Development of machine learning: derived mhealth composite biomarkers for trial@home clinical trials

Zhuparris, A.

CHAPTER 2

Machine learning techniques for developing remotely monitored central nervous system biomarkers using wearable sensors: a narrative literature review

Ahnjili ZhuParris,[1,2,3] Annika A. de Goede,[1] Iris E. Yocarini,[2] Wessel Kraaij,[2,4] Geert Jan Groeneveld,[1,2] and Robert-Jan Doll [1]

Sensors. 2023;23(11):5243. doi:10.3390/s23115243

1 Centre for Human Drug Research (CHDR), Leiden, NL

2 Leiden Institute of Advanced Computer Science (LIACS), Leiden, NL

3 Leiden University Medical Center (LUMC), Leiden, NL

4 The Netherlands Organisation for Applied Scientific Research (TNO), Den Haag, NL

Abstract

Background: Central nervous system (CNS) disorders benefit from ongoing monitoring to assess disease progression and treatment efficacy. Mobile health (MHEALTH) technologies offer a means for the remote and continuous symptom monitoring of patients. Machine Learning (ML) techniques can process and engineer MHEALTH data into a precise and multidimensional biomarker of disease activity. **Objective:** This narrative literature review aims to provide an overview of the current landscape of biomarker development using MHEALTH technologies and ML. Additionally, it proposes recommendations to ensure the accuracy, reliability, and interpretability of these biomarkers. **Methods:** This review extracted relevant publications from databases such as PubMed, IEEE, and CTTI. The ML methods employed across the selected publications were then extracted, aggregated, and reviewed. **Results:** This review synthesized and presented the diverse approaches of 66 publications that address creating MHEALTH-based biomarkers using ML. The reviewed publications provide a foundation for effective biomarker development and offer recommendations for creating representative, reproducible, and interpretable biomarkers for future clinical trials. **Conclusion:** MHEALTH-based and ML-derived biomarkers have great potential for the remote monitoring of CNS disorders. However, further research and standardization of study designs are needed to advance this field. With continued innovation, MHEALTH-based biomarkers hold promise for improving the monitoring of CNS disorders.

Introduction

MOTIVATION

Disorders that are affected by the Central Nervous System (CNS), such as Parkinson's Disease (PD) and Alzheimer's Disease (AD), have a significant impact on the quality of life of patients. These disorders are often progressive and chronic, making long-term monitoring essential for assessing disease progression and treatment effects. However, the current methods for monitoring disease activity are often limited by accessibility, cost, and patient compliance.[1,2] Limited accessibility to clinics or disease monitoring devices may hinder the regular and consistent monitoring of a patient's condition, especially for patients living in remote areas or for those who have mobility limitations. Clinical trials incur costs related to personnel, infrastructure, and equipment. A qualified healthcare team, including clinical raters, physicians, and nurses, contributes to personnel costs through salaries, training, and administrative support. Trials involving specialized equipment for measuring biomarkers can significantly impact the budget due to costs associated with procurement, maintenance, calibration, and upgrades. Furthermore, infrastructure costs may increase as suitable facilities are required for data collection during patient visits and equipment storage. Patient compliance poses challenges for disease monitoring, as some methods require patients to adhere to strict protocols, collect data at specific time intervals, or perform certain tasks that can be challenging for patients to execute. Low or no compliance can lead to incomplete or unreliable monitoring results, which in turn can hinder the reliability of the assessments. Given these limitations, there is a growing interest in exploring alternative approaches to monitoring CNS disorders that can overcome these challenges. The increasing adoption of smartphones and wearables among patients and researchers offers a promising avenue for remote monitoring.

Patient-generated data from smartphones, wearables, and other remote monitoring devices can potentially complement or supplement clinical visits by providing data during evidence gaps between visits. As

the promise of mobile Health (MHEALTH) technologies is to provide more sensitive, ecologically valid, and frequent measures of disease activity, the data collected may enable the development and validation of novel biomarkers. The development of novel 'digital biomarkers' using data collected from electronic Health (EHEALTH) and MHEALTH device sensors (such as accelerometers, GPS, and microphones) offers a scalable opportunity for the continuous collection of data regarding behavioral and physiological activity under free-living conditions. Previous clinical studies have demonstrated the benefits of smartphone and wearable sensors to monitor and estimate symptom severity associated with a wide range of diseases and disorders, including cardiovascular diseases,[3] mood disorders,[4] and neurodegenerative disorders.[5,6] These sensors can capture a range of physiological and behavioral data, including movement, heart rate, sleep, and cognitive function, providing a wealth of information that can be used to develop biomarkers for CNS disorders in particular. These longitudinal and unobtrusive measurements are highly valuable for clinical research, providing a scalable opportunity for measuring behavioral and physiological activity in real-time. However, these approaches may carry potential pitfalls as the data sourced from these devices can be large, complex, and highly variable in terms of availability, quality, and synchronicity, which can therefore complicate analysis and interpretation.[7,8] Machine Learning (ML) may provide a solution to processing heterogenous and large datasets, identifying meaningful patterns within the datasets, and predicting complex clinical outcomes from the data. However, the complexities involved in developing biomarkers using these new technologies need to be addressed. While these tools can aid the discovery of novel and important digital biomarkers, the lack of standardization, validation, and transparency of the ML pipelines used can pose challenges for clinical, scientific, and regulatory committees.

WHAT IS MACHINE LEARNING

In clinical research, one of the primary objectives is to understand the relationship between a set of observable variables (features) and one or more outcomes. Building a statistical model that captures the relationship between these variables and the corresponding outputs facilitates the attainment of this understanding.[9] Once this model is built, it can be used to predict the value of an output based on the features.

ML is a powerful tool for clinical research as it can be used to build statistical models. A ML model consists of a set of tunable parameters and a ML algorithm that enables the generation of outputs based on given inputs and selected parameters. Although ML algorithms are fundamentally statistical learning algorithms, ML and traditional statistical learning algorithms can differ in their objectives. Traditional statistical learning aims to create a statistical model that represents causal inference from a sample, while ML aims to build generalizable predictive models that can be used to make accurate predictions on previously unseen data.[10,11] However, it is essential to recognize that while ML models can identify relationships between variables and outcomes, they may not necessarily identify a causal link between them. This is because even though these models may achieve good performances, it is crucial to ensure that their predictions are based on relevant features rather than spurious correlations. This enables the researchers to gain meaningful insights from ML models while also being aware of their inherent limitations.

While ML is not a substitute for the clinical evaluation of patients, it can provide valuable insights into a patient's clinical profile. ML can help to identify relevant features that clinicians may not have considered, leading to better diagnosis, treatment, and patient outcomes. Additionally, ML can help to avoid common pitfalls observed in clinical decision making by removing bias, reducing human error, and improving the accuracy of predictions.[12–15] As the volume of data generated for clinical trials and outside clinical settings continues to grow, ML's support in processing data and informing the decision-making process becomes necessary. ML can help to uncover insights from large and complex datasets that would be difficult or impossible to identify manually.

To develop an effective ML model, it is necessary to follow a rigorous and standardized procedure. This is where ML pipelines come in. Table 1

showcases an exemplary ML pipeline, which serves as a systematic framework for automating and standardizing the model generation process. The pipeline encompasses multiple stages, as defined by the authors, to ensure an organized and efficient approach to model development. First, defining the study objective guides the subsequent stages and ensures the final model meets the desired goals. Second, raw data must be preprocessed to remove errors, inconsistencies, missing data, or outliers. Third, feature extraction and selection identify quantifiable characteristics of the data relevant to the study objective and extracts them for use in the ML model. Fourth, ML algorithms are applied to learn patterns and relationships between features, with optimal configurations identified through iterative processes until desired performance metrics are achieved. Finally, the model is validated against a new dataset that is not used in training to ensure generalizability. Effective reporting and assessment of ML procedures must be established to ensure transparency, reliability, and reproducibility.

OBJECTIVES

The objective of this narrative literature review is to provide an overview of the ML practices used in studies that use MHEALTH technologies and ML to develop novel biomarkers for clinical trials. In this review, each component of the ML pipeline has a dedicated section. Based on the results obtained from the review process, each ML component section provides a comprehensive analysis and discussion of the most common and notable practices. These sections delve into the motivations behind these practices, their limitations, and their overall impact on the ML pipeline. This review will not provide precise recommendations for best practices, as much of the research in this area is new and quickly evolving. Rather, the recommendation section discusses the approaches for standardization and validation procedures that are necessary for the development of ML biomarkers to ensure the effectiveness and acceptance of these biomarkers by clinical, scientific, and regulatory committees.

Methods

INFORMATION SOURCES AND SEARCH STRATEGY

Given the wide range of study designs and clinical populations that use smartphones and wearables to collect data, we used the Joanna Briggs Institute (JBI) guidelines to develop a search strategy.[16] Based on an initial limited search of online databases for clinical trials that report using MHEALTH devices and ML, we developed a custom keyword strategy and performed an in-depth search in PubMed, IEEE Xplore, and CTTI (Table 2). The search terms for the CNS disorder terms were based on the National Library of Medicine's CNS MeSH descriptor data.[17] The relevant papers were selected based on the title and abstract. Finally, other literature review studies that explore the same questions were reviewed; the references cited by these studies were then identified and reviewed if they met our criteria. The date range for the search was between 1 January 2012 and 31 December 2022. The search was conducted on 7 January 2023.

INCLUSION CRITERIA

The authors adopted the Population, Intervention, Comparator, Outcomes, Study type (PICOS) framework to define the inclusion and exclusion criteria (Table 3).[13] The studies included were restricted to those involving participants diagnosed with CNS disorders who were remotely monitored under free-living conditions. The intervention and device criteria were limited to passive data collected from smartphones and other non-invasive remote monitoring sensors, whereas data collected using active engagement from participants, such as disposable blood tests or small scales, were excluded. As we chose to focus on ML pipelines, we selected studies in which a statistical model was used to analyze a dataset and could potentially be used to generate future predictions using an independent dataset. Therefore, traditional statistical models such as linear or logistic regression were included, but statistical models such as ANOVA and correlation analyses were not included. Further, as the focus

was on the development and validation of ML models, we did not include studies that did not report on model performance.

DATA EXTRACTION

Two authors conducted the data extraction following the inclusion criteria, and the results were reviewed by the remaining authors. Data relating to the database source, title, DOI, publication year, trial setting or scenario, objective, devices used, data collection period, number of participants, inclusion of healthy controls, data processing steps, feature engineering, feature selection, machine learning models used, hyperparameters and hyperparameter optimization, model performance, and validation procedure were extracted. The comprehensive data extraction and review conducted by the authors encompassed various essential aspects of the studies, ensuring a thorough analysis of the database source, trial details, data processing steps, machine learning models, and validation procedures.

Results

STUDY SELECTION

Our initial keyword search revealed a total of 2310 articles that utilized digital phenotyping devices, such as smartphones and wearables, in a clinical study and applied ML techniques. After screening the titles and abstracts based on our predefined criteria, we narrowed down the articles to 66 studies, which were used for our analysis. Figure 1 provides an overview of the complete selection process.

STUDY CHARACTERISTICS

For each of the 66 studies, we extracted information about the clinical population and the ML pipeline that was used to develop the digital biomarkers. We found that only half of the studies included healthy controls (N = 34). As seen in Figure 2, Parkinson's disease (PD) (N = 27) was the most prevalent disorder identified in our search, followed by Bipolar Disorder (BD) (N = 11), and Unipolar Depression or Major Depressive Disorder (MDD) (N = 9). The sample size of the selected studies was heterogenous, ranging from 7 to 6221 participants (Figure 3). Overall, our review provides a comprehensive overview of the characteristics of studies that have utilized MHEALTH devices and ML techniques, which can help inform future research in this field. In the following sections, we addressed how the selected studies approached the construction of their ML pipelines.

Missing and Outlier Data

Missing and outlier data are commonly encountered problems for remote sensing clinical trials. Missing data can be the result of device charging frequency, device robustness, and participant compliance.[18] Outliers can be the result of sensor or device dysfunction or malfunction, incorrect data entry, and incorrect classifications.[19] Data preprocessing, which refers to the dropping or manipulation of data, is required for identifying and removing redundant or irrelevant data and for cleaning the data prior to analysis. Without preprocessing, learning from an imperfect dataset can influence the prediction accuracy of the models.[20] In this section, we address how the selected studies preprocessed their raw data by treating their missing data and outliers, and the limitations of doing so.

HANDLING OF MISSING DATA

Missing data can be Missing Completely at Random (MDD), Missing at Random (MDD), and Missing Not at Random (MNAR).[21] MDD assumes that each observation has the same probability of being included or being missed; therefore, there is no difference in the characteristics between participants or observations without missing data and those with missing data. For example, data may be missed due to the battery of the smartphone running out. MDD assumes that missing data may have systematic differences between the missing and non-missing data; however, the cause of the missing data can be explained by the non-missing data. For example, a smartphone may have more missing values when the smartphone

battery is low. If the battery percentage is known during the data acquisition, researchers can verify the probability of acquiring missing data depending on the battery percentage. MNAR assumes that missing data are caused by unknown reasons. For example, smartphone sensors may be gradually worn down, which therefore creates more missing data over time. The type of missing data present in the dataset influences whether a researcher should ignore, exclude, or impute the missing data.

Among the selected studies, we found that only 21 of the studies reported the quantity of missing data acquired. Only 29 studies reported how they handled their missing data. We found that complete-case analysis and imputation were the most popular. We identified 14 studies that report using complete-case analysis.[22–36] Complete-case analysis (otherwise known as listwise deletion) is the deletion of an observation involving one or multiple elements of missing data.[26,37,38] While complete-case analysis is the simplest approach to handle missing data, it does reduce the sample size and statistical power of the analysis [39] and can potentially lead to bias if the data are not MDD.[40] Imputation is the statistical process of replacing missing data with substituted inferred values.[41] We identified studies that imputed their missing data using linear interpolation,[29,42,43] forward filling,[44–1,45] zeros, median, means, and the most frequent value in the column.[24,46] The advantage of imputation is that it enables researchers to use all observations in the dataset. However, the inclusion of imputed values can lead to a false impression of the number of complete cases and reduce the variance in the dataset.[47–49]

IDENTIFICATION OF OUTLIERS

Aggarwal's Data mining: the textbook states that it is the subjective definition of the researcher that defines an outlier.[50] In cases where the outlier data were discussed in the selected studies, we found that researchers customized their definition of outliers by either defining a range of acceptable values [32] or by defining a threshold based on the mean and standard deviation.[51–53] Visual inspection by the researchers or the optimization of different threshold mechanisms can both be used to define the boundaries of normal or outlier data.[54,55] Maleki et al. defined outliers as observations that were most likely the result of measurement errors.[36] In terms of the handling of outliers, we only identified six studies that explicitly stated that outliers were excluded.[26,30,51–53,56]

Feature Engineering

FEATURE SCALING

Feature scaling is used to normalize the ranges of the features in a dataset.[57] Several feature engineering techniques and ML models (such as Principal Component Analysis and Linear Regression) calculate the distances between two observations. If one feature has a broader range of values compared to the other features, the calculated distances will be heavily influenced by this feature.[58] Therefore, the ranges of all the features should be normalized or standardized so that each feature is appropriately and proportionally considered with respect to the estimated distances.[57] Feature normalization is a common scaling method for rescaling the features into a bounding range using the minimum and maximum values, for example, between 0 and 1. Normalization is an ideal approach when the distribution of the data is not Gaussian, as normalization preserves the original distribution of the data. However, normalization uses minimum and maximum values to define ranges. This makes the method sensitive to outliers.[57,59] Alternatively, feature standardization, also known as z-score normalization, is a method for rescaling the data to fit a standard normalized distribution by using the mean and standard deviation and does not define a bounding range. Consequently, the standardization approach is not sensitive to outliers as it has no bounding range.[57,59] Normalization, log-transformation, and standardization have been reported in a small selection of the selected studies.[26,27,36,60,61]

EXPERT FEATURE ENGINEERING

Feature engineering is the process of constructing (new) features from the raw data or existing features while maintaining the original patterns and information in the data.[62] The newly engineered features can be added to or replace features in the original dataset. Engineering of the features

can speed up the model performance, improve learning accuracy, and ease the interpretability of the model. The latter is particularly important for clinical trials.[63] Features can be engineered manually by relying on domain-knowledge or automatically by using statistical models, such as Principal Component Analysis (PCA) and Deep Learning (DL).[62–64] All features aim to increase the separability between the classes or signals, which in turn reduces noise in the dataset. While expert engineered features are easy to interpret and explain and have been widely used in the development of digital biomarkers, these features are typically task- or population-dependent. Due to intra-class variability, some clinically relevant characteristics may be exhibited differently by different individuals (such as different symptom profiles among patients with the same diagnosis). Furthermore, expert engineered features may not be sufficient for representing the most important characteristics of complex patterns and can be time-consuming to acquire, especially when handling large-scale datasets.[65,66] As clinical data has expanded in terms of diversity, availability, and complexity, the aforementioned techniques may be insufficient for developing generic features. In the following sections, we address the notable and generic procedures used to perform feature engineering.

