Journal of the Brazilian Computer Society, 2026, 32:1, doi: 10.5753/jbcs.2026.5939

A Machine Learning Classification Model for Identifying College Students with Depression Based on Digital Phenotyping

Evandro Y. A. Ribeiro [**Federal University of São Carlos** | ***evandro@estudante.ufscar.br***]
Franco E. Garcia [**Federal University of São Carlos** | ***francogarcia@protonmail.com***]
Conrado dos S. Alves Saud [**Federal University of São Carlos** | ***conradosaud@gmail.com***]
Helena de M. Caseli [**Federal University of São Carlos** | ***helenacaseli@ufscar.br***]
Vivian G. Motti [**George Mason University** | ***vmotti@gmu.edu***]
Taís Bleicher [**Federal University of São Carlos** | ***tbleicher@ufscar.br***]
Jair B. Neto [**Federal University of São Carlos** | ***jairbneto@ufscar.br***]
Heloisa C. Figueiredo Frizzo [**University of São Paulo** | ***heloisa.frizzo@usp.br***]
Larissa C. Martini [**Federal University of São Carlos** | ***larissacmb@ufscar.br***]
Luciano de O. Neris [**Federal University of São Carlos** | ***lneris@ufscar.br***]
Anderson Ara [**Federal University of Paraná** | ***ara@ufpr.br***]
Alan D. Baria Valejo [**Federal University of São Carlos** | ***alanvalejo@ufscar.br***]
Vânia P. de Almeida Neris [**Federal University of São Carlos** | ***vania.neris@ufscar.br***]

Departamento de Computação, Universidade Federal de São Carlos, Rod. Washington Luís, km 235, Jardim Guanabara, São Carlos, SP, 13565-905, Brazil.

Received: 17 April 2025 • **Accepted:** 24 November 2025 • **Published:** 16 April 2026

Abstract Depression is a serious global mental health illness that causes significant suffering to the individual and social impairment in their lives. Compared to the general population, depression shows a higher prevalence among college students. With recent advancements in digital phenotyping data analysis to infer depressive symptoms, machine learning (ML) techniques have been increasingly employed to indicate behaviors related to potential depressive profiles (PDP). However, despite the growing body of work on ML usage to detect depression, few studies have focused on data preprocessing approaches to handle missing values in datasets that go beyond common data imputation. In this study, we conducted a series of experiments to evaluate the combination of data preprocessing methods and ML algorithms for effectively classifying PDP and non-PDP students using data from the Amive project. The primary challenges were implementing a data processing workflow to address missing values and class imbalance, common issues in digital phenotyping datasets, and selecting algorithms capable of handling such data. The experimental results showed promising outcomes, with individual classification models, including Random Forest, XGBoost, and SVM(rbf), achieving accuracies of 77%, 75%, and 76%, respectively. The best performance was obtained by training on datasets that went through outlier filtering, specifically removing rows with four or more missing values. This combination of data preprocessing approaches and ML algorithms resulted in a Random Forest classification model with the best performance ranging between 77% of accuracy and with mean errors metrics of AUC and MCC above 0.5.

Keywords: College students, depression, digital phenotyping, machine learning, mobile sensors

1 Introduction

Major depression is a serious global mental health illness that causes suffering for the individuals affected as well as social impairments. Unlike regular mood changes, depression significantly affects all aspects of personal life, including relationships, work, study, and social functioning [World Health Organization, 2024]. It is estimated that around 4% of the world population experience depression or depressive symptoms World Health Organization [2024]. Pre-pandemic data indicate that approximately 7% to 26% of the American population suffered from depression [Eichstaedt *et al*., 2018]. With the onset of the COVID-19 pandemic, there was a significant increase in depression and anxiety cases worldwide, around 25% [World Health Organization, 2022]. Furthermore, according to Meleiro *et al*. [2023], Brazil has the highest number of depression cases compared to all other countries, with depression affecting 9.3% of the population.

In the university environment, which is the focus of this study, students appear to carry a disproportionate burden of depression compared to the general population [Lauckner *et al*., 2020]. In this context, Ibrahim *et al*. [2013] conducted a systematic review of studies on the prevalence of depression in college students around the world, finding that the prevalence of depressive symptoms was around 30%. Similar surveys of depression prevalence among Brazilian students have found a prevalence of approximately 30–51% [Pacheco *et al*., 2017; Lima *et al*., 2025]. During the pandemic, the prevalence of depression among university students was around 39.2% [Schuch *et al*., 2023].

The diagnosis of major depressive disorder in an individual can be made by identifying five or more symptoms, as defined

by the Diagnostic and Statistical Manual of Mental Disorders (DSM-V). These symptoms include weight loss or gain, changes in appetite, insomnia or hypersomnia, psychomotor agitation or retardation, fatigue or low energy, feelings of worthlessness or guilt, decreased concentration or indecisiveness, and recurrent suicidal thoughts. For a conclusive diagnosis, at least one of the symptoms must be associated with a depressed mood or loss of interest or pleasure [American Psychiatric Association, 2013].

Treatment for mental health illnesses requires time and close monitoring by domain experts, as the diagnosis typically occurs through observation of the individual's behavior and emotions along time, rather than based on a single symptomatic episode [American Psychiatric Association, 2013]. Therefore, depression requires a comprehensive model of care and monitoring, not limited to pharmacological treatments. In response to this need, there has been a significant increase in research aimed at supporting the diagnosis and intervention of depression and other mental health issues [Wang *et al*., 2018; Doryab *et al*., 2019; Alves *et al*., 2023; Hu *et al*., 2023]. In this context, research in computing, particularly in the fields of Human-Computer Interaction (HCI), mobile Health (mHealth), and Machine Learning (ML), plays a significant role in the development of computational systems to support the detection and monitoring of mental health symptoms.

A systematic review of HCI research works in ML and mental health detection, conducted by Thieme *et al*. [2020], revealed a prominence of studies that aimed at supporting and detecting depression symptoms, often through behavioral monitoring using analyses of social network posts, mobile device data, and evaluation questionnaires. This review also highlighted that some studies have explored the use of pervasive technologies to passively collect data for building ML models.

This pervasive data collection, often physiological, is commonly referred to as Digital Phenotyping [Davidson, 2022], it is characterized by the moment-by-moment quantification and analysis of the human phenotype using data from digital devices, like smartphones and smartwatches [Torous *et al*., 2016]. Since most of these devices are present in individuals' daily lives, digital phenotyping is an efficient method to measure and monitor human social and physiological behavior. In general, digital phenotyping is divided into active and passive data categories [Melcher *et al*., 2020]. Within the scope of passive data, daily actions of individuals can be monitored to identify behaviors and other indicators that may correlate with symptoms of depressive disorders [Akbarova *et al*., 2023; Alves *et al*., 2023].

Santos *et al*. [2024], presented a literature review and taxonomy of the uses of digital phenotyping data in ML research which demonstrated that most works are categorized in the health disorders domain, with 96 studies focusing on this area. As revealed by the authors, the most explored disorders are mental disorders, particularly related to mood and depression issues. In the data domain of these studies, passive sensing from smartphones and wearable devices was also highlighted as the most prominent source.

Thus, a growing number of research works are exploring the use of behavioral data collected from digital phenotyping to detect depression symptoms that correlate with those established in the DSM-5, such as poor sleep quality, agitation, fatigue, and lethargy. An example of the relationship between this type of data and the depressive symptoms that can be inferred is presented in Table 1, based on Saud [2023].

Table 1. Relationship between mobile sensor data and symptoms described by the DSM-V

DSM-V Symptoms	Corresponding Sensor Data
Depressed Mood	Text messages, heart rate
Loss of Pleasure/Interest	Usage time GPS, step count, text messages
Weight Loss/Gain	Food intake monitoring
Insomnia or Hypersomnia	Sleep monitoring Accelerometer Gyroscope
Agitation, Lethargy, and Fatigue	GPS Step count Physical activities Heart rate
Feelings of Worthlessness or Guilt	Text messages
Suicidal Thoughts	Text messages

Source: Adapted from Saud [2023]

Another example is presented by Wang *et al*. [2017], in which a smartphone app called StudentLife was developed to automatically and continuously collect data from college students. The primary objective of their research was to infer the correlation between students' behaviors and their Grade Point Average (GPA). They discovered that the collected data strongly correlated with a broad set of mental well-being measures, such as the Patient Health Questionnaire-9 (PHQ-9) and the Perceived Stress Scale (PSS).

Even though recent ML studies have have shown promising findings in depression detection using sensor data, the use of digital phenotyping data with ML has its limitations and challenges, as highlighted in Santos *et al*. [2024], digital phneotyping datasets often presents noisy data, missing values, incomplete data and small sets. In this scenery, most of the recent studies deal with these limitations relying only on data imputation methods to handle missing values or class imbalance. Considering the nature of the problem of depression identification with human data collected by mobile sensors, training ML models with artifically imputed data points may not be suitable for the sensitivity of the problem.

Although recent ML studies have shown promising results in depression detection using sensor data, the use of digital phenotyping data with ML presents several limitations and challenges. As highlighted in Santos *et al*. [2024], digital phenotyping datasets often contain noisy data, missing values, class imbalance, and small sample sizes. In this scenario, most recent studies address these limitations primarily by relying on data imputation techniques to handle missing values or class imbalance. However, considering the sensitive nature of depression identification using human data collected from mobile sensors, training ML models on artificially imputed

data points may not be and ideal approach.

