class_hours | 0.036251256 | 0.034354408 |
| 24 | Administration response_appreciative | 0.036075483 | 0.035237284 |
| 25 | Lab_duration | 0.0310954 | 0.06955524 |
| 26 | Exercise_schedule | −0.029959267 | 0.080384652 |
| ... | ... | ... | ... |
| 76 | Dartmouth now_saddened | −0.001650002 | 0.92330252 |
| 77 | Behavior_critical | −0.00147211 | 0.931549995 |
| 78 | Dartmouth now_frustrated | $3.92106 \times 10^{-5}$ | 0.99817455 |

**Passive sensing features:**

1. Dark duration indicates the length of the time period for which the screen remains dark for more than one hour.
2. Audio inference continuously detected audio signals using the microphone (0: silence, 1: voice, 2: noise, 3: unknown).
3. Phone charge duration records the time spent charging the smartphone.
4. Activity was continuously analyzed using the accelerometer (0: stationary, 1: walking, 2: running, 3: unknown).

**EMA features:**

1. PAM_pictur_idx: A participant selected a picture that best represented their mood from a grid of 16 pictures provided randomly from a library of 48 photos designed for a photographic affect meter (PAM) [61].
2. mood_happy_or_not question: "Do you feel at all happy right now?" (1: yes, 2: no).

3. sleep_social question: "How often did you have trouble staying awake yesterday while in class, eating meals, or engaging in social activity?" (1: none, 2: once, 3: twice, 4: three or more times).
4. mood_sad question: "How sad do you feel?" (1: a little bit, 2: somewhat, 3: very much, 4: extremely).
5. sleep_rate question: "How would rate your overall sleep last night?" (1: very good, 2: fairly good, 3: fairly bad, 4: very bad).
6. stress_level question: "Stress level?" (1: a little stressed, 2: definitely stressed, 3: stressed out, 4: feeling good, 5: feeling great).

In this paper, we randomly split 80% and 20% of the dataset into the training and test sets, respectively. We trained the ML models and evaluated their detection performance by using the data in the test set, which was unseen during the training. We used scikit-learn [62] to evaluate the 10 ML models described in Section 3.4, while we used Keras [63] for the DNN and CNN. To calculate the $r$ and $p$ values of the Pearson correlation coefficient, we used scipy [64].

### *4.2. Tuning the Model Parameters*

To enhance the accuracy of depressed mood detection and reduce possible overfitting, we used feature selection and data augmentation, as described in Sections 3.2, 3.3, 4.1.3, and 4.1.4. In addition, we tuned the model parameters carefully to further minimize the depressed mood detection error and overfitting as follows:

- LR: We set L-BFGS (Limited-Memory Broyden–Fletcher–Goldfarb–Shanno) algorithm [65] as the optimizer to tune the logistic regression parameters in a gradient-descent manner with an L2 norm regularization penalty, which is known as ridge regression, to reduce overfitting.
- SVM: We used the radial basis function (RBF) kernel to support the nonlinear classification. To avoid possible overfitting, we set the generalization penalty parameter to C = 1.0 and the kernel coefficient to $\gamma = \frac{1}{N_f \times var(X)}$, where $N_f$ is the number of features and $var(X)$ is the variance of the training data.
- kNN: In kNN, the number of neighbors $k$ affects the performance. We used the Euclidean distance to measure the error. We set $k = 5$ to minimize the error by trial and error.
- DT: We used the Gini impurity as the classification criterion for splitting the nodes to build the decision tree by using the training data.
- AB, GBDT, and RF: These ensemble models are designed to be more robust to noise than a simple decision tree by aggregating multiple weak learners. We used a decision tree classifier as the weak learner/base estimator for all three of these ensemble algorithms. We set the numbers of trees to 50, 100, and 10 for AB, GBDT, and RF, respectively. A more detailed description of the process for tuning AB, GBDT, and RF follows.
  - For AB, we set the learning rate to 1.0, and used the boosting algorithm of SAMME.R. AB built the model by adding up the weights of each tree, which were learned and updated via iterations.
  - For GBDT, we set the learning rate to 0.1, the maximum depth of a tree to 3, and the minimum sample leaf to 1. GBDT calculated the residual from the previous tree and aggregated all of them as the whole model.
  - To configure RF, we set the square root of the total number of the features as the maximum number of features for each individual tree, and we set the minimum sample leaf to 1. RF randomly selected subsets of features from the training set to build the decision trees and make the final decision with the majority vote.
- GNB: The GNB model applied the Bayes theorem in order to calculate the conditional probability of each label under every selected feature of the training data.

