INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

ORIGINAL ARTICLE

Acute Stress Detection through Typing Behaviors, Smartphone Sensors and Machine Learning Techniques

Purba Banerjee[1], Soumen Roy[2], Sandip Dutta[3]*, Utpal Roy[1]

1 Department of Computer and System Sciences, Visva-Bharati, Santiniketan, 731235, West Bengal, India
2 Department of Computer Science, Bagnan College, Khalore, Bagnan, Howrah, 711303, West Bengal, India
3 Department of Computer Science and Engineering, School of Engineering and Technology, Adamas University, Kolkata, West Bengal, India

OPEN ACCESS

Received: 26/08/2026
Accepted: 07/10/2026
Published: 25/10/2026

Citation: Banerjee P, Roy S, Dutta S, Roy U (2026) Acute Stress Detection through Typing Behaviors, Smartphone Sensors and Machine Learning Techniques. Indian Journal of Science and Technology 19(38): 2624-2638. https://doi.org/10.17485/IJST/v19i38.1223

* **Corresponding author**.

sandip.dutta2@adamasuniversity.ac.in

Funding: None

Competing Interests: None

Copyright: © 2026 Banerjee et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: To examine whether acute stress can be recognised from smartphone typing behaviour and device motion without wearable sensors, and to identify the classifier family best suited to such data. **Method:** A browser-based application recorded keystroke events together with accelerometer and gyroscope readings (2 Hz) while 288 participants typed simple number series (calm condition) and mental-arithmetic sequences (stress condition), yielding 2,016 samples. Five keystroke timing features and the inertial features were combined, and ten supervised classifiers were trained and tested on the same stratified 80:20 split using accuracy, AUC and weighted and macro precision, recall and F1-score. **Findings:** Ensemble methods outperformed the linear, probabilistic, distance-based and neural models. XGBoost achieved the best result, with 89.30% accuracy and an AUC of 95.42%, and Random Forest was close behind (88.72%, 95.32%); Naïve Bayes was weakest (57.98%). Weighted and macro F1 scores of the leading models differed by less than 0.3 points, indicating balanced detection of both classes. The accuracy exceeds the 87.56% reported by the closest smartphone motion-sensor study on 46 participants, and the ranking of XGBoost ahead of Random Forest agrees with recent wearable-sensor studies. Unlike deep models applied to dense physiological signals, the multilayer perceptron performed poorly on these compact behavioural features, which indicates that tree ensembles suit this type of data better. **Novelty:** Keystroke dynamics and smartphone inertial sensing are combined in one installation-free, synchronised protocol and benchmarked with ten classifiers on a cohort six times larger than the closest smartphone study.

Keywords: Acute Stress Detection; Keystroke Dynamics; Smartphone Sensing; Machine Learning; Behavioural Biometrics

1 Introduction

Acute stress lowers cognitive performance, disturbs psychomotor coordination and weakens decision making, and repeated episodes contribute to burnout in demanding professions [1]. Its early detection is therefore relevant to digital mental healthcare, workplace well-being and adaptive human–computer interfaces [2]. Recent reviews of wearable-based detection agree that physiological signals such as ECG, EDA and HRV can measure stress accurately, but they also show that dedicated sensors, controlled recording and continuous user compliance keep these systems in the laboratory [3], [4]. Pooling open physiological datasets does not remove this problem either: models depend strongly on the sensors and protocols used to collect the data [5]. Accuracy is thus no longer the main obstacle; practical deployment is.

The smartphone is a more practical source of data, because the keyboard and the built-in sensors record behaviour during normal use. Keystroke timing has been proposed as a digital biomarker of neurocognitive state [6], and facial expressions recorded during smartphone use have been shown to track perceived stress [7]. Keystroke dynamics has also been used with ensemble classifiers to detect neurological and stress-related changes on touch devices [8]. The closest work to ours used only the accelerometer and gyroscope while 46 participants typed and reached 87.56% accuracy with k-NN [9], and StresSense likewise relied on motion data alone [10]. A critical reading of these studies shows three weaknesses. Samples are small and mostly homogeneous; typing and motion are studied separately although stress affects both at the same moment; and studies change the device, the task and the classifier at the same time, so it is difficult to say whether a gain comes from an added modality or from a changed protocol [2]. One question therefore remains open: does adding inertial sensing to keystroke dynamics give reliable stress detection when participants, tasks and classifiers are kept the same?

