scientific reports

OPEN

Robust remote detection of depressive tendency based on keystroke dynamics and behavioural characteristics

Ruba Fadul[1]✉, Aamna AlShehhi[1,2] & Leontios Hadjileontiadis[1,2,3]

Depressive Disorder (DD) is a leading cause of disability worldwide. Screening tools for detecting DD symptoms are essential for monitoring and efficient managing. Remarkably, individuals' kinetic activities, including their interaction with touchscreen smartphones, can be a proxy for their mental status. Therefore, studying these typing patterns can assist in developing passive screening tools for detecting even the early stage of DD, i.e., the depressive tendency (DT). Here we extend a previous study by exploring different machine learning models with various feature engineering approaches to detect the subjects' DT, as indicated by the self-administered Patient Health Questionnaire-9 (PHQ-9) score, via keystroke digital biomarkers. The keystroke timing sequences were unobtrusively collected from 24 subjects during routine interaction with touchscreen smartphones, resulting in 23,264 typing sessions. The proposed framework was investigated under two keystroke feature combinations—hold-time and flight-time variables—and validated using nested cross-validation scheme. Different feature selection (FS) techniques were employed to select informative features from the keystroke sequences. The best-performing gradient boosting classifier with features selected by the mutual information FS method achieved an improved Area Under Curve (AUC) of 0.98 [95% confidence interval: 0.91-1.00]. The proposed DT pipeline, which surpasses the state-of-the-art models, could effectively capture DT, considering users' behavioural characteristics. This would potentially provide users with information regarding the evolution of their mental health, simultaneously contributing to improving digital tools for objectively screening mental disorders in-the-wild.

Keywords Depression disease, Keystroke dynamics, Typing, Remote screening, Detection, Machine learning

Mental and neurological disorders exert a significant disease burden globally, substantially contributing to health loss throughout life[1]. These disorders account for 12% of the global Disability Adjusted Life Years (DALY) and 35% of Years Lived with Disability (YLD)[2]. However, a considerable number of cases of mental and neurological disorders remain undetected despite their high prevalence[3]. Mental disorders have a profound impact on individuals' cognition, emotions, and behavior, leading to a diminished quality of life for affected individuals and their communities[4]. These disorders not only affect mental well-being, but also have implications for physical health[5]. Despite the staggering number of individuals (approximately 792 million worldwide, affected by various forms of mental disorders), these disorders remain significantly under-reported[6]. Among them, depression stands out as the most prevalent mental disorder and a major contributor to global disability, affecting over 264 million people globally[7]. It also accounts for approximately 10 million suicide attempts annually[4]. Furthermore, social stigma, limited resources, and a shortage of specialized healthcare providers hinder effective care and early diagnosis[8].

Clinical diagnosis and assessment of mental health is an ongoing challenge characterized by considerable uncertainty[9]. To shed light on this complexity, researchers have harnessed the power of qualitative clinician-administered and self-administered scales/questionnaires. Notably, the widely employed Patient Health Questionnaire-9 (PHQ-9) serves as a diagnostic tool for depression[10]. Despite their utility, these scales may fall short in detecting early manifestations of depression, thus beckoning the exploration of novel, finely-

[1]Department of Biomedical Engineering and Biotechnology, Khalifa University of Science and Technology, Abu Dhabi, UAE. [2]Healthcare Engineering Innovation Group (HEIG), Khalifa University of Science and Technology, Abu Dhabi, UAE. [3]Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Thessaloniki, Greece. ✉email: ruba.fadul@ku.ac.ae

tuned biomarkers. Nevertheless, digital biomarkers captured through digital devices such as smartphones have the potential to provide valuable insights in the healthcare research[11]. In recent years, the proliferation of smartphones in parallel with the advancement in Artificial Intelligence and Machine Learning (ML) technologies resulted in a new domain of behavioral sensing[12]. Data available in everyday life, including audio and visual recordings[13], text-context[13], and typing-related characteristics[14], are utilized for the recognition of the mental state. Keystroke timing data, also known as keystroke dynamics, can reflect the users' emotional state, which is related to movement, particularly finger kinetics during typing[15]. Notably, due to the sophisticated relationship between affection and kinetic expression, keystroke dynamics and their behavioural characteristics represent a promising digital biomarker that can monitor individuals' mental mood and passively capture related typing data due to the users' interaction with technology devices, such as computers and smartphones.

The current work provides improvements upon a previous study[14], aiming to examine the diagnostic ability of the keystroke dynamics digital biomarkers and investigate whether a machine learning pipeline with feature extraction and selection techniques can efficiently detect the existence of depression and act as a screening tool for detecting even the early stage of the onset of depression, i.e., the depressive tendency (DT). We utilized a different feature engineering approach that derives an extensive feature set from keystroke sequences followed by applying various feature selection techniques to select the most descriptive features in order to develop more accurate and efficient depression diagnostic model which can be utilized easily in the future and run on touchscreen mobile devices. Hold time (HT) and flight time (FT) keystroke sequences, produced by multiple passively collected typing sessions during routine typing activities, are used as means that contribute to improving digital tools for objectively screening mental disorders in the wild. Here, 23,264 typing sessions provided by 10 depressive disorder (DD) patients and 14 healthy controls (HC) were analyzed for DT detection under a pipeline that included various ML models, namely Decision Tree, Random Forest (RF), AdaBoost, Gradient Boosting (GB), Gaussian Naive Bayes, Logistic Regression, K-Nearest Neighbors, and Multi-Layer Perceptron (MLP), in combination with different feature selection (FS) methods: chi-square (CHI2), mutual information (MI), and lasso. The performance of each proposed diagnostic model is evaluated and reported. Furthermore, features importance analysis of the best-performing pipeline was conducted to identify the keystroke features that can serve as potential digital biomarkers and show how they can reflect the fine motor skills and be interpreted in the context depressive disorder.

Results

Dataset

The proposed DT detection models utilize keystroke dynamics data collected from 24 subjects who composed matched groups in terms of age and gender and provided a total of 23,264 typing sessions during natural interaction with touchscreen android smartphones to predict the PHQ-9 class. Demographic and clinical characteristics of the keystroke dynamics dataset are tabulated in Table 1. The characteristics are summarized for healthy users and users with depressive tendency. The two groups are categorized based on the depression severity level indicated by the PHQ-9 score.

Classification outcomes

Table 2 compares the evaluation metrics of the different machine learning classifiers implemented with the various feature selection techniques and hyperparameter optimization using nested cross-validation scheme. Among the proposed models, the GB classifier outperformed the other ML models, and the MI method achieved better performance compared to the other explored feature selection techniques. The GB classifier with the features selected by the MI method achieved an Area Under the receiver operating characteristic Curve (AUC) of 0.98, Sensitivity/Specificity (SE)/(SP) of 1.00/0.93, respectively, F1-score 0.95, and accuracy of 95.83%, for the corresponding probabilities of the subjects' status. The performance of the different implemented classification pipelines is analyzed by visualizing the Receiver Operating Characteristics (ROC) curves along with 95% Confidence Interval (CI) in Figure 1. As can be noticed from Table 2, the ensemble and boosting classifiers, including Random Forest, AdaBoost, and Gradient Boosting models, generally achieved better results compared to the other proposed ML classifiers. Additionally, it can be observed from Figure 1, the filter feature selection methods, namely chi-square and mutual information, have relatively better results than the embedded methods—lasso technique. This indicates that simple filter methods are quite sufficient to perform the feature selection task when selecting informative features from the hold time and flight time.