- LDA and QDA: LDA and QDA both used the Bayes rule to fit the class-conditional probability densities to the training data and build the decision boundary. LDA assumed that there was a common covariance of both classes, while QDA considered a separate covariance for each class.
- DNN and CNN: To train both the DNN and CNN, we randomly picked one-third of the training set as a validation set and performed cross-validation to tune the model parameters carefully, thus avoiding overfitting. We used the stochastic gradient descent algorithm to tune the parameters by using the mean squared error (MSE) as the loss function. We also used the Adam optimizer [66]. Specifically, the batch size was 32. We had a total of 200 epochs with a learning rate of 0.001.

### *4.3. Evaluation Results*

In this section, we evaluate the performance of HADD in comparison with that of several state-of-the-art baselines. In particular, we assess if HADD can considerably enhance the detection accuracy of ML models even when the available dataset, e.g., the StudentLife dataset, is small and imbalanced. First, we evaluate the effectiveness of the multimodal data analysis using passive sensing and EMA data in terms of the detection of a depressive mood. After that, we evaluate the overall accuracy improvement achieved by HADD in comparison with a comprehensive set of popular ML models without feature selection, data augmentation, or two-stage detection.

#### 4.3.1. Effectiveness of the Multimodal Approach in Detecting Depressed Mood

Most of the related work discussed in Section 2 is not open source. In addition, many of these studies used proprietary datasets to which we do not have access. As a result, we cannot reproduce them and directly compare them with HADD. Thus, in this subsection, we compare the performance of the multimodal analysis method of HADD, which analyzes the 10 key objective and subjective features selected in the previous subsection, with that of the following baselines (B1–B3) that represent the state of the art:

1. **B1 (all features)**: This baseline simply uses all 78 features in the StudentLife dataset to predict depression.
2. **B2 (sensing only)**: This baseline only uses the four selected passive sensing features, similarly to [12,17,19,22,36,37].
3. **B3 (EMA only)**: This baseline only uses the six selected EMA features, similarly to [31,35].

As shown in Table 3a, the mean accuracy of the baseline B1 is only 0.75; detecting depressed mood using all 78 features is not a good design choice, since certain features with weak correlations may even disrupt the identification of relevant patterns.

As shown in Table 3b, the baseline B2 provided the lowest accuracy. Thus, B2's approach of only using sensor data was the least effective.

As shown in Table 3c, the baseline B3 outperformed B1 and B2, and it had an average accuracy of 0.8 because the EMA features directly reflected participants' status, e.g., mood and stress levels.

Passive sensing is essential for the recognition of depressed mood in daily life, and it is less obtrusive and burdensome than questionnaires. From the results of B2 and B3, however, we observed that the sensor data were weaker indicators of depression than self-reports, e.g., the EMA or a questionnaire.

By selecting key subjective and objective features and analyzing them, HADD considerably enhanced the detection performance; as shown in Table 3d, HADD's mean accuracy was higher than the average accuracies of B1–B3, which are shown in Table 3a–c, by 8–28% (16.3% on average). It also outperformed B1–B3 in terms of the other performance metrics, as summarized in the table.

**Table 3.** Effectiveness of the multimodal approach of HADD based on feature selection. By analyzing key sensor and EMA features, HADD enhanced the accuracy by 8–28% in comparison with the state-of-the-art baselines.