This study addresses that question, and its novelty lies in three points. We present a web-based framework that records keystroke dynamics and smartphone accelerometer and gyroscope data in the same session, under a controlled calm-versus-stress protocol, and we compare ten supervised classifiers on the resulting multimodal features. The contributions are threefold. First, a browser-based acquisition platform captures typing and inertial signals in parallel without any dedicated app or hardware. Second, a multimodal behavioural dataset of 2,016 samples was collected from 288 participants under matched calm and induced-stress conditions. Third, ten classifiers are compared on identical data and splits, which separates the effect of the classifier from the differences in device, protocol and cohort that make earlier single-modality studies hard to compare.

1.1 Stress detection modalities

Stress detection has traditionally relied on physiological measurements of the autonomic and neural response to psychological stress. The modalities most often used are electrocardiography (ECG), electroencephalography (EEG), electromyography (EMG), electrodermal activity (EDA), photoplethysmography (PPG), respiration, skin temperature and heart-rate variability (HRV). They are usually recorded with wearables or laboratory instruments, and under controlled conditions they respond reliably to stress [4], [5], [11], [12].

Behavioural stress detection developed alongside these methods as a less intrusive option. Instead of measuring a body signal, it infers stress from what people are already doing: typing, touching the screen, walking, speaking, facial expression, body movement, phone usage and other context gathered during ordinary device use. Much of this can be recorded passively and without a wearable sensor, which is more comfortable for the user and better suited to long-term monitoring outside the laboratory [3], [4], [13].

Several recent systems combine physiological and behavioural modalities to make recognition more reliable. Figure 1 shows the body locations and interaction channels from which stress-related information has been obtained; each of them captures a different part of the response. Multimodal physiological systems often perform well, yet they still depend on specialised hardware, and this restricts how widely they can be used for continuous monitoring outside controlled settings. This limitation is one of the main reasons for the growing interest in smartphone-based behavioural sensing, discussed in the next subsection.

1.2 Stress detection using smartphones

Research on stress detection with smartphones has moved from physiological sensing towards the analysis of everyday phone behaviour. Rather than using a wearable, recent studies examine touchscreen interaction, typing, motion patterns, app usage and context, and try to infer the cognitive and motor effects of acute stress from these signals without additional hardware [7], [9], [14].

Existing behavioural approaches can be grouped by their main sensing modality. The first group uses touchscreen interaction, for example swipes, taps, scrolling and gestures. The second uses inertial sensing: the accelerometer and gyroscope register stress-related changes in how the phone is held and moved and in postural stability. The third uses context such as app usage, communication and mobility to follow behavioural change over longer periods. A fourth, more recent group combines several

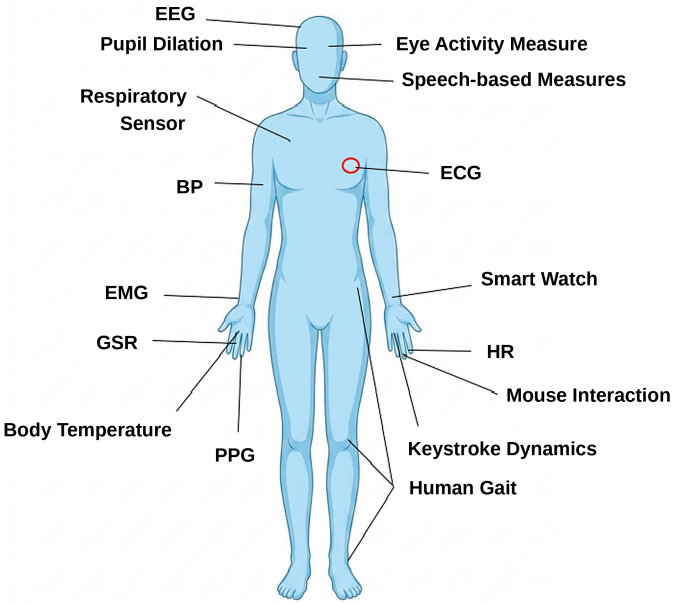


Fig 1. Representative physiological and behavioural sensing modalities reported for stress detection.