Beyond the challenges related to data processing, there is also an issue concerning data representativeness. Most related studies conduct their analyses and train their models using data from Global North populations, leaving individuals from the Global South underrepresented in this type of research. In this context, this paper proposes the development of an ML classifier model designed to identify potential depressive profiles (PDP) in college students from a university in a global south country, using digital phenotyping data from smartphones and wearable devices while also investigating data preprocessing approaches to minimize minimize the quantity of imputed data. This work is part of the Amive Project [1], a computational solution aimed at developing an interactive system capable of real-time identification and intervention for PDP among college students. The contributions of this paper are as follows:

- Feature selection and extraction from the Amive Database, focusing on building a set of variables that can be used to infer the depressive symptoms described in Table 1.
- Data preprocessing workflow with three experimental approaches to handle missing values, including an outlier filtering step and testing different subsets combined with data imputation for training.
- ML classifier development experiments combining commonly used ML algorithms, evaluating the performance of models trained with each subset constructed in the data preprocessing step, and also testing both single models and ensemble model approaches.
- An ML model developed to classify PDP based on digital phenotyping data from Brazilian college students.

The structure of this paper is organized as follows: Section 2 presents previous research that has explored the use of ML techniques for the prediction and identification of depression and mental health issues. Section 3 introduces the Amive Project, detailing its computational infrastructure, integrated systems and the context in which this research is situated. Section 4 presents the experimental proposal, describing the data collection, data preprocessing method, labeling and feature selection for ML model's training, the choice of classifier algorithms to be tested and the classification workflow for model training. Section 5 shows the results and discussion about the model's performance and the impact of data preprocessing approaches on the predictive task. Section 6 concludes the paper by describing the main findings and possible directions for future works.

2 Related Works

The convenience of passive data collection from wearable devices has supported research to explore this type of data for monitoring college students' behaviors that can indicate depression symptoms. Melcher *et al*. [2020] conducted a clinical review of 25 studies from 2014 to 2020 that utilized digital phenotyping data to assess college students' mental health. Some of the findings from their research are of particular interest to this paper, such as the number of participants, the duration of the study, and the types of data, whether passive or active. In general, most studies had an average of 81 participants, with a duration of 6 weeks. Most of them used some form of active data collection, such as ecological momentary assessments (EMAs) and health-based questionnaires like PHQ-8 and 9, PSS, and The Depression, Anxiety and Stress Scale (DASS). Meanwhile, passive data included smartphone sensor data, such as location, screen time and usage, and accelerometer data.

Wang *et al*. [2018] report a study conducted with 83 students from Dartmouth College over 9 weeks, aiming to analyze the correlation between the behavioral data collected and the PHQ depression scale. The passive behavioral data were collected through the StudentLife smartphone app and the Microsoft Band 2. The sensing data included GPS location, accelerometer, audio, screen usage, and heart rate, and these data were categorized into DSM-V symptom features such as sleep changes, diminished ability to concentrate, interest and pleasure, and depressed mood. As ground truth, the PHQ-8 was administered at the beginning and end of the study, while the PHQ-4 was administered once a week through the smartphone app. To evaluate the sensing data in comparison with the PHQ-8 and PHQ-4 scales, a Lasso Linear Regression model was used. The performance achieved by the linear model was around 81% precision, and the reults also revealed a strong correlation between the digital phenotyping data and the depression scale questionnaires.

Doryab *et al*. [2019] explored the potential of passive sensing data from smartphones and wearable devices to detect digital phenotypes associated with loneliness among college students. In their research, data were collected from 160 college students over a semester, using smartphones and the Fitbit Flex 2 to capture activity levels, mobility, communication, phone usage, and sleep data. Similar to other studies, they also collected active data from the University of California, Los Angeles (UCLA) Loneliness Questionnaire at the beginning and end of the study, using the questionnaire's scores as metrics for classification. For analysis, they employed a combination of statistical, data mining, and ML algorithms. The ML classifier was used to infer the level of loneliness using an ensemble of linear regression and gradient boosting algorithms. With the combination of these methods, they were able to classify students on the UCLA scale with an accuracy of 80.2% and detect fluctuations in loneliness levels with an accuracy of 88.4%. These findings suggest some efficiency in using a combination of health questionnaires and ML techniques to identify and classify loneliness and other issues. The authors also highlighted the risks of loneliness and social isolation, noting that loneliness is associated with high rates of depression and other mental health issues.

The use of ML with digital phenotyping data has been explored for the detection of depression symptoms. For example, in Ware *et al*. [2020], the authors utilized GPS and Wi-Fi data from the smartphones of 182 students to identify depressive symptoms using the Support Vector Machine (SVM) algorithm. Similar to the studies discussed previously, they collected data through a smartphone app and used questionnaires (PHQ-9 and QIDS) to assess the participants'

[1] Amive Project: `https://amive.ufscar.br/english`

depression levels, achieving F1-scores of 0.86 and 85% precision.

Asare *et al*. [2022] studied whether self reported mood and feelings, along with data from passive sensing devices, could be used to classify individuals as depressed or non-depressed. They conducted a longitudinal study collecting data from smartphones and the Oura ring (which tracks physical activity and mood). Their work gathered information about GPS mobility, physical activities, sleep quality, and phone usage to delineate the differences between the two groups, to develop machine learning classification models for depression. In their study, they used the Depression Anxiety Stress Scale (DASS) for data labeling and algorithms such as XGBoost and SVM for classifier modeling, achieving accuracy of 77% and 81%, respectively.

As cited in Santos *et al*. [2024], some studies have focused on monitoring mood changes in participants diagnosed with Major Depressive Disorder (MDD) to evaluate the feasibility of ML models for classifying, analyzing, and predicting depressive behavior using digital phenotyping data from smartphones and wearable devices. For instance, Pedrelli *et al*. [2020] conducted a study assessing depressive symptoms using behavioral and physiological data from mobile devices. They employed ensemble boosting and Random Forest models trained on three data combinations: smartphone and wristband data, smartphone-only data, and wristband-only data, achieving the lowest Root Mean Squared Error (RMSE) of 4.08%. Similarly, Bai *et al*. [2021] evaluated the use of passive smartphone and wristband data with ML models to monitor mood changes in MDD patients. Their app, Mood Mirror, collected phone usage, GPS, app logs, heart rate, sleep, and step data. Data labeling relied on PHQ-9 questionnaire results administered three times during the study, segmenting the dataset into specific intervals. The tested ML models included Random Forest, SVM, Decision Tree, Logistic Regression, Naive Bayes, and KNN. The most effective features were sleep, steps, heart rate, and phone call logs, with Naive Bayes, Decision Trees, and Random Forest achieving the best performance, ranging between 76 to 81%.

Furthermore, as there is no model that fits every type of predictive task [Ahmed and Ahmed, 2023], several studies have investigated the combination of various ML techniques and algorithms to infer depression symptoms, with many including those that are commonly used in the mental health domain. For example, Srividya *et al*. [2018] have explored SVM, decision trees, K-nearest neighbor (KNN), and Logistic Regression algorithms to identify mental health issues in diverse target groups based on participants' questionnaire responses, with the SVM and KNN algorithms achieving 89% accuracy in classification. Hu *et al*. [2023] also used an ensemble of various ML algorithms (SVM, KNN, Logistic Regression, case-based reasoning, and decision trees) to classify depression based only on sleep data from wearable devices. The ensembles employed in classification task achieved 87% of accuracy. A similar set of algorithms selected for classification tasks can also be seen in research works like Sultana *et al*. [2020].

Some recent works have not only focused on the combination of machine learning algorithms but also on data preprocessing techniques. For instance, in Tate *et al*. [2020], the authors explored depression classification using Random Forest, XGBoost, Logistic Regression, and SVM algorithms. In addition to that, they conducted a statistical analysis of the target dataset and noted an imbalance in the training set, where positive cases of depression were the minority predictive class. To address this issue, they artificially inflated the dataset using the Synthetic Minority Over-Sampling Technique (SMOTE) in R, resulting in the Random Forest model achieving the best performance in the predictive task, with an AUC of 0.7399. Similarly, to handle missing values, Hu *et al*. [2023] also employed an artificial method of data imputation known as the Generative Adversarial Imputation Network (GAIN), both. Both works resulted in promising improvements in classification, with the authors attributing the good models' performance to the data imputation methods.

Overall, several similarities can be noted in related work, particularly the frequent use of well-known validated algorithms, such as Logistic Regression, SVM, KNN, and XGBoost, for depression classification tasks. Most studies also rely on mental health assessment questionnaires (PHQ-9, DASS, UCLA, EMA, etc.) as the ground truth and data labeling, and many propose strategies and techniques to address data preprocessing challenges, such as class imbalance and missing values.

However, there is still a lack of studies focusing on data preprocessing experiments for handling missing values in datasets used for ML model training, going beyond the common reliance on data imputation. This paper addresses this problem by proposing a series of experiments in data preprocessing, exploring different subsets of the original dataset to identify which one provides the best construction for training effective ML models. Additionally, in the context of ML classifier construction, this study explores the combination of widely used algorithms for depression classification tasks, while also investigating different approaches, such as the use of single models and ensemble learning models. These contributions set this study apart from existing works. Moreover, this research represents one of the first studies on addressing depression identification using ML in college students, conducted in Brazil.

3 Backgroud

3.1 Amive

Seeking to support the implementation of mental health care, the current interaction methods with mobile devices can be explored as channels for personalized interventions, enabling computational solutions to actively engage with individuals who show symptoms of depression. In this context, to aid in identifying depression symptoms by combining data from wearable sensors with the analysis of ML algorithms and models, the Amive Project (Amive stands for three Portuguese words: Friend, Virtual and Specialized) was developed. This project offers a computational infrastructure endorsed by mental health experts, that aims to be capable of real-time autonomous identification and intervention for college students with PDP.