To further compare the different machine learning classifiers based on the AUC results, analysis of variance (ANOVA) statistical test is performed on the different classification models with the top FS method, i.e. MI method, and followed by Tukey post-hoc test since the ANOVA test indicates that at least one ML model is significantly different from the other ML models ($p < 0.05$). The Tukey test is applied to identify which specific

Classes	Healthy control	Depressed class	Total
No. of users (female)	14 (6)	10 (3)	24
Avg. age in years (std)	23.86 (4.44)	23.6 (3.24)	23.73 (3.95)
Avg. PHQ-9 (std)	2.29 (1.73)	10.9 (3.47)	–
No. of sessions	13069	10195	23264
Mean ± std of no. sessions per user	46±39	36±30	42±30

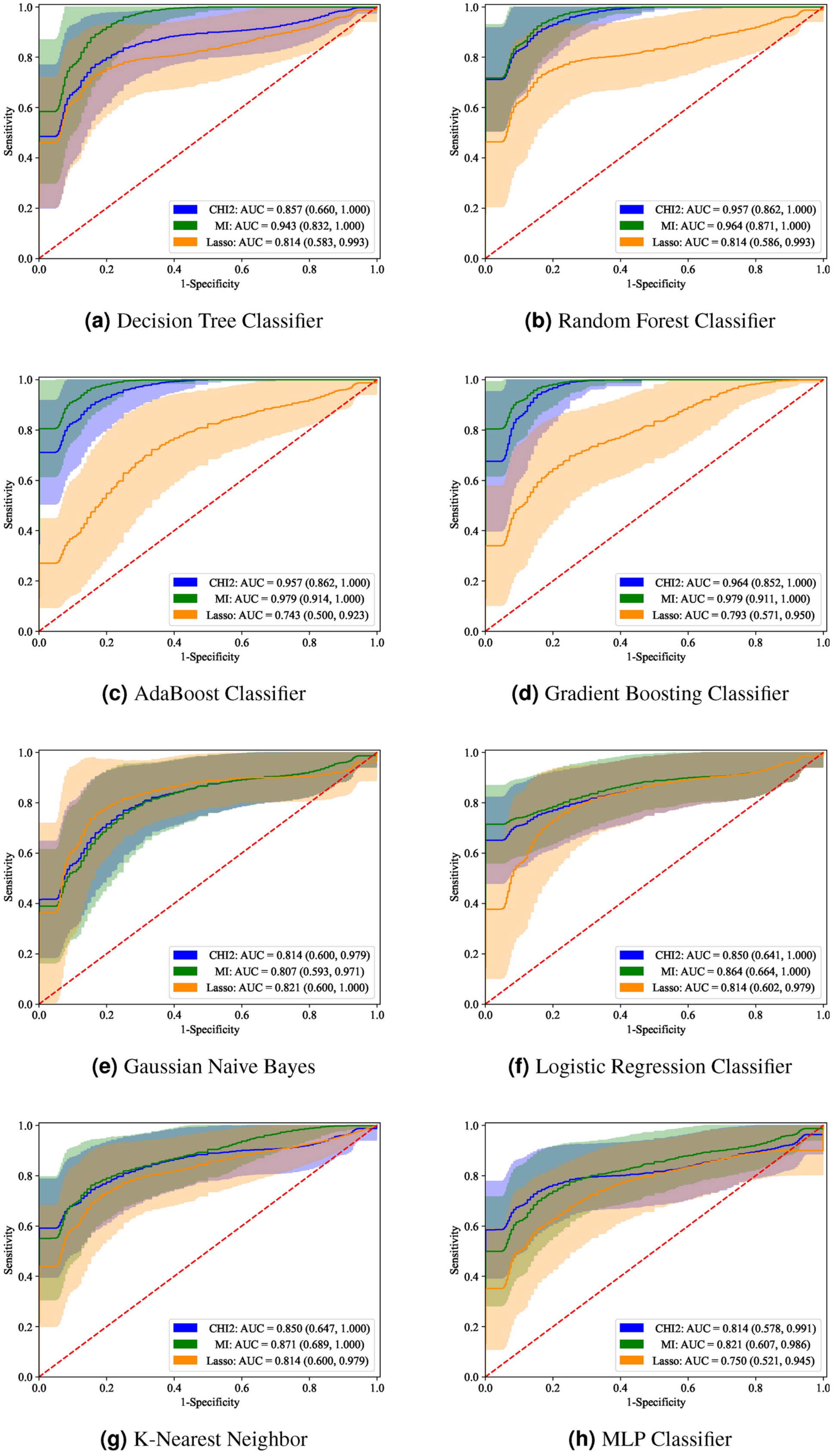
Table 1. Dataset demographic and clinical characteristics for healthy users and users with depressive tendency.

Model	Metrics	Different features selection methods		
		CHI2	MI	Lasso
Decision tree classifier	AUC	0.86 (0.66, 1.00)	0.94 (0.83, 1.00)	0.81 (0.58, 0.99)
	SE	0.80 (0.50, 1.00)	0.90 (0.67, 1.00)	0.80 (0.50, 1.00)
	SP	0.86 (0.64, 1.00)	0.93 (0.75, 1.00)	0.71 (0.50, 0.93)
	F1-score	0.80 (0.55, 0.96)	0.90 (0.71, 1.00)	0.73 (0.47, 0.90)
	Accuracy	0.83 (0.67, 0.96)	0.92 (0.79, 1.00)	0.75 (0.58, 0.92)
Random forest classifier	AUC	0.96 (0.86, 1.00)	0.96 (0.87, 1.00)	0.81 (0.59, 0.99)
	SE	0.80 (0.50, 1.00)	0.80 (0.50, 1.00)	0.80 (0.50, 1.00)
	SP	0.93 (0.75, 1.00)	0.93 (0.75, 1.00)	0.71 (0.50, 0.93)
	F1-score	0.84 (0.60, 1.00)	0.84 (0.60, 1.00)	0.73 (0.47, 0.90)
	Accuracy	0.88 (0.75, 1.00)	0.88 (0.75, 1.00)	0.75 (0.58, 0.92)
AdaBoost classifier	AUC	0.96 (0.85, 1.00)	0.98 (0.91, 1.00)	0.74 (0.50, 0.92)
	SE	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)	0.80 (0.50, 1.00)
	SP	0.86 (0.64, 1.00)	0.86 (0.64, 1.00)	0.71 (0.46, 0.92)
	F1-score	0.91 (0.74, 1.00)	0.91 (0.74, 1.00)	0.73 (0.46, 0.90)
	Accuracy	0.92 (0.79, 1.00)	0.92 (0.79, 1.00)	0.75 (0.58, 0.92)
Gradient boosting classifier	AUC	0.96 (0.85, 1.00)	**0.98 (0.91, 1.00)**	0.79 (0.57, 0.95)
	SE	1.00 (1.00, 1.00)	**1.00 (1.00, 1.00)**	0.70 (0.38, 1.00)
	SP	0.93 (0.75, 1.00)	**0.93 (0.75, 1.00)**	0.71 (0.46, 0.92)
	F1-score	0.95 (0.80, 1.00)	**0.95 (0.80, 1.00)**	0.67 (0.38, 0.87)
	Accuracy	0.96 (0.83, 1.00)	**0.96 (0.83, 1.00)**	0.71 (0.54, 0.88)
Gaussian Naive Bayes	AUC	0.81 (0.60, 0.98)	0.81 (0.59, 0.97)	0.82 (0.60, 1.00)
	SE	0.80 (0.50, 1.00)	0.80 (0.50, 1.00)	0.80 (0.50, 1.00)
	SP	0.71 (0.47, 0.93)	0.79 (0.56, 1.00)	0.71 (0.47, 0.93)
	F1-score	0.73 (0.47, 0.91)	0.76 (0.50, 0.94)	0.73 (0.47, 0.91)
	Accuracy	0.75 (0.58, 0.92)	0.79 (0.62, 0.96)	0.75 (0.58, 0.92)
Logistic regression classifier	AUC	0.85 (0.64, 1.00)	0.86 (0.66, 1.00)	0.81 (0.60, 0.98)
	SE	0.80 (0.50, 1.00)	0.80 (0.50, 1.00)	0.80 (0.50, 1.00)
	SP	0.79 (0.54, 1.00)	0.71 (0.50, 0.93)	0.71 (0.50, 0.93)
	F1-score	0.76 (0.50, 0.93)	0.73 (0.47, 0.90)	0.73 (0.47, 0.90)
	Accuracy	0.79 (0.62, 0.92)	0.75 (0.58, 0.92)	0.75 (0.58, 0.92)
K-nearest neighbors	AUC	0.85 (0.65, 1.00)	0.87 (0.69, 1.00)	0.81 (0.60, 0.98)
	SE	0.80 (0.50, 1.00)	0.80 (0.50, 1.00)	0.80 (0.50, 1.00)
	SP	0.71 (0.46, 0.93)	0.71 (0.46, 0.93)	0.71 (0.50, 0.93)
	F1-score	0.73 (0.44, 0.91)	0.73 (0.44, 0.91)	0.73 (0.47, 0.90)
	Accuracy	0.75 (0.58, 0.92)	0.75 (0.58, 0.92)	0.75 (0.58, 0.92)
MLP classifier	AUC	0.81 (0.58, 0.99)	0.82 (0.61, 0.99)	0.75 (0.52, 0.94)
	SE	0.70 (0.38, 1.00)	0.80 (0.50, 1.00)	0.60 (0.29, 0.89)
	SP	0.86 (0.67, 1.00)	0.71 (0.50, 0.93)	0.71 (0.50, 0.93)
	F1-score	0.74 (0.46, 0.93)	0.73 (0.47, 0.90)	0.60 (0.31, 0.82)
	Accuracy	0.79 (0.62, 0.96)	0.75 (0.58, 0.92)	0.67 (0.46, 0.83)