SIGNAL PROCESSING

To monitor changes in the physical activity of study participants using time series data collected from wearable sensors, signal processing is necessary to detect, clean, and analyze the components of interest. The feature extraction technique used is influenced by the sensor type, study objectives, and signal quality. Typically, signal features are extracted from the frequency, time, or cepstrum domain.[67] Frequency domain features show the prominence of a signal within a given frequency, whereas time-domain features show the changes in the signal of time. Cepstrum domain features represent the rate of change in the different frequency bands. The analysis of the frequency, time, or cepstrum domain features is not mutually exclusive. We identified studies that use both time- and frequency-based features for the estimation of gait speed,[68] speech-tasks,[69] seizure detection,[70] tremor detection,[71] and FOG detection.[72] In particular, Tougui et al. built 138 voice related features extracted from the cepstral, frequency, and time domains.[24] In sum, time series data collected from wearable sensors can be used to monitor the physical activity of study participants, but signal processing is necessary to extract meaningful features. Different feature extraction techniques can be used depending on the sensor type, signal quality, and study objectives. The analysis of these features is not mutually exclusive, and studies that use multiple domains for different clinical applications have been identified.

PRINCIPAL COMPONENT ANALYSIS

A common linear dimensionality reduction technique for feature engineering and selection is Principal Component Analysis (PCA).[28,73,74] PCA is used to sufficiently explain a high-dimensional dataset through a few principal components and, therefore, to reduce a high-dimensional dataset to one of fewer dimensions.[75] To this purpose, PCA converts a set of correlated features into a set of uncorrelated features by utilizing orthogonal transformation.[75] The principal components enable a reduction in the feature space by creating a linear combination of the original features, which consequently reduces the storage space and reduces the learning time. Therefore, the periodic components within a concurrent time series dataset can be isolated using PCA, which can subsequently be used to identify any underlying patterns within the dataset. It is important to note that PCA assumes that the data are normally distributed and is sensitive to feature variance.[75,76] Consequently, features with larger ranges will dominate features with smaller ranges. To make the variables comparable, transformation of the data prior to PCA is required.[75,76] Of the studies selected, PCA was used to engineer and select features from times series data sourced from waist-worn triaxial accelerometers and wearable activity trackers.[28,73,74] However, the limitations of PCA are its sensitivity to missing data and outliers and the limited interpretation of the original features. Hence, this observation highlights the need for thorough data preprocessing prior to using PCA.

CLUSTERING

A clustering algorithm is a common feature engineering method that assigns similar observations to a single cluster and assigns dissimilar observations to another.[77] While PCA compresses the features into principal components, clustering compresses the individual observations into clusters. The grouping of similar observations can improve the model's ability to discriminate between classes.[78] Clustering algorithms, more specifically DBSCAN and K-means clustering, have been deployed in smartphone GPS systems and Wi-Fi-network sensors to extract meaningful location features such as frequented location clusters,[79] location patterns,[80] and mobility patterns.[81] These studies demonstrate that clustering algorithms are a powerful method for reducing the number of observations into a smaller number of artificial variables that account for the variance within the dataset.

DEEP LEARNING

The performance of ML models can be limited by the development of manual and arbitrary features, and this potential obstacle can be overcome by DL algorithms. DL algorithms eliminate the need for manual feature engineering, as the DL layers can translate the data into more compact and intermediate abstractions of the data, which in turn can be used as features to predict the final output.[82] While DL can reduce the need for manual data preprocessing and feature extraction, which can potentially improve the generalizability and robustness of a model, the interpretation of the DL model is difficult, as the abstracted features may not be explainable by clinicians. However, it is important to note that the discriminative power of the DL-derived abstractions is strongly influenced by the architecture of the DL algorithm, which is also dependent on the trial-and-error process.[59] Due to DL's representation learning, DL is data-hungry, and therefore requires more data than other ML algorithms.[83,84] For clinical trial data, because of technological limitations and small sample sizes, there may not be enough data to train a sufficiently representative DL model.[76,83]

Four studies used DL to engineer features using time series data.[23,85–87] These models were used to extract gait features from accelerometer data [85,87] and tremor characteristics from IMU data.[23,86] However, it should be noted that the DL models do not always outperform the 'shallow learning' models, as shown in a study by Juen et al. in which smartphone accelerometers were used to predict natural walking speed and distance during a six-minute walk test.[85]

Feature Selection

In recent decades, high-dimensional clinical datasets have relied on feature selection.[88] Feature selection is the process of selecting a subset of the most informative features that will be processed by the ML algorithm.[89] Reducing the features for analysis has both computational and practical benefits. Selecting features can limit storage requirements, increase the algorithm processing speed, increase the interpretability of a model, and improve model performance.

OVERFITTING AND UNDERFITTING

Overfitting and underfitting are common pitfalls for ML models. Overfitting refers to when a ML model fits too well to its training dataset and is unable to generalize its patterns to unseen data. This problem can occur when the training dataset is small and not representative of the overall potential data distribution. Additionally, if the training dataset contains many outliers, the ML model may also fit the outlier data. Underfitting occurs when the trained ML model is too simple; therefore, it cannot identify the relationship between the features and the outputs. Underfitted models will perform poorly for both the train and validation datasets. To address overfitting, reducing the number of features considered by the model or updating the model architecture to include fewer features can be effective.[90] Underfitting can be improved by adding more features considered by the model or by updating the model architecture to increase the complexity of the feature space.[90]

Feature selection identifies the most important features in the dataset and eliminates the irrelevant ones, which thereby reduces noise. However, it is important to strike a balance, as strict feature selection may remove important signals from the data. Therefore, selecting the optimal set of features is important for preventing over- and underfitting. In the following sections, we will elaborate on the three general methods of feature selection that are suitable for ML models.[75]

FILTER METHODS

Filter methods are used during preprocessing prior to training the ML model. Filtering involves removing features based on domain knowledge, missing data, low variance, or correlation.[89,91,92] As filter methods are independent of any model that is to be used in later steps, they are typically faster to implement and reduce the need for repeating feature selection for different ML models. In our selected studies, we found five studies that used Analysis of Variance (ANOVA), Pearson's Correlation, or Spearman's Correlation to identify features that were statistically significant predictors of the outcomes.[24,93–96] p-value based feature selection, while commonly used in clinical studies, is not always suitable for training a ML model. The use of p-values to identify statistically significant features was a popular approach that relied on the belief that insignificant features were not informative. However, important features can be missed when sample sizes are small. Furthermore, p-values can be biased towards low values due to the increased risk of type 1 errors during multiple comparisons, which in turn increases the probability of random variables being included into the final statistical model.[97,98] Additionally, p-value based feature selection methods may be based on certain assumptions that may not be applicable to ML models, such as assuming that the distribution of scores for the groups among the independent variables are the same.[99]

We wanted to highlight one filtering method identified in our selected studies: Relief.[100] Relief is a feature selection technique that also ranks features and selects only the top-scoring features; however, it is notably sensitive to feature interactions.[101,102] Yaman et al. first obtained 177 speech-related features and used Relief to select 66 most predictive vocal biomarkers for the classification of PD.[103] Rodriguez-Molinero used Relief to select frequency features that were subsequently used to predict gait disturbances among PD patients.[104] Overall, Relief has demonstrated its effectiveness in selecting relevant features in various studies related to the prediction of PD using high-dimensional clinical datasets.

EMBEDDED METHODS

The embedded method is a feature selection technique integrated into the ML algorithm itself and is commonly seen in penalized regression.[105] Penalized regression algorithms aim to learn the optimal coefficients for each feature by minimizing its loss function. Regularization (also known as penalization) limits the learning process of the model by increasing the penalty of the loss function.[106] The two common penalized regression methods, identified in the selected studies, are LASSO (also known as L1 penalization) (N = 9) [22,24,29,33,42,95,100,101,107,108] and Ridge (L2 penalization) (N = 2).[109,110] An advantage of LASSO is that it eliminates non-informative features by reducing their coefficients to zero. The first limitation of LASSO is that, if the number of features f is greater than the number of observations o, LASSO will select a maximum of o predictors as non-zeros, regardless of the relevance of other features. The second limitation is that LASSO also suffers from collinearity; hence, if two or more variables are highly correlated, then LASSO will randomly select one feature and penalize the other correlated features. A disadvantage of Ridge is that it only reduces the weights of the non-informative features by reducing their coefficients towards zero, but it never reduces the number of variables. Therefore, all predictors are included in the final model. However, because of this approach, Ridge protects ML models from overfitting.[111]

WRAPPER METHODS

Wrapper methods rely on a stand-alone model to select features, but the performance of the selected features is reflected in the performance of the trained model.[112] The wrapper method algorithms tend to be greedy

search algorithms that aim to select the optimal feature subset by iteratively selecting the features based on ML performance. As the wrapper method is an iterative process and the model must be evaluated on each feature subset combination, this method is computationally expensive. Wrapper-based feature selection can be completed by ranking the features in terms of relative importance using a ML model (such as decision trees or random forests).[88,101,113] We identified a handful of feature ranking methods that include two stepwise regression techniques: Forward Selection and Backwards Elimination,[29,36,52,114–116] as well as Recursive Feature Selection (RFE).[30,117] Forward selection starts the modelling process with zero features and adds a new feature to the model incrementally, each time testing for statistical significance. Backwards elimination starts the modelling process with all features and incrementally removes each feature to evaluate its relative importance in predicting the model output.[97,118] RFE fits a model, ranks the features, and removes the least informative features and continues to remove features until a predefined number of features is met.[64,119,120] Senturk et al. illustrated that RFE-based feature selection increased the prediction accuracy of ANN, CART, and SVM when using vocal data to classify a PD diagnosis.[121]

Machine learning algorithms

ML algorithms build a statistical model based on a training dataset, which can subsequently be used to make predictions about a new, unseen dataset. ML algorithms have been used in a wide variety of clinical trial applications, such as the classification of a diagnoses, classification of physical or mental state (such as a seizure or mood), and the estimation of symptom severity. Within the realm of clinical research, ML algorithms can be broadly divided into two learning paradigms: supervised and unsupervised learning.[122] In this section, we will discuss the model objectives of supervised and unsupervised learning and the specific ML models used to achieve these model objectives.

Supervised ML algorithms use labeled data to map the patterns within a dataset to a known label, while unsupervised ML algorithms do not.[123] Rather, the unsupervised ML algorithms learn the structure present within a dataset without relying on annotations. Supervised learning can be used to automate the labelling process, detect disease cases, or predict clinical outcomes (such as treatment outcomes). There are scenarios when experts or participants can provide labelled data; however, it can become labor-intensive or time-consuming to label every observation. For example, a supervised learning algorithm trained to classify human sounds can be used to automatically annotate and quantify hours of coughs [124] and instances of crying.[125] These algorithms can also be used to differentiate between clinical populations and control participants [95] to identify known clinical population subtypes [23] or classify a clinical event (such as a seizure or tremor).[126] The majority of our selected studies (N = 38) used a clinician to provide the label data. Some studies (N = 22) used a combination of a clinician and self-reported label data, and six studies solely relied on self-reported assessments. Unsupervised ML algorithms can be used to investigate the similarities and differences within a dataset without human intervention. This makes it the ideal solution for exploratory data analysis, subgroup phenotype identification, and anomaly detection. Among digital phenotyping studies, unsupervised learning has been used to identify location patterns [81] and classify sleep disturbance subtypes using wrist-worn accelerometer data.[127]

It is important to recognize that unsupervised and supervised methods are not mutually exclusive, and they can be effectively combined. For instance, unsupervised methods can be employed to extract a meaningful latent representation of the input data. Subsequently, these latent vectors, along with the original inputs, can be used as inputs for a supervised model. This type of approach is commonly observed when applying techniques such as PCA, clustering, or other dimensionality reduction methods.[29,73,74,128] By combining unsupervised and supervised methods, valuable information can be extracted from the data and used to enhance the performance and interpretability of the overall model.

In clinical research, supervised ML algorithms have been used to classify class labels or estimate scores. Classification algorithms learn to map a new observation to a predefined class label. These algorithms can be used to classify patient populations and patient population subtypes and identify clinical events. Regression algorithms learn to map an observation to a continuous output. These algorithms are commonly used to estimate symptom severity,[129] quantify physical activity, and forecast future events.[130] Among the selected papers that were focused on the classification of a diagnosis or state, the four most common algorithms were Random Forest, Support Vector Machine, Logistic Regression, and k-Nearest Neighbors (Figure 4). Some additional classification algorithm families identified were Naïve Bayes, Ensemble-based methods (including Decision Trees, Bagging, and Gradient Boosting), and Neural Networks (such as Convolutional, Artificial, and Recurring Neural Networks). The three most common algorithms for the regression focused papers were Linear Regression (including linear mixed effects models), Support Vector Machine, and k-Nearest Neighbors (Figure 4). We found that most studies only considered or reported a single ML algorithm (N = 32). Additionally, 29 of the studies considered or reported two to five ML algorithms, and the remaining 5 studies considered six or more. The following section provides an overview of the most widely used machine learning models, their properties, advantages, and disadvantages. In addition, we discuss some notable off-the-shelf ML approaches and some custom-built ML methods such as transfer learning, multi-task learning, and generalized and personalized models.

TREE-BASED MODELS

A Decision Tree (DT) is a supervised non-parametric algorithm that is used for both classification and regression. A DT algorithm has a hierarchical structure in which each node represents a test of a feature, each branch represents the result of that test, and each leaf represents the class label or class distribution.[131,132] A Random Forest (RF) algorithm is a supervised ensemble learning algorithm consisting of multiple DTs that aims to predict a class or value.[133] Ensemble learning algorithms use multiple ML algorithms to obtain a prediction.[134] Tree-based models have several benefits. As each tree is only based on a subset of features and data and because they make no assumptions about the relationship between the features and distribution, they are not sensitive to collinearity between features, can ignore missing data, and are less susceptible to overfitting (for multiple trees), making the model more generalizable.[135] Another advantage of RF and DT models is that they can support linear and nonlinear relationships between the dependent and independent variables.[136] Further, as the design of the RF models can be interpreted in terms of feature importance and proximity plots, the interpretability of the RF model is feasible. However, a limitation of using tree-based models is that small changes in the data can lead to drastically different models. Additionally, the more complicated a tree-based model becomes, the less explainable a model becomes. However, pruning the trees can help to reduce the complexity of the model.

According to the selected studies, RF is a versatile and powerful model used for classification and regression tasks across multiple datatypes and populations. RF models have been used for the classification of diagnoses among PD patients,[107,110] Multiple Sclerosis,[34,118] and BD and unipolar depressed patients.[45,61] It is also a popular classification model for the classification of states or episodes, such as the detection of flares among Rheumatoid Arthritis or Axial Spondylarthritis patients [32] and tremor detection among PD patients,[137] to quantify physical activity among cerebral palsy patients [138] and detect the moods of BD patients.[69,139] RF regression algorithms have also been used to predict anxiety deterioration among patients who suffer with anxiety.[140]

SUPPORT VECTOR MACHINES

A Support Vector Machine (SVM) is a supervised algorithm that is used for classification and regression tasks. The objective of a SVM is to identify the optimal hyperplane based on the individual observations, also known as the support vectors. For SVM regression, the optimal hyperplane

represents the minimal distance between the hyperplane and the support vectors. Whereas for SVM classification, the objective is to find the hyperplane that represents the maximum distance between two classes.[141] The hyperplanes can separate the classes in either a linear or non-linear fashion.[136] Given that SVM are influenced by the support vectors closest to the hyperplanes, SVM are less influenced by outliers, making them more suitable for extreme case binary classification. The performance of a SVM can be relatively poor when the classes are overlapping or do not have clear decision boundaries. This makes SVM less appealing for classification tasks as inter class similarity is low. SVM are computationally demanding models as they compute the distance between each support vector; hence, SVM do not scale well for large datasets.[142]

SVM classifiers have been used to classify clinical populations (e.g., facial nerve palsy and their control participants).[143] SVM classifiers have also been used to classify events or states, such as detecting gait among PD patients [104] and classifying seizures among epileptic children.[144] We identified studies that used SVM regression to estimate motor fluctuations and gait speed among PD and Multiple Sclerosis patients, respectively.[74,145]

K-NEAREST NEIGHBORS

A k-Nearest Neighbor (K-NN) algorithm is a non-parametric supervised learning approach that can be used for multi-class classification and regression tasks. Classification K-NN algorithms determine class membership by the plurality vote of its nearest neighbors. They can estimate the continuous value of an output by calculating the average value of its nearest neighbors.[136] Given this, the quality of predictions is not only dependent on the amount of data but also on the density of the data (the number of points per unit). K-NN is simple to implement, intuitive to understand, and robust to noisy training data. However, the disadvantage is that K-NN is computationally slow when it is faced with large multi-dimensional datasets. Further, K-NN does not work well with imbalanced datasets, as under- or over-represented datapoints will influence the classification.[146]

The most popular application for K-NN algorithms is for wearable-based time series data. K-NN classification models have been used to classify PD and healthy controls,[24] classify tremor severity,[147] predict acute exacerbations of chronic obstructive pulmonary disease (AECOPD),[44] and identify mood stability among BD and MDD patients.[33,69,148] Using wearable data, K-NN regression models have been used to predict the deterioration of symptoms associated with anxiety disorder.[140]

NAÏVE BAYES

A Naïve Bayes (NB) classifier is a supervised multi-class classification algorithm. NB classifiers calculate the class conditional probability—the probability that a datapoint belongs to a given class in the data.[141,149] NB classifiers are computational efficient algorithms; thus, they are suitable for real-time predictions, scale well for larger datasets, and can handle missing values. A limitation of NB is that it assumes that all features are conditionally independent; hence, it is recommended that collinear features are removed in advance. Another limitation is that when new feature-observation pairs do not resemble the data in the training data, the NB assigns a probability of zero to that observation. This approach is particularly harsh, especially when dealing with a smaller dataset. Hence, the training data should represent the entire population.