The Amive computational infrastructure consists of a mo-

bile app, installed on students' smartphones, allowing the data collection from smartwatch, along with systems running on a server to process the user's data. Its app contains features to: (1) collect user's digital phenotyping and textual data; (2) offer interaction via chatbot; (3) provide virtual journaling to record the user's sentiment and (4) apply periodic questionnaires about subjective feelings. The text data are collected from the user's posts on social networks (such as Facebook or X, previous Twitter posts) and the app's virtual journal, while the digital phenotyping data are collected from the user's wearable sensors (smartphone and smartwatch). All data collection was made with the student's knowledge and explicit consent.

In Amive's current state, there are two separate ML models: an NLP (Natural Language Processing) model to process textual data, and the classification model proposed in this paper to process sensor data. The first version of the Amive app was used to collect data to the model's training. The platform also allows students to evaluate the collected data through the app's interface. Validated data is sent to the database, while unapproved data is excluded from the training set. This step ensures trust in the collected data by employing a Human-in-the-Loop (HITL) [Tomaszewski, 2021] method, involving human interaction with the ML models' inputs and outputs. Amive's HITL approach was proposed in Alves *et al*. [2023].

3.2 Introduction to selected ML Algorithms

A brief introduction to the algorithms selected for this study is provided below:

LR : The Logistic Regression classifier is a linear probabilistic model that estimates the probability of an instance belonging to a specific class, often used in binary classification problems. It uses the logit transformation to convert the output of a linear regression calculus into probabilities, resulting in values between 0 and 1 as class labels for classification [Srividya *et al*., 2018]. This algorithm was used for classification in two different ways: in default mode and with the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) method, an optimization technique that adjusts the parameters of the regression model, storing fewer vectors and using less memory.

DT : The decision tree classifier is a technique that uses a hierarchical structure of nodes and branches, similar to a tree representation that starts from the root and descends to some leaf node, to provide classification of an instance. Each node specifies a test on an attribute of the instance, followed by a descending division where each branch from that node corresponds to one of the possible values for that attribute. This process is recursively repeated at each sub-node until a stopping criterion is met, commonly when a predefined depth is reached or when the final classification of the instance is made [Mitchell, 1997].

RF : The Random Forest algorithm is a technique that works by creating multiple DTs. Each tree uses a random subset of the dataset and evaluates a random subset of features at each tree node. The algorithm then aggregates the results of all the trees, for classification problems, this aggregation is done through a voting process [Lorena *et al*., 2021].

SVM : The Support vector machines is a supervised technique that works by locating the hyperplane that best separates instances into different classes. This hyperplane is chosen based on the distance between it and the closest data points from each side, the optimal hyperplane is the one that maximizes this distance. A new data point is classified based on each side of the hyperplane they're assigned [Srividya *et al*., 2018]. The SVM was also used in two ways: the default implementation on scikit-learn library, called Support Vector Classifier (SVC), and the SVM using Radial Basis Function (RBF) as kernel.

KNN : The K-Nearest Neighbors is an instance-based classification method. It assumes that all data can be represented as points in an n-dimensional space. The algorithm calculates the Euclidean distance between a new data point and all previously classified points. Then, it selects the k closest neighbors and assigns the most frequent class among those neighbors to the new point [Mitchell, 1997; Lorena *et al*., 2021].

N.B : The Naive Bayes classifier is a probabilistic algorithm that bases its classification on the assumption that the attributes of an instance are conditionally independent given the class. It calculates the probability of an instance belonging to a determined class based on the products of the probabilities of its attributes [Mitchell, 1997; Lorena *et al*., 2021].

XGBoost : The XGBoost is an ensemble technique that combines multiple DTs models. The modesl are created sequentially, with each new tree aiming to reduce the error made by the previous trees. In each iteration, the algorithm uses gradient descent to optimize the loss function, adjusting the parameters of the trees.

4 Experimental Proposal

4.1 Data Collection

An initial study of the Amive project was carried out between August and October of 2022 with the purpose of data collection for the ML models training. This first study included the participation of 89 student volunteers from the Federal University of São Carlos, who were divided into two labeled groups: students with depression and students without depression, considering the outcomes of the PHQ-9.

The data collection was conducted in a case-control type study, divided into two phases. In the first phase, participants answered three questionnaires: (1) the PHQ-9, adapted and validated for the Brazilian Portuguese language [Santos *et al*., 2013], (2) a sociodemographic questionnaire, and (3) the WHODAS 2.0. In the second phase, the students used the Amive app for five weeks. During this phase, data was collected from Facebook Social Network (SN) and smartphones and smartwatch sensors (see Table 3). Additionally, phase 2 also included the application of questionnaires at specific pe-

riods: the PHQ-9 was applied bi-weekly, the WHODAS 2.0 was applied monthly, and the subjective feeling questionnaire was answered each time the app was accessed.

For sensor data collection, the Amive project used the Galaxy Watch 4 smartwatch in combination with the Samsung Health app[2]. Due to the limited availability of smartwatches, 28 devices were distributed among participants by an expert in psychiatry at UFSCAR. The distribution prioritized students who expressed interest in borrowing the device, considering their location across different university campuses, as well as the sociodemographic information like gender, social class, and field of study, to ensure an equitable allocation and diversity in the sample, a better visualization of the sociodemographic data collected can be seen in Table 2. Students who did not receive a smartwatch generated sensor data only through their smartphones.

Table 2. Sociodemographic information collected

Category	Information
Personal characteristics	Date of birth, gender identity, sexual orientation, marital status, children
Family background	Parents' education levels, family income per capita, family structure, family provider
Academic profile	Current academic level, field of study, progress in course, campus location, time dedicated to study, prior education
Mental health	Mental health professional follow-up, use of psychiatric medication

4.2 Features and Labeling

To gather the training set, the features of interest were selected from the collected data. The selection aimed to correlate the sensor data, from smartphone and smartwatch, with depression symptoms, a possible correlation is presented in Table **??**. The correlation between sensor data and depression symptoms can be established using heart rate data, sleep duration and quality, activity information, and self-reported subjective feelings.

The Samsung Health app collects data from Galaxy Watch devices and smartphones, processes this information, and allows users to export their daily personal data as Comma Separated Values (CSV) files. To simplify the features for the training set, an inclusion and exclusion process was applied to the sensor data, selecting only the most relevant features for the model's training, including the subjective feeling report collected by Amive's app. The relationship between sensor data and depression symptoms and the derived features for dataset generation is defined in Table 3.

[2]Samsung Health: `https://www.samsung.com/br/apps/samsung-health/`

Table 3. Relationship of data domain and features

Data	Features selected
Heart rate	average_heart_rate average_heart_rate_when_awake variance_of_heart_rate
Sleep	bedtime average_nightly_sleep_score number_of_sleep_interruptions wake_up_time sleep_duration nap_sleep_score
Activity information	average_steps distance_traveled exercised (boolean) which_exercise_was_done_the_longest_in_a_day
Subjective feeling	predominant_subjective_feeling

An important step in training ML models using supervised learning is data labeling. For this reason, students were categorized as either PDP or non-PDP based on the results of the last administered PHQ-9 questionnaire during the data collection period. A score of 9 points or higher indicated PDP, while a score of 8 points or lower indicated non-PDP.

4.3 Dataset and Data preprocessing

One of the most common problems when dealing with digital phenotyping data is that the heterogeneity of devices and sensors used during collection can cause data gaps in the dataset [Santos *et al*., 2024]. In addition, the data collection relied on users commitment by manually sending and collecting daily data, as well as the full-time use of devices, which resulted in many missing values.

These data inconsistencies may have occurred because some participants did not effectively use the app, submitted little or no data, or because the data was excluded during the HITL (validation) process. To deal with the outliers, users who either did not submit any data or provided insufficient data were filtered out. As a consequence, our refined dataset contains data from 32 students. This filtering process is detailed in Saud [2023].

However, even after the filtering phase, there were still some data gaps, and given that many ML algorithms are not able to handle missing values, the dataset had to go through a preprocessing step, resulting in three versions of the same dataset. The preprocessing steps are: (1) data aggregation in daily intervals; (2) user's outliers filtering; (3) imputation; and (4) features subsets combination for each dataset. Those steps are illustrated in Figure 1.

The Dataset #1 is the standard dataset that went through data preprocessing step. It contains features from all data sources in the database and only users with data points for all the features. Consequently, its final set consists of only 78 rows of data from 10 students, with approximately 82% of the students labeled in the non-PDP group (false label),

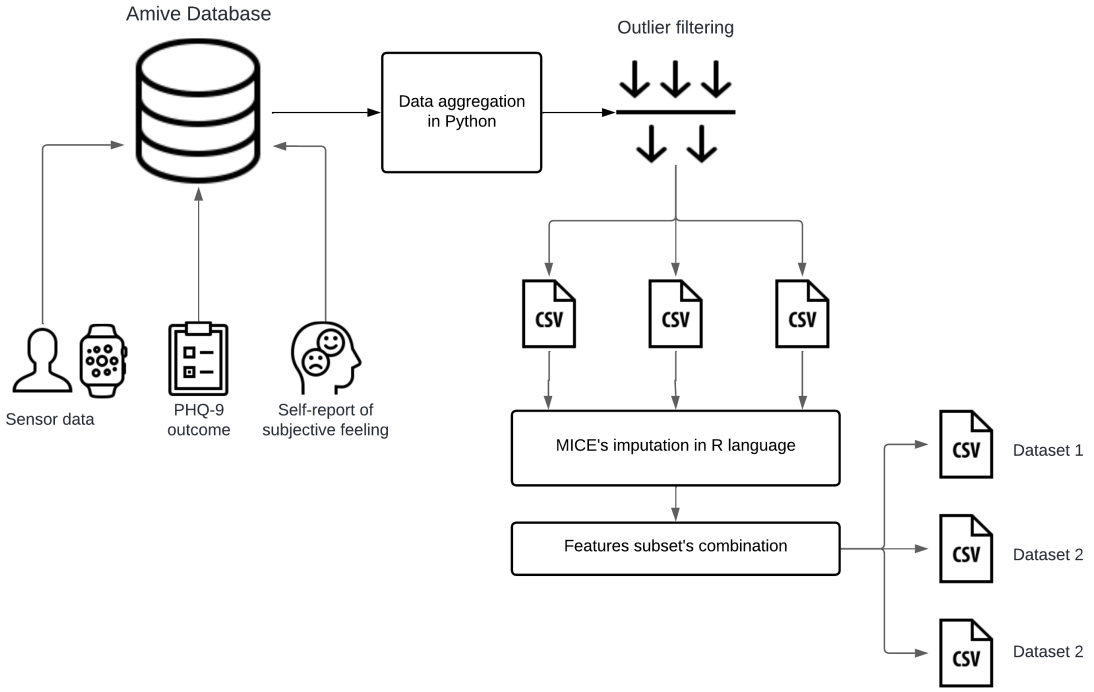

Figure 1. Data preprocessing workflow

resulting in a very imbalanced dataset.