Table 2. Results of the machine learning classification models showing the evaluation metrics with 95% CI. Significant values are in bold.

ML model is significantly different from the others, and it shows that most of the proposed ML models are significantly different except for four models which show significant association between the GB classifier and AdaBoost classifier, as well as K-Nearest Neighbors and Logistic Regression classifiers.

Figure 2, which is inspired by the previous study[14], illustrates representative instances of the day-to-day variation in estimated probabilities for randomly selected individuals with depressive tendencies and HC. Each sub-plot corresponds to the complete dataset contributed by the individual, where the y-coordinate represents the daily average prediction probability of the individual, produced by the best-performing GBC-MI classification pipeline and input features representing the day's typing sessions. The opacity of each circle's color denotes the number of typing sessions conducted by the individual during the day, while the black diamond marker (⧫) indicates the number of typing sessions exceeding 50 sessions. In both cases, these typing sessions were used in the classification pipeline to generate the daily probability. Additionally, the dashed horizontal line represents the Youden index threshold for equal misclassification cost, as computed from the overall prediction probabilities of all participants. It can be seen from Figure 2, that despite the variations in the prediction probabilities of the



(a) Decision Tree Classifier

(b) Random Forest Classifier

(c) AdaBoost Classifier

(d) Gradient Boosting Classifier

(e) Gaussian Naive Bayes

(f) Logistic Regression Classifier

(g) K-Nearest Neighbor

(h) MLP Classifier

Figure 1. ROC curves for the machine learning classification models showing AUC with 95% CI.

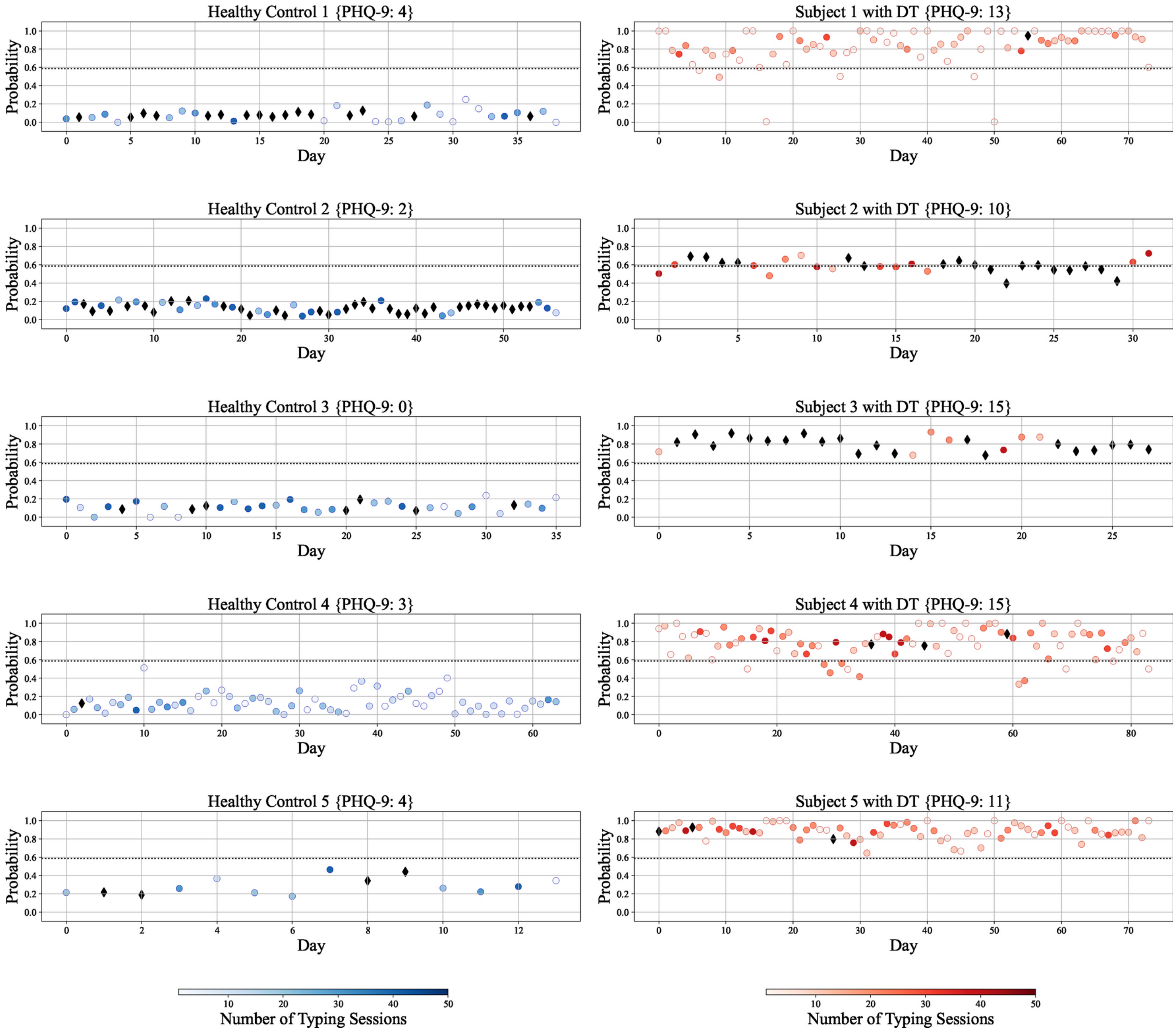


Figure 2. Day to day depression variation for random healthy controls and subjects with depressive tendency, respectively, using best-performing pipeline.

subjects' daily typing sessions, the overall classification output is not negatively affected, for both subjects with DT and HC.