As NB classifiers help form classification models, we found that NB classifiers have been used for the classification of tremors or for freezing gait among PD patients,[52] as well as to classify flares among Rheumatoid Arthritis and Axial Spondylarthritis patients [32] and classify bipolar episodes and mood stability among BD and MDD patients.[33,69,148]

LINEAR AND LOGISTIC REGRESSION

A Linear Regression model is a supervised regression model that predicts a continuous output. It finds the optimal hyperplane that minimizes the sum of squared difference between the true data points and the hyperplane. A Logistic Regression model is a supervised classification model that can be used for binomial, multinominal, and ordinal classification

tasks. Logistic Regression classifies observations by examining the outcome variables on the extreme ends and determines a logistic line that divides two or more classes.[136] Linear and Logistic Regression are popular in algorithms as they are easy to implement, efficient to train, and easy to interpret. However, a limitation of both models is that they make multiple assumptions, e.g., that a solution is linear, the input residuals are normally distributed, and that all features are mutually independent.[150] Multicollinearity, the correlation between multiple features, and outliers will inflate the standard error of the model and may undermine the significance of significant features.[151] Further, outliers that deviate from the expected range of the data can skew the extreme bounds of the probability, making both algorithms sensitive to outliers in the dataset.[150]

Linear Regression has been used to quantify tremors among Essential Tremor (ET) patients [116] and to estimate motor-related symptom severity among PD patients.[31,93] It has also been used to forecast convergence between body sides for Hemiparetic patients.[130] Logistic Regression was a popular approach for classifying PD diagnosis,[107,110] Post-Traumatic Stress Disorder,[109] and distinguishing fallers and non-fallers.[152] Logistic Regression has been used to classify drug effects, such as predicting the pre- and post-medication states among PD patients.[22]

NEURAL NETWORKS

Neural Networks (NN), also known as Artificial Neural Networks (ANN), can be used for unsupervised and supervised classification and regression tasks.[153] NN consists of a collection of artificial neurons (or nodes). Each artificial neuron receives, processes, and sends the signal to the artificial neuron connected to it. The neurons are aggregated into multiple layers, and each layer performs different transformations on the signal. The signal first travels from the input layer into the output layer while possibly traversing multiple hidden layers in between. NN offer several advantages, such as the ability to detect complex non-linear relationships between features and outcomes and work with missing data, while it also requires less preprocessing of the data and offers the availability of multiple training algorithms. However, the disadvantages of NN include increased computational burden, reduced explainability and interpretability (as NN are 'black box' in nature), and the fact that NN are prone to overfitting.[154] However, it is important to highlight the growing number of studies that specifically explore explainable deep learning approaches for biomarker discovery and development. Studies utilizing methodologies such as LIME (LIME Tabular Explainer), SHAP (SHAPley Additive exPlanations), and other visual inspections of feature distribution and importance have aided clinicians in understanding the model mechanisms. These approaches also provide patient-specific insights by describing the importance of each feature, which may, in turn, facilitate personalized treatment opportunities.[90,155–157]

The most popular applications for neural networks were for the classification of a diagnosis or classification of a state or event. The most popular application is the detection of tremors among PD patients.[23,52,86,137,158] NN have been used to classify unipolar and bipolar depressed patients based on motor activity,[45,159] estimate depression severity,[159] forecast seizures,[160] and classify a treatment response using keyboard patterns among PD patients.[161]

TRANSFER LEARNING

Transfer learning (also known as domain adaption) refers to the act of deriving the representations of a previously trained ML model to extract meaningful features from another dataset for an inter-related task.[162] One applicable scenario is the training of a supervised ML model on data collected in a controlled setting (such as in a lab or clinic). The performance of the model may suffer when applied to a dataset collected under free-living conditions. Rather than developing a new model trained solely on a free-living condition dataset, transfer learning can use patterns learned from the controlled setting dataset to improve the learning of the patterns from the free-living conditions dataset.

Transfer learning can also be a valuable technique for enhancing the utilization of limited or rare data.[163] One practical application is to employ

pretraining on abundant control data and subsequently finetune the model on the specific population of interest to improve the model's performance.[163–165] This approach not only optimizes the efficiency of utilizing scarce data but also facilitates model personalization. By adapting a pretrained model to individual characteristics or preferences, it becomes possible to create personalized models that better cater to unique needs or circumstances. Transfer learning thus offers a powerful means to leverage existing knowledge and make the most of available data resources, enhancing both the efficiency and personalization of biomarkers.

Given its application, transfer learning reduces the amount of labeled data and computational resources required to train new ML models,[162] thus making this method advantageous when the sensor modalities, sensor placements, and populations differ between studies. While we only identified two studies that applied transfer learning to estimate PD disease severity using movement sensor data,[166,167] we predict that the application of transfer learning will enable future researchers to overcome the challenges of a limited dataset and develop more sensitive and effective ML models.

MULTI-TASK LEARNING

Multi-task learning (MTL) enables the learning of multiple tasks simultaneously.[168] Learning the commonalities and differences between multiple tasks can improve both the learning efficiency and the prediction accuracy of the ML models.[168] A traditional single-task ML model can have a performance ceiling effect, given the limitations of the dataset size and the model's ability to learn meaningful representations. MTL uses all available data across multiple datasets and can learn to develop generalized models that are applicable to multiple tasks. To use MTL, there should be some degree of information shared between or across all tasks. The correlation allows MTL to exploit the underlying shared information or principles within tasks. Sometimes MTL models can perform worse than single-task models because of 'negative transfers'. This occurs when different tasks share no mutual information or if the information of tasks are contradictory.[169] MTL models have been used to simultaneously model data sourced from two separate sources or to model multiple outcomes.[170,171] For example, Lu ET AL. explored the use of MTL to jointly model data collected from two different smartphone platforms (iPhone and Android) to jointly predict two different types of depression assessments (QIDS and a DSM-5 survey).[79] They illustrated that the classification accuracy of the MTL approach outperformed the single-task learning approach by 48%; thus, the classification model benefited from learning from observations sourced from multiple devices.

GENERALIZED VERSUS PERSONALIZED

ML algorithms can be trained on population data or individual subject data. Generalized models, which are trained on population data, are fed data from all participants for the purpose of general knowledge learning. Conversely, personalized models are trained on an individual's data and take into consideration individual factors such as biological or lifestyle-related variations.[172] We have adopted these terms from Kahdemi et al.'s study, in which they developed generalized and personalized models for sleep-wake prediction.[173] The heterogenous nature of each population or individual can be a potential hinderance for generalizable models. A single individual's deviation from the 'norm' may be viewed as a source of 'noise' in a generalized model. For example, patients with mood disorders such as MDD and BD have large inter-individual symptom variability. Abdullah ET AL., reliably predicted the social rhythms of BD patients with personalized models using smartphone activity data.[30] Cho et al. compared the mood prediction accuracy of personalized and generalized models based on the circadian rhythms of MDD and BD participants.[38] Their studies illustrated that their personalized model predictions were, on average, 24% more accurate than the generalized models. These studies lay the groundwork for developing personalized models that are more sensitive to individual differences.

MODEL HYPERPARAMETERS

The process of building an effective ML model consists of two main steps: selecting the appropriate ML algorithm and optimizing the model performance by tuning its parameters. Each model consists of two types of parameters:

- The parameters that are initialized and continuously updated throughout the learning process (e.g., the weights of neurons of a neural networks).
- The hyperparameters that must be set prior to the learning process as they define the model architecture (e.g., the regularization parameters of a Linear Regression model, and the learning rates of a neural network).[174]

Every combination of the selected hyperparameters will have a direct influence on the performance of the learned model. For example, as the number of trees in a RF increases, the more features tend to be selected by the model, which may not always be relevant for the development of biomarkers.[175] Similarly, the number of layers, number of neurons per layer, activation functions, and the regularization techniques used for NN can each influence the model performance.[176] While most ML algorithms come with default values for the hyperparameters, these may not be optimal for the dataset at hand, and even tuned hypermeters are at risk of being non-optimal for a different dataset. The process of selecting the optimal hyperparameter configurations is known as hyperparameter tuning.[177]

To identify the optimal hyperparameters for a model, researchers must define the hyperparameter space and the hyperparameter search strategy. When defining the hyperparameter space, the distribution of the hyperparameter ranges can be either uniform or logarithmic. The uniform distribution assigns equal probability to all hyperparameter values within a manually defined range. The log-uniform distribution samples hyperparameter values uniformly between the logarithmic transformations of the lower and upper thresholds. We argue that log-uniform distribution is particularly useful when exploring values that vary over several orders of magnitude. Consider the example of tuning a linear regression model with the hyperparameter alpha, which determines the strength of regularization. To efficiently explore a wide range of alpha values, such as between 0.001 and 10, the log-uniform distribution allows for an evenly distributed search space over different orders of magnitude. Log-uniform distribution can be used for the initial exploration of a large range of hyperparameter values. The range can then be narrowed down to explore with a uniform-distribution to determine the optimal hyperparameters for the respective models.

The manual tuning of hyperparameters is impractical due to the large number of available hyperparameters, hyperparameter configurations, and time-consuming model evaluations. Automated tuning approaches are preferred, and there are a wide variety of approaches available, including GridSearch, RandomSearch, and Bayesian Optimization.[177] GridSearch uses brute force to test a finite combination of hyperparameters to identify the optimal hyperparameter configuration.[178] This approach can suffer from the effects of dimensionality, as more potential hyperparameter configurations can be time-consuming and computationally expensive. An alternative to GridSearch is RandomSearch. RandomSearch only samples a subset of all possible hyperparameter configurations within a specific time or computational budget.[179] While RandomSearch only relies on a subsample of configurations, it has been shown to outperform the GridSearch method.[179] As GridSearch and RandomSearch do not consider previous performance evaluations for their hyperparameter optimization strategy, they are inefficient in exploring the hyperparameter search space. Bayesian Optimization, which uses Bayes Theorem, is a powerful approach. It considers previous hyperparameter evaluations to choose which hyperparameters to evaluate next and disregards potential hyperparameter combinations that are deemed irrelevant.[178] This approach reduces the time and computations required for hyperparameter tuning. The benefit of using these more automated

approaches to hyperparameter tuning is three-fold. First, it reduces the time effort required to optimize a ML model. Next, the performance of the ML models is improved as the hyperparameters explore different optimal model configurations for different datasets. Finally, when the hyperparameters and their ranges (together also referred to as the hyperparameter space) and the hyperparameter tuning methods are reported, the models and the findings become reproducible.[180] When similar hyperparameter tuning processes can be used for different ML algorithms for different datasets, researchers can then identify the optimal ML model.

Among the selected studies, 25 discussed which hyperparameters were considered for their models,[23,24,34,43,44,46,53,69,73,86,87,94,95,107–110,114,138,158,159,181–184] of which one stated they used the default hyperparameters of the models.[69] Only nine studies discussed how they selected or optimized their hyperparameters. We identified four studies that stated GridSearch was used for the hyperparameter tuning.[36,46,95,110] We did not identify any studies that used RandomSearch or Bayesian Optimization. The limited reporting of hyperparameters and the hyperparameter tuning process poses a problem for the transparency, reproducibility, and comparison of ML models.

Model evaluation

Assessing a ML model's performance is an essential component for determining the usability and reliability of the model. Depending on the objective of the research, it is often necessary to try to compare the performance of multiple ML models to identify the optimal model.[185,186] In ML, the terms metric and measure are often used interchangeably, but they do have slightly different meanings. A metric is a function used to evaluate the performance of a model, while a measure is a numerical summary of the performance of a model obtained using one or more metrics. It is best practice to use multiple metrics and model performance visualizations for the model evaluation, as a model may perform well for one evaluation metric and poorly for another. Using multiple evaluation metrics ensures that the model is operating optimally and correctly. The following sections provide more details about the performance metrics used for classification and regression models. Table 4 provides an overview of the most common performance metrics used in the selected studies, their respective calculations, and their clinical interpretations.

CLASSIFICATION MEASURES

Classification models have discrete outcomes; thus, a metric must reflect how often an observation belongs to the correct label or class.[187] There are three categories of classification measures: Threshold Metrics, Ranking Metrics, and Error Metrics. Threshold Metrics (such as accuracy and F1 score) quantify the prediction errors of the classification model as a ratio or rate. Ranking Metrics (such as the Receiver Operating Characteristics (ROC) and Area Under the Curve (AUC)) focus on evaluating classification models based on how effective they can discern separate classes. Error Metrics (such as Root Mean Square Error) quantify the uncertainty of the classification model's predictions. While the Threshold and Ranking Metrics are focused on correct and incorrect predictions, the Error Metrics quantify the proportion of classification errors.

As ML models are increasingly being used to perform high-impact tasks pertaining to clinical assessments, an evaluation metric must be selected based on what the stakeholders find to be important regarding the model prediction, which can make the selection of the model metrics challenging. As seen in Table 4, accuracy, sensitivity, specificity, and precision are calculated based on four test results. The True Positive (TP) and True Negative (TN) indicate the presence or absence of a diagnostic or characteristic. The False Positive (FP) and False Negative (FN) indicate the opposite of the true condition.

Binary classification models typically involve a decision threshold hyperparameter that determines how the model assigns labels based on the predicted probabilities. The default threshold is typically 0.5, meaning that if the predicted probability is greater than 0.5, the positive label is assigned, and vice versa. However, it is important to note that this

threshold can be adjusted to accommodate specific needs or domain considerations. To evaluate the performance of binary classification models across different decision thresholds, the ROC curve is commonly used. The ROC curve provides an overview of the model's performance by illustrating the trade-off between TP and FP rates at various threshold values. ROC can aid the assessment of the model's performance across a range of decision thresholds and enable the selection of the threshold that aligns with a specific objective.

It is worth noting that many classification metrics, including accuracy, precision, recall, and F1 score, assume binary labels. However, when dealing with multiclass classification problems, another approach is to use one-vs-rest or one-vs-one strategies, wherein the problem is decomposed into multiple binary classification tasks. The performance of the model on each task can then be evaluated using the binary classification metrics, and the results can be aggregated or averaged to provide an overall assessment of the model's performance on the multiclass problem.

Class imbalance can be an obstacle for assessing model performance. In particular, accuracy, AUC, ROC, may be sensitive to such imbalances.[188] Hence, when facing class imbalance, there are two approaches to consider: one can choose a metric that accounts for class imbalance or one can choose to balance the classes. Metrics such as balanced accuracy, F1-score, or Matthews Correlation Coefficient (MCC) are common metrics for handling class imbalance, as identified by 15 studies.[23,24,29,36,44,60,61,107,108,110,114,140,159,161,189] Balanced accuracy represents the mean of the sensitivity and specificity, while the F1-score represents the mean of the precision and recall.[190] The MCC measures the correlation coefficient of the binary and even multiclass classes. Therefore, the MCC score is high only if the classification model correctly predicts both the positive and negative predictions.[190,191]

The other approach to handling class imbalances is adjusting the class distribution using oversampling or undersampling. We identified eight studies that used random over/under sampling or SMOTE.[29,44–46,61,95,109,192] Oversampling techniques duplicate the samples of the minority class, while undersampling removes samples of the majority class. However, these techniques also have their disadvantages, as the duplication of multiple samples can lead to overfitting of a model, while undersampling reduces the diverse representation of the majority class. Thus, we would specifically recommend using the Synthetic Minority Oversampling Technique (SMOTE) with Tomek Links or Edited Nearest Neighbor (ENN)—two undersampling techniques.[193,194] SMOTE is first applied to create an artificial minority class to minimize the class imbalance. Next, Tomek Links or ENN can be used to remove samples that are close to the boundaries between the classes, which would further separate the classes.[193,194]

REGRESSION MEASURES

As regression models generate predictions on a continuous scale, the objective is to estimate how close the predictions were to the true values.[195] Among the studies selected, we found that regression models used Distance Metrics and Error Metrics to estimate the strength of the association or the distance between the predicted values and the true values.[29,42,87,93,96,128,152] We would like to emphasize that these metrics are used to compare the performance of the composite biomarkers rather than the performance of the individual features. The most common Distance Metrics were the correlation (also known as R) and the percentage of the variance explained (R2). Both were used to assess the strength of the association between the predicted and true values.[196] There is no rule of thumb for interpreting the strength of R2. While an R2 closer to 1 can be obtained in clinical trials, a low R2 can still be useful with respect to trends in the data. We would like to address two points of caution when using the R2.[185,187] First, it is not always suitable to compare R2 across different datasets, as different clinical populations are likely to differ in their feature variance. Second, the R2 will increase with the number of features. To compensate for this, one may use the adjusted R2 to account for the number of features.[197,198]

The Error Metrics included the Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).[133] The MAE measures the average absolute difference between the true and predicted values. The MAE is easy to interpret and robust to outliers. The absolute difference accounts for negative differences. The MSE squares the error instead of providing the absolute error, which gives more weight to the bigger errors. The MSE is sensitive to outliers and not easy to interpret, as the results will not have the same unit as the output. However, the RMSE provides an estimation of the error in the same units as the output while maintaining the properties of the MSE.[199]

Model validation

In ML, model validation refers to the process of evaluating the generalizability of a trained model on an unseen dataset. Selecting the most appropriate model validation approach depends on the size and characteristics of the datasets. Three datasets are required for model validation: the training, test, and validation datasets. In most cases, the validation dataset can be a subset of the original dataset; however, this can lead to data leakage, which could produce overly optimistic results. Another approach is to create a validation dataset from an independent (but comparable) dataset, which ensures an unbiased and independent evaluation of the ML model. However, a limitation is that the performance evaluation may reflect high variance due to the limited size of the dataset.[200] Moreover, it is crucial to highlight that a participant should only be present in a single dataset, such as the training dataset, and should not simultaneously appear in other datasets such as the testing or validation datasets. When a participant's observations are distributed across multiple datasets, data leakage can occur, compromising the accuracy estimation and its applicability to new participants.[183] As a result, cross-validation on the observation level rather than the participant level is methodologically flawed. Unfortunately, this is a common issue and needs to be accounted for in future studies.[201]

Cross-validation is a popular validation method that uses resampling to train, test, and validate a model using different subsets of the data. The training dataset is used to train the ML model to learn the patterns within a dataset. The validation dataset is used to tune the hyperparameters of the model based on the performance of the ML model trained on the training dataset. The test dataset provides an unbiased estimate of the performance of the final ML model after training and validation. In the scenario when both validation and test datasets are used, the test datasets are only used to assess the model once (via hold-out validation) or multiple times (via nested cross-validation). In general, datasets need to meet two main requirements. The datasets should not have shared or overlapping observations to ensure that data leakage does not lead to bias in the estimates, and all observations must be statistically independent.[202] When applying feature engineering or feature selection with cross-validation, any transformation or selection steps should be performed within each fold of the cross-validation to prevent biasing in the training of the prediction model with information from the test dataset.[203] The overall performance of the prediction models, obtained by averaging across each iteration of the cross-validation, evaluates the effectiveness of the combined feature reduction and learning methods in estimating the label for a given dataset.