Dataset #2 went through the same preprocessing steps as the first one but included a combination of sensor data and self-reported feelings. Additionally, outliers were removed by filtering out rows with six or more missing values per feature. This resulted in a final dataset consisting of 384 rows from 32 students, with approximately 45% of the students labeled in the PDP group (true label).

Dataset #3 was generated as a variation of Dataset #2. It followed the same preprocessing steps, but in this case, rows with four or more missing values were filtered out. As a result, this subset contains 319 rows of data from 26 students, with approximately 45% of the participants labeled in the PDP group.

These versions and preprocessing steps were implemented with the intent to experiment and test which approach in the data preparation stage could result in a dataset more generalizable by the ML algorithms. However, it is important to note that all datasets are imbalanced to some extent, with dataset #1 being the most imbalanced, despite containing only real data, whereas datasets #2 and #3 are the most balanced. The data preparation steps were implemented in Python and the statistical language R. The imputation, when needed, was implemented using R's package, MICE [3] (Multivariate Imputation by Chained Equations).

4.4 Classifier Selection

The classifier was built using supervised learning techniques, specifically employing the fixed-effect statistical modeling approach. The model uses an ensemble learning method, combining traditional ML algorithms, including *Logistic Regression (LR), Decision Trees (DT), Support Vector Machines (SVM), Naive Bayes, K-Nearest Neighbors (KNN), XGBoost*, and *Random Forest*. The classification algorithms were chosen based on the nature of the intended classification type, which is binary classification (PDP or non-PDP). Furthermore, our dataset is limited in size, which is a common issue in digital phenotyping datasets, as noted in Santos *et al*. [2024]. In this context, traditional ML techniques are generally more reliable for training on small datasets. Recent studies in the ML field have employed deep learning for classification problems. However, this approach would not be suitable for our research problem, as deep neural networks often involve a considerable number of parameters that require substantially more data to achieve reliable estimations and to avoid overfitting [Goodfellow *et al*., 2016]. In this context, we decided to employ traditional ML algorithms in the experiments. Additionally, these algorithms are also commonly used in similar and related works (see Section 2).

The ensemble learning method is a machine learning technique that combines the predictions of multiple classifier algorithms. This combination is performed through a voting process, where each algorithm casts a "vote" for the predicted class of a data point. The voting can be done by means of standard voting, where the class with the most votes is chosen as the final classification and all classifiers' votes carry equal weight, or via weighted voting, where each classifier is assigned a specific weight Lorena *et al*. [2021]. All of the algorithms used were implemented using the Python's scikit-learn library [4].

[3] R-project. "mice: Multivariate Imputation by Chained Equations". `https://cran.r-project.org/web/packages/mice/index.html`

[4] `https://scikit-learn.org/stable/supervised_learning.`

4.5 Experimental Design and Performance Evaluation

Following the definition of the classifiers algorithms and the training sets, the classification and training process involved the following steps: (1) extracting the input data and labels for supervised learning from the processed datasets; (2) extracting the selected features from the input data; (3) training the classifier algorithms using cross validation approach, using k-folds, dividing the train and test in 5 splits of 80% (train) and 20% (test); and (4) evaluating the performance of each algorithm and the ensemble of classifiers. This process of training and testing was applied to all three datasets. Figure 2 illustrates this workflow of the model's construction.

As some ML algorithms can not handle non-standardized data, an additional data processing step was included in the input extraction step for training those algorithms (Logistic Regression (lbfgs), SVM and KNN). This step involved data standardization, that was implemented using a scikit-learn's method, called StandardScaler[5].

The ensemble classifier was implemented in six different ways, combining voting of sets of algorithms that can handle both standardized and non-standardized data, sets that require standardized data, a set including all the algorithms, and weighted voting using standardized and non-standardized data. Table 4 shows which algorithms were included in each version of the Ensemble classifiers.

For the evaluation of the models, we collected metrics such as Accuracy, F1-score, Area Under the Curve (AUC), and Matthews Correlation Coefficient (MCC) for each individual model, as well as for the ensemble voting across all metrics. Accuracy and F1-score are traditional performance metrics for ML models. However, they only allow a superficial analysis. This is why metrics that take into account the statistical relationships in the confusion matrix are also important, as they provide additional insights and help identify potential issues in classification.

5 Results and Discussion

The results obtained from training and test showed variations, which seem to be related to the way the datasets were preprocessed. The performance of the algorithms when fed with Datasets #2 and #3 was relatively similar when compared to the metrics of the algorithms on Dataset #1.

The performance of each single algorithm, in terms of accuracy, was higher during the classification of Dataset #1, where the percentage of participants labeled in the non-PDP group was about 82%. In this set, the KNN and Random Forest classifiers outperformed the other algorithms, with accuracy rates of 81% and 83%, respectively. However, it is important to note that these results should not be evaluated based only on the achieved accuracy. An overall view of the algorithms' metrics in this set is given in Table 5.

As for the performance of the Ensemble Learning with Dataset #1, the results of the Voting classifiers showed relatively higher accuracies, even with the classification combining votes from all algorithms that received both standardized and non-standardized data, ranging from 82% to 84%, respectively. The overall metric performance of the Ensemble voting classifiers is presented in Table 8.

Despite the promising accuracy metrics of the classifiers for the first set (Dataset #1), the F1-scores and MCC are significantly low, suggesting that the models may not effectively identify and classify the minority class, specifically the PDP class in the case of dataset #1, highlighting the performance of the SVM (rbf) that achieved a negative result on the MCC metric. The high accuracy achieved is primarily driven by the algorithm's performance in classifying the majority class (non-PDP), which highlights a significant issue of class and dataset imbalance. Figure 3 and Figure 6 provide a clearer visualization of this discrepancy.

As mentioned, the algorithms metrics in classification of datasets 2 and 3 are relatively close, in both cases, the same issue observed in the classification of dataset #1 was not noted, suggesting a more reliable performance in depression classification. Moreover, the MCC results are slightly higher for the models trained on dataset 3, indicating a somewhat better ability to correctly distinguish between the two classes (PDP and non-PDP). Regarding the overall accuracy, in both datasets, the algorithms that performed best were respectively Random Forest, XGBoost, and SVM (in dataset 3). When trained with dataset #2, the XGBoost algorithm outperformed the others, achieving accuracies around 75% and F1-scores around 72%. Meanwhile, with dataset #3, the best performance was achieved by SVM(rbf), with accuracies around 77% and F1-scores around 74%. Even though the performance of the best algorithms in each dataset is very similar, the Random Forest trained on dataset 3 achieved better performance in terms of AUC and MCC, with values around 0.77 and 0.53. This results suggests that, in addition to solid overall performance, the model also has a considerably better ability to accurately distinguish between the classes. An overview and comparisons of the performance metrics can be seen in Figure 4 and Figure 7, and in Table 6 and Table 9, respectively.

Establishing a relationship with the results of the tested ensembles in datasets 2 and 3, the algorithms that performed best, individually, were precisely the classifiers capable of handling non-standardized data. The performances of the ensembles are demonstrated in Table 7 and Table 10. A comparison of accuracy and F1-scores of each Ensemble can be visualized in Figure 5 and Figure 8.

In summary, the performance evaluations across the ML algorithms and dataset preprocessing techniques indicate that the filtering step for users in datasets #2 and #3 allows for a better generalization by the selected algorithms. This is especially evident in dataset #3, where the Random Forest algorithm achieved the highest accuracy, AUC and MCC scores among all classifiers tested on the more balanced datasets, particularly regarding class balance.

These results suggest that the output models trained on dataset #3 could be suitable for the classification task proposed in this research: identifying and classifying potential depressive profiles among college students. This provides a possible approach for further research in digital phenotyping

`html`

[5] Scikit-learn: "StandardScaler". `scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html`

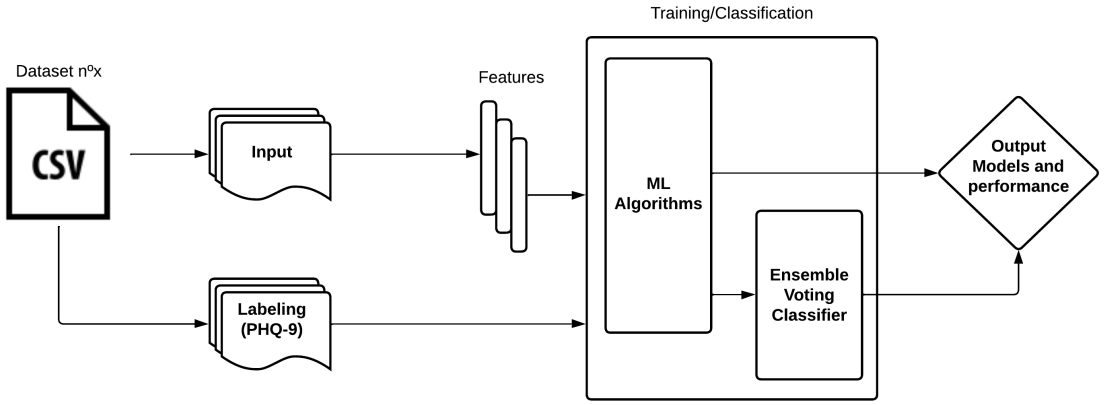

Figure 2. Training and Classification

Table 4. Algorithms included in each Ensemble Voting Classifier

Ensemble	Algorithms Included
All | non-std.	LR (default), DT, RF, XGBoost, NB
Only non-std | non-std	LR (default), DT, RF, XGBoost
All | std.	LR (lbfgs), SVM, SVM (RBF), NB, KNN
Only std | std.	LR (lbfgs), SVM, SVM (RBF), NB, KNN
Weighted | non-std	LR (default), DT, RF, XGBoost
Weighted | std.	LR (lbfgs), NB, KNN

and mental health assessment.