After reporting and comparing the performance of the proposed diagnostic models, we further analyzed the best-performing model, gradient boosting classifier with MI feature selection technique (GBC-MI), to give clue on the features selected by the model using the average and standard deviation (STD) of the feature importance variable of the GB classifier across all the subjects, as shown in Figure 3. The top ten most informative features are ordered based on their importance from top to bottom, nine out of the top ten selected features are extracted from the hold time sequences, which show clear discrimination between HC and subjects with DT, while one feature is extracted from the flight time keystroke variable, particularly, flight time mean change.

Table 3 provides a performance comparison of our best-performing model with the previously published work[14]. The proposed GB model with MI feature selection technique outperformed the RF classifier published by Mastoras et al.[14] for detecting depressive tendency which employed the same keystroke dynamics dataset. Mastoras et al.[14] reported the median and STD of the HT, STD of the speed (SP), and median and STD of the press flight rate (PFR) variable as the best selected features with the best performing classifier. On the other hand, as Figure 3 shows, the HT mean change which refers to the average over the differences between subsequent HT series values, the different quantiles, and STD of the HT, along with the mean change of the FT sequences are reported as the top features with our proposed classification pipeline. The current work shows that extracting the mean change feature from the HT and FT sequences besides the quantiles of the HT provides more accurate classification results. Overall, the feature engineering techniques we applied to the keystroke time sequences before feeding the machine learning models considerably improved our classification performance compared to Mastoras et al.[14] model.

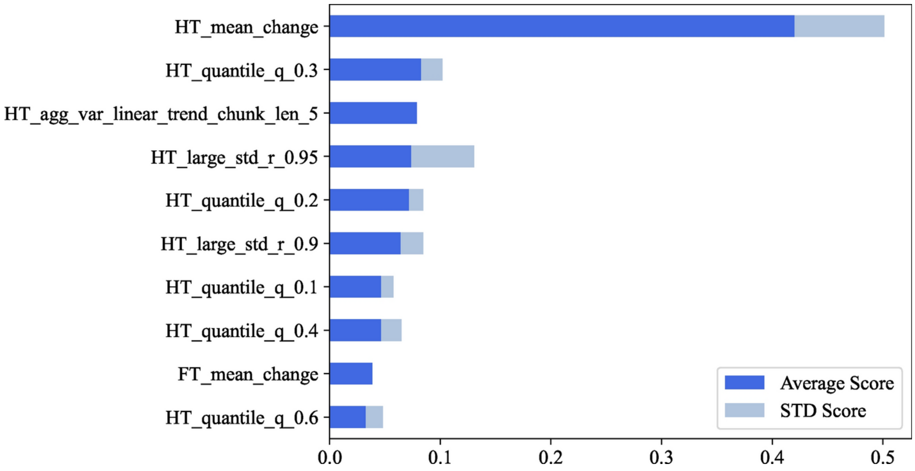


Figure 3. Feature importance results for the best-performing model showing top ten most informative features affecting significantly the model performance. Nine features are extracted from the hold time (HT) keystroke variable, while one feature is extracted from the flight time (FT) variable.

Study	Model	AUC	SE	SP
Mastoras et al.[14]	RF classifier	0.89	0.82	0.86
This work	GB classifier with MI FS	0.98	1.00	0.93

Table 3. Performance comparison with a previously published study on utilizing keystroke dynamics for classification of HC and DT subjects.

Discussion

Depressive Disorder remains one of the significant causes of disability across the globe. To manage and contain the spread of this illness, passive screening tools must be used to identify its symptoms. An approach to identifying depressive tendencies is through analyzing an individual's physical movements and behaviors, such as their interaction with touchscreen smartphones. These everyday activities can provide information on keystroke dynamics and their various characteristics. By analyzing keystroke dynamics data collected unobtrusively during natural interaction with touchscreen devices, patterns and indicators of depression disease can potentially be uncovered. In this work, new learning-based perspectives for processing touchscreen typing patterns to detect the existence of depressive disorder disease are presented. The passively collected keystroke dynamics are assessed as digital biomarkers for diagnosing depressive disorders. The mental health clinical assessment utilized to evaluate depressive disorder disease is the PHQ-9 depression assessment. The proposed machine learning classification pipeline utilized features extracted from the keystroke dynamics variables using various feature extraction and selection methods. The experimental findings have justified the efficiency of the Gradient Boosting model with the mutual information feature selection method in discriminating subjects with depressive tendency from healthy controls. The GB classifier can capture complex, non-linear relationships through iterative improvements, while MI selects the most relevant features by measuring their dependency on the target variable. Thus, the GB model combined with the MI feature selection method could reduce the noise and enhance the model performance leading to optimal results. The mean change and the different quantiles of the HT sequences that have been reported as the most important features for the best performing GB-MI model, as presented in Figure 3, reflect the variability that can be observed in the hold time sequences which can be linked to the fine motor skills and psychomotor impairment[16]. The promising results indicate the potentiality of the keystroke dynamic in capturing depressive disorders pathology and providing a remote screening tool for psychomotor impairment.

Concerning the observations and limitations of the proposed research work, this study utilized the PHQ-9 composite depressive score as an evaluation criterion to detect and predict the depressive tendencies of the subjects. The PHQ-9 assessment is a self-reported tool where the subjects fill out the questionnaire themselves in the wild without the involvement of clinicians. The use of the PHQ-9 as an indicator of depression tendencies in this study resulted in labeling noise in the collected data due to the self-reporting nature of the assessment. Future research could introduce additional clinician-based depression assessment tools to mitigate the labeling noise imposed in the collected data due to the employment of the self-reported PHQ-9 assessment alone as an evaluation metric. Alternatively, the study participants could also be asked to undertake the PHQ-9 depression assessment periodically to ensure that the assigned severity index is valid during the whole study period and reduce the opportunity of collecting noisy labeled data. Another limitation of this research work is that the procedure followed to collect the data from the Android touchscreen smartphone users is not unified with the iOS data collection tool, which impedes the utilization of all the data collected across the different platforms in the proposed framework. However, the large number of typing sessions collected from the android users addresses this limitation. A point that could be considered for future improvements of this study is the scalability of the study participants as well as the data acquisition procedure, where a unified methodology would be applied for both Android and iOS touchscreen smartphone users to collect the keystroke dynamics data from the typing sessions. Moreover, additional data decomposition approaches could be investigated to extract informative features and generate input data with higher dimensional. Overall, quantitatively and qualitatively, dedicating more efforts toward enhancing the keystroke dynamics data representation is crucial for further improvement of the proposed framework.

Methods

Figure 4 provides the workflow of the classification pipeline. Further details of the workflow are discussed in the following sections.