| Model | Non-Depression | | | Depression | | | Accuracy |
|---|---|---|---|---|---|---|---|
| | Precision | Recall | F1-Score | Precision | Recall | F1-Score | |
| (**a**) B1 (All Features): Average accuracy = 0.75 | | | | | | | |
| LR | 1.00 | 0.83 | 0.91 | 0.50 | 1.00 | 0.67 | 0.86 |
| SVM | 0.00 | 0.00 | 0.00 | 0.14 | 1.00 | 0.25 | 0.14 |
| kNN | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| DT | 1.00 | 0.83 | 0.91 | 0.50 | 1.00 | 0.67 | 0.86 |
| GBDT | 1.00 | 0.83 | 0.91 | 0.50 | 1.00 | 0.67 | 0.86 |
| AB | 1.00 | 0.83 | 0.91 | 0.50 | 1.00 | 0.67 | 0.86 |
| RF | 0.80 | 0.67 | 0.73 | 0.00 | 0.00 | 0.00 | 0.57 |
| GNB | 0.80 | 0.67 | 0.73 | 0.00 | 0.00 | 0.00 | 0.57 |
| LDA | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| QDA | 0.83 | 0.83 | 0.83 | 0.00 | 0.00 | 0.00 | 0.71 |
| DNN | 1.00 | 0.83 | 0.91 | 0.50 | 1.00 | 0.67 | 0.86 |
| CNN | 0.83 | 0.83 | 0.83 | 0.00 | 0.00 | 0.00 | 0.71 |
| *mean* | *0.86* | *0.76* | *0.81* | *0.39* | *0.67* | *0.47* | *0.75* |
| (**b**) B2 (Sensing Only): Average accuracy = 0.60 | | | | | | | |
| LR | 0.00 | 0.00 | 0.00 | 0.14 | 1.00 | 0.25 | 0.14 |
| SVM | 0.00 | 0.00 | 0.00 | 0.14 | 1.00 | 0.25 | 0.14 |
| kNN | 1.00 | 0.83 | 0.91 | 0.50 | 1.00 | 0.67 | 0.86 |
| DT | 1.00 | 0.67 | 0.80 | 0.33 | 1.00 | 0.50 | 0.71 |
| GBDT | 1.00 | 0.83 | 0.91 | 0.50 | 1.00 | 0.67 | 0.86 |
| AB | 1.00 | 0.83 | 0.91 | 0.50 | 1.00 | 0.67 | 0.86 |
| RF | 0.80 | 0.67 | 0.73 | 0.00 | 0.00 | 0.00 | 0.57 |
| GNB | 0.83 | 0.83 | 0.83 | 0.00 | 0.00 | 0.00 | 0.71 |
| LDA | 0.00 | 0.00 | 0.00 | 0.14 | 1.00 | 0.25 | 0.14 |
| QDA | 1.00 | 0.83 | 0.91 | 0.50 | 1.00 | 0.67 | 0.86 |
| DNN | 0.75 | 0.50 | 0.60 | 0.00 | 0.00 | 0.00 | 0.43 |
| CNN | 1.00 | 0.83 | 0.91 | 0.50 | 1.00 | 0.67 | 0.86 |
| *mean* | *0.70* | *0.57* | *0.63* | *0.27* | *0.75* | *0.38* | *0.60* |
| (**c**) B3 (EMA Only): Average accuracy = 0.80 | | | | | | | |
| LR | 0.00 | 0.00 | 0.00 | 0.14 | 1.00 | 0.25 | 0.14 |
| SVM | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| kNN | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| DT | 1.00 | 0.83 | 0.91 | 0.50 | 1.00 | 0.67 | 0.86 |
| GBDT | 1.00 | 0.83 | 0.91 | 0.50 | 1.00 | 0.67 | 0.86 |
| AB | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| GNB | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| RF | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| LDA | 0.00 | 0.00 | 0.00 | 0.14 | 1.00 | 0.25 | 0.14 |
| QDA | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| DNN | 1.00 | 0.83 | 0.91 | 0.50 | 1.00 | 0.67 | 0.86 |
| CNN | 0.83 | 0.83 | 0.83 | 0.00 | 0.00 | 0.00 | 0.71 |
| *mean* | *0.82* | *0.78* | *0.80* | *0.65* | *0.92* | *0.71* | *0.80* |
| (**d**) HADD (EMA+Sensing): Average accuracy = 0.88 | | | | | | | |
| LR | 0.00 | 0.00 | 0.00 | 0.14 | 1.00 | 0.25 | 0.14 |
| SVM | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| kNN | 1.00 | 0.83 | 0.91 | 0.50 | 1.00 | 0.67 | 0.86 |
| DT | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| GBDT | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| AB | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| RF | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| GNB | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| LDA | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| QDA | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| DNN | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| CNN | 0.80 | 0.67 | 0.73 | 0.00 | 0.00 | 0.00 | 0.57 |
| *mean* | *0.90* | *0.86* | *0.88* | *0.76* | *0.92* | *0.80* | *0.88* |

#### 4.3.2. Overall Accuracy Improvement by HADD

In this section, HADD is fully optimized via feature selection, data augmentation, and two-stage detection of depressed mood. Its overall performance is compared to that of

vanilla state-of-the-art ML models that use all features and do not apply data augmentation or two-stage outlier detection.

Figure 3 depicts the accuracy of HADD and vanilla ML models that use all features without feature selection. They also do not apply our data augmentation or two-stage detection methods. HADD improved the accuracy by up to 57% (GNB) and 17%, on average, in comparison with the other models. (In Figure 3, the CNN is omitted because we were not able to train it properly without the feature selection, data augmentation, and two-stage detection supported by HADD due to the small size of the StudentLife dataset and the data imbalance therein).