However, it is important to acknowledge that even with the decent performance achieved by the models trained with the balanced datasets, there are still some challenges and limitations when dealing with digital phenotyping data to build robust and reliable ML classifiers. In the tests for the development of the proposed model, some of the common issues noted in similar works were highlighted. As mentioned by Santos *et al*. [2024] in their literature review, most research involving ML and digital phenotyping faces challenges related to class imbalance in prediction, as well as missing values and small data samples.

As cited, this research work, some of the mentioned problems may introduce significant risks to the effectiveness of the model's predictions. The first issue is that, although the data collection phase of the Amive project involved 89 university students, the final raw dataset contained a small sample of valid data from only 33 participants, which still required data imputation to address missing values and enable model training. Furthermore, as demonstrated in this study's experiments, dataset #1 was the most imbalanced in terms of predictive classes, which introduces some risks, because classifiers trained on imbalanced data often achieve high accuracy but tend to exhibit biases in predictive tasks [Huang *et al*., 2024].

Regarding these challenges related to data processing, specifically missing data, this research proposes an experimental design using three subsets of the dataset as a method for addressing these issues. In the context of this study, the main goal is to develop a model that can be integrated into the AMIVE project solution and effectively applied to classify college students' data and perform real-time identification. This real-world data scenario raises important concerns, as real-world datasets are often imperfect and is susceptible to have noise, missing values, and incomplete entries. To address this challenge, our experiments were designed to identify a preprocessing strategy aimed at reducing missing values, primarily by eliminating data points based on the proportion of missing values per feature.

In comparison with other related works, this study performs experiments with data from Brazilian college students and does not rely solely on data imputation to address the common missing data issues. Instead, it explores different strategies to reduce the need for imputation while investigating the limits of removing data rows in order to achieve a balance between imputed and real data. As discussed in the experiments, dataset #1 represents the dataset with no imputation. During the preprocessing steps, of this dataset, row exclusion was performed so that only rows with complete data were included. Therefore, dataset #1 represents an extreme case of row removal to eliminate missing values, whereas datasets #2 and #3 represent more balanced approaches to row exclusion while maintaining the dataset relatively balanced.

Table 5. Overall performance metrics (Dataset #1)

Algorithms	Accuracy	F1-score	AUC	MCC
Logistic Regression (default)	0.78 ± 0.03	0.29 ± 0.15	0.58 ± 0.08	0.19 ± 0.15
Logistic Regression (lbfgs)	0.80 ± 0.09	0.40 ± 0.25	0.66 ± 0.19	0.31 ± 0.32
Decision Tree Classifier	0.76 ± 0.07	0.33 ± 0.19	0.59 ± 0.14	0.23 ± 0.24
SVM / SVC	0.78 ± 0.07	0.42 ± 0.14	0.69 ± 0.13	0.35 ± 0.21
SVM / RBF	0.81 ± 0.07	0.06 ± 0.13	0.49 ± 0.01	-0.02 ± 0.04
Naive Bayes	0.79 ± 0.09	0.40 ± 0.26	0.65 ± 0.18	0.29 ± 0.29
KNN	0.81 ± 0.07	0.18 ± 0.22	0.55 ± 0.10	0.13 ± 0.25
Random Forest	0.83 ± 0.06	0.08 ± 0.16	0.53 ± 0.05	0.09 ± 0.18
XGBClassifier	0.82 ± 0.08	0.15 ± 0.18	0.54 ± 0.05	0.10 ± 0.13

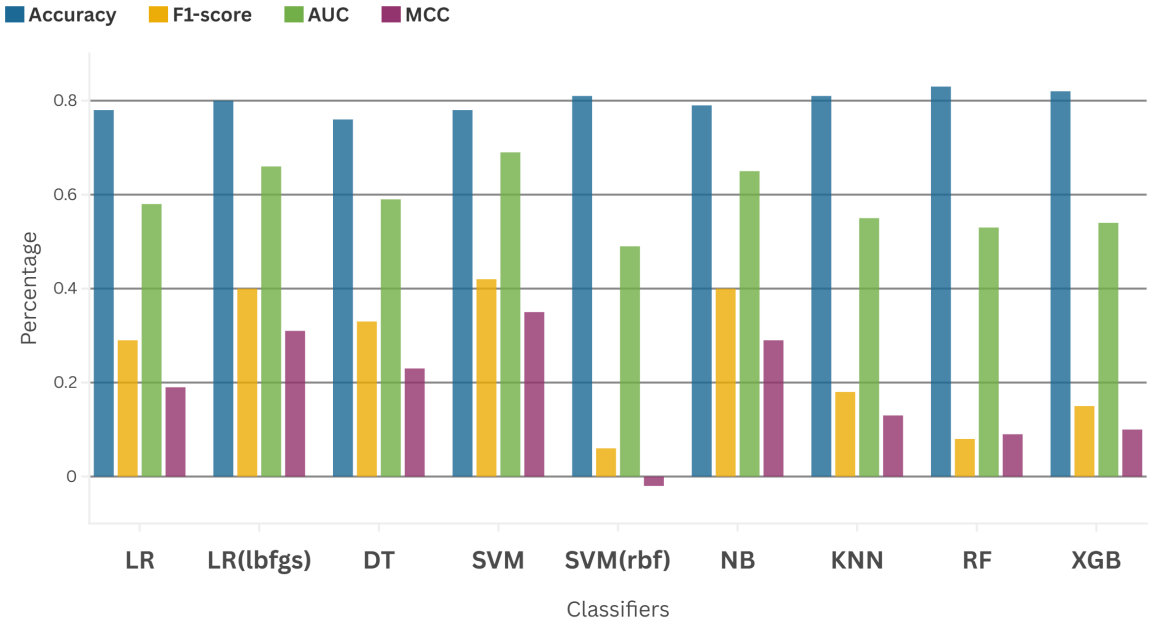

Figure 3. Overall Performance Comparison across Classifiers (Dataset #1)

6 Conclusion and Future Work

Starting from the relationship between digital phenotyping data from smartphones and wearable devices with depression symptom signals, this paper conducted an experimental methodology approach in the development of a machine learning model capable of predicting the identification of PDPS in college students. Additionally, this work investigated different approaches in the dataset preprocessing steps, including data outlier filtering, feature selection, data aggregation and imputation, as well as three different levels of rows removal and subsets variations.

The conducted experiments have demonstrated that the combination of the training subset, which underwent the removal of rows containing four or more missing values, along with the ML algorithms of Random Forest, Logistic Regression, XGBoost and SVM(rbf), as well as the ensemble (voting) classifier, performed better in the predictive task of PDP classification. This resulted in models with performance ranging between 73-77% accuracy and MCC metrics above 0.5. These results suggest a possible valid approach to classification models and dataset treatment.

Expanding on the findings, the performance of the ensemble classifiers combining algorithms that can handle non-standardized data shows some correlations with the ML algorithms that outperformed the others. The best accuracies come from classifiers that accept and can deal with non-standardized data. For this reason, in terms of the ensemble learning technique, a possible direction is to also build and test ensembles that combine only the algorithms that performed well individually, such as the XGBoost and Random Forest.

However, there is still room for improvement in the proposed model, as well as potential risks that deserve further investigation. A major concern is how the model would perform when presented with new, unseen data. This is particularly important to consider, given that the dataset generated from Amive's data collection represents a relatively small sample size for a ML model training. Managing these uncertainties related to model performance is essential for enhancing the model robustness and reliability, in addition to ensuring its applicability in real world scenarios, which is the main goal of the Amive project.