Among the selected studies, we found that the most popular cross-validation methods were k-fold cross-validation (N = 27), Leave-One-Out cross-validation (N = 16), and custom validation (N = 8). Overall, 15 studies did not report the use a validation method. K-fold cross-validation randomly splits the datasets in 'k' folds; one-fold is used for testing and the remaining folds are used for training. This step is repeated until every unique fold has been used as the test dataset, and the overall performance is based on the average of the performance of each model in each fold.[204] Leave-one-out cross-validation is a specific type of k-fold cross-validation, wherein individual observations (or participants) are the test datasets, and the remaining cases are used for training. Leave-one-out cross validation prevents data leakage across datasets, as repeated

measurements of the same subjects can lead to the violation of independence assumption for ordinary cross-validation.[204–206]

We would like to highlight the advantages of the nested cross-validation approach. While nested cross-validation was the least popular approach, we would argue that nested cross-validation is a more robust approach for selecting and evaluating a ML model.[207] Currently, the model section without the nested cross-validation approach uses the same data to both tune the model hyperparameters and evaluate its performance. Therefore, information is 'leaked' between the training and validation of the model, which can lead to overfitting.[207] Nested cross-validation consists of an inner loop and an outer loop. The outer loop assesses the model performance, while the inner loop assesses the hyperparameter selection.[207] Each iteration of the outer loop is split into a different combination of training and test sets. The outer loop training set is used in the inner loop, which is further split into a training and validation dataset. The inner loop split is repeated over k-folds, and the best performing model across the k-folds is evaluated in the outer loop. This ensures that different data are used to optimize the models' hyperparameters and evaluate the model's performance. The final model performance represents the average and standard deviation of the model performance as selected by each of the outer loops. Without the standard deviation or confidence intervals, it is not possible to evaluate the spread or stability of the prediction error of the given models.[208,209]

It is important to highlight that cross-validation is only used to approximate the generalization error of the models built and not to build the final model that will be used for making predictions.[205,210] The average prediction error across the folds gives an expected error for a single model built on the single dataset. If the variance of the prediction error is too high, then the model is considered unstable. To select a single model, it is recommended that researchers rebuild the model using the full dataset.[211] If an external validation set is available, then this validation set can be used to evaluate and compare the single prediction error to that of the cross-validation prediction error.

Recommendations

In this recommendation section, we address the main issues consistently identified in the selected studies and how to amend these issues for future trials (see Figure 5 for a simplified overview of these recommendations). It is important to bear in mind the regulatory implications for developing ML-derived biomarkers. Within the European Union, AI medical systems and devices are considered high risk; therefore, they are subject to stringent reviews prior to being made available on the market.[212] These review requirements emphasize the importance of achieving high levels of performance, transparency, and minimal risk in ML-derived biomarker development.[213] High performance implies that the developed ML models must be accurate, robust, and capable of reliably and consistently predicting the target outcome variable. Furthermore, transparency in ML-derived biomarker development refers to the provision of clear and adequate information to the user, including appropriate human-readable measures to minimize risks associated with the use of the system. The development of ML-derived biomarkers must also aim to minimize risks and discriminatory outcomes, which can be achieved by training the ML model on high-quality datasets that are representative of the target population and by conducting adequate risk assessment checks.[214] These considerations are critical for ensuring the safe and effective use of ML-derived biomarkers in clinical practice.

INCLUSION OF HEALTHY CONTROLS

When conducting a study focused on disease classification or estimation, the inclusion of control data can serve several purposes. By comparing the data from individuals with the condition of that of the healthy controls, researchers can discern whether the observed differences are specific to the condition or a result of unrelated factors. Moreover, analyzing the performance of a model on control subjects can shed light on the biomarker's effectiveness and reliability. By evaluating how well the model distinguishes between healthy controls and patients with the

condition, researchers can gain a better understanding of its predictive capabilities. This evaluation can provide insights into potential false positives or false negatives that may occur when using the model in real-world settings.

It is worth noting that, when including control data, the control data should be appropriately matched with the patient population data. Having age- and gender-matched control subjects can help minimize confounding variables, improving the accuracy of the analysis. This matching process allows researchers to draw more robust conclusions about the relationship between the identified features or patterns and the disease activity while also reducing the potential impact of demographic factors on the results.

The finding that only half of the studies included healthy controls is significant as it highlights a potential gap or limitation in the existing body of research. Without the inclusion of controls, it becomes challenging to attribute identified features or patterns solely to the CNS disorder or the severity of the condition. Further, if the dataset only contains a relatively homogeneous population, it calls the reliability and predictive capabilities of the models into question. We encourage future researchers to include control subjects in their studies, as it would improve the strength of their biomarkers and the validity of their findings.

DATA QUALITY AND PREPROCESSING

The remote monitoring of clinical trials can generate large and complex datasets that include longitudinal data from multiple subjects and data sourced from multiple sensors, resulting in a multi-dimensional data structure. To this point, we recommend using the WHO MHEALTH Technical Evidence Review Groups' MHEALTH evidence and evidence reporting and assessment (MERA) 16-item checklist to provide transparency on which MHEALTH invention was used, where, and how it was implemented to support the reproducibility of the MHEALTH data collection.[215] To ensure the quality and reliability of the data, it is important to assess the quality of the data. This assessment includes examining the data for missing and outlier data and understanding how these factors might affect the generalizability and reproducibility of the ML model. While most studies provide detailed information on patient populations, the devices used, and the data collected, they often underreport information related to data quality and preprocessing steps. Therefore, it is important to provide sufficient details on the methods used to preprocess the data, including the quantity of missing and outlier data and the strategies employed to handle such data. This information can ensure that the data collection and preprocessing process can be reproduced, which, in turn, can enhance the credibility and generalizability of the ML model.

FEATURE ENGINEERING AND SELECTION

There is a wide variety of manual or automated techniques used for engineering and selecting features to feed a model. ML models perform best when feature engineering and selection are leveraged to formulate potentially clinically relevant features from existing data. In addition, the performance of the ML model can be optimized, and the computational time can be reduced when the redundancy across the features is reduced. While only selecting the most informative features can remove noise (therefore reducing the likelihood of overfitting), selecting too few features may reduce the strength of the (combined) signal in the dataset, making the ML model vulnerable to underfitting. Feature engineering and selection can be guided by domain expertise and/or automated statistical models, where multiple features are evaluated by their importance in predicting the outcome. While automated feature engineering techniques, such as clustering, PCA, and DL, can be used to extract a reduced set of representative features, this risks a potential decline in interpretability, which may limit its clinical application.

MODEL CONFIGURATION AND OPTIMIZATION

When selecting the ML models, there are several factors that should be considered, such as model objectives, model types, model hyperparameters, and model evaluation. Poor design choices and lenient

hyperparameter tuning and validation in these steps can lead to poor model performance. We recommend that researchers carefully consider each step of building their ML pipeline by comparing multiple ML algorithms, using automated methods for assessing multiple hyperparameter configurations, and using nested cross validation to both optimize and validate the ML models.

MODEL VALIDATION

We would recommend using a minimum of three datasets to validate a ML model and train, validate, and test a dataset. At no point should the test set be used for the model configuration, which includes the data transformation, feature engineering, and selection, or the tuning of the hyperparameters. The test dataset could either be a subset of the original data (with no overlapping subjects or observations) or a separate external dataset. The use of an external dataset is ideal as this ensures that there is no influence of bias during the data collection period and that there is no data leakage between the datasets. If an external dataset is not available or if the dataset is not sufficiently large, we recommend nested cross-validation. This resampling method supports model hyperparameter tuning and performance evaluation without the risk of data leakage across the dataset.

It is crucial to report the evaluation metric results for each dataset. In the case of cross-validation reporting, we recommend that researchers report the distribution of the performance measures (e.g., the mean and standard deviation or median and 95% confidence interval) across the folds to show the average and variability of the performance of the models. As cross-validation evaluates the prediction error across multiple ML models, we would also recommend reporting the performance of the final model selected. This is achieved by re-training a ML model on the full dataset and evaluating the performance on an external dataset.[207,210] This would give insight into how well the model would perform under different circumstances. We also highly recommend using multiple evaluation metrics for assessing the model's performance. Seeing as a model might excel for one metric and fail for another, this underscores the need for comprehensive evaluation. Employing multiple metrics ensures optimal operation and reduces the likelihood of blind spots.

Once the final model has been trained, there are three approaches to choose from to apply the model to a new target dataset. The first approach is to test the model 'as-is', implying that the ready-made model can be used in its original state without modifications.[216] In the second scenario, the train data and the target data may have different characteristics, which may lead to a distribution shift. The type of distribution shift between the two datasets can occur for many reasons, including different MHEALTH devices used for data collection, environmental noise, and sampling bias.[217] When this occurs, transfer learning can be used to fine-tune the ready-made model and update its weights to better suit the target dataset.[216] In the third scenario, the target dataset may have different requirements than the original training dataset.[216] As a result, the decision boundary of the classification model can be altered, such as optimizing the model for a sensitivity of 90% instead of accuracy. Whether testing the model as-is, employing transfer learning, or adjusting the decision boundary, these strategies offer flexibility in adapting the model to different settings and improving its performance for validation purposes.

MODEL REPRODUCIBILITY AND INTERPRETABILITY

Equally important as the model performance are the ML models' reproducibility and interpretability. Reproducibility is a core component for ensuring that a ML model can be validated and reused by clinical researchers. Technical reproducibility involves using the same computational procedures to produce consistent model outcomes. Statistical reproducibility ensures that the model demonstrates similar statistical performance across different subsets of data. Conceptual reproducibility refers to achieving consistent results under new conditions, such as data collected from different settings.[216] Transparency regarding data quality, feature engineering and selection methods, the hyperparameters considered and selected, and the model validation protocol can help ease the

ability of the scientific community to recreate the work in the published literature. Best practices for reproducibility include publishing the code on GitHub or by publishing FAIR metadata.[211,218,219]

Given the potential clinical application of ML models, prior to modeling, researchers should determine the model's interpretability requirement. While ML models provide researchers with what was predicted, interpretability requires that the model can explain why it made the prediction.[185] Interpretability enables us to understand the causal relationships between the data and the ML model's predictions. There are two situations in which the interpretability of a model is required: when an inaccurate prediction can have severe or even fatal consequences for the patients (such as a misclassified diagnosis [220]) and when the interpretability can be used to identify novel relationships between clinical factors and the predicted outcome (such as factors influencing treatment outcomes [221]). There can be two situations in which interpretability is not required: situations in which incorrect predictions do not have severe consequences (such as counting the number of coughs [222]) or situations in which the ML model has been sufficiently validated in real clinical applications, even if the predictions are not perfect.[223] While black box models may offer more accurate predictions than an interpretable model, they only provide limited insight into how the predictions were made. Therefore, both interpretable and black box models have their respective merits.

There are two broad approaches towards achieving interpretability.[224] One approach is to use easy-to-interpret models, such as Linear or Logistic Regression, where the coefficients of the features can provide insight into the features' associations with the predicted outcome. The other approach is to use explanation methods for explaining complex or black box models, such as SHAPley Additive exPlanations plots (SHAP), Local Interpretable Model-agnostic Explanations (LIME), or Anchors.[224] We recommend that researchers report whether their final selected model was an interpretable model or a black box.[225] If it was interpretable, we recommend discussing what interpretations can be derived from the models.

Conclusions

The rise and breadth of ML applications in clinical trials highlight the increasing reliance and importance of ML in the development of novel biomarkers.[226] While the advances in ML applications have demonstrated great potential for innovative biomarker development, the process of its development is not well documented, which, in turn, limits the reproducibility of these findings. This review has illustrated the steps taken to translate raw data from MHEALTH technologies into meaningful clinical biomarkers using ML. Given the lack of consistent reporting in the ML methods, the present review cannot provide a complete or detailed picture of the notable and generic practices. However, the authors have provided an overview of the status quo of the development and translation of ML-derived biomarkers in MHEALTH-focused clinical trials. The recommended checklist provided in the review could serve as a foundation for the design of future ML-derived biomarkers in conventional ML practices. By encouraging consistent and transparent reporting, researchers can accelerate the integration of novel biomarkers derived from MHEALTH sensors and ML pipelines into future clinical trials.

REFERENCES

1 Au, R.; Lin, H.; Kolachalama, V.B. Tele-Trials, Remote Monitoring, and Trial Technology for Alzheimer's Disease Clinical Trials. In Alzheimer's Disease Drug Development; Cambridge University Press: Cambridge, UK, 2022; pp. 292–300.

2 Inan, O.T.; Tenaerts, P.; Prindiville, S.A.; Reynolds, H.R.; Dizon, D.S.; Cooper-Arnold, K.; Turakhia, M.; Pletcher, M.J.; Preston, K.L.; Krumholz, H.M.; ET AL. Digitizing clinical trials. NPJ Digit. Med. 2020, 3, 101.

3 Teo, J.X.; Davila, S.; Yang, C.; Hii, A.A.; Pua, C.J.; Yap, J.; Tan, S.Y.; Sahlén, A.; Chin, C.W.-L.; Teh, B.T.; ET AL. Digital phenotyping by consumer wearables identifies sleep-associated markers of cardiovascular disease risk and biological aging. bioRxiv 2019.

4 Brietzke, E.; Hawken, E.R.; Idzikowski, M.; Pong, J.; Kennedy, S.H.; Soares, C.N. Integrating digital phenotyping in clinical characterization of individuals with mood disorders. Neurosci. Biobehav. Rev. 2019, 104, 223–230.

5 Kourtis, L.C.; Regele, O.B.; Wright, J.M.; Jones, G.B. Digital biomarkers for Alzheimer's disease: The mobile/wearable devices opportunity. NPJ Digit. Med. 2019, 2, 9.

6 Bhidayasiri, R.; Mari, Z. Digital phenotyping in Parkinson's disease: Empowering neurologists for measurement-based care. Park. Relat. Disord. 2020, 80, 35–40.

7 Prosperi, M.; Min, J.S.; Bian, J.; Modave, F. Big data hurdles in precision medicine and precision public health. BMC Med. Inform. Decis. Mak. 2018, 18, 139.

8 Torres-Sospedra, J.; Ometov, A. Data from Smartphones and Wearables. Data 2021, 6, 45.

9 García-Santıllán, A.; del Flóres-Serrano, S.; López-Morales, J.S.; Rios-Alvarez, L.R. Factors Associated that Explain Anxiety toward Mathematics on Undergraduate Students. (An Empirical Study in Tierra Blanca Veracruz-México). Mediterr. J. Soc. Sci. 2014, 5.

10 Iniesta, R.; Stahl, D.; Mcguffin, P. Machine learning, statistical learning and the future of biological research in psychiatry. Psychol. Med. 2016, 46, 2455–2465.