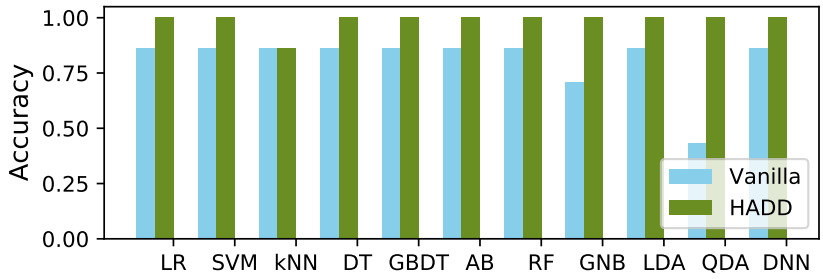


**Figure 3.** Accuracy of HADD and the vanilla state-of-the-art ML methods that do not exploit the feature selection, data augmentation, or two-stage detection techniques supported by HADD. Instead, they used all features, performed no data augmentation, and detected depressed mood in a single stage by using trained ML models

.

The average accuracy of the 12 ML models supported by HADD (Table 4) was higher than that reported in recent work on depressed mood classification (Table 1) by 2–37%. Among these related studies, that of Kumar et al. [33] showed the highest accuracy of 0.96 by using questionnaires consisting of 42 and 21 questions. This was mainly because questionnaires are effective in screening depression, the authors' (proprietary) dataset was nearly complete and had few missing data, and they handpicked and optimized features based on their correlations with depression. In the context of mHealth, however, frequent (e.g., several times per day) administration of a survey with lengthy questionnaires in daily life can be burdensome and intrusive to participants. In addition, manual feature selection may not be as scalable as an automated wrapper method, such as the PLMR method used in our approach, when there are many features in a dataset, such as the StudentLife dataset, due to a potential combinatorial explosion.

In comparison with related work that used both EMA and sensor data for depressed mood classification [26–28], as shown in Table 1, most of the 12 ML models extended by HADD showed a higher accuracy, precision, recall, and F1-score than those from [26–28]. In addition, our AUC (area under the curve) was close to 1, except for kNN and CNN. kNN showed a relatively low performance. It made classification decisions via a majority vote, which could be swayed considerably by data imbalances. Therefore, it was suboptimal for depressed mood classification, where the distribution was heavily tailed. The CNN achieved a relatively low performance due to the small and imbalanced dataset, which was potentially insufficient for training a sophisticated deep learning model, such as a CNN. Its performance could be improved further if more data would become available to train the model. A thorough investigation is reserved for future work.

**Table 4.** After extending the 12 ML models by applying all of the proposed optimizations (multimodal analysis, feature selection, data augmentation, and two-stage detection), HADD showed an average accuracy of 0.98.

| Model | Non-Depression | | | Depression | | | Accuracy |
|---|---|---|---|---|---|---|---|
| | Precision | Recall | F1-Score | Precision | Recall | F1-Score | |
| LR | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| SVM | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| kNN | 1.00 | 0.83 | 0.91 | 0.50 | 1.00 | 0.67 | 0.86 |
| DT | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| GBDT | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| AB | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| RF | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| GNB | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| LDA | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| QDA | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| DNN | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| CNN | 1.00 | 0.83 | 0.91 | 0.50 | 1.00 | 0.67 | 0.86 |
| *mean* | *1.00* | *0.97* | *0.99* | *0.92* | *1.00* | *0.95* | *0.98* |

## 5. Conclusions and Future Work

Depression is a serious mood disorder that is under-recognized and under-treated. Recent advances in mobile/wearable technology and machine learning have provided opportunities for detecting depressed moods in participants in their daily lives. For ubiquitous detection of depressed mood with high accuracy, even when only a small/imbalanced dataset is available, we propose a new framework called HADD. In particular, HADD analyzes not only objective sensor data, but also subjective data, such as EMA (ecological momentary assessment) data, in order to enhance the accuracy of ubiquitous depressed mood detection. In addition, HADD enhances the accuracy of state-of-the-art machine learning algorithms that are widely used for depressed mood detection via several optimizations: effective feature selection, data augmentation, and two-stage classification for outlier detection. Notably, these methods are not tied to specific machine learning methods, but are generally applicable to various models for depressed mood detection, as demonstrated in this paper. In our evaluation, HADD enhanced the accuracy of the 12 baseline machine learning models by 17%, on average, via feature selection, data augmentation, and two-stage detection. In the future, we will apply HADD to more datasets and investigate more advanced approaches to detecting depressed moods. We will also investigate if HADD is applicable to other mood disorders, such as anxiety.