Table 6. Overall performance metrics (Dataset #2)

Algorithms	Accuracy	F1-score	AUC	MCC
Logistic Regression (default)	0.73 ± 0.05	0.71 ± 0.05	0.73 ± 0.05	0.46 ± 0.11
Logistic Regression (lbfgs)	0.73 ± 0.04	0.69 ± 0.05	0.72 ± 0.04	0.45 ± 0.07
Decision Tree Classifier	0.67 ± 0.04	0.60 ± 0.06	0.66 ± 0.04	0.33 ± 0.09
SVM / SVC	0.71 ± 0.04	0.66 ± 0.04	0.70 ± 0.04	0.41 ± 0.07
SVM / RBF	0.72 ± 0.03	0.67 ± 0.04	0.72 ± 0.03	0.44 ± 0.06
Naive Bayes	0.64 ± 0.02	0.64 ± 0.03	0.65 ± 0.03	0.29 ± 0.06
KNN	0.71 ± 0.03	0.65 ± 0.06	0.70 ± 0.03	0.40 ± 0.06
Random Forest	0.75 ± 0.02	0.70 ± 0.04	0.74 ± 0.02	0.48 ± 0.04
XGBClassifier	0.76 ± 0.04	0.72 ± 0.08	0.75 ± 0.05	0.50 ± 0.09

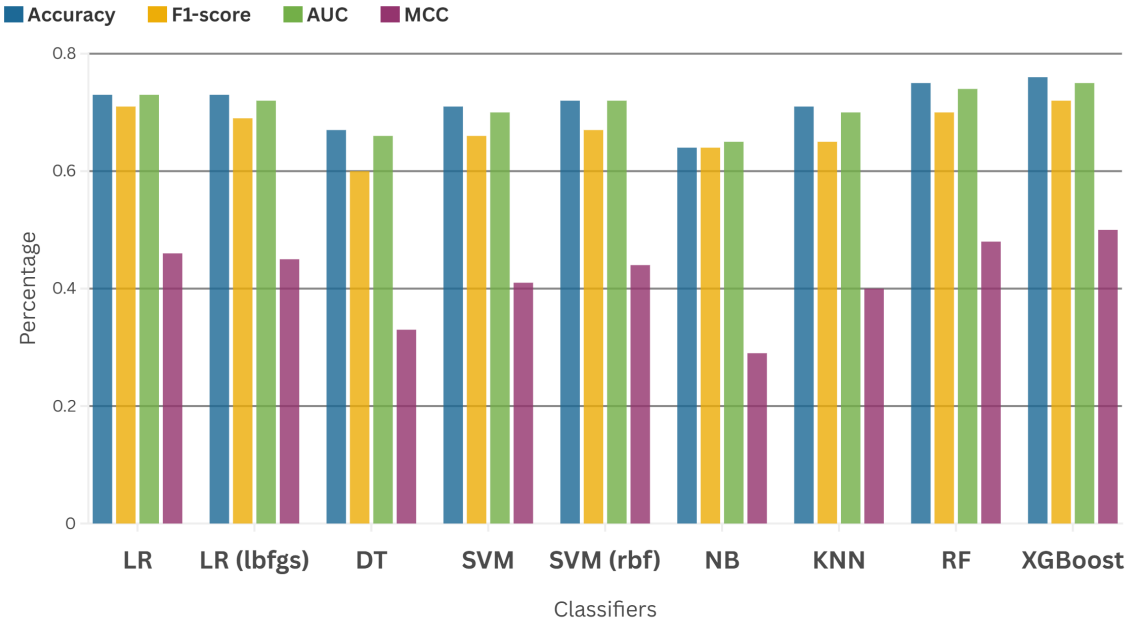

Figure 4. Overall Performance Comparison across Classifiers (Dataset #2)

In summary, for future work, we plan to expand further on our data analysis and experiments of the dataset, our plan includes testing different approaches and techniques to handle missing values. In this study, the MICE input method was employed, however, recent research has achieved interesting results, in depression detection, using data imputation techniques like in Tate *et al*. [2020] where they used the SMOTE technique, which is a form of artificial data augmentation, to address minority classes in the imbalanced dataset. Additionally, in Hu *et al*. [2023] the GAIN, a method of data imputation that utilizes generative networks, has also been applied to handle missing values in training sets.

Furthermore, new tests and experiments are essential to validate the capacity of the proposed model to generalize new data beyond the training set. To enhance the model's robustness, combining the classification of digital phenotyping data with textual data, from SN and journaling collected by the Amive app, can extend the model contextualization ability. As stated in Khoo *et al*. [2024], data from multiple sources and modalities can achieve complementary effects, improving ML models performance.

Table 7. Overall performance metrics (Dataset #3)

Algorithms	Accuracy	F1-score	AUC	MCC
Logistic Regression (default)	0.73 ± 0.03	0.72 ± 0.04	0.73 ± 0.02	0.47 ± 0.05
Logistic Regression (lbfgs)	0.74 ± 0.05	0.69 ± 0.09	0.74 ± 0.06	0.47 ± 0.12
Decision Tree Classifier	0.67 ± 0.03	0.63 ± 0.08	0.66 ± 0.05	0.33 ± 0.09
SVM / SVC	0.73 ± 0.05	0.69 ± 0.07	0.73 ± 0.06	0.45 ± 0.11
SVM / RBF	0.76 ± 0.03	0.72 ± 0.03	0.76 ± 0.03	0.53 ± 0.04
Naive Bayes	0.65 ± 0.05	0.67 ± 0.07	0.66 ± 0.03	0.33 ± 0.09
KNN	0.74 ± 0.05	0.70 ± 0.05	0.74 ± 0.04	0.48 ± 0.09
Random Forest	0.77 ± 0.04	0.74 ± 0.04	0.77 ± 0.04	0.53 ± 0.08
XGBClassifier	0.75 ± 0.03	0.72 ± 0.05	0.75 ± 0.03	0.50 ± 0.06

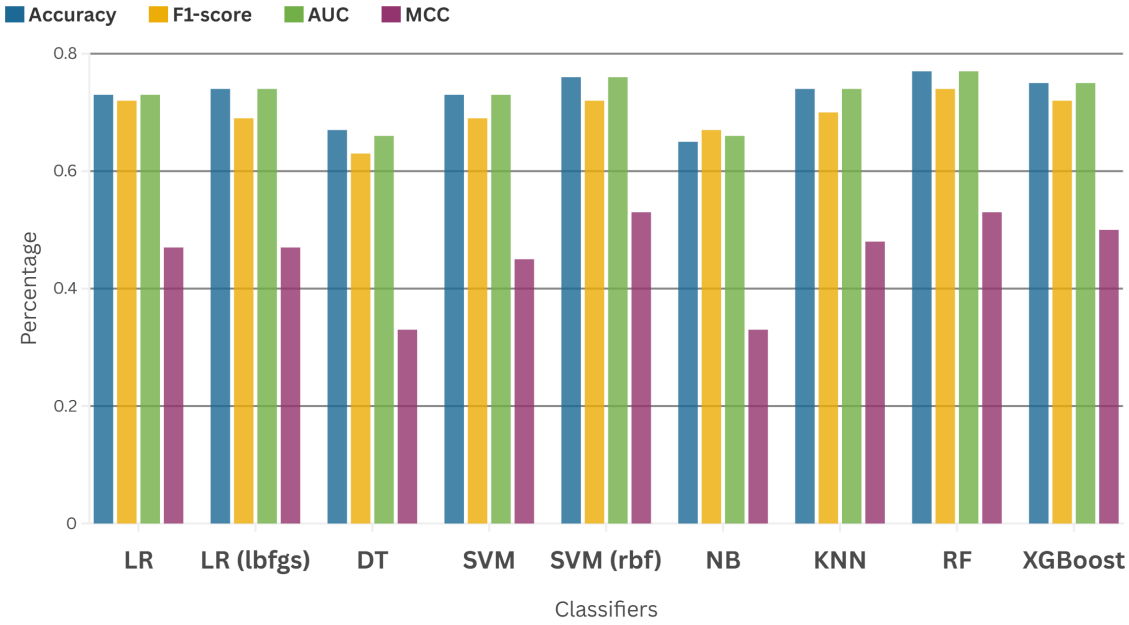

Figure 5. Overall Performance Comparison across Classifiers (Dataset #3)

Table 8. Performance of Ensemble Voting Classifier (Dataset #1)

Algorithms | data type	Accuracy	F1-score	AUC
All | non-std.	0.82 ± 0.07	0.17 ± 0.24	0.53 ± 0.06
Only non-std | non-std.	0.82 ± 0.05	0.42 ± 0.25	0.66 ± 0.14
All | std.	0.84 ± 0.05	0.41 ± 0.25	0.64 ± 0.13
Only std | std.	0.81 ± 0.07	0.06 ± 0.13	0.49 ± 0.02
Weighted | non-std.	0.84 ± 0.08	0.42 ± 0.30	0.68 ± 0.17
Weighted | std.	0.79 ± 0.08	0.52 ± 0.10	0.78 ± 0.09

Table 9. Performance of Ensemble Voting Classifier (Dataset #2)

Algorithms | data type	Accuracy	F1-score	AUC
All | non-std.	0.71 ± 0.06	0.67 ± 0.07	0.70 ± 0.07
Only non-std | non-std.	0.69 ± 0.04	0.65 ± 0.04	0.69 ± 0.04
All | std.	0.69 ± 0.04	0.64 ± 0.04	0.69 ± 0.04
Only std | std.	0.67 ± 0.03	0.59 ± 0.03	0.66 ± 0.03
Weighted | non-std.	0.76 ± 0.04	0.73 ± 0.04	0.75 ± 0.04
Weighted | std.	0.66 ± 0.04	0.71 ± 0.03	0.68 ± 0.03

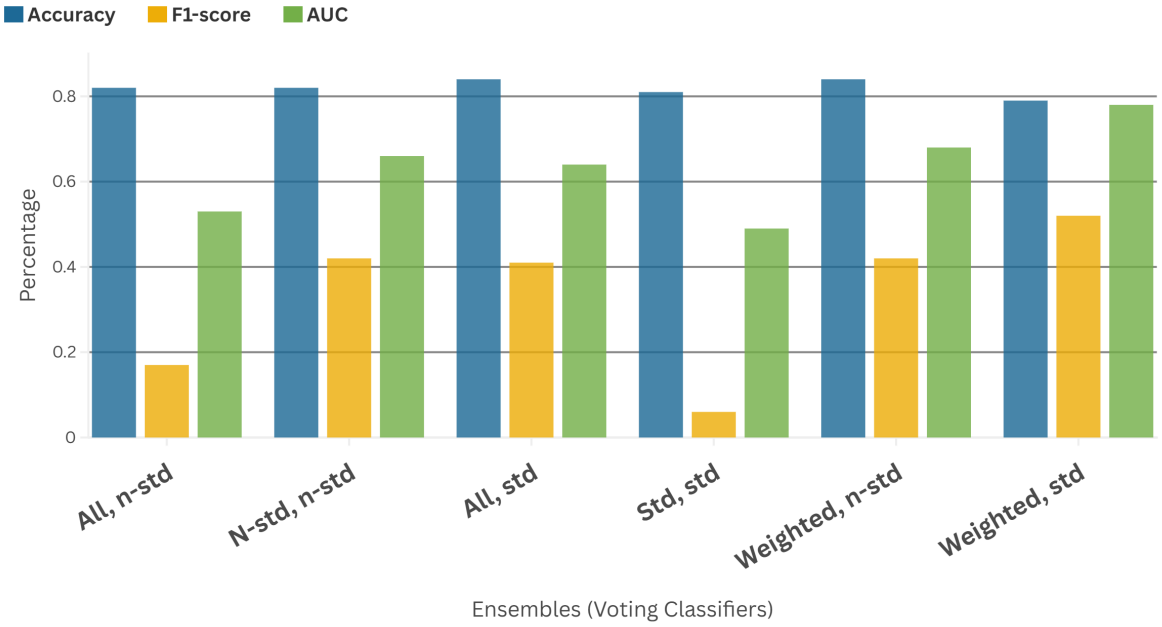

Figure 6. Overall Comparison of the Ensemble Voting Classifiers (Dataset #1)