11 Rajula, H.S.R.; Verlato, G.; Manchia, M.; Antonucci, N.; Fanos, V. Comparison of Conventional Statistical Methods with Machine Learning in Medicine: Diagnosis, Drug Development, and Treatment. Medicina 2020, 56, 455.

12 Getz, K.A.; Rafael, A.C. Trial watch: Trends in clinical trial design complexity. Nat. Rev. Drug. Discov. 2017, 16, 307.

13 Getz, K.A.; Stergiopoulos, S.; Marlborough, M.; Whitehill, J.; Curran, M.; Kaitin, K.I. Quantifying the Magnitude and Cost of Collecting Extraneous Protocol Data. Am. J. Ther. 2015, 22, 117–124.

14 Getz, K.A.; Wenger, J.; Campo, R.A.; Seguine, E.S.; Kaitin, K.I. Assessing the Impact of Protocol Design Changes on Clinical Trial Performance. Am. J. Ther. 2008, 15, 450–457.

15 Globe Newswire. Rising Protocol Design Complexity Is Driving Rapid Growth in Clinical Trial Data Volume, According to Tufts Center for the Study of Drug Development. Available online: https://www.globenewswire.com/news-release/2021/01/12 /2157143/0/en/Rising-Protocol-Design-Complexity-Is-Driving-Rapid-Growth-in-Clinical-Trial-Data-Volume-According-toTufts-Center-for-the-Study-of-Drug-Development.html (accessed on 12 January 2021).

16 Santos, W.M.D.; Secoli, S.R.; de Araújo Püschel, V.A. The Joanna Briggs Institute approach for systematic reviews. Rev. Lat. Am. Enferm. 2018, 26, e3074.

17 Central Nervous System Diseases—MeSH—NCBI. 2023. Available online: https://www.ncbi.nlm.nih.gov/mesh?Db=mesh& Cmd=DetailsSearch&Term=%22Central+Nervous+System+Diseases%22%5BMeSH+Terms%5D (accessed on 5 January 2023).

18 Martinez, G.J.; Mattingly, S.M.; Mirjafari, S.; Nepal, S.K.; Campbell, A.T.; Dey, A.K.; Striegel, A.D. On the Quality of Real-world Wearable Data in a Longitudinal Study of Information Workers. In Proceedings of the 2020 IEEE International Conference on Pervasive Computing and Communications Workshops, PerCom Workshops 2020, Austin, TX, USA, 23–27 March 2020.

19 Ruiz Blázquez, R.R.; Muñoz-Organero, M. Using Multivariate Outliers from Smartphone Sensor Data to Detect Physical Barriers While Walking in Urban Areas. Technologies 2020, 8, 58.

20 Poulos, J.; Valle, R. Missing Data Imputation for Supervised Learning. Appl. Artif. Intell. 2018, 32, 186–196.

21 Schafer, J.L.; Graham, J.W. Missing data: Our view of the state of the art. Psychol. Methods 2002, 7, 147–177.

22 Evers, L.J.; Raykov, Y.P.; Krijthe, J.H.; de Lima, A.L.S.; Badawy, R.; Claes, K.; Heskes, T.M.; Little, M.A.; Meinders, M.J.; Bloem, B.R. Real-life gait performance as a digital biomarker for motor fluctuations: The Parkinson@Home validation study. J. Med. Internet Res. 2020, 22, e19068.

23 Papadopoulos, A.; Kyritsis, K.; Klingelhoefer, L.; Bostanjopoulou, S.; Chaudhuri, K.R.; Delopoulos, A. Detecting Parkinsonian Tremor from IMU Data Collected In-The-Wild using Deep Multiple-Instance Learning. IEEE J. Biomed. Health Inform. 2019, 24, 2559–2569.

24 Tougui, I.; Jilbab, A.; El Mhamdi, J. Analysis of smartphone recordings in time, frequency, and cepstral domains to classify Parkinson's disease. Healthc. Inform. Res. 2020, 26, 274–283.

25 Meyerhoff, J.; Liu, T.; Kording, K.P.; Ungar, L.H.; Kaiser, S.M.; Karr, C.J.; Mohr, D.C. Evaluation of Changes in Depression, Anxiety, and Social Anxiety Using Smartphone Sensor Features: Longitudinal Cohort Study. J. Med. Internet Res. 2021, 23, e22844.

26 Dinesh, K.; Snyder, C.W.; Xiong, M.; Tarolli, C.G.; Sharma, S.; Dorsey, E.R.; Sharma, G.; Adams, J.L. A Longitudinal Wearable Sensor Study in Huntington's Disease. J. Huntingt. Dis. 2020, 9, 69–81.

27 Cho, C.-H.; Lee, T.; Lee, H.-J. Mood Prediction of Patients with Mood Disorders by Machine Learning Using Passive Digital Phenotypes Based on the Circadian Rhythm: Prospective Observational Cohort Study. 2019. Available online: https://www.ncbi. nlm.nih.gov/pmc/articles/PMC6492069/ (accessed on 5 January 2023).

28 Tanaka, T.; Kokubo, K.; Iwasa, K.; Sawa, K.; Yamada, N.; Komori, M. Intraday activity levels may better reflect the differences between major depressive disorder and bipolar disorder than average daily activity levels. Front. Psychol. 2018, 9, 2314.

29 Palmius, N.; Tsanas, A.; Saunders, K.E.A.; Bilderbeck, A.C.; Geddes, J.R.; Goodwin, G.M.; De Vos, M. Detecting bipolar depression from geographic location data. IEEE Trans. Biomed. Eng. 2017, 64, 1761–1771.

30 Abdullah, S.; Matthews, M.; Frank, E.; Doherty, G.; Gay, G.; Choudhury, T. Automatic detection of social rhythms in bipolar disorder. J. Am. Med. Inform. Assoc. 2016, 23, 538–543.

31 Ramsperger, R.; Meckler, S.; Heger, T.; van Uem, J.; Hucker, S.; Braatz, U.; Graessner, H.; Berg, D.; Manoli, Y.; Serrano, J.A.; ET AL. Continuous leg dyskinesia assessment in Parkinson's disease -clinical validity and ecological effect. Park. Relat. Disord. 2016, 26, 41–46.

32 Gossec, L.; Guyard, F.; Leroy, D.; Lafargue, T.; Seiler, M.; Jacquemin, C.; Molto, A.; Sellam, J.; Foltz, V.; Gandjbakhch, F.; ET AL. Detection of Flares by Decrease in Physical Activity, Collected Using Wearable Activity Trackers in Rheumatoid Arthritis or Axial Spondyloarthritis: An Application of Machine Learning Analyses in Rheumatology. Arthritis Care Res. 2019, 71, 1336–1343.

33 Bai, R.; Xiao, L.; Guo, Y.; Zhu, X.; Li, N.; Wang, Y.; Chen, Q.; Feng, L.; Wang, Y.; Yu, X.; ET AL. 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 2021, 9, e24365.

34 Schwab, P.; Karlen, W. A Deep Learning Approach to Diagnosing Multiple Sclerosis from Smartphone Data. IEEE J. Biomed. Health Inform. 2021, 25, 1284–1291.

35 Aghanavesi, S. Smartphone-Based Parkinson's Disease Symptom Assessment. Licentiate Dissertation, Dalarna University, Falun, Sweden, 2017.

36 Maleki, G.; Zhuparris, A.; Koopmans, I.; Doll, R.J.; Voet, N.; Cohen, A.; van Brummelen, E.; Groeneveld, G.J.; De Maeyer, J. Objective Monitoring of Facioscapulohumeral Dystrophy During Clinical Trials Using a Smartphone App and Wearables: Observational Study. JMIR Form. Res. 2022, 6, e31775.

37 Twose, J.; Licitra, G.; McConchie, H.; Lam, K.H.; Killestein, J. Early-warning signals for disease activity in patients diagnosed with multiple sclerosis based on keystroke dynamics. Chaos 2020, 30, 113133.

38 Cho, C.H.; Lee, T.; Kim, M.G.; In, H.P.; Kim, L.; Lee, H.J. Mood prediction of patients with mood disorders by machine learning using passive digital phenotypes based on the circadian rhythm: Prospective observational cohort study. J. Med. Internet Res.2019, 21, e11029.

39 Little, R.J.A.; Rubin, D.B. Complete-Case and Available-Case Analysis, Including Weighting Methods; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2014; pp. 41–58.

40 Demissie, S.; LaValley, M.P.; Horton, N.J.; Glynn, R.J.; Cupples, L.A. Bias due to missing exposure

data using complete-case analysis in the proportional hazards regression model. Stat. Med. 2003, 22, 545–557.
41 Enders, C.K.; London, N.Y. Applied Missing Data Analysis; Guilford Press: New York, NY, USA, 2010.
42 Zhang, Y.; Folarin, A.A. Predicting Depressive Symptom Severity Through Individuals' Nearby Bluetooth Device Count Data Collected by Mobile Phones: Preliminary Longitudinal Study. Available online: https://www.ncbi.nlm.nih.gov/pmc/articles/ PMC8367113/ (accessed on 5 January 2023).
43 Creagh, A.P.; Dondelinger, F.; Lipsmeier, F.; Lindemann, M.; De Vos, M. Longitudinal Trend Monitoring of Multiple Sclerosis Ambulation using Smartphones. IEEE Open J. Eng. Med. Biol. 2022, 3, 202–210.
44 Wu, C.-T.; Li, G.-H.; Huang, C.-T.; Cheng, Y.-C.; Chen, C.-H.; Chien, J.-Y.; Kuo, P.-H.; Kuo, L.-C.; Lai, F. Acute exacerbation of a chronic obstructive pulmonary disease prediction system using wearable device data, machine learning, and deep learning:Development and cohort study. JMIR mHealth Uhealth 2021, 9, e22591.
45 Jakobsen, P.; Garcia-Ceja, E.; Riegler, M.; Stabell, L.A.; Nordgreen, T.; Torresen, J.; Fasmer, O.B.; Oedegaard, K.J. Applying machine learning in motor activity time series of depressed bipolar and unipolar patients compared to healthy controls. PLoS ONE 2020, 15, e0231995.
46 Lekkas, D.; Jacobson, N.C. Using artificial intelligence and longitudinal location data to differentiate persons who develop posttraumatic stress disorder following childhood trauma. Sci. Rep. 2021, 11, 10303.
47 Richman, M.B.; Trafalis, T.B.; Adrianto, I. Missing data imputation through machine learning algorithms. In Artificial Intelligence Methods in the Environmental Sciences; Springer: Dordrecht, The Netherlands, 2009; pp. 153–169.
48 Jerez, J.M.; Molina, I.; García-Laencina, P.J.; Alba, E.; Ribelles, N.; Martín, M.; Franco, L. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artif. Intell. Med. 2010, 50, 105–115.
49 Lakshminarayan, K.; Harp, S.A.; Goldman, R.P.; Samad, T. Imputation of Missing Data Using Machine Learning Techniques.In KDD Proceedings 1996; AAAI Press: Palo Alto, CA, USA, 1996; Volume 96.
50 Aggarwal, C.C. Data Mining; Springer International Publishing: Cham, Switzerland, 2015.
51 Ledolter, J.; Kardon, R.H. Does Testing More Frequently Shorten the Time to Detect Disease Progression? Transl. Vis. Sci. Technol. 2017, 6, 1.
52 Bazgir, O.; Habibi, S.A.H.; Palma, L.; Pierleoni, P.; Nafees, S. A classification system for assessment and home monitoring of tremor in patients with Parkinson's disease. J. Med. Signals Sens. 2018, 8, 65–72.
53 Williamson, J.R.; Telfer, B.; Mullany, R.; Friedl, K.E. Detecting Parkinson's Disease from Wrist-Worn Accelerometry in the U.K. Biobank. Sensors 2021, 21, 2047.
54 Buda, T.S.; Khwaja, M.; Matic, A. Outliers in Smartphone Sensor Data Reveal Outliers in Daily Happiness. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 2021, 5, 1–19.
55 Buda, T.S.; Caglayan, B.; Assem, H. DeepAD: A generic framework based on deep learning for time series anomaly detection. In Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Berlin/Heidelberg, Germany, 2018; pp. 577–588.
56 Arora, S.; Venkataraman, V.; Zhan, A.; Donohue, S.; Biglan, K.; Dorsey, E.; Little, M. Detecting and monitoring the symptoms of Parkinson's disease using smartphones: A pilot study. Park. Relat. Disord. 2015, 21, 650–653.
57 Guyon, I.; Elisseeff, A. An Introduction to Feature Extraction. In Feature Extraction; Springer: Berlin/ Heidelberg, Germany, 2006; pp. 1–25.
58 Raju, V.N.G.; Lakshmi, K.P.; Jain, V.M.; Kalidindi, A.; Padma, V. Study the Influence of Normalization/ Transformation process on the Accuracy of Supervised Classification. In Proceedings of the 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 20–22 August 2022; pp. 729–735.
59 Dara, S.; Tumma, P. Feature Extraction by Using Deep Learning: A Survey. In Proceedings of the 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 29–31 March 2018; pp. 1795–1801.
60 Tizzano, G.R.; Spezialetti, M.; Rossi, S. A Deep Learning Approach for Mood Recognition from Wearable Data. In Proceedings of the IEEE Medical Measurements and Applications, MeMeA 2020—Conference Proceedings, Bari, Italy, 1 June–1 July 2020; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2020.
61 Garcia-Ceja, E.; Riegler, M.; Jakobsen, P.; Torresen, J.; Nordgreen, T.; Oedegaard, K.J.; Fasmer, O.B. Motor Activity Based Classification of Depression in Unipolar and Bipolar Patients. In Proceedings of the 2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS), Karlstad, Sweden, 18–21 June 2018; pp. 316–321.
62 Liu, H. Feature Engineering for Machine Learning and Data Analytics. In Feature Engineering for Machine Learning and Data Analytics; Taylor & Francis Group: Boca Raton, FL, USA, 2018.
63 Nargesian, F.; Samulowitz, H.; Khurana, U.; Khalil, E.B.; Turaga, D. Learning feature engineering for classification. IJCAI Int. Jt. Conf. Artif. Intell. 2017, 2529–2535.
64 Kuhn, M.; Johnson, K. Feature Engineering and Selection: A Practical Approach for Predictive Models; Chapman and Hall/CRC: Boca Raton, FL, USA, 2019.
65 Ronao, C.A.; Cho, S.-B. Human activity recognition with smartphone sensors using deep learning neural networks. Expert Syst. Appl. 2016, 59, 235–244.
66 Nweke, H.F.; Teh, Y.W.; Al-garadi, M.A.; Alo, U.R. Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges. Expert Syst. Appl. 2018, 105, 233–261.
67 Zdravevski, E.; Lameski, P.; Trajkovik, V.; Kulakov, A.; Chorbev, I.; Goleva, R.; Pombo, N.; Garcia, N. Improving Activity Recognition Accuracy in Ambient-Assisted Living Systems by Automated Feature Engineering. IEEE Access 2017, 5, 5262–5280.
68 McGinnis, R.S.; Mahadevan, N.; Moon, Y.; Seagers, K.; Sheth, N.; Wright, J.A., Jr.; Dicristofaro, S.; Silva, I.; Jortberg, E.; Ceruolo, M.; et al. A machine learning approach for gait speed estimation using skin-mounted wearable sensors: From healthy controls to individuals with multiple sclerosis. PLoS ONE 2017, 12, e0178366.
69 Maxhuni, A.; Muñoz-Meléndez, A.; Osmani, V.; Perez, H.; Mayora, O.; Morales, E.F. Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients. Pervasive Mob. Comput. 2016, 31, 50–66.
70 Yamakawa, T.; Miyajima, M.; Fujiwara, K.; Kano, M.; Suzuki, Y.; Watanabe, Y.; Watanabe, S.; Hoshida, T.; Inaji, M.; Maehara, T. Wearable epileptic seizure prediction system with machine-learning-based anomaly detection of heart rate variability. Sensors 2020, 20, 3987.
71 Fuchs, C.; Nobile, M.S.; Zamora, G.; Degeneffe, A.; Kubben, P.; Kaymak, U. Tremor assessment using smartphone sensor data and fuzzy reasoning. BMC Bioinform. 2021, 22, 57.
72 Aich, S.; Pradhan, P.M.; Park, J.; Sethi, N.; Vathsa, V.S.S.; Kim, H.C. A validation study of freezing of gait (FOG) detection and machine-learning-based FOG prediction using estimated gait characteristics with a wearable accelerometer. Sensors 2018, 18, 3287.
73 Rodríguez-Martín, D.; Samà, A.; Pérez-López, C.; Català, A.; Arostegui, J.M.M.; Cabestany, J.; Bayés, À.; Alcaine, S.; Mestre, B.; Prats, A.; et al. Home detection of freezing of gait using Support Vector Machines through a single waist-worn triaxial accelerometer. PLoS ONE 2017, 12, e0171764.
74 Supratak, A.; Datta, G.; Gafson, A.R.; Nicholas, R.; Guo, Y.; Matthews, P.M. Remote monitoring in the home validates clinical gait measures for multiple sclerosis. Front. Neurol. 2018, 9, 561.
75 Bro, R.; Smilde, A.K. Principal component analysis. Anal. Methods 2014, 6, 2812–2831.
76 Kim, J.; Lim, J. A Deep Neural Network-Based Method for Prediction of Dementia Using Big Data. Int. J. Environ. Res. Public Health 2021, 18, 5386.
77 Clustering. In Principles of Data Mining; Springer: London, UK, 2007; pp. 221–238.
78 Arabie, P.; Hubert, L.J. An Overview of Combinatorial Data Analysis. In Clustering and Classification; World Scientific: Singapore, 1996; pp. 5–63.
79 Lu, J.; Shang, C.; Yue, C.; Morillo, R.; Ware, S.; Kamath, J.; Bamis, A.; Russell, A.; Wang, B.; Bi, J. Joint Modeling of Heterogeneous Sensing Data for Depression Assessment via Multi-task Learning. In Proceedings of the Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies; Association for Computing Machinery: New York, NY, USA, 2018; Volume 2, pp. 1–21.
80 Sabatelli, M.; Osmani, V.; Mayora, O.; Gruenerbl, A.; Lukowicz, P. Correlation of significant places with self-reported state of bipolar disorder patients. In Proceedings of the 2014 4th International Conference on Wireless Mobile Communication and Healthcare-Transforming Healthcare through Innovations in Mobile and Wireless Technologies (MOBIHEALTH), Athens, Greece, 3–5 November 2014; pp. 116–119.
81 Faurholt-Jepsen, M.; Busk, J.; Vinberg, M.; Christensen, E.M.; HelgaÞórarinsdóttir; Frost, M.; Bardram, J.E.; Kessing, L.V. Daily mobility patterns