Study protocol and data acquisition

The data used in this work were collected unobtrusively during the activities of daily living via an earlier remote study[14] using a mobile application, namely TypeOfMood[17]. The study protocol used to collect this data was reviewed and authorized by the American Center of Psychiatry and Neurology of Abu Dhabi (Bioethics Committee, approval IRB reference 0022 ACPN). Recruitment and study procedures were performed in accordance with institutional and international guidelines on research involving adult human beings. All subjects provided informed consent to share their data processing within the application before participating in the study. This study considered privacy issues and collected the timing sequences of keystrokes without using or logging the actual content of the typed text.

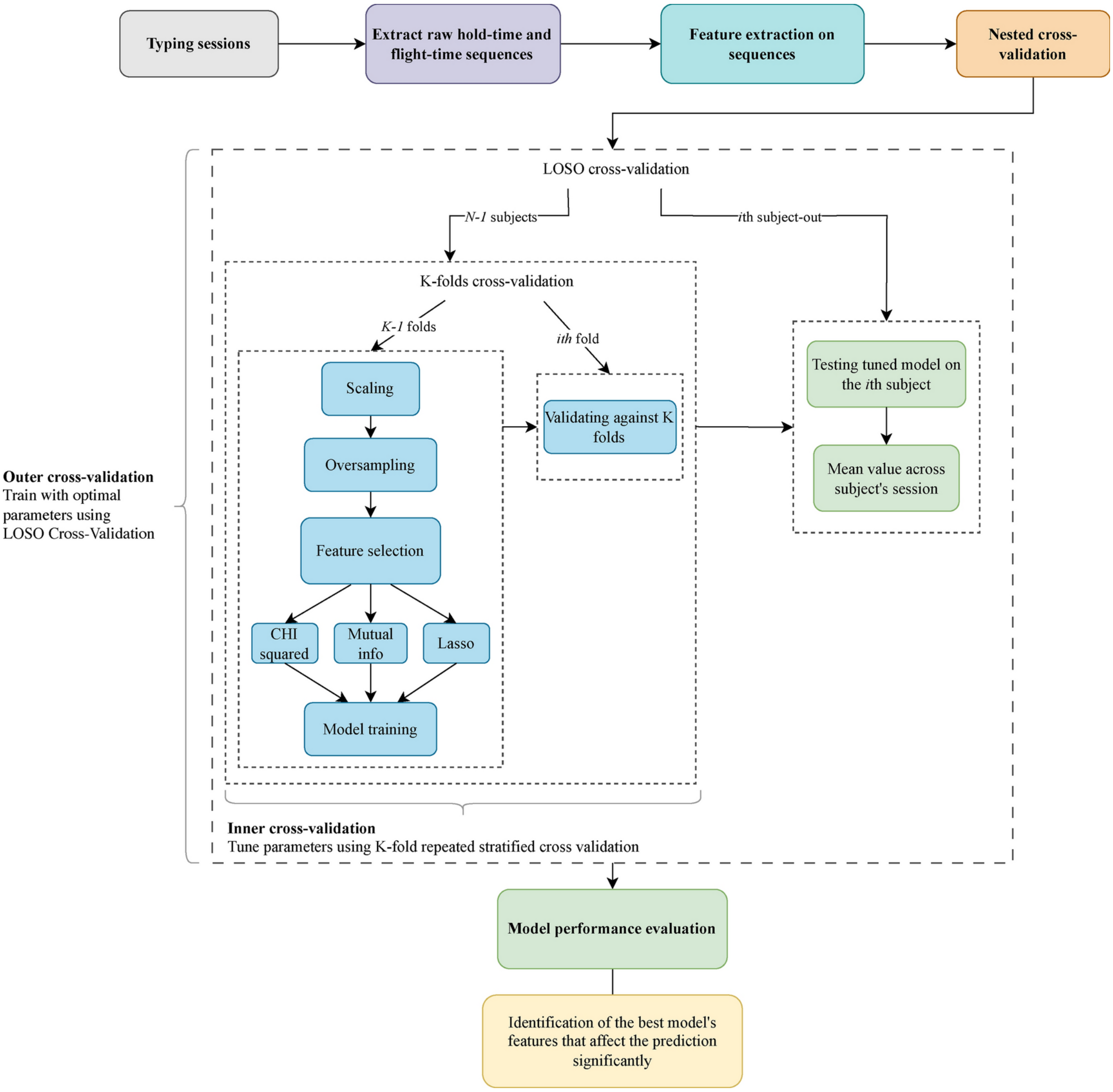

Typing sessions
Extract raw hold-time and flight-time sequences
Feature extraction on sequences
Nested cross-validation
LOSO cross-validation
N-1 subjects
ith subject-out
K-folds cross-validation
K-1 folds
ith fold
Scaling
Oversampling
Feature selection
CHI squared
Mutual info
Lasso
Model training
Validating against K folds
Testing tuned model on the ith subject
Mean value across subject's session
Outer cross-validation
Train with optimal parameters using LOSO Cross-Validation
Inner cross-validation
Tune parameters using K-fold repeated stratified cross validation
Model performance evaluation
Identification of the best model's features that affect the prediction significantly

Figure 4. Depressive disorder classification pipeline.

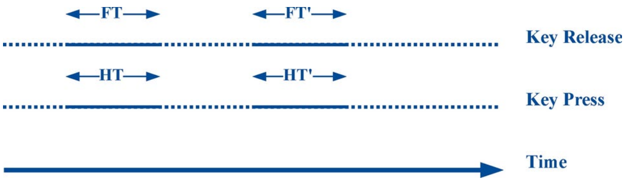


Figure 5. Keystroke dynamics variables.

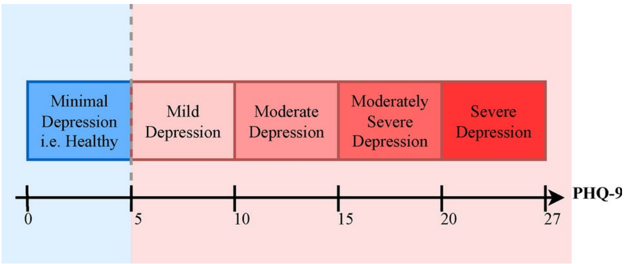


Figure 6. Depressive disorder severity classes based on PHQ-9 score.

To analyze the correlation between DD and keystroke dynamics, TypeOfMood android mobile application was utilized to collect the meta-data of typing characteristics. It included a custom software keyboard with all modern functionalities, such as auto-correction and word prediction. The custom keyboard supports several languages, such as English and Greek. The data collected by the application refers to the emotional state that is self-reported by the user using the emotions application and meta-data of keystroke dynamics captured unobtrusively by the application's custom keyboard, including keystrokes timing information, key switching speed, and other typing meta-data such as delete rate, pauses, and typing sessions' duration. However, the collected keystroke data are independent of the keyboard layout. This study utilized 23,264 typing sessions provided by 24 subjects with various depression severity levels.

Keystroke dynamics definition

The typing-related characteristics passively captured during natural interaction with touchscreen smartphones are utilized in this research as digital biomarkers to detect the existence of the depressive disorder. This study employs two keystroke variables, the hold time and flight time sequences, as raw data input for the proposed classification pipeline. The hold time is the duration between pressing and releasing a key, while the flight time refers to the duration between releasing one key and pressing the next one. Figure 5 illustrates the defined keystroke dynamics variables as derived from[14]. The values of the hold time HT_n and flight time FT_n in the HT and FT sequences are defined, respectively, by the following equations:

$$\begin{aligned} HT_n &= t_n^r - t_n^p, \quad n = 1, 2, \cdots, N \\ FT_n &= t_{n+1}^p - t_n^r, \quad n = 1, 2, \cdots, N-1 \end{aligned}$$

Where t_n^p is the touchscreen press event timestamp, t_n^r is the timestamp of the release event, and N is the total number of keystrokes in the session.