of these modalities, on the assumption that heterogeneous phone sensors together give more reliable results than any single one[6],[15],[16].

The results are promising, but several problems stand in the way of deployment. Behaviour differs widely between people, situations and devices, so inter-subject variability is high. Studies also differ in sensing modality, feature engineering, experimental protocol and evaluation method, which makes direct comparison difficult. Reviews therefore call for standardised behavioural datasets, common evaluation frameworks and multimodal sensing designed for realistic conditions[3],[10],[13].

1.3 Keystroke dynamics and multimodal behavioural sensing

Among phone-based behavioural modalities, keystroke dynamics has received particular attention because typing involves both cognitive processing and fine motor control. Timing measures such as key hold time, inter-key latency, rhythm, flight time, typing speed and the frequency of errors and corrections reflect workload and psychological state. The field began with behavioural authentication but has since been applied to cognitive state estimation, neurological assessment, emotion recognition and acute stress detection[8],[17],[18].

A consistent finding is that stress changes typing. Response latencies shift, rhythm becomes less regular, correction behaviour changes and motor coordination deteriorates, and these changes are large enough for a machine learning model to separate stressed from non-stressed typing without physiological sensors. For this reason keystroke dynamics is regarded as one of the more useful behavioural markers for unobtrusive stress monitoring on ordinary phones[8],[18],[19].

Typing alone, however, does not capture every effect of stress. Stress also shows in hand movement, device orientation, touch interaction and general body motion, and a typing-only model ignores this information. Multimodal behavioural sensing therefore combines typing with accelerometer, gyroscope and touchscreen data. Such systems are usually more reliable than single-modality ones because they describe more of how behaviour changes during phone use[10],[15],[16].

Multimodal studies are still few, and many use small cohorts, laboratory-only settings or data from a single task, so it is unclear how far their results generalise. Because the studies also differ in modality and evaluation method, their machine learning results cannot be compared easily. There is thus still a need for standardised behavioural sensing frameworks that are reproducible and can be deployed in practice.

Table 1. Representative studies (2020–2026) on physiological and behavioural stress detection.

Study	Application	Platform / Sensor	Primary Modality	Learning Approach	Key Limitation
Pepa et al. [19]	Computer stress	Keyboard + mouse	User interaction	RF + MIL	Desktop environment only.
Vargas-Lopez et al. [12]	Driver stress	EMG	Physiological	XAI-SVM	EMG-specific sensing.
Sağbaş et al. [9]	Smartphone stress	Accelerometer + gyroscope	Motion during typing	k-NN	No keystroke timing; 46 participants.
Yaacob et al. [18]	Review	Keyboard	Keystrokes	Review	Authentication-focused review.
Roy et al. [17]	Typing analytics	Keyboard	Keystrokes	Ensemble ML	General typing prediction.
Hoeijmakers et al. [14]	MS monitoring	Smartphone	Keystrokes	ML	Disease monitoring, not stress.
Nguyen et al. [6]	Review	Smartphone	Keystrokes	Review	Focus on cognitive assessment.
Naegelin et al. [20]	Office stress	Keyboard + mouse + HRV	Interaction + physiology	Gradient Boosting	Needs heart-rate sensor; 3-class task.
da Silva et al. [21]	Computer stress	Pressure-sensing keyboard/mouse	Keystroke + pressure	SVM/RF	Requires instrumented hardware.
Taskasaplidis et al. [3]	Survey	Multiple wearables	Physiological	Review	Lack of standardised systems.
Haque et al. [16]	HRV review	HRV	Physiological	AI/DL review	HRV-specific analysis only.
Fontes et al. [15]	Stress detection	rPPG (camera)	Physiological	Hybrid DL	Camera-based acquisition required.
Saddaf Khan et al. [10]	Smartphone stress	Motion sensors	Accelerometer, gyroscope	DNN	No keystroke integration.
Finnegan et al. [13]	Behavioural biometrics	Multiple	Behavioural signals	Review	Limited clinical validation.
Shahapur et al. [22]	Student stress	Student data	Physiological/survey	ML	No smartphone behavioural sensing.
Roy et al. [8]	Neural stress	Smartphone	Keystrokes	Ensemble	Neurological focus; typing only.
Banerjee et al. [2]	Acute stress	Smartphone	Keystrokes + sensors	ML	Preliminary; no ten-classifier benchmark.
Rupp et al. [7]	Perceived stress	Smartphone camera	Facial expression	ML	Camera-based pilot study.
Ladakis et al. [5]	Stress prediction	Open datasets	Physiological	ML/DL	Physiological only; protocol-dependent.
Pataca et al. [4]	Systematic review	Wearables	Physiological	ML review	Wearable-dependent evidence.
Shahin and Deshpande [11]	Stress prediction	Wearable sensors	Physiological	ML	Requires body-worn hardware.