in patients with bipolar disorder and healthy individuals. J. Affect. Disord. 2021, 278, 413–422.
82 Miotto, R.; Wang, F.; Wang, S.; Jiang, X.; Dudley, J.T. Deep learning for healthcare: Review, opportunities and challenges. Brief. Bioinform. 2018, 19, 1236–1246.
83 Marx, V. The big challenges of big data. Nature 2013, 498, 255–260.
84 Li, Y.; Ding, L.; Gao, X. On the decision boundary of deep neural networks. arXiv 2018, arXiv:1808.05385.
85 Juen, J.; Cheng, Q.; Schatz, B. A Natural Walking Monitor for Pulmonary Patients Using Mobile Phones. IEEE J. Biomed. Health Inform. 2015, 19, 1399–1405.
86 Cole, B.T.; Roy, S.H.; De Luca, C.J.; Nawab, S.H. Dynamical learning and tracking of tremor and dyskinesia from wearable sensors. IEEE Trans. Neural Syst. Rehabil. Eng. 2014, 22, 982–991.
87 Peraza, L.R.; Kinnunen, K.M.; McNaney, R.; Craddock, I.J.; Whone, A.L.; Morgan, C.; Joules, R.; Wolz, R. An automatic gait analysis pipeline for wearable sensors: A pilot study in parkinson's disease. Sensors 2021, 21, 8286.
88 Saeys, Y.; Abeel, T.; Van De Peer, Y. Robust feature selection using ensemble feature selection techniques. In Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Berlin/Heidelberg, Germany, 2008; pp. 313–325.
89 Cai, J.; Luo, J.; Wang, S.; Yang, S. Feature selection in machine learning: A new perspective. Neurocomputing 2018, 300, 70–79.
90 Jabar, H.; Khan, R.Z. Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study). Comput. Sci. Commun. Instrum. Devices 2015, 70, 163–172.
91 Hall, M.A. Correlation-based Feature Selection for Machine Learning. Ph.D. Thesis, The University of Waikato, Hamilton, New Zealand, 1999.
92 Hall, M.A.; Smith, L.A. Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper. In Proceedings of the FLAIRS Conference 1999, Orlando, FL, USA, 1–5 May 1999; Volume 1999, pp. 235–239.
93 Galperin, I.; Hillel, I.; Del Din, S.; Bekkers, E.M.; Nieuwboer, A.; Abbruzzese, G.; Avanzino, L.; Nieuwhof, F.; Bloem, B.R.; Rochester, L.; ET AL. Associations between daily-living physical activity and laboratory-based assessments of motor severity in patients with falls and Parkinson's disease. Park. Relat. Disord. 2019, 62, 85–90.
94 Dong, C.; Ye, T.; Long, X.; Aarts, R.M.; van Dijk, J.P.; Shang, C.; Liao, X.; Chen, W.; Lai, W.; Chen, L.; ET AL. A Two-Layer Ensemble Method for Detecting Epileptic Seizures Using a Self-Annotation Bracelet with Motor Sensors. IEEE Trans. Instrum. Meas. 2022, 71, 4005013.
95 Creagh, A.P.; Simillion, C.; Bourke, A.K.; Scotland, A.; Lipsmeier, F.; Bernasconi, C.; van Beek, J.; Baker, M.; Gossens, C.; Lindemann, M.; ET AL. Smartphone-and Smartwatch-Based Remote Characterisation of Ambulation in Multiple Sclerosis during the Two-Minute Walk Test. IEEE J. Biomed. Health Inform. 2021, 25, 838–849.
96 Chen, O.Y.; Lipsmeier, F.; Phan, H.; Prince, J.; Taylor, K.I.; Gossens, C.; Lindemann, M.; de Vos, M. Building a Machine-Learning Framework to Remotely Assess Parkinson's Disease Using Smartphones. IEEE Trans. Biomed. Eng. 2020, 67, 3491–3500.
97 Steyerberg, E.W.; Eijkemans, M.J.C.; Habbema, J.D.F. Stepwise selection in small data sets: A simulation study of bias in logistic regression analysis. J. Clin. Epidemiol. 1999, 52, 935–942.
98 Austin, P.C.; Tu, J.V. Bootstrap Methods for Developing Predictive Models. Am. Stat. 2004, 58, 131–137.
99 Zimmerman, D.W. Power Functions of the Test and Mann-Whitney Test Under Violation of Parametric Assumptions. Percept. Mot. Skills 1985, 61, 467–470.
100 Urbanowicz, R.J.; Meeker, M.; la Cava, W.; Olson, R.S.; Moore, J.H. Relief-based feature selection: Introduction and review. J. Biomed. Inform. 2018, 85, 189–203.
101 Kira, K.; Rendell, L.A. A Practical Approach to Feature Selection; Elsevier: Amsterdam, The Netherlands, 1992; pp. 249–256.
102 Verma, N.K.; Salour, A. Feature selection. Stud. Syst. Decis. Control 2020, 256, 175–200.
103 Yaman, O.; Ertam, F.; Tuncer, T. Automated Parkinson's disease recognition based on statistical pooling method using acoustic features. Med. Hypotheses 2020, 135, 109483.
104 Rodriguez-Molinero, A.; Samà, A.; Pérez-Martínez, D.A.; López, C.P.; Romagosa, J.; Bayes, A.; Sanz, P.; Calopa, M.; Gálvez-Barrón, C.; De Mingo, E.; ET AL. Validation of a portable device for mapping motor and gait disturbances in Parkinson's disease. JMIR mHealth Uhealth 2015, 3, e9.
105 Chandrashekar, G.; Sahin, F. A survey on feature selection methods. Comput. Electr. Eng. 2014, 40, 16–28.
106 Goldsmith, J.; Bobb, J.; Crainiceanu, C.M.; Caffo, B.; Reich, D. Penalized functional regression. J. Comput. Graph. Stat. 2011, 20, 830–851.
107 Prince, J.; Andreotti, F.; De Vos, M. Multi-Source Ensemble Learning for the Remote Prediction of Parkinson's Disease in the Presence of Source-Wise Missing Data. IEEE Trans. Biomed. Eng. 2019, 66, 1402–1411.
108 Motin, M.A.; Pah, N.D.; Raghav, S.; Kumar, D.K. Parkinson's Disease Detection Using Smartphone Recorded Phonemes in Real World Conditions. IEEE Access 2022, 10, 97600–97609.
109 Cakmak, A.S.; Alday, E.A.P.; Da Poian, G.; Rad, A.B.; Metzler, T.J.; Neylan, T.C.; House, S.L.; Beaudoin, F.L.; An, X.; Stevens, J.S.; ET AL. Classification and Prediction of Post-Trauma Outcomes Related to PTSD Using Circadian Rhythm Changes Measured via Wrist-Worn Research Watch in a Large Longitudinal Cohort. IEEE J. Biomed. Health Inform. 2021, 25, 2866–2876.
110 Tracy, J.M.; Özkanca, Y.; Atkins, D.C.; Ghomi, R.H. Investigating voice as a biomarker: Deep phenotyping methods for early detection of Parkinson's disease. J. Biomed. Inform. 2020, 104, 103362.
111 Abdulhafedh, A. Comparison between Common Statistical Modeling Techniques Used in Research, Including: Discriminant Analysis vs Logistic Regression, Ridge Regression vs LASSO, and Decision Tree vs Random Forest. OAlib 2022, 9, 1–19. 112. Sánchez-Maroño, N.; Alonso-Betanzos, A.; Tombilla-Sanromán, M. Filter methods for feature selection—A comparative study. In Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Berlin/Heidelberg, Germany, 2007; pp. 178–187.
113 Porter, B.W.; Bareiss, R.; Holte, R.C. Concept learning and heuristic classification in weak-theory domains. Artif. Intell. 1990, 45, 229–263.
114 Wu, C.-T.; Wang, S.-M.; Su, Y.-E.; Hsieh, T.-T.; Chen, P.-C.; Cheng, Y.-C.; Tseng, T.-W.; Chang, W.-S.; Su, C.-S.; Kuo, L.-C.; ET AL. A Precision Health Service for Chronic Diseases: Development and Cohort Study Using Wearable Device, Machine Learning, and Deep Learning. IEEE J. Transl. Eng. Health Med. 2022, 10, 2700414.
115 de Lima, A.L.S.; Evers, L.J.; Hahn, T.; de Vries, N.M.; Daeschler, M.; Boroojerdi, B.; Terricabras, D.; Little, M.A.; Bloem, B.R.; Faber, M.J. Impact of motor fluctuations on real-life gait in Parkinson's patients. Gait Posture 2018, 62, 388–394.
116 Pulliam, C.; Eichenseer, S.; Goetz, C.; Waln, O.; Hunter, C.; Jankovic, J.; Vaillancourt, D.; Giuffrida, J.; Heldman, D. Continuous in-home monitoring of essential tremor. Park. Relat. Disord. 2014, 20, 37–40.
117 Goni, M.; Eickhoff, S.B.; Far, M.S.; Patil, K.R.; Dukart, J. Smartphone-Based Digital Biomarkers for Parkinson's Disease in a Remotely-Administered Setting. IEEE Access 2022, 10, 28361–28384.
118 Livingston, E.; Cao, J.; Dimick, J.B. Tread carefully with stepwise regression. Arch. Surg. 2010, 145, 1039–1040.
119 Li, F.; Yang, Y. Analysis of recursive feature elimination methods. In Proceedings of the the 28th ACM/SIGIR International Symposium on Information Retrieval 2005, Salvador, Brazil, 15–19 August 2005.
120 Kuhn, M.; Johnson, K.; Kuhn, M.; Johnson, K. An Introduction to Feature Selection. In Applied Predictive Modeling; Springer: New York, NY, USA, 2013; pp. 487–519.
121 Senturk, Z.K. Early diagnosis of Parkinson's disease using machine learning algorithms. Med. Hypotheses 2020, 138, 109603.
122 Zhang, X.D. Machine Learning. In A Matrix Algebra Approach to Artificial Intelligence; Springer: Singapore, 2020. 123. Russell, S.J.; Norvig, P. Artificial Intelligence: A Modern Approach, 4th ed.; Prentice Hall: Hoboken, NJ, USA, 2020.
124 Tinschert, P.; Rassouli, F.; Barata, F.; Steurer-Stey, C.; Fleisch, E.; Puhan, M.; Kowatsch, T.; Brutsche, M.H. Smartphone-Based Cough. Detection Predicts Asthma Control—Description of a Novel, Scalable Digital Biomarker; European Respiratory Society (ERS): Lausanne, Switzerland, 2020; p. 4569.
125 ZhuParris, A.; Kruizinga, M.D.; van Gent, M.; Dessing, E.; Exadaktylos, V.; Doll, R.J.; Stuurman, F.E.; Driessen, G.A.; Cohen, A.F. Development and Technical Validation of a Smartphone-Based Cry Detection Algorithm. Front. Pediatr. 2021, 9, 262.
126 Fatima, M.; Pasha, M. Survey of Machine Learning Algorithms for Disease Diagnostic. J. Intell. Learn. Syst. Appl. 2017, 9, 1–16.
127 Ensari, I.; Caceres, B.A.; Jackman, K.B.; Suero-Tejeda, N.; Shechter, A.; Odlum, M.L.; Bakken,