Mental health clinical assessment

Considerable attention has been directed toward detecting depression, yet the severity assessment is also essential for providing professional treatment. The mental health measures for diagnosing the pathology and monitoring the response to therapy outside clinical settings are subject to multiple factors, including sensitivity to change, number of items, the time required for completion, mode of administration (self-rating vs. interviewer-administrative scale), and complexity of scoring, along with the specific items included in the scale. Many comparable measures are available for identifying depression, including several self-administrated measures. The PHQ-9 is one of the most valuable and easy-to-use self-reported measures for depressive disorder diagnosis. The PHQ is a criteria-based instrument for diagnosing depressive disorder and other common mental disorders. It is a self-administrative scale[18]. The PHQ-9 depressive questionnaire is a part of the full PHQ module, consisting of nine depressive items. These items represent the diagnostic criteria for DSM-V depressive disorders. The PHQ-9 can be used for assessing the severity of the DD, ranging from 0 to 27, where each of the nine depressive items can be scored with 0 (not at all), 1 (several days), 2 (more than half the days), or 3 (nearly every day) for the past two weeks. This depressive score is further divided into five categories of increasing severity, based on the 27-scale as follows: 0-4, 5-9, 10-14, 15-19, and 20-27, for minimal, mild, moderate, moderately severe, and severe depression, respectively[10], as depicted in Figure 6. The PHQ-9 score is employed as an evaluation criterion in this research to reveal the depressive tendency of the subjects based on the PHQ-9 score. The symptoms of mild depressive disorder with a cut-off threshold of five on the PHQ-9 scale are used to distinguish between healthy subjects and subjects with depressive tendencies.

Data preprocessing

The TypeOfMood dataset is used in this work to train and evaluate the depression detection models. Firstly, the data files were read and prepared with a unified definition. A preliminary exclusion procedure is applied to the collected data. Out of all the participants in the study, we only considered adult subjects, less than 40 years old, who composed a matched group in terms of age and gender and did not suffer from any psych-motor impairments. All subjects used smartphones and text messaging for over a year before the study.

After preparing and cleaning the dataset, a conditional filtering scheme is applied to the time sequences, where long presses that might correspond to special events are discarded, and hold time and flight time sequences with more than 300 milliseconds and three seconds, respectively, are removed. Moreover, the sessions are selected based on length, where a valid session has at least nine keystrokes. Fixed-length vectors with a length of 100 are also generated for HT and FT sequences by applying zero padding on sequences with less than this threshold and resizing sequences with more than the defined length. Furthermore, the subjects are included in the study based on the number of typing sessions for each subject, where valid subjects provide at least 50 sessions each.

Feature engineering

Feature engineering, which includes feature extraction, scaling, oversampling, and feature selection techniques, significantly affect building accurate and reliable ML models. By carefully selecting and transforming the input features, feature engineering can improve the model's ability to learn and generalize from the data[19]. This is particularly important for keystroke timing data, where the temporal dynamics and patterns are often complex and difficult to capture using raw keystroke data. Moreover, feature engineering can reduce noise and redundancy in the data, identify relevant information, and enhance the discriminatory power of the model. The proposed machine learning models for addressing the problem of DT detection utilized various informative features extracted from the hold time and flight time sequences. We applied Min-max scaling on the extracted HT and FT features for data normalization. Then, we employed the Synthetic Minority Oversampling Technique (SMOTE)[20] to generate synthetic typing sessions for the minority class, namely the depressed class, in the keystroke dataset. Further, we examined different feature selection techniques with different ML classifiers, and their performance is evaluated, reported, and analyzed.

Feature extraction

Time series data analysis involves the extraction of useful features that can be used to model and predict future trends. The TSFRESH (Time Series Feature extraction based on Scalable Hypothesis tests) Python package is a popular tool for automatically extracting a broad range of features from time series data[21,22]. These features are extracted using various mathematical and statistical techniques, including signal processing, information theory, and machine learning.

The extracted features can be broadly categorized into the following categories: Statistical Features: this category includes various statistical measures such as mean, median, standard deviation, variance, skewness, kurtosis, maximum, minimum, range, interquartile range, the absolute sum of changes, and others. These measures are useful for providing information about the distribution, central tendency, and variability of the time series data. Information-Theoretic Features: this category includes measures based on information theory, such as entropy, cross-entropy, and mutual information. These features quantify the amount of information present in the time series data and can provide insights into the structure, complexity, and predictability of the data. Fourier Features: these measures are extracted using Fourier analysis, which involves decomposing a signal into its frequency components, such as amplitude, phase, frequency, and spectral density. These features are useful for describing the frequency content and identifying the periodic patterns in the time series data. Time-Domain Features: this category includes measures based on the time domain, such as autocorrelation, number of crossings, linear trend, and others. These measures capture the temporal structure of the time series data. Value Count Features: this category includes measures based on the frequency of different values in the time series data, such as the number of values above or below a threshold, the number of distinct values, and others. These measures provide information about the occurrence and distribution of different values in the time series data. Moment Features: this category includes measures of higher-order moments of the distribution, such as skewness, kurtosis, and others. These measures provide information about the shape and symmetry of the distribution of the time series data. Peak Detection Features: this category includes measures related to the detection of peaks in the time series data, such as the location of peaks, amplitude of peaks, number of peaks, and others. These measures provide information about the occurrence and magnitude of significant events in the time series data. Change Detection Features: this category includes measures related to the detection of changes in the time series data, such as the time and magnitude of change in the data. These measures provide information about the stability and evolution of the time series data over time. Additional features, such as energy, root mean square, and quantiles, provide additional information about the data that can be useful for modeling and prediction tasks.

Overall, we utilized the TSFRESH package to provide a comprehensive set of features that can be used for the task of depressive disorder detection using the keystroke time series data. All the 63 feature extraction characterization methods offered by the TSFRESH package[21] are used to extract the various features from the hold time and flight time sequences, which resulted in a total of 639 features. These features have been further used with different feature selection approaches to select the most informative features based on their relevance with the PHQ-9 depression classes and employed in the machine learning pipeline for the depression detection task.

Feature selection

Feature selection improves the overall performance of a predictive model by eliminating irrelevant and redundant features from the input data. The primary objective of feature selection is to identify the most relevant features to the class output that contribute the most to the classification task by reducing the noise in the data, which results in decreasing the computational cost involved in building the model[23].

Feature selection methods can be categorized into different categories, including filter methods and embedded methods. In filter FS approaches, the features are selected based on various statistical tests that correlate the features with the dependent variable[24]. Filter methods do not involve predictive models, which results in a fast feature selection approach that is less prone to overfitting[25]. One of the filter methods is the Chi-square method[26]. It is a statistical test that uses the frequency distribution of the categorical features to evaluate the association between the groups of these features. The Mutual Information FS technique, which belongs to the filter methods, is a non-parametric and computationally efficient approach that can capture complex relationships between features[27]. In the MI method, the mutual information score is calculated between each feature and the target variable. The score represents the reduction in uncertainty of the target variable when the feature is known. Features with higher MI scores are considered more informative and are more likely to improve the performance of a machine learning model. Additionally, embedded methods combine the advantages of filter and other FS methods; by including their own built-in feature selection methods that implement regularization terms to avoid overfitting. This makes embedded methods less computationally expensive and less prone to overfitting. The feature selection process is embedded in the model learning. Lasso regression is a linear model and one example of an embedded method that performs L1 regularization. In this research, Lasso regression is utilized as a feature selection method because the extracted features from the hold time and flight time sequences showed some linearity with the PHQ9 class. Overall, the different types of feature selection methods provide different tradeoffs between performance, computational cost, and overfitting risk. Choosing the appropriate method depends on the specific task and the characteristics of the data.