1.4 Research gap

Table 1 lists representative studies on physiological and behavioural stress detection. Early work relied almost entirely on physiological measures such as ECG, EMG, respiration, EDA and HRV, recorded with dedicated wearable or laboratory sensors[4],[5],[11],[12],[16]. These methods often estimate stress reliably, but they need special hardware, sensors attached to the body and controlled recording conditions, and none of this fits unobtrusive everyday monitoring.

More recent studies have therefore turned to behavioural sensing on smartphones, using keystroke dynamics, typing pressure, touchscreen interaction and the built-in inertial sensors[7],[8],[9],[10],[14]. They show that stress-related changes in behaviour can be recorded passively during normal phone use, without a wearable. Most of them, however, study one modality at a time, either typing or motion. Since acute stress changes typing rhythm, hand movement, device handling and motor coordination together, a model built on one source is likely to lose reliability once it leaves the controlled setting.

The variety of experimental protocols is a further problem. Participant groups, stress-induction procedures, devices, features and evaluation metrics all differ between studies, so published methods cannot be compared on equal terms[3],[4]. Reviews repeatedly note that no standardised framework for behavioural stress detection exists and recommend a common benchmarking method.

Machine learning is now the usual analysis tool, yet few studies compare several classifiers on the same synchronised behavioural dataset. Without such a comparison, a reported gain cannot be assigned with confidence to the features, the data collection or the choice of classifier. In addition, several of the studies were not designed for acute stress detection at all. Their aim was emotion recognition, cognitive assessment, authentication, disease monitoring or typing analytics[6],[14],[17], so it is uncertain how well their findings apply to acute stress monitoring on smartphones.

Four gaps follow from this review. First, physiological approaches still depend on wearables or laboratory equipment, which makes continuous real-world monitoring impractical. Second, most smartphone studies use a single behavioural modality, and the combined information in synchronised keystroke and inertial data has hardly been examined. Third, there is no standard acquisition framework for collecting synchronised multimodal phone data under one protocol, which harms reproducibility and comparison between studies. Fourth, broad comparisons of classifiers on the same multimodal behavioural data are rare, so the relative contribution of the features and of the classifier to overall performance is unclear.

This work is designed around these gaps. We propose a web-based behavioural sensing framework for smartphones that records keystroke dynamics together with accelerometer and gyroscope data, synchronised in time, while participants use the phone normally. A single feature extraction pipeline turns these recordings into multimodal behavioural descriptors, and several machine learning algorithms are then compared under identical conditions for acute stress detection. The aim is to develop a method for smartphone stress monitoring that is unobtrusive, independent of the device and reproducible.