S. Digital phenotyping of sleep patterns among heterogenous samples of Latinx adults using unsupervised learning. Sleep. Med. 2021, 85, 211–220.
128 Ko, Y.-F.; Kuo, P.-H.; Wang, C.-F.; Chen, Y.-J.; Chuang, P.-C.; Li, S.-Z.; Chen, B.-W.; Yang, F.-C.; Lo, Y.-C.; Yang, Y.; ET AL. Quantification Analysis of Sleep Based on Smartwatch Sensors for Parkinson's Disease. Biosensors 2022, 12, 74.
129 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.
130 Derungs, A.; Schuster-Amft, C.; Amft, O. Longitudinal walking analysis in hemiparetic patients using wearable motion sensors: Is there convergence between body sides? Front. Bioeng. Biotechnol. 2018, 6, 57.
131 Freedman, D.A. Statistical Models. In Statistical Models: THeory and Practice; Cambridge University Press: Cambridge, UK, 2009. 132. Ahmed, S.T.; Basha, S.M.; Arumugam, S.R.; Kodabagi, M.M. Pattern Recognition: An Introduction, 1st ed.; MileStone Research Publications: Bengaluru, India, 2021.
133 Ruppert, D. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. J. Am. Stat. Assoc. 2004, 99, 567.
134 Opitz, D.; Maclin, R. Popular Ensemble Methods: An Empirical Study. J. Artif. Intell. Res. 1999, 11, 169–198. 135. Kosasi, S. Perancangan Prototipe Sistem Pemesanan Makanan dan Minuman Menggunakan Mobile Device. Indones. J. Netw. Secur. 2015, 1, 1–10.
136 Hastie, T.; Tibshirani, R.; Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction; Springer Science & Business Media: New York, NY, USA, 2013.
137 San-Segundo, R.; Zhang, A.; Cebulla, A.; Panev, S.; Tabor, G.; Stebbins, K.; Massa, R.E.; Whitford, A.; de la Torre, F.; Hodgins, J. Parkinson's disease tremor detection in the wild using wearable accelerometers. Sensors 2020, 20, 5817.
138 Ahmadi, M.N.; O'neil, M.E.; Baque, E.; Boyd, R.N.; Trost, S.G. Machine learning to quantify physical activity in children with cerebral palsy: Comparison of group, group-personalized, and fully-personalized activity classification models. Sensors 2020, 20, 3976.
139 Faurholt-Jepsen, M.; Busk, J.; Frost, M.; Vinberg, M.; Christensen, E.M.; Winther, O.; Bardram, J.E.; Kessing, L.V. Voice analysis as an objective state marker in bipolar disorder. Transl. Psychiatry 2016, 6, e856.
140 Jacobson, N.C.; Lekkas, D.; Huang, R.; Thomas, N. Deep learning paired with wearable passive sensing data predicts deterioration in anxiety disorder symptoms across 17–18 years. J. Affect. Disord. 2021, 282, 104–111.
141 Hastie, T.; Tibshirani, R.; Friedman, J. Statistics the Elements of Statistical Learning. Math. Intell. 2009, 27, 83–85.
142 Patle, A.; Chouhan, D.S. SVM kernel functions for classification. In Proceedings of the 2013 International Conference on Advances in Technology and Engineering, ICATE 2013, Mumbai, India, 23–25 January 2013.
143 Kim, H.S.; Kim, S.Y.; Kim, Y.H.; Park, K.S. A smartphone-based automatic diagnosis system for facial nerve palsy. Sensors 2015, 15, 26756–26768.
144 Luca, S.; Karsmakers, P.; Cuppens, K.; Croonenborghs, T.; Van de Vel, A.; Ceulemans, B.; Lagae, L.; Van Huffel, S.; Vanrumste, B. Detecting rare events using extreme value statistics applied to epileptic convulsions in children. Artif. Intell. Med. 2014, 60, 89–96.
145 Ghoraani, B.; Hssayeni, M.D.; Bruack, M.M.; Jimenez-Shahed, J. Multilevel Features for Sensor-Based Assessment of Motor Fluctuation in Parkinson's Disease Subjects. IEEE J. Biomed. Health Inform. 2020, 24, 1284–1295.
146 Kramer, O. K-Nearest Neighbors. In Dimensionality Reduction with Unsupervised Nearest Neighbors. Intelligent Systems Reference. Library; Springer: Berlin/Heidelberg, Germany, 2013; Volume 51.
147 Jeon, H.; Lee, W.; Park, H.; Lee, H.J.; Kim, S.K.; Kim, H.B.; Jeon, B.; Park, K.S. Automatic classification of tremor severity in Parkinson's disease using awearable device. Sensors 2017, 17, 2067.
148 Grunerbl, A.; Muaremi, A.; Osmani, V.; Bahle, G.; Ohler, S.; Troster, G.; Mayora, O.; Haring, C.; Lukowicz, P. Smartphone-based recognition of states and state changes in bipolar disorder patients. IEEE J. Biomed. Health Inform. 2015, 19, 140–148.
149 Pranckevičius, T.; Marcinkevičius, V. Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification. Balt. J. Mod. Comput. 2017, 5, 221–232.
150 Worster, A.; Fan, J.; Ismaila, A. Understanding linear and logistic regression analyses. Can. J. Emerg. Med. 2007, 9, 111–113.
151 Morrow-Howell, N. The M word: Multicollinearity in multiple regression. Soc. Work. Res. 1994, 18, 247–251.
152 Schwenk, M.; Hauer, K.; Zieschang, T.; Englert, S.; Mohler, J.; Najafi, B. Sensor-derived physical activity parameters can predict future falls in people with dementia. Gerontology 2014, 60, 483–492.
153 Lecun, Y.; Bengio, Y.; Hinton, G. Deep learning. Nature 2015, 521, 436–444.
154 Tu, J.V. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. J. Clin. Epidemiol. 1996, 49, 1225–1231.
155 Mudiyanselage, T.K.B.; Xiao, X.; Zhang, Y.; Pan, Y. Deep Fuzzy Neural Networks for Biomarker Selection for Accurate Cancer Detection. IEEE Trans. Fuzzy Syst. 2020, 28, 3219–3228.
156 Yagin, F.H.; Cicek, I.B.; Alkhateeb, A.; Yagin, B.; Colak, C.; Azzeh, M.; Akbulut, S. Explainable artificial intelligence model for identifying COVID-19 gene biomarkers. Comput. Biol. Med. 2023, 154, 106619.
157 Wang, Y.; Lucas, M.; Furst, J.; Fawzi, A.A.; Raicu, D. Explainable Deep Learning for Biomarker Classification of OCT Images. In Proceedings of the 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE), Cincinnati, OH, USA, 26–28 October 2020; pp. 204–210.
158 Fisher, J.M.; Hammerla, N.Y.; Ploetz, T.; Andras, P.; Rochester, L.; Walker, R.W. Unsupervised home monitoring of Parkinson's disease motor symptoms using body-worn accelerometers. Park. Relat. Disord. 2016, 33, 44–50.
159 Frogner, J.I.; Noori, F.M.; Halvorsen, P.; Hicks, S.A.; Garcia-Ceja, E.; Torresen, J.; Riegler, M.A. One-dimensional convolutional neural networks on motor activity measurements in detection of depression. In Proceedings of the HealthMedia 2019—Proceedings of the 4th International Workshop on Multimedia for Personal Health and Health Care, Co-Located with MM 2019, Nice, France, 21–25 October 2019; pp. 9–15.
160 Meisel, C.; el Atrache, R.; Jackson, M.; Schubach, S.; Ufongene, C.; Loddenkemper, T. Machine learning from wristband sensor data for wearable, noninvasive seizure forecasting. Epilepsia 2020, 61, 2653–2666.
161 Matarazzo, M.; Arroyo-Gallego, T.; Montero, P.; Puertas-Martín, V.; Butterworth, I.; Mendoza, C.S.; Ledesma-Carbayo, M.J.; Catalán, M.J.; Molina, J.A.; Bermejo-Pareja, F.; ET AL. Remote Monitoring of Treatment Response in Parkinson's Disease: The Habit of Typing on a Computer. Mov. Disord. 2019, 34, 1488–1495.
162 Weiss, K.; Khoshgoftaar, T.M.; Background, D.W. A survey of transfer learning. J. Big Data 2016, 3, 1345–1459.
163 Kamishima, T.; Hamasaki, M.; Akaho, S. TrBagg: A Simple Transfer Learning Method and its Application to Personalization in Collaborative Tagging. In Proceedings of the 2009 Ninth IEEE International Conference on Data Mining, Miami, FL, USA, 6–9 December 2009; pp. 219–228.
164 Fu, Z.; He, X.; Wang, E.; Huo, J.; Huang, J.; Wu, D. Personalized Human Activity Recognition Based on Integrated Wearable Sensor and Transfer Learning. Sensors 2021, 21, 885.
165 Chen, Y.; Qin, X.; Wang, J.; Yu, C.; Gao, W. FedHealth: A Federated Transfer Learning Framework for Wearable Healthcare. IEEE Intell. Syst. 2020, 35, 83–93.
166 Goschenhofer, J.; Pfister, F.M.J.; Yuksel, K.A.; Bischl, B.; Fietzek, U.; Thomas, J. Wearable-Based Parkinson's Disease Severity Monitoring Using Deep Learning. In Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Berlin/Heidelberg, Germany, 2020; Volume 11908 LNAI, pp. 400–415.
167 Hssayeni, M.D.; Jimenez-Shahed, J.; Burack, M.A.; Ghoraani, B. Ensemble deep model for continuous estimation of Unified Parkinson's Disease Rating Scale III. Biomed. Eng. Online 2021, 20, 1–20.
168 Zhang, Y.; Yang, Q. Special Topic: Machine Learning An overview of multi-task learning. Natl. Sci. Rev. 2018, 5, 30–43.
169 Lee, G.; Yang, E.; Hwang, S. Asymmetric multi-task learning based on task relatedness and loss. In Proceedings of the International Conference on Machine Learning 2016, New York, NY, USA, 19–24 June 2016; pp. 230–238.
170 Xin, W.; Bi, J.; Yu, S.; Sun, J.; Song, M. Multiplicative Multitask Feature Learning. J. Mach. Learn. Res. JMLR 2016, 17, 1–33.
171 Zhang, Z.; Jung, T.P.; Makeig, S.; Pi, Z.; Rao,

B.D. Spatiotemporal sparse Bayesian learning with applications to compressed sensing of multichannel physiological signals. IEEE Trans. Neural Syst. Rehabil. Eng. 2014, 22, 1186–1197.
172 Schneider, J.; Vlachos, M. Personalization of deep learning. In Data Science–Analytics and Applications: Proceedings of the 3rd International Data Science Conference–iDSC2020; Springer: Wiesbaden, Geramny, 2021; pp. 89–96.
173 Khademi, A.; El-Manzalawy, Y.; Buxton, O.M.; Honavar, V. Toward personalized sleep-wake prediction from actigraphy. In Proceedings of the 2018 IEEE EMBS International Conference on Biomedical and Health Informatics, BHI 2018, Vegas, NV, USA, 4–7 March 2018; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2018; pp. 414–417.
174 Kuhn, M.; Johnson, K. Applied Predictive Modeling; Springer: New York, NY, USA, 2013.
175 Pal, M. Random forest classifier for remote sensing classification. Int. J. Remote Sens. 2005, 26, 217–222.
176 Putin, E.; Mamoshina, P.; Aliper, A.; Korzinkin, M.; Moskalev, A.; Kolosov, A.; Ostrovskiy, A.; Cantor, C.; Vijg, J.; Zhavoronkov, A. Deep biomarkers of human aging: Application of deep neural networks to biomarker development. Aging 2016, 8, 1021–1033.
177 Yang, L.; Shami, A. On hyperparameter optimization of machine learning algorithms: Theory and practice. Neurocomputing 2020, 415, 295–316.
178 Waring, J.; Lindvall, C.; Umeton, R. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. Artif. Intell. Med. 2020, 104, 101822.
179 Bergstra, J.; Ca, J.B.; Ca, Y.B. Random Search for Hyper-Parameter Optimization Yoshua Bengio. 2012. Available online: http://scikit-learn.sourceforge.net (accessed on 5 January 2023).
180 Beam, A.L.; Manrai, A.K.; Ghassemi, M. Challenges to the Reproducibility of Machine Learning Models in Health Care. JAMA 2020, 323, 305.
181 Ahlrichs, C.; Samà, A.; Lawo, M.; Cabestany, J.; Rodríguez-Martín, D.; Pérez-López, C.; Sweeney, D.; Quinlan, L.R.; Laighin, G.Ò.; Counihan, T.; ET AL. Detecting freezing of gait with a tri-axial accelerometer in Parkinson's disease patients. Med. Biol. Eng. Comput. 2016, 54, 223–233.
182 Rosenwein, T.; Dafna, E.; Tarasiuk, A.; Zigel, Y. Detection of Breathing Sounds during Sleep Using Non-Contact Audio Recordings; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2014.
183 Pérez-López, C.; Samà, A.; Rodríguez-Martín, D.; Moreno-Aróstegui, J.M.; Cabestany, J.; Bayes, A.; Mestre, B.; Alcaine, S.; Quispe, P.; Laighin, G.; ET AL. Dopaminergic-induced dyskinesia assessment based on a single belt-worn accelerometer. Artif. Intell. Med. 2016, 67, 47–56.
184 Bernad-Elazari, H.; Herman, T.; Mirelman, A.; Gazit, E.; Giladi, N.; Hausdorff, J.M. Objective characterization of daily living transitions in patients with Parkinson's disease using a single body-fixed sensor. J. Neurol. 2016, 263, 1544–1551.
185 Carvalho, D.V.; Pereira, E.M.; Cardoso, J.S. Machine Learning Interpretability: A Survey on Methods and Metrics. Electronics 2019, 8, 832.
186 Zhou, J.; Gandomi, A.H.; Chen, F.; Holzinger, A. Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. Electronics 2021, 10, 593.
187 Hossin, M.; Sulaiman, M.N. A Review on Evaluation Metrics for Data Classification Evaluations. Int. J. Data Min. Knowl. Manag. Process 2015, 5, 1–11.
188 He, H.; Ma, Y. Imbalanced Learning; Wiley: Hoboken, NJ, USA, 2013.
189 Wan, S.; Liang, Y.; Zhang, Y.; Guizani, M. Deep Multi-Layer perceptron classifier for behavior analysis to estimate Parkinson's disease severity using smartphones. IEEE Access 2018, 6, 36825–36833.
190 Chicco, D.; Tötsch, N.; Jurman, G. The matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Min. 2021, 14, 1–22.
191 Jurman, G.; Riccadonna, S.; Furlanello, C. A Comparison of MCC and CEN Error Measures in Multi-Class Prediction. PLoS ONE 2012, 7, e41882.
192 Faurholt-Jepsen, M.; Busk, J.; HelgaÞórarinsdóttir; Frost, M.; Bardram, J.E.; Vinberg, M.; Kessing, L.V. Objective smartphone data as a potential diagnostic marker of bipolar disorder. Aust. N. Z. J. Psychiatry 2019, 53, 119–128.
193 Xu, Z.; Shen, D.; Nie, T.; Kou, Y. A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data. J. Biomed. Inform. 2020, 107, 103465.
194 Zeng, M.; Zou, B.; Wei, F.; Liu, X.; Wang, L. Effective prediction of three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data. In Proceedings of the 2016 IEEE International Conference of Online Analysis and Computing Science, ICOACS 2016, Chongqing, China, 28–29 May 2016; pp. 225–228.
195 Botchkarev, A. Performance Metrics (Error Measures) in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology. Interdiscip. J. Inf. Knowl. Manag. 2018, 14, 45–76.
196 di Bucchianico, A. Coefficient of Determination. In Encyclopedia of Statistics in Quality and Reliability; Wiley: Hoboken, NJ, USA, 2007.
197 Piepho, H. A coefficient of determination (R2) for generalized linear mixed models. Biom. J. 2019, 61, 860–872.
198 Gelman, A.; Pardoe, I. Bayesian Measures of Explained Variance and Pooling in Multilevel (Hierarchical) Models. Technometrics 2006, 48, 241–251.
199 Hodson, T.O. Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. Geosci. Model. Dev. 2022, 15, 5481–5487.
200 Mezzadri, G.; Laloë, T.; Mathy, F.; Reynaud-Bouret, P. Hold-out strategy for selecting learning models: Application to categorization subjected to presentation orders. J. Math. Psychol. 2022, 109, 102691.
201 Gholamiangonabadi, D.; Kiselov, N.; Grolinger, K. Deep Neural Networks for Human Activity Recognition with Wearable Sensors: Leave-One-Subject-Out Cross-Validation for Model Selection. IEEE Access 2020, 8, 133982–133994.
202 Little, M.A.; Varoquaux, G.; Saeb, S.; Lonini, L.; Jayaraman, A.; Mohr, D.C.; Kording, K.P. Using and understanding crossvalidation strategies. Perspectives on Saeb ET AL. Gigascience 2017, 6, 1–6.
203 Peterson, R.A.; Cavanaugh, J.E. Ordered quantile normalization: A semiparametric transformation built for the cross-validation era. J. Appl. Stat. 2020, 47, 2312–2327.
204 Zhang, Y.; Yang, Y. Cross-validation for selecting a model selection procedure. J. Econom. 2015, 187, 95–112.
205 Refaeilzadeh, P.; Tang, L.; Liu, H. Cross-Validation. In Encyclopedia of Database Systems; Springer: Berlin/Heidelberg, Germany, 2016; pp. 1–7.
206 Browne, M.W. Cross-validation methods. J. Math. Psychol. 2000, 44, 108–132.
207 Wainer, J.; Cawley, G. Nested cross-validation when selecting classifiers is overzealous for most practical applications. Expert Syst. Appl. 2021, 182, 115222.
208 Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. 1995. Available online: http://robotics.stanford.edu/~ronnyk (accessed on 5 January 2023).
209 Vanwinckelen, G.; Blockeel, H. On estimating model accuracy with repeated cross-validation. In BeneLearn 2012: Proceedings of the 21st Belgian-Dutch Conference on Machine Learning; Benelearn 2012 Organization Committee: Ghent, Belgium, 2012; pp. 39–44.
210 Parvandeh, S.; Yeh, H.-W.; Paulus, M.P.; McKinney, B.A. Consensus Features Nested Cross-Validation. bioRxiv 2020.
211 Goble, C.; Cohen-Boulakia, S.; Soiland-Reyes, S.; Garijo, D.; Gil, Y.; Crusoe, M.; Peters, K.; Schober, D. Fair computational workflows. Data Intell. 2020, 2, 108–121.
212 Muehlematter, U.J.; Daniore, P.; Vokinger, K.N. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): A comparative analysis. Lancet Digit. Health 2021, 3, e195–e203.
213 Beckers, R.; Kwade, Z.; Zanca, F. The EU medical device regulation: Implications for artificial intelligence-based medical device software in medical physics. Phys. Med. 2021, 83, 1–8.
214 van Oirschot, J.; Ooms, G. Interpreting the EU Artificial Intelligence Act for the Health Sector; Health Action International: Amsterdam, The Netherlands, February 2022.
215 Agarwal, S.; LeFevre, A.; Lee, J.; L'engle, K.; Mehl, G.; Sinha, C.; Labrique, A. Guidelines for reporting of health interventions using mobile phones: Mobile health (MHEALTH) evidence reporting and assessment (mERA) checklist. BMJ 2016, 352, i1174.
216 Yang, J.; Soltan, A.A.S.; Clifton, D.A. Machine learning generalizability across healthcare settings: Insights from multi-site COVID-19 screening. NPJ Digit. Med. 2022, 5, 69.
217 Petersen, E.; Potdevin, Y.; Mohammadi, E.; Zidowitz, S.; Breyer, S.; Nowotka, D.; Henn, S.; Pechmann, L.; Leucker, M.; Rostalski, P.; ET AL. Responsible and Regulatory Conform Machine Learning for Medicine: A Survey of Challenges and Solutions. IEEE Access 2022, 10, 58375–58418.
218 FAIR Principles—GO FAIR. Available online: https://www.go-fair.org/fair-principles/ (accessed on 16 December 2021).
219 Fletcher, R.R.; Nakeshimana, A.; Olubeko, O. Addressing Fairness, Bias, and Appropriate Use

of Artificial Intelligence and Machine Learning in Global Health. Front. Artif. Intell. 2021, 3, 116.
220 Mei, J.; Desrosiers, C.; Frasnelli, J. Machine Learning for the Diagnosis of Parkinson's Disease: A Review of Literature. Front. Aging Neurosci. 2021, 13, 633752.
221 Chekroud, A.M.; Bondar, J.; Delgadillo, J.; Doherty, G.; Wasil, A.; Fokkema, M.; Cohen, Z.; Belgrave, D.; DeRubeis, R.; Iniesta, R.; ET AL. The promise of machine learning in predicting treatment outcomes in psychiatry. World Psychiatry 2021, 20, 154–170.
222 Kruizinga, M.D.; Zhuparris, A.; Dessing, E.; Krol, F.J.; Sprij, A.J.; Doll, R.; Stuurman, F.E.; Exadaktylos, V.; Driessen, G.J.A.; Cohen, A.F. Development and technical validation of a smartphone-based pediatric cough detection algorithm. Pediatr. Pulmonol. 2022, 57, 761–767.
223 Doshi-Velez, F.; Kim, B. Towards A Rigorous Science of Interpretable Machine Learning. arXiv 2017, arXiv:1702.08608.
224 Ignatiev, A. Towards Trustable Explainable AI. 2020. Available online: https://www.kaggle.com/uciml/zoo-animal-classification (accessed on 5 January 2023).
225 Walsh, I.; Fishman, D.; Garcia-Gasulla, D.; Titma, T.; Pollastri, G.; Capriotti, E.; Casadio, R.; Capella-Gutierrez, S.; Cirillo, D.; Del Conte, A.; ET AL. DOME: Recommendations for supervised machine learning validation in biology. Nat. Methods 2021, 18, 1122–1127.
226 Zippel, C.; Bohnet-Joschko, S. Rise of Clinical Studies in the Field of Machine Learning: A Review of Data Registered in ClinicalTrials.gov. Int. J. Environ. Res. Public Health 2021, 18, 5072.