**Author Contributions:** Y.L. designed the machine learning models and HADD and undertook the performance evaluation. K.-D.K. advised Y.L. in formulating the research problem investigated in this paper. He also helped Y.L. in designing and analyzing the proposed approach and in writing the paper while coordinating the project by communicating with the other authors. M.J.D. helped Y.L. and K.-D.K. in understanding mood disorders and screening instruments, designing the high-level concept of the multimodal data analysis, and writing/revising the paper from the perspective of mental health. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Science Foundation: CNS-2007854.

**Data Availability Statement:** The StudentLife dataset used in this paper is openly available in [12]. Our source code is available at: https://github.com/Real-Time-Lab/Depressed-Mood-Detection (accessed on 28 November 2022).

**Conflicts of Interest:** The authors declare no conflict of interest. The funding sponsor had no role in the design of the study, in the collection, analyses, or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

## References


1. World Health Organization. Depression. Available online: https://www.who.int/news-room/fact-sheets/detail/depression (accessed on 27 September 2022).
2. National Alliance on Mental Health. Mental Health By the Numbers. Available online: https://www.nami.org/mhstats (accessed on 27 September 2022).
3. American Psychiatric Association Foundation. Quantifying the Cost of Depression. Available online: https://www.workplacementalhealth.org/Mental-Health-Topics/Depression/Quantifying-the-Cost-of-Depression (accessed on 27 September 2022).
4. Asare, K.O.; Visuri, A.; Ferriera, D.S. Towards early detection of depression through smartphone sensing. In Proceedings of the UbiComp/ISWC 2019-Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers, London, UK, 9–13 September 2019.
5. Moshe, I.; Terhorst, Y.; Asare, K.O.; 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] [PubMed]
6. National Network of Depression Centers. Get the Facts. Available online: https://nndc.org/facts/ (accessed on 27 September 2022).
7. Marshall, J.M.; Dunstan, D.A.; Bartik, W. The Digital Psychiatrist: In Search of Evidence-Based Apps for Anxiety and Depression. *Front. Psychiatry* **2019**, *10*, 831. [CrossRef] [PubMed]
8. Bardram, J.E.; Matic, A. A Decade of Ubiquitous Computing Research in Mental Health. *IEEE Pervasive Comput.* **2020**, *19*, 62–72. [CrossRef]
9. Vahratian, A.; Blumberg, S.J.; Terlizzi, E.P.; Schiller, J.S. Symptoms of Anxiety or Depressive Disorder and Use of Mental Health Care among Adults during the COVID-19 Pandemic—United States, August 2020–February 2021. *MMWR Morb Mortal Wkly Rep.* **2021**, *70*, 490–494. [CrossRef] [PubMed]
10. Kroenke, K.; Spitzer, R.L.; Williams, J.B. The PHQ-9: Validity of a brief depression severity measure. *J. Gen. Intern. Med.* **2001**, *16*, 606–613. [CrossRef] [PubMed]
11. Beck, A.T.; Steer, R.A.; Brown, G.K. *Beck Depression Inventory-II*; Psychological Corporation: San Antonio, TX, USA, 1996
12. Wang, R.; Chen, F.; Chen, Z.; Li, T.; Harari, G.; Tignor, S.; Zhou, X.; Ben-Zeev, D.; Campbell, A.T. StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. In Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, Seattle, WA, USA, 13–17 September 2014; pp. 3–14.
13. O'Brien, J.T.; Gallagher, P.; Stow, D.; Hammerla, N.; Ploetz, T.; Firbank, M.; Ladha, C.; Ladha, K.; Jackson, D.; McNaney, R.; et al. A study of wrist-worn activity measurement as a potential real-world biomarker for late-life depression. *Psychol. Med.* **2017**, *47*, 93–102. [CrossRef] [PubMed]
14. da Estrela, C.; McGrath, J.; Booij, L.; Gouin, J.P. Heart rate variability, sleep quality, and depression in the context of chronic stress. *Ann. Behav. Med.* **2021**, *55*, 155–164. [CrossRef] [PubMed]
15. Shatte, A.B.; Hutchinson, D.M.; Teague, S.J. Machine learning in mental health: A scoping review of methods and applications. *Psychol. Med.* **2019**, *49*, 1426–1448. [CrossRef]
16. Thieme, A.; Belgrave, D.; Doherty, G. Machine learning in mental health: A systematic review of the HCI literature to support the development of effective and implementable ML systems. *ACM Trans. Comput. Hum. Interact. (TOCHI)* **2020**, *27*, 1–53. [CrossRef]
17. Chikersal, P.; Doryab, A.; Tumminia, M.J.; 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. Interactation* **2021**, *28*, 1–41. [CrossRef]
18. 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]
19. Cai, H.; Han, J.; Chen, Y.; Sha, X.; Wang, Z.; Hu, B.; Yang, J.; Feng, L.; Ding, Z.; Chen, Y.; et al. A pervasive approach to EEG-based depression detection. *Complexity* **2018**, *2018*, 5238028. [CrossRef]
20. Booij, S.H.; Bos, E.H.; Bouwmans, M.E.; van Faassen, M.; Kema, I.P.; Oldehinkel, A.J.; de Jonge, P. Cortisol and $\alpha$-amylase secretion patterns between and within depressed and non-depressed individuals. *PLoS ONE* **2015**, *10*, e0131002. [CrossRef] [PubMed]
21. Sadeque, F.; Xu, D.; Bethard, S. Measuring the latency of depression detection in social media. In Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, Los Angeles, CA, USA, 5–9 February 2018; pp. 495–503.
22. Xu, X.; Chikersal, P.; Dutcher, J.M.; Sefidgar, Y.S.; Seo, W.; Tumminia, M.J.; Villalba, D.K.; Cohen, S.; Creswell, K.G.; Creswell, J.D.; et al. Leveraging Collaborative-Filtering for Personalized Behavior Modeling: A Case Study of Depression Detection among College Students. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2021**, *5*, 1–27. [CrossRef]
23. Nasir, M.; Jati, A.; Shivakumar, P.G.; Nallan Chakravarthula, S.; Georgiou, P. Multimodal and multiresolution depression detection from speech and facial landmark features. In Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, Amsterdam, The Netherlands, 16 October 2016; pp. 43–50.
24. Shen, J.; Zhao, S.; Yao, Y.; Wang, Y.; Feng, L. A novel depression detection method based on pervasive EEG and EEG splitting criterion. In Proceedings of the 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Kansas City, MO, USA, 13–16 November 2017; pp. 1879–1886.
25. Estrin, D. A survey on image data augmentation for deep learning. *Commun. ACM* **2014**, *57*, 32–34. [CrossRef]