In this work, we utilized the CHI2, MI, and lasso FS methods to provide a balanced approach that combines statistical relevance, the ability to capture complex relationships, and regularization to prevent overfitting, and avoids incorporating FS techniques that are computationally expensive and prone to overfitting. We implemented the CHI2 method using the select-k-best algorithm, with an upper limit of the selected features set to twenty ($k = 20$). In the MI method, the MI score is utilized to select the features based on their importance with a lower limit of 0.2, where the features with an MI score greater than 0.2 are selected for the later stage. Finally, the Lasso FS method is set with a threshold of 0.1.

Machine learning models

The most common machine learning classifiers are adopted in this research to evaluate and analyze the performance of the binary classification pipeline. We employed ML classifiers that are capable to handle sophisticated relationships between keystroke features and DT status. The proposed classifiers can be categorized into tree-based classifiers, boosting classifiers, probabilistic classifiers, distance-based classifiers, and neural network-based classifiers. The tree-based classifiers build a tree structure that recursively partitions the data into subsets based on the most informative features, resulting in a set of decision rules that can be used for classification. The Decision Tree and Random Forest classifiers[28,29] are tree-based classifiers used in the proposed depression detection pipeline. The boosting classifiers presented in AdaBoost and Gradient Boosting classifiers[30,31] are also explored. These classifiers are ensemble methods that combine multiple weak learners to create a more robust and accurate model. Moreover, the proposed classification pipeline is implemented using probabilistic classifiers such as Gaussian Naive Bayes and Logistic Regression[32], which model the probability distribution of the input data to predict the class labels. Further, the K-Nearest Neighbors distance-based classifier[33] and the multi-layer perceptron neural network classifier[34] are introduced to detect depressive tendencies based on touchscreen typing characteristics. The scenarios that would be examined for the data input to feed the machine learning classifiers are combinations of the informative features extracted from the raw HT and FT time sequences obtained from the typing sessions using various feature selection techniques.

The binary classification pipeline proposed to discriminate subjects with depressive tendencies from healthy controls is evaluated and optimized on a subject level. The typing sessions of each subject are inputted into the particular trained detector model, which generates a probability score for each typing session. These scores are then aggregated for each subject using an aggregator function to decide on the overall subject status and determine his class, whether depressed or healthy control. The statistical mean function is employed as an aggregator in this work.

Models tuning with nested cross-validation

The performance of machine learning algorithms is significantly dependent on tuning the model's hyperparameters. Hyper-parameters are used in the learning process of the algorithm and control how it behaves. They should be tuned to maximize the algorithm performance for the addressed problem. The number of filters and kernel size in a convolutional layer, the number of hidden layers, the activation function, and the number of estimators are examples of the network structure definition hyper-parameters that define the model structure. On the other hand, the training-related hyper-parameters are those such as the optimizer function and its learning rate, the batch size, and the number of training epochs. The hyperparameters tuning process is a time-consuming task, especially when performed manually, that depends on trial and error.

Therefore, automated hyperparameter optimization (HPO) frameworks are emerging and have become essential for an efficient hyperparameter tuning process[35,36]. These optimization frameworks aim to find the set of optimal hyper-parameters that maximize -or minimize- the evaluation metric. Optuna is an automatic hyper-parameter optimization framework with a define-by-run API, enabling users to dynamically construct the hyper-parameters search space. It also combines sampling and pruning algorithms which significantly improves the effectiveness of the hyper-parameters optimization. It sequentially evaluates each trial using Tree-structured Parzen estimators (TPE) as the default sampler[37]. In this study, we employed the Optuna HPO framework to automate the hyperparameter tuning process for the proposed diagnostic models. The hyperparameters search space used for fine-tuning the different machine learning classification models is presented in Table 4.

Another critical step in the model-tuning process and evaluation of the model performance is the cross-validation scheme. The proposed depression tendency detectors are trained and validated using a nested cross-validation scheme[38]. In the nested cross-validation, the hyperparameter optimization is performed in the inner cross-validation, while the outer cross-validation computes an unbiased estimate of the model performance. Nested cross-validation can help improve the proposed framework's reliability and generalizability.

The performance of the binary classification pipeline is evaluated in the outer cross-validation using the leave-one-subject-out (LOSO) scheme, where the model is trained *N* times. On each iteration, the typing sessions of one subject are left out for testing, and the remaining subjects' sessions are used for tuning the model hyperparameters and training the model. The model tuning process in the inner cross-validation is performed by training and validating the model under a stratified *k*-fold cross-validation scheme using five-folds on the $N - 1$ subjects. On each iteration of the five folds, four folds of the data are used to train the model for the hyperparameters tuning procedure, and the fifth fold is used to validate the tuned model. The objective function of the HPO framework selects the best model hyperparameters values based on the average of the evaluation metric across the five folds. Then, the chosen hyperparameters build the classification model, which is trained with the $N - 1$ subjects and tested on the typing sessions of the left-out subject. This approach is repeated for each of the *N* subjects in the outer cross-validation.

Model	Parameters	Parameters space
Decision Tree Classifier	Criterion	{gini, entropy}
	Max depth	{None, 3, 5, 7, 9}
	Max features	{None, sqrt, log2}
	Min samples split	{0.5, 2, 4, 8}
	Min samples leaf	{0.25, 1, 2, 4}
Random Forest Classifier	Number of estimators	{10, 50, 100, 500}
	Criterion	{gini, entropy}
	Max depth	{None, 3, 5, 7, 9}
	Max features	{None, sqrt, log2}
	Min samples split	{0.5, 2, 4, 8}
	Min samples leaf	{0.25, 1, 2, 4}
AdaBoost Classifier	Number of estimators	{10, 50, 300, 750}
	Learning rate	{0.001, 0.01, 0.1, 1}
Gradient Boosting Classifier	Number of estimators	{10, 50, 100, 500}
	Min samples split	{0.5, 2, 4, 8}
	Min samples leaf	{0.25, 1, 2, 4}
	Max depth	{None, 3, 5, 7, 9}
	Learning rate	{0.001, 0.01, 0.1, 1}
Gaussian Naive Bayes	variance smoothing	{1, 1e-1, 1e-2, 1e-3, 1e-4, 1e-5, 1e-6, 1e-7, 1e-8, 1e-9, 1e-10}
Logistic Regression Classifier	Penalty	{l2, none}
	Regularization strength inverse	{0.1, 1, 10, 100}
	Solver	{lbfgs, sag, saga}
K-Nearest Neighbors	Number of neighbors	{2, 3, 4, 5, 6}
	Weights	{uniform, distance}
	Algorithm	{auto, ball_tree, kd_tree, brute}
MLP Classifier	Hidden layer sizes	{(100,), (100,100), (50,100), (50,50), (50,100,100), (50,50,50)}
	Activation	{'relu', 'logistic', 'tanh'}
	Solver	{adam, sgd}
	Alpha	{0.0001, 0.001, 0.01, 0.1}
	Learning rate	{constant, adaptive}

Table 4. Hyperparameters space for machine learning depression detection models.