TABLE 1 Representation of a standard machine learning pipeline.

Stage	Objective	Example
STUDY DESIGN	The ML pipeline is provided with a study objective in which the features and corresponding outputs are defined. The ML model aims to identify the associations between the features and outputs.	The study objective is to classify Parkinson's Disease patients and control groups using smartphone-based features.
DATA PREPROCESSING	Data preprocessing filters and transforms raw data to guarantee or enhance the ML training process.	To improve the model performance, one may identify and exclude any missing or outlier data.
FEATURE ENGINEERING AND SELECTION	Feature engineering uses raw data to create new features that are not readily available in the dataset. Feature selection selects the most relevant features for the model objective by removing redundant or noisy features. Together, the goal is to simplify and accelerate the computational process while also improving the model process. For deep learning methods, the concept of 'feature engineering' is typically embedded within the model architecture and training process, although substantial preprocessing steps may occur prior to that.	An interaction of two or more predictors (such as a ratio or product) or re-representation of a predictor are examples of feature engineering. Removing highly correlated or non-informative features are examples of feature selection. Note: The feature selection step can occur during model training
MODEL TRAINING AND VALIDATION	During training, the ML model(s) iterates through all the examples in the training dataset and optimizes the parameters of the mathematical function to minimize the prediction error. To evaluate the performance of the trained ML model, the predictions of an unseen test set are compared with a known ground truth label.	Cross-validation can be used to optimize and evaluate model performance. Classification models may be evaluated based on their prediction accuracy, sensitivity, and specificity, while regression models may be evaluated using variance explained (R2) and Mean Absolute Error.

TABLE 2 An overview of the keyword strategy used for this study.

Domain	Search String
TECHNOLOGY	((‘smartphone’[tiab] OR ‘wearable’[tiab] OR ‘remote + monitoring’[tiab] OR ‘home + monitoring’[tiab] OR ‘mobile + sensors’[tiab] OR ‘mobile + montoring’[tiab] OR ‘behavioral + sensing’[tiab] OR ‘geolocation’[tiab] OR ‘mHealth’[tiab] OR ‘passive + monitoring’[tiab] OR ‘digital + phenotype’[tiab] OR ‘digital + phenotyping’[tiab] OR ‘digital + biomarker’[tiab])
ANALYSIS	AND (‘machine + learning’[tiab] OR ‘deep + learning’[tiab] OR ‘random + forest’[tiab] OR ‘neural + network’[tiab] OR ‘time + series’[tiab] OR ‘regression’[tiab] OR ‘SVM’[tiab] OR ‘knn’[tiab] OR ‘dynamics + model’[tiab] OR ‘decision + tree’[tiab] OR ‘discriminant + analysis’[tiab] OR ‘feature + engineering’[tiab] OR ‘feature + selection’[tiab] OR ‘data + mining’[tiab] OR ‘model’[tiab] OR ‘classification’[tiab] OR ‘diagnostic’[tiab] OR ‘prognostic’[tiab] OR ‘symptom + severity’[tiab] OR ‘prediction’[tiab] OR ‘monitoring’[tiab])
POPULATION	AND (‘disease’[tiab] OR ‘disorder’[tiab] OR ‘diagnosis’[tiab] OR ‘prognosis’ OR ‘alzheimer’[tiab] OR ‘parkinson’[tiab] OR ‘Huntington’[tiab] OR ‘neurodegenerative’[tiab] OR ‘degenerative’ OR ‘tremor’[tiab] OR ‘bipolar’[tiab] OR ‘depression’[tiab] OR ‘manic’[tiab] OR ‘anxiety’[tiab] OR ‘vocal + biomarker’[tiab] OR ‘amyotrophic + lateral + sclerosis’[tiab] OR ‘central + nervous + system’[tiab] OR ‘symptom’[tiab] OR ‘psychosis’[tiab] OR ‘stroke’[tiab] OR ‘muscular dystrophy’[tiab] OR ‘Facioscapulohumeral Dystrophy’[tiab] OR ‘autoimmune’[tiab] OR ‘seizure’[tiab] OR ‘multiple + sclerosis’[tiab])
DATE	AND (‘2012/01/01’[PDAT]:’2022/12/31’[PDAT])
LANGUAGE	AND (English[lang])
EXCLUSION CRITERIA	NOT(‘animals’[tiab] OR ‘implant’[tiab] OR ‘hospital’[tiab] OR ‘caregiver’[tiab] OR ‘telemedicine’[tiab] OR ‘telerehabilitation’[tiab] OR ‘smartphone + addiction’[tiab] OR ‘nursing’[tiab] OR’screening’[tiab] OR ‘recruitment’[tiab] OR ‘diabetes’[tiab] OR ‘malaria’[tiab] OR ‘self-care’[tiab] OR ‘self-management’[tiab] OR ‘self-help’[tiab])
ARTICLE TYPE	AND (clinicalstudy[Filter] OR clinicaltrial[Filter] OR clinicaltrialphasei[Filter] OR clinicaltrialphaseii[Filter] OR clinicaltrialphaseiii[Filter] OR clinicaltrialphaseiv[Filter] OR controlledclinicaltrial[Filter] OR meta-analysis[Filter] observationalstudy[Filter] OR randomizedcontrolledtrial[Filter] OR systematicreview[Filter])

TABLE 3 Table of the inclusion and exclusion criteria used for study selection.

Category	Criteria
POPULATION	The study must be initiated by a research organization and not by the participants. The participants must have a clinical diagnosis that is affected by the CNS. Hence, studies that collected data from participants with no clinically confirmed diagnosis were not considered.
INTERVENTION	The study must include the use of smartphone or non-invasive wearables to remotely monitor and quantify passive biomarkers under free-living conditions.
COMPARATOR	A ground truth comparator for digital phenotyping such as clinical assessment, medical records, or self-reported outcomes.
OUTCOMES	A ML model that is used to classify a clinical label (such as a diagnosis, or clinical event), estimate symptom severity, or to detect treatment effects.
STUDY TYPE	The paper must be about a human-centered observational study (cohort or longitudinal) where the data were collected outside the clinic, lab, or hospital (free-living conditions). Hence, studies that use smartphones or wearables as a form of intervention or as screening tools are not of interest. The study must show if the ML models had ecological validity by validating the models using free-living data. The study has to have been written or translated into English and published within the last 10 years (2012 onwards).

TABLE 4 Clinical interpretations of common ML performance metrics.

Term	Equation	Objective
ACCURACY	$\frac{TP}{TP+TN}$	Out of all the predictions, how many predictions were correctly identified as positive or negative?
PRECISION	$\frac{2\times Precision\times Recall}{Precision+Recall}$	How many predictions were correctly labeled as patients out of all correctly classified patients and misclassified healthy controls?
SPECIFICITY	$\frac{1}{N}\sum Actual - Predicted$	How many predictions were correctly labeled as healthy controls out of all healthy controls? In other words, of all healthy controls, who were correctly identified as such?
RECALL/ SENSITIVITY	$\sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}}$	Of all the patients, who were correctly classified/identified as such?
F1-SCORE	$\frac{2\times Precision\times Recall}{Precision+Recall}$	How many predictions were correctly labeled as patients (recall) and what was the accuracy with regards to correctly predicted patients (precision)?
MEAN SQUARE ERROR	$\frac{1}{N}\sum Actual - Predicted$	What is the absolute difference between the true scores and the predicted scores?
ROOT MEAN SQUARE ERROR	$\sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}}$	What is the average difference between the true and the predicted scores (in the same unit of the true scores)?
R2	$1 - \frac{RSS}{TSS}$	What fraction of the variance in the data is captured by the model?

True Positive = TP, True Negative = TN, False Positives = FP, False Negatives = FN, Sum of Squares of Residuals = RSS, Total Sum of Squares = TSS, Number of Observations = N

FIGURE 1 Flow diagram illustrating the paper selection process for this review.

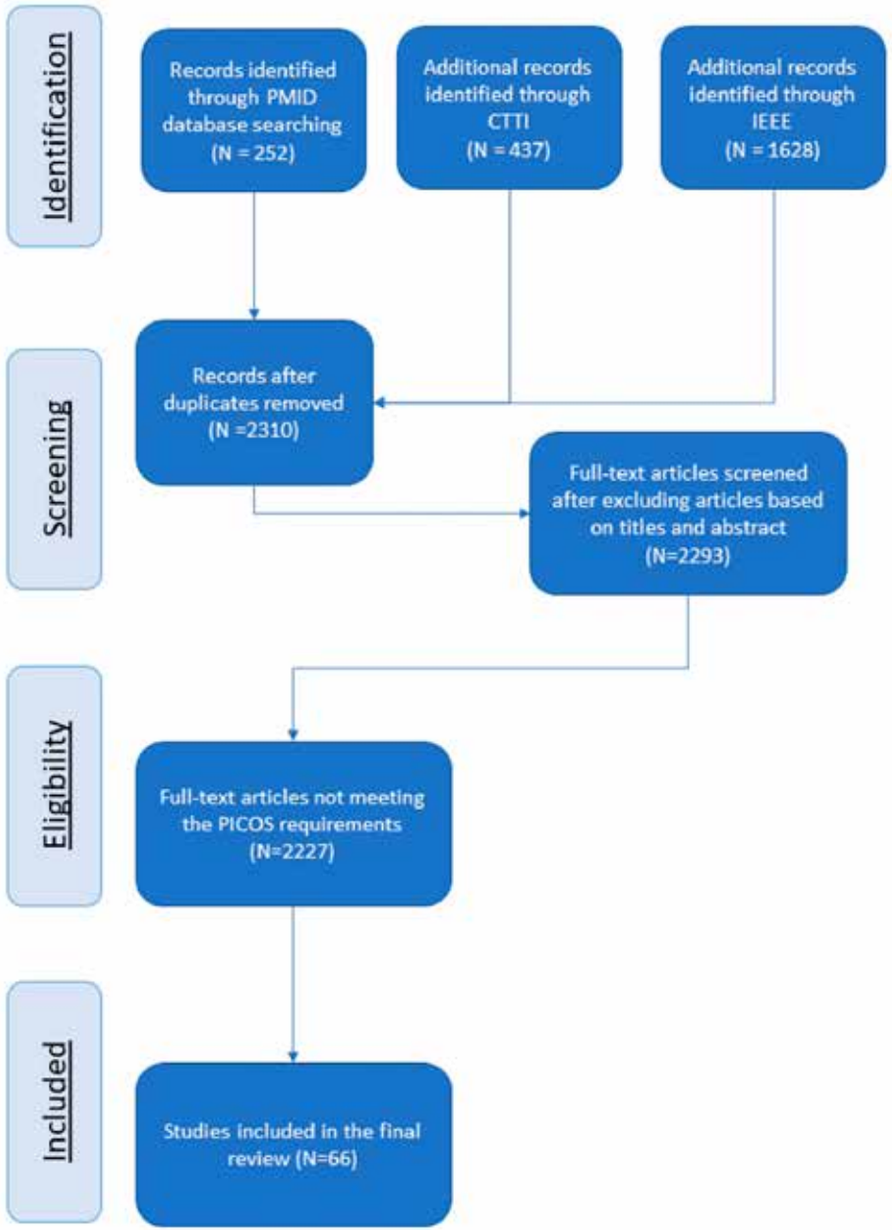

FIGURE 2 Clinical populations and the use of healthy controls in the selected studies.

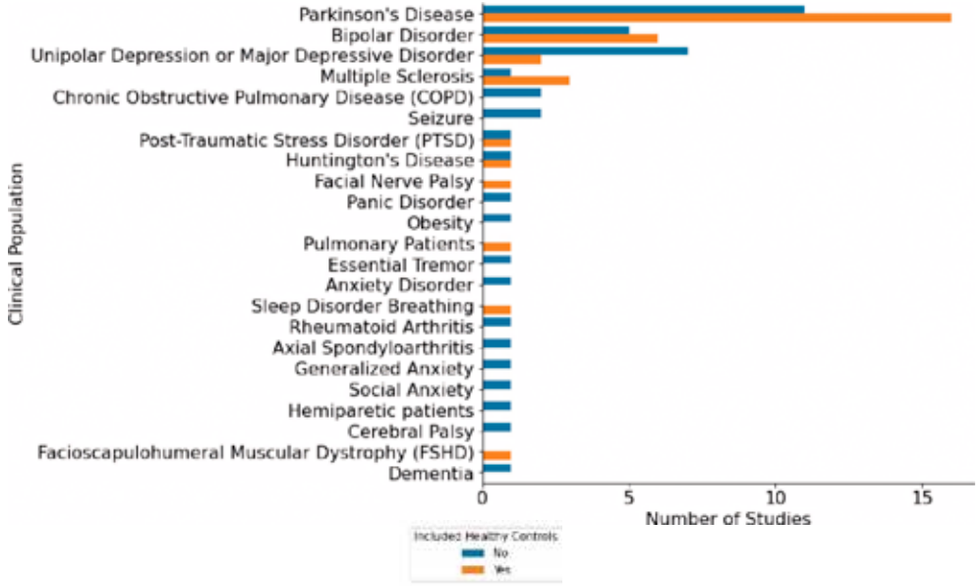

FIGURE 3 Sample sizes of clinical populations included in selected studies, with x-axis (sample size) presented on a logarithmic scale.

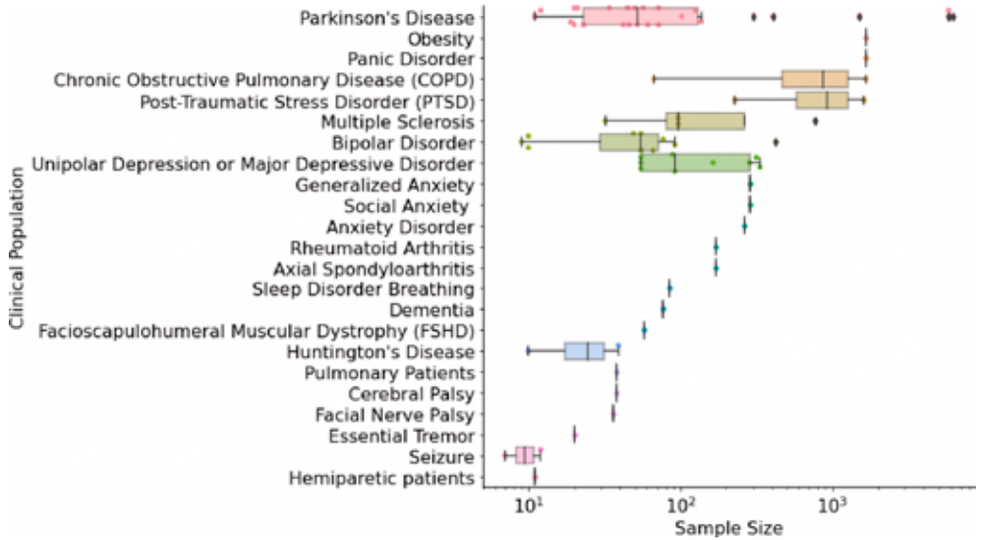

FIGURE 4 Machine learning algorithms and their respective objectives in the selected studies.

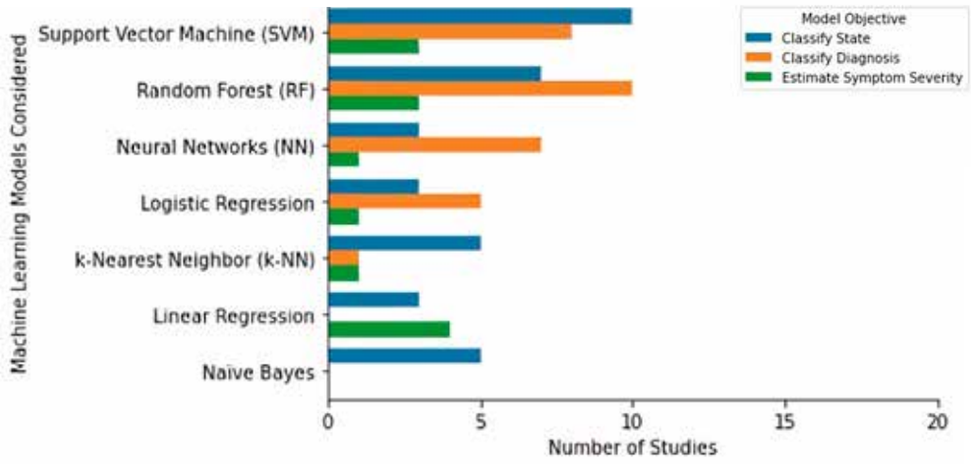

FIGURE 5 General recommendations for building an effective and reproducible ML pipeline.