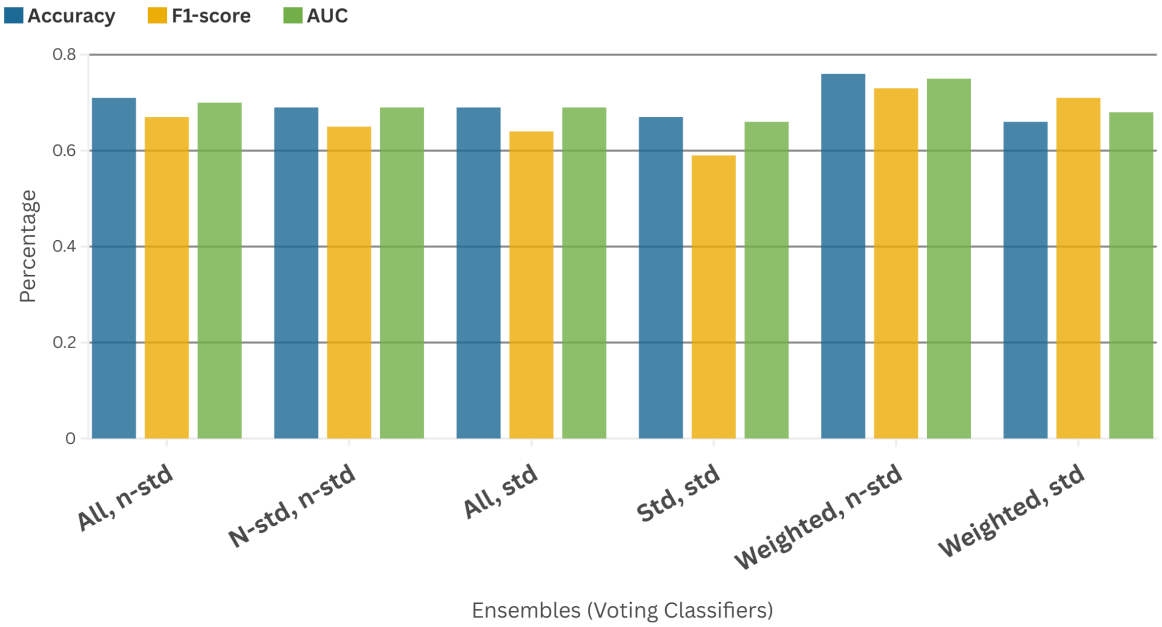

Figure 7. Overall Comparison of the Ensemble Voting Classifiers (Dataset #2)

Table 10. Performance of Ensemble Voting Classifier (Dataset #3)

Algorithms | data type	**Accuracy**	**F1-score**	**AUC**
All | non-std.	0.68 ± 0.03	0.61 ± 0.06	0.67 ± 0.04
Only non-std | non-std.	0.67 ± 0.03	0.62 ± 0.06	0.66 ± 0.04
All | std.	0.70 ± 0.05	0.64 ± 0.08	0.69 ± 0.05
Only std | std.	0.69 ± 0.04	0.61 ± 0.07	0.68 ± 0.04
Weighted | non-std.	0.74 ± 0.06	0.71 ± 0.08	0.73 ± 0.06
Weighted | std.	0.66 ± 0.04	0.70 ± 0.03	0.67 ± 0.03

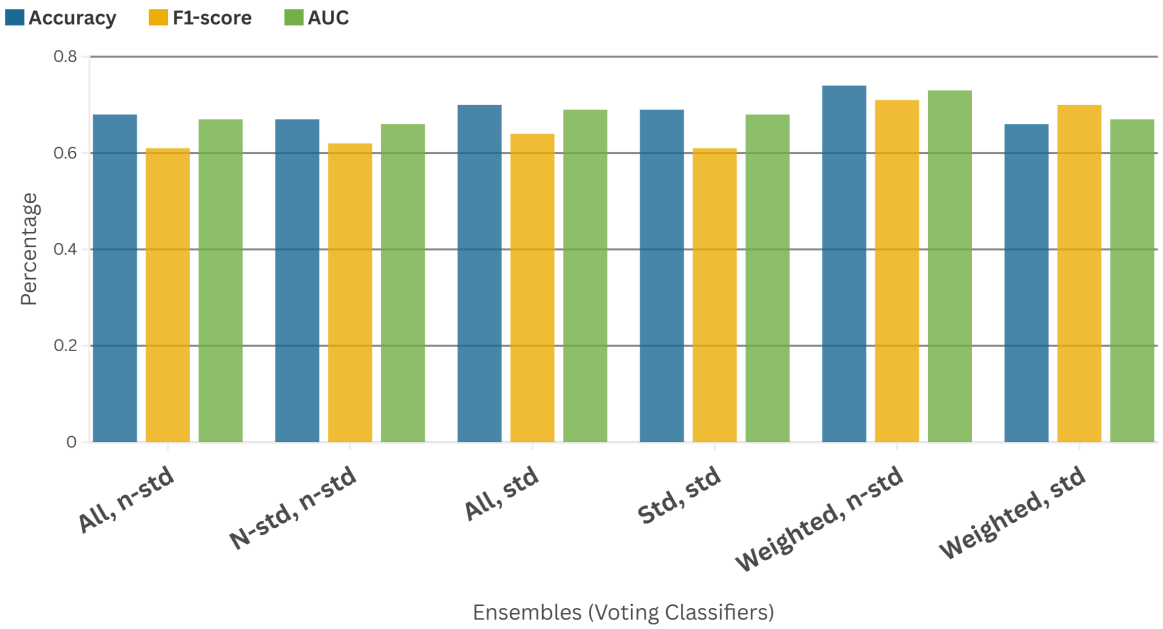

Figure 8. Overall Comparison of the Ensemble Voting Classifiers (Dataset #3)

Declarations

Authors' Contributions

Evandro Ribeiro: Methodology, Software, Formal analysis, Investigation, Writing - Original Draft. **Franco Garcia**: Methodology, Software, Data Curation, Investigation. **Conrado Saud**: Software, Data Curation. **Helena de M. Caseli**: Conceptualization, Validation, Writing- Reviewing and Editing. **Vivian Motti**: Conceptualization, Methodology, Writing-Reviewing and Editing. **Tais Bleicher**: Conceptualization, Resources, Writing-Reviewing and Editing. **Jair Borges**: Conceptualization, Resources, Data Curation, Supervision. **Heloisa Frizzo**: Conceptualization, Resources. **Larissa Martini**: Conceptualization, Resources. **Luciano Neris**: Conceptualization, Resources. **Anderson Ara**: Formal Analysis, Investigation. **Alan Valejo**: Formal analysis, Supervision. **Vania Neris**: Methodology, Supervision, Writing - Review and Editing, Project administration, Funding acquisition.

Competing interests

The authors declare that they have no conflicts of interest.

Funding

This work was partially supported by Coordination for the Improvement of Higher Education Personnel (CAPES) - Finance Code 001, The São Paulo Research Foundation (FAPESP) under grant numbers 22/03090-0, 20/05157-9 and 2024/12772-2, and the National Council for Scientific and Technological Development (CNPq) under grant number 420025/2023-5.

Ethics and Consent to Participate

The data used in this study was collected and used under ethical guidelines defined by the Human Research Ethics Committee (CEP) of the University of São Carlos (CAAE 57554722.6.0000.5504). Informed consent was obtained from all participants prior to their involvement in the study.

The participants signed an informed consent form outlining the risks and benefits of the study. Participation was voluntary and could be withdrawn at any time without consequences. All data were anonymized, including within the research team, and were stored in separate databases: (1) one containing anonymized data for research use; and (2) another storing the participants' private and personal information, which is not used for research purposes.

Availability of Data and Materials

The datasets used in this study was collected by the project and is available for research purposes upon request. Interested researchers may contact the authors, through the access form[6], to obtain access to the data, subject to ethical and institutional approval where applicable.

The training pipeline source-code is available on a dedicated GitHub repository[7].

[6] `https://forms.gle/37rxfdeL46VPTWSn8`

[7] `https://github.com/projeto-amive/Amive-ML-Pipeline-Notebook.git`

References

Ahmed, M. S. and Ahmed, N. (2023). A fast and minimal system to identify depression using smartphones: Explainable machine learning–based approach. *JMIR Form Res*, 7:e28848. DOI: 10.2196/28848.

Akbarova, S., Im, M., Kim, S., Toshnazarov, K., Chung, K.-M., Chun, J., Noh, Y., and Kim, Y.-A. (2023). Improving depression severity prediction from passive sensing: Symptom-profiling approach. *Sensors*, 23(21). DOI: 10.3390/s23218866.