26. Minaeva, O.; Riese, H.; Lamers, F.; Antypa, N.; Wichers, M.; Booij, S.H. Screening for depression in daily life: Development and external validation of a prediction model based on actigraphy and experience sampling method. *J. Med. Internet. Res.* **2020**, *22*, e22634. [CrossRef] [PubMed]
27. Kim, H.; Lee, S.H.; Lee, S.E.; Hong, S.; Kang, H.J.; Kim, N. Depression prediction by using ecological momentary assessment, actiwatch data, and machine learning: Observational study on older adults living alone. *JMIR mHealth uHealth* **2019**, *7*, e14149. [CrossRef]
28. Narziev, N.; Goh, H.; Toshnazarov, K.; Lee, S.A.; Chung, K.M.; Noh, Y. STDD: Short-term depression detection with passive sensing. *Sensors* **2020**, *20*, 1396. [CrossRef] [PubMed]
29. Van Breda, W.; Pastor, J.; Hoogendoorn, M.; Ruwaard, J.; Asselbergs, J.; Riper, H. Exploring and comparing machine learning approaches for predicting mood over time. In Proceedings of the International Conference on Innovation in Medicine and Healthcarem, Puerto de la Cruz, Spain, 15–17 June 2016; pp. 37–47.
30. Ghandeharioun, A.; Fedor, S.; Sangermano, L.; Ionescu, D.; Alpert, J.; Dale, C.; Sontag, D.; Picard, R. Objective assessment of depressive symptoms with machine learning and wearable sensors data. In Proceedings of the 2017 Seventh International Conference On Affective Computing And Intelligent Interaction (ACII), San Antonio, TX, USA, 23–26 October 2017, pp. 325–332.
31. Asselbergs, J.; Ruwaard, J.; Ejdys, M.; Schrader, N.; Sijbrandij, M.; Riper, H. Mobile phone-based unobtrusive ecological momentary assessment of day-to-day mood: An explorative study. *J. Med. Internet. Res.* **2016**, *18*, e5505. [CrossRef] [PubMed]
32. Becker, D.; Bremer, V.; Funk, B.; Asselbergs, J.; Riper, H.; Ruwaard, J. How to predict mood? delving into features of smartphone-based data. In Proceedings of the 22nd Americas Conference on Information Systems, San Diego, CA, USA, 11–14 August 2016.
33. Kumar, P.; Garg, S.; Garg, A. Assessment of anxiety, depression and stress using machine learning models. *Procedia Comput. Sci.* **2020**, *171*, 1989–1998. [CrossRef]
34. Open-Source Psychometrics Project. Available online: https://openpsychometrics.org/_rawdata/ (accessed on 27 September 2022).
35. Suhara, Y.; Xu, Y.; Pentland, A. Deepmood: Forecasting depressed mood based on self-reported histories via recurrent neural networks. In Proceedings of the 26th International Conference on World Wide Web, Perth, Australia, 3–7 April 2017; pp. 715–724.
36. 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*, e5960. [CrossRef] [PubMed]
37. Farhan, A.A.; Yue, C.; Morillo, R.; Ware, S.; Lu, J.; Bi, J.; Kamath, J.; Russell, A.; Bamis, A.; Wang, B. Behavior vs. introspection: refining prediction of clinical depression via smartphone sensing data. In Proceedings of the 2016 IEEE Wireless Health (WH), Bethesda, MD, USA, 25–27 October 2016; pp. 1–8.