2 Methodology

2.1 Web-based behavioural data acquisition system

A web-based acquisition application was written in HTML, CSS and JavaScript to record keystroke dynamics and smartphone inertial data through a single interface (Figure 2). It runs in any current mobile browser and requires no installation, so participants could use a wide range of devices. Participant interaction, keystroke logging and sensor acquisition all run within the same page. Every keyboard event and every sensor reading receives a timestamp, which keeps typing and phone motion aligned in time. The accelerometer and gyroscope were sampled at 2 Hz. This rate was a compromise between temporal resolution, processing load and the rates that mobile browsers deliver reliably. The interface also collected metadata in structured fields: demographics, typing habits, handedness, education and a session ID. These fields were kept out of the numerical feature pipeline. They serve traceability and later analysis, and keeping them separate prevents them from confounding the behavioural measurements.

2.2 Experimental protocol and stress induction

Data were recorded in two controlled conditions. The baseline condition captured natural typing under low cognitive load, and the acute stress condition raised cognitive demand during phone use. Both conditions used the same web application (Figure 3), so the acquisition settings were identical and differences in software could not affect the data.

The seven typing tasks were divided into two groups according to cognitive demand. The first four are simple, predictable arithmetic series and were used to record baseline typing under low mental load. The other three demand more mental arithmetic, more working memory and frequent switching between rules; they raise cognitive effort and were used to induce acute stress.



Fig 2. Web interface for keystroke data collection. Application: https://computerscience.shinyapps.io/datacollect/.

Low Cognitive Load Tasks (Minimal Stress)

Task 1: Series Typing (Increment by 2). Participants started with the number 2 and repeatedly added 2, typing each resulting value until reaching 50.

Passphrase 1: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50

Task 2: Series Typing (Increment by 5). Participants started with the number 5 and repeatedly added 5, typing each resulting value until reaching 100.

Passphrase 2: 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100

Task 3: Series Typing (Increment by 10). Participants started with the number 10 and repeatedly added 10, typing each resulting value until reaching 200.

Passphrase 3: 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200

Task 4: Backward Series Typing (Decrement by 10). Participants started from 100 and repeatedly subtracted 10, typing each resulting value until reaching 10.

Passphrase 4: 100, 90, 80, 70, 60, 50, 40, 30, 20, 10

High Cognitive Load Tasks (Induced Stress)

Task 5: N-Back Arithmetic Task. Participants started with the number 15 and followed a repeating arithmetic pattern: add 3, subtract 2, and then add 5. Each computed value was typed sequentially until the value reached 100.

Passphrase 5: 15, 21, 27, 33, 39, 45, 51, 57, 63, 69, 75, 81, 87, 93, 99

Task 6: Skip Counting Task. Participants started with the number 5 and alternated between skipping two numbers and skipping three numbers while typing the resulting sequence until reaching 100.

Passphrase 6: 5, 8, 12, 15, 19, 22, 26, 29, 33, 36, 40, 43, 47, 50, 54, 57, 61, 64, 68, 71, 75, 78, 82, 85, 89, 92, 96, 99

Task 7: Serial Subtraction Task (Subtract by 7). Participants started from 100 and repeatedly subtracted 7, typing each resulting value until no further subtraction of 7 was possible (last value 2).

Passphrase 7: 100, 93, 86, 79, 72, 65, 58, 51, 44, 37, 30, 23, 16, 9, 2

The stress condition used mental arithmetic: serial subtraction, N-back arithmetic and skip counting. These tasks are common in cognitive stress research because they load working memory, attention and executive control enough to change behaviour measurably while the task is being done; mental demand is also the workload dimension most strongly linked to cognitive fatigue[23]. In both conditions the application recorded keystroke events and accelerometer and gyroscope readings continuously, without interrupting the participant.

All participants completed the same sequence of tasks under the same conditions and on the same acquisition framework, which kept environmental variation low. As typing and inertial data were recorded together, every behavioural sample can be linked to the cognitive state of the participant at that moment of the session.

2.3 Participant recruitment and data collection

In total, 288 people took part voluntarily. To obtain