Models performance evaluation

The detection and prediction performance of the proposed classification models is evaluated for different settings of hyperparameters optimization with nested cross-validation scheme. The ROC analysis[39] is employed as a performance evaluation tool for the proposed DT detection models by estimating the AUC. Moreover, SE, SP, F1-score, and accuracy metrics are used to assess the performance of the DT detectors. The classification performance of each optimization setting is evaluated using the average value and 95% CI of the AUC, SE, SP, F1-score, and accuracy, over 1,000 bootstraps to enhance the results' statistical reliability. The DT detection models are evaluated across the typing sessions of the subjects.

Data availability

All data generated and analyzed during the current study are available on request from the corresponding author.

Received: 23 July 2024; Accepted: 31 October 2024
Published online: 14 November 2024

References

1. Collaborators, G. M. D. et al. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019. *Lancet Psychiatry* **9**, 137–150 (2022).
2. Whiteford, H. A., Ferrari, A. J., Degenhardt, L., Feigin, V. & Vos, T. The global burden of mental, neurological and substance use disorders: an analysis from the global burden of disease study 2010. *PLoS ONE* **10**, e0116820 (2015).
3. Wang, P. S. et al. Delay and failure in treatment seeking after first onset of mental disorders in the world health organization's world mental health survey initiative. *World Psychiatry* **6**, 177 (2007).
4. Guilbert, J. J. The world health report 2002 - reducing risks, promoting healthy life. *Educ. Health* **16**, 230 (2002).
5. Marcus, M. *et al.* Depression: a global public health concern. In *Depression: A Global Public Health Concern* (2012).
6. Saloni Dattani, H. R. & Roser, M. Mental health. *Our World in Data* (2021).
7. GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the global burden of disease study 2017. *Lancet* **392**, 1789–1858 (2017).
8. Reddy, M. S. Depression: The disorder and the burden. *Indian J. Psychol. Med.* **32**, 1–2 (2010).
9. Schumann, I., Schneider, A., Kantert, C., Löwe, B. & Linde, K. Physicians attitudes, diagnostic process and barriers regarding depression diagnosis in primary care: a systematic review of qualitative studies. *Fam. Pract.* **29**, 255–263 (2011).
10. Kroenke, K., Spitzer, R. L. & Williams, J. B. W. The PHQ-9. *J. Gen. Intern. Med.* **16**, 606–613 (2001).
11. Macias Alonso, A. K., Hirt, J., Woelfle, T., Janiaud, P. & Hemkens, L. G. Definitions of digital biomarkers: a systematic mapping of the biomedical literature. *BMJ Health Care Inform* **31**, e100914 (2024).
12. Pentland, A. *Honest Signals: How they Shape our World* (MIT Press, 2010).
13. Goncalves, L. & Busso, C. Robust audiovisual emotion recognition: Aligning modalities, capturing temporal information, and handling missing features. *IEEE Trans. Affect. Comput.* **13**, 2156–2170 (2022).
14. Mastoras, R.-E. et al. Touchscreen typing pattern analysis for remote detection of the depressive tendency. *Sci. Rep.* **9**, 13414 (2019).
15. Deng, Y. & Zhong, Y. Keystroke dynamics user authentication based on gaussian mixture model and deep belief nets. *ISRN Signal Process*. 565183 (2013).
16. Giancardo, L., SÃ¡nchez-Ferro, A., Butterworth, I., Mendoza, C. S. & Hooker, J. M. Psychomotor impairment detection via finger interactions with a computer keyboard during natural typing. *Sci. Rep.* **5**, 9678 (2015).
17. Type of mood (2018). Publisher i-PROGNOSIS EU H2020 funded project (GA No. 690494).
18. Spitzer, R. L., Kroenke, K. & Williams, J. B. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. primary care evaluation of mental disorders. patient health questionnaire. *JAMA* **282**, 1737–1744 (1999).
19. Agrawal, S. & Sharma, D. K. Feature extraction and selection techniques for time series data classification: A comparative analysis. In *2022 9th International Conference on Computing for Sustainable Global Development (INDIACom)*, 860–865 (2022).
20. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).
21. Christ, M., Braun, N., Neuffer, J. & Kempa-Liehr, A. W. Time series FeatuRe extraction on basis of scalable hypothesis tests (tsfresh-a python package). *Neurocomputing* **307**, 72–77 (2018).
22. Barandas, M. et al. TSFEL: Time series feature extraction library. *SoftwareX* **11**, 100456 (2020).
23. Deng, X., Li, Y., Weng, J. & Zhang, J. Feature selection for text classification: A review. *Multimed. Tools Appl.* **78**, 3797–3816 (2019).
24. Too, J. & Mirjalili, S. General learning equilibrium optimizer: A new feature selection method for biological data classification. *Appl. Artif. Intell.* **35**, 247–263 (2021).
25. Hu, L., Gao, W., Zhao, K., Zhang, P. & Wang, F. Feature selection considering two types of feature relevancy and feature interdependency. *Expert Syst. Appl.* **93**, 423–434 (2018).
26. Uysal, A. K. & Gunal, S. A novel probabilistic feature selection method for text classification. *Knowl.-Based Syst.* **36**, 226–235 (2012).
27. Chandrashekar, G. & Sahin, F. A survey on feature selection methods. *Comput. Electr. Eng.* **40**, 16–28 (2014).
28. Song, Y.-Y. & Lu, Y. Decision tree methods: applications for classification and prediction. *Shanghai Arch. Psychiatry* **27**, 130–135 (2015).
29. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
30. Cao, Y., Miao, Q.-G., Liu, J.-C. & Gao, L. Advance and prospects of AdaBoost algorithm. *Acta Automatica Sinica* **39**, 745–758 (2013).
31. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001).
32. Dreiseitl, S. & Ohno-Machado, L. Logistic regression and artificial neural network classification models: a methodology review. *J. Biomed. Inform.* **35**, 352–359 (2002).
33. Peterson, L. E. K-nearest neighbor. *Scholarpedia* **4**, 1883 (2009).
34. Windeatt, T. Accuracy/diversity and ensemble MLP classifier design. *IEEE Trans. Neural Netw.* **17**, 1194–1211 (2006).
35. Yu, T. & Zhu, H. Hyper-parameter optimization: A review of algorithms and applications. *arXiv*:2003.05689 [cs, stat] (2020).
36. Huang, C., Yuan, B., Li, Y. & Yao, X. Automatic parameter tuning using bayesian optimization method. In *2019 IEEE Congress on Evolutionary Computation (CEC)*, 2090–2097 (2019).
37. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, 2623–2631 (Association for Computing Machinery, 2019).
38. Wainer, J. & Cawley, G. Nested cross-validation when selecting classifiers is overzealous for most practical applications. *Expert Syst. Appl.* **182**, 115222 (2021).

39. Hajian-Tilaki, K. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian J. Intern. Med.* **4**, 627–635 (2013).

Acknowledgements

This work was financially supported by ASPIRE Award for Research Excellence (AARE-20), Abu Dhabi, UAE, under grant number AARE20-293.

Author contributions

L.H. conceived the data acquisition experiments. R.F. and A.A. designed the machine learning pipelines. R.F. implemented the machine learning models and applied them to the experimental data. R.F., A.A., and L.H. analyzed the results. R.F. wrote the first draft, and A.A. and L.H. contributed to the writing and editing. All authors reviewed, discussed, and agreed on submitting the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to R.F.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.