38. Wang, R.; Wang, W.; DaSilva, A.; Huckins, J.F.; Kelley, W.M.; Heatherton, T.F.; Campbell, A.T. Tracking depression dynamics in college students using mobile phone and wearable sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2018**, *2*, 1–26. [CrossRef]
39. Zanella-Calzada, L.A.; Galván-Tejada, C.E.; Chávez-Lamas, N.M.; Gracia-Cortés, M.; Magallanes-Quintanar, R.; Celaya-Padilla, J.M.; Galván-Tejada, J.I.; Gamboa-Rosales, H. Feature extraction in motor activity signal: Towards a depression episodes detection in unipolar and bipolar patients. *Diagnostics* **2019**, *9*, 8. [CrossRef] [PubMed]
40. Garcia-Ceja, E.; Riegler, M.; Jakobsen, P.; Tørresen, J.; Nordgreen, T.; Oedegaard, K.J.; Fasmer, O.B. Depresjon: A Motor Activity Database of Depression Episodes in Unipolar and Bipolar Patients. In Proceedings of the 9th ACM on Multimedia Systems Conference, Amsterdam, The Netherlands, 12–15 June 2018; pp. 474–477.
41. Pedrelli, P.; Fedor, S.; Ghandeharioun, A.; Howe, E.; Ionescu, D.F.; Bhathena, D.; Fisher, L.B.; Cusin, C.; Nyer, M.; Yeung, A.; et al. Monitoring changes in depression severity using wearable and mobile sensors. *Front. Psychiatry* **2020**, 1413. [CrossRef]
42. Penninx, B.W.; Beekman, A.T.; Smit, J.H.; Zitman, F.G.; Nolen, W.A.; Spinhoven, P.; Cuijpers, P.; De Jong, P.J.; Van Marwijk, H.W.; Assendelft, W.J.; et al. The Netherlands Study of Depression and Anxiety (NESDA): Rationale, objectives and methods. *Int. J. Methods Psychiatr. Res.* **2008**, *17*, 121–140. [CrossRef] [PubMed]
43. Torous, J.; Staples, P.; Shanahan, M.; Lin, C.; Peck, P.; Keshavan, M.; Onnela, J.P. Utilizing a personal smartphone custom app to assess the patient health questionnaire-9 (PHQ-9) depressive symptoms in patients with major depressive disorder. *JMIR Ment. Health* **2015**, *2*, e3889. [CrossRef]
44. Hung, S.; Li, M.S.; Chen, Y.L.; Chiang, J.H.; Chen, Y.Y.; Hung, G.C.L. Smartphone-based ecological momentary assessment for Chinese patients with depression: An exploratory study in Taiwan. *Asian J. Psychiatry* **2016**, *23*, 131–136. [CrossRef] [PubMed]
45. Targum, S.D.; Sauder, C.; Evans, M.; Saber, J.N.; Harvey, P.D. Ecological momentary assessment as a measurement tool in depression trials. *J. Psychiatr. Res.* **2021**, *136*, 256–264. [CrossRef] [PubMed]
46. Deady, M.; Johnston, D.; Milne, D.; Glozier, N.; Peters, D.; Calvo, R.; Harvey, S. Preliminary effectiveness of a smartphone app to reduce depressive symptoms in the workplace: Feasibility and acceptability study. *JMIR mHealth uHealth* **2018**, *6*, e11661. [CrossRef] [PubMed]
47. Adams, L.; Igbinedion, G.; DeVinney, A.; Azasu, E.; Nestadt, P.; Thrul, J.; Joe, S. Assessing the real-time influence of racism-related stress and suicidality among black men: Protocol for an ecological momentary assessment study. *JMIR Res. Protoc.* **2021**, *10*, e31241. [CrossRef]
48. Burns, M.N.; Begale, M.; Duffecy, J.; Gergle, D.; Karr, C.J.; Giangrande, E.; Mohr, D.C. Harnessing context sensing to develop a mobile intervention for depression. *J. Med. Internet. Res.* **2011**, *13*, e55. [CrossRef] [PubMed]