Alves, V. d. C., Garcia, F. E., Saud, C., Mendes, A., Medeiros Caseli, H., Genaro Motti, V., de Oliveira Neris, L., Blecher, T., and Almeida Neris, V. P. (2023). College students-in-the-loop for their mental health: a case of ai and humans working together to support well-being. *Interaction Design and Architecture(s)*, (59):79–94. DOI: 10.55612/s-5002-059-003.

American Psychiatric Association, A. (2013). *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Association. DOI: 10.1176/appi.books.9780890425596.

Asare, K. O., Moshe, I., Terhorst, Y., Vega, J., Hosio, S., Baumeister, H., Pulkki-Råback, L., and Ferreira, D. (2022). Mood ratings and digital biomarkers from smartphone and wearable data differentiates and predicts depression status: A longitudinal data analysis. *Pervasive and Mobile Computing*, 83:101621. DOI: 10.1016/j.pmcj.2022.101621.

Bai, R., Xiao, L., Guo, Y., Zhu, X., Li, N., Wang, Y., Chen, Q., Feng, L., Wang, Y., Yu, X., Wang, C., Hu, Y., Liu, Z., Xie, H., and Wang, G. (2021). Tracking and monitoring mood stability of patients with major depressive disorder by machine learning models using passive digital data: Prospective naturalistic multicenter study. *JMIR Mhealth Uhealth*, 9(3):e24365. DOI: 10.2196/24365.

Davidson, B. I. (2022). The crossroads of digital phenotyping. *General Hospital Psychiatry*, 74:126–132. DOI: 10.1016/j.genhosppsych.2020.11.009.

Doryab, A., Villalba, D. K., Chikersal, P., Dutcher, J. M., Tumminia, M., Liu, X., Cohen, S., Creswell, K., Mankoff, J., Creswell, J. D., *et al*. (2019). Identifying behavioral phenotypes of loneliness and social isolation with passive sensing: statistical analysis, data mining and machine learning of smartphone and fitbit data. *JMIR mHealth and uHealth*, 7(7):e13209. DOI: 10.2196/13209.

Eichstaedt, J. C., Smith, R. J., Merchant, R. M., Ungar, L. H., Crutchley, P., Preoţiuc-Pietro, D., Asch, D. A., and Schwartz, H. A. (2018). Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208. DOI: 10.1073/pnas.1802331115.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. DOI: 10.1038/nature14539.

Hu, Y., Chen, J., Chen, J., Wang, W., Zhao, S., and Hu, X. (2023). An ensemble classification model for depression based on wearable device sleep data. *IEEE Journal of Biomedical and Health Informatics*. DOI: 10.1109/jbhi.2023.3258601.

Huang, C.-M., Hung, C.-S., Hsu, Y.-Y., Zheng, Y.-C., Yu,

C.-H., Lin, C.-H. R., and Chen, S.-H. (2024). A k-means clustering based under-sampling method for imbalanced dataset classification. *International Conference on Information Networking*, page 708 – 713. DOI: 10.1109/ICOIN59985.2024.10572133.

Ibrahim, A. K., Kelly, S. J., Adams, C. E., and Glazebrook, C. (2013). A systematic review of studies of depression prevalence in university students. *Journal of psychiatric research*, 47(3):391–400. DOI: 10.1016/j.jpsychires.2012.11.015.

Khoo, L. S., Lim, M. K., Chong, C. Y., and McNaney, R. (2024). Machine learning for multimodal mental health detection: A systematic review of passive sensing approaches. *Sensors*, 24(2). DOI: 10.3390/s24020348.

Lauckner, C., Hill, M., and Ingram, L. A. (2020). An exploratory study of the relationship between social technology use and depression among college students. *Journal of college student psychotherapy*, 34(1):33–39. DOI: 10.1080/87568225.2018.1508396.

Lima, J. D., Plácido, J., Andrade, B., Abend, L. D., Waclawovsky, A. J., Pires, D. A., *et al*. (2025). Intersectionality and mental health in university students: a jeopardy index approach. *Revista de Saúde Pública*, 59:e3. DOI: 10.11606/s1518-8787.2025059006197.

Lorena, A., Faceli, K., Almeida, T., de Carvalho, A., and Gama, J. (2021). *Inteligência Artificial: uma abordagem de Aprendizado de Máquina (2nd edition)*. LTC. Book.

Melcher, J., Hays, R., and Torous, J. (2020). Digital phenotyping for mental health of college students: a clinical review. *BMJ Ment Health*, 23(4):161–166. DOI: 10.1136/ebmental-2020-300180.

Meleiro, A., Teng, C. T., Demetrio, F. N., Batista, V. C., Vieira, L. F., and Elorza, P. M. (2023). Understanding the journey of patients with depression in brazil: A systematic review. *Clinics*, 78:100192. DOI: 10.1016/j.clinsp.2023.100192.

Mitchell, T. M. (1997). *Machine learning*, volume 1. McGraw-hill New York. DOI: 10.1007/978-1-4613-2279-5.

Pacheco, J. P., Giacomin, H. T., Tam, W. W., Ribeiro, T. B., Arab, C., Bezerra, I. M., and Pinasco, G. C. (2017). Mental health problems among medical students in brazil: a systematic review and meta-analysis. *Brazilian Journal of Psychiatry*, 39:369–378. DOI: 10.1590/1516-4446-2017-2223.

Pedrelli, P., Fedor, S., Ghandeharioun, A., Howe, E., Ionescu, D. F., Bhathena, D., Fisher, L. B., Cusin, C., Nyer, M., Yeung, A., Sangermano, L., Mischoulon, D., Alpert, J. E., and Picard, R. W. (2020). Monitoring changes in depression severity using wearable and mobile sensors. *Frontiers in Psychiatry*, 11. DOI: 10.3389/fpsyt.2020.584711.

Santos, I. S., Tavares, B. F., Munhoz, T. N., Almeida, L. S. P. d., Silva, N. T. B. d., Tams, B. D., Patella, A. M., and Matijasevich, A. (2013). Sensibilidade e especificidade do patient health questionnaire-9 (phq-9) entre adultos da população geral. *Cadernos de Saúde Pública*, 29(8):1533–1543. DOI: 10.1590/0102-311X00144612.

Santos, M. P. d., Heckler, W. F., Bavaresco, R. S., and Barbosa, J. L. V. (2024). Machine learning applied to digital phenotyping: A systematic literature review and taxonomy. *Computers in Human Behavior*, 161:108422. DOI: 10.1016/j.chb.2024.108422.

Saud, C. d. S. A. (2023). Uma infraestrutura computacional para a identificação de estudantes universitários com possível perfil depressivo usando dados de sensores móveis. Available at:`https://repositorio.ufscar.br/handle/ufscar/20186`.

Schuch, H. S., CADEMARTORI, M. G., DIAS, V. D., LEVANDOWSKI, M. L., MUNHOZ, T. N., HALLAL, P. C., and DEMARCO, F. F. (2023). Depression and anxiety among the university community during the covid-19 pandemic: a study in southern brazil. *Anais da Academia Brasileira de Ciências*, 95(1). DOI: 10.1590/0001-3765202320220100.

Srividya, M., Mohanavalli, S., and Bhalaji, N. (2018). Behavioral modeling for mental health using machine learning algorithms. *Journal of medical systems*, 42:1–12. DOI: 10.1007/s10916-018-0934-5.

Sultana, M., Al-Jefri, M., and Lee, J. (2020). Using machine learning and smartphone and smartwatch data to detect emotional states and transitions: Exploratory study. *JMIR Mhealth Uhealth*, 8(9):e17818. DOI: 10.2196/17818.

Tate, A. E., McCabe, R. C., Larsson, H., Lundström, S., Lichtenstein, P., and Kuja-Halkola, R. (2020). Predicting mental health problems in adolescence using machine learning techniques. *PLOS ONE*, 15(4):1–13. DOI: 10.1371/journal.pone.0230389.

Thieme, A., Belgrave, D., and Doherty, G. (2020). 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.*, 27(5). DOI: 10.1145/3398069.

Tomaszewski, J. E. (2021). Overview of the role of artificial intelligence in pathology: the computer as a pathology digital assistant. In *Artificial intelligence and deep learning in pathology*, pages 237–262. Elsevier. DOI: 10.1016/b978-0-323-67538-3.00011-7.

Torous, J., Kiang, M. V., Lorme, J., and Onnela, J.-P. (2016). New tools for new research in psychiatry: A scalable and customizable platform to empower data driven smartphone research. *JMIR Mental Health*, 3(2):e16. DOI: 10.2196/mental.5165.

Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., Zhou, X., Ben-Zeev, D., and Campbell, A. T. (2017). Studentlife: Using smartphones to assess mental health and academic performance of college students. *Mobile Health: Sensors, Analytic Methods, and Applications*, pages 7–33. DOI: 10.1007/978-3-319-51394-2_2.

Wang, R., Wang, W., DaSilva, A., Huckins, J. F., Kelley, W. M., Heatherton, T. F., and Campbell, A. T. (2018). Tracking depression dynamics in college students using mobile phone and wearable sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1):1–26. DOI: 10.1145/3191775.

Ware, S., Yue, C., Morillo, R., Lu, J., Shang, C., Bi, J., Kamath, J., Russell, A., Bamis, A., and Wang, B. (2020). Predicting depressive symptoms using smartphone data. *Smart Health*, 15:100093. DOI: 10.1016/j.smhl.2019.100093.

World Health Organization (2022). Mental health and covid-19: scientific brief, 2 march 2022. Available at:https://www.who.int/publications/i/item/WHO-2019-nCoV-Sci_Brief-Mental_health-2022.1.

World Health Organization (2024). Depression. DOI: 10.1037/e303202003-001.