49. Boonstra, T.W.; Nicholas, J.; Wong, Q.J.; Shaw, F.; Townsend, S.; Christensen, H. Using mobile phone sensor technology for mental health research: Integrated analysis to identify hidden challenges and potential solutions. *J. Med. Internet. Res.* **2018**, *20*, e10131. [CrossRef] [PubMed]
50. Dang, M.; Mielke, C.; Diehl, A.; Haux, R. Accompanying Depression with FINE-A Smartphone-Based Approach. In Proceedings of the MIE, Munich, Germany, 28 August–2 September 2016; pp. 195–199.
51. Porras-Segovia, A.; Molina-Madueño, R.M.; Berrouiguet, S.; Lopez-Castroman, J.; Barrigón, M.L.; Pérez-Rodríguez, M.S.; Marco, J.H.; Díaz-Oliván, I.; de León, S.; Courtet, P.; et al. Smartphone-based ecological momentary assessment (EMA) in psychiatric patients and student controls: A real-world feasibility study. *J. Affect. Disord.* **2020**, *274*, 733–741. [CrossRef] [PubMed]
52. Schueller, S.M.; Begale, M.; Penedo, F.J.; Mohr, D.C. Purple: A modular system for developing and deploying behavioral intervention technologies. *J. Med. Internet. Res.* **2014**, *16*, e3376. [CrossRef] [PubMed]
53. Chandrashekar, G.; Sahin, F. A survey on feature selection methods. *Comput. Electr. Eng.* **2014**, *40*, 16–28. [CrossRef]
54. Permutation Test. Available online: https://www.sciencedirect.com/topics/mathematics/permutation-test (accessed on 27 September 2022).
55. Witten, I.H.; Frank, E. Data mining: Practical machine learning tools and techniques with Java implementations. *ACM Sigmod Rec.* **2002**, *31*, 76–77. [CrossRef]
56. Zhai, Y.; Ma, N.; Ruan, D.; An, B. An effective over-sampling method for imbalanced data sets classification. *Chin. J. Electron.* **2011**, *20*, 489–494.
57. Yen, S.J.; Lee, Y.S. Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Syst. Appl.* **2009**, *36*, 5718–5727. [CrossRef]
58. Ghojogh, B.; Crowley, M. Linear and quadratic discriminant analysis: Tutorial. *arXiv* **2019**, arXiv:1906.02590.
59. Chandola, V.; Banerjee, A.; Kumar, V. Anomaly detection: A survey. *ACM Comput. Surv. (CSUR)* **2009**, *41*, 1–58. [CrossRef]
60. Boslaugh, S.; Watters, D.P.A. *Statistics In a Nutshell*, 2nd ed.; O'Reilly & Associates, Inc.: Sebastopol, CA, USA, 2014.
61. Pollak, J.P.; Adams, P.; Gay, G. PAM: A photographic affect meter for frequent, in situ measurement of affect. In Proceedings of the Sigchi Conference on Human Factors in Computing Systems, Vancouver, BC, Canada, 7–12 May 2011; pp. 725–734.
62. Buitinck, L.; Louppe, G.; Blondel, M.; Pedregosa, F.; Mueller, A.; Grisel, O.; Niculae, V.; Prettenhofer, P.; Gramfort, A.; Grobler, J.; et al. API design for machine learning software: Experiences from the scikit-learn project. In Proceedings of the ECML PKDD Workshop: Languages for Data Mining and Machine Learning, Prague, Czech Republic, 23–27 September 2013; pp. 108–122.
63. Chollet, F. Keras. 2015. Available online: https://github.com/fchollet/keras (accessed on 27 September 2022).
64. scipy.stats.pearsonr. Available online: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html (accessed on 27 September 2022).
65. Liu, D.C.; Nocedal, J. On the limited memory BFGS method for large scale optimization. *Math. Program.* **1989**, *45*, 503–528. [CrossRef]
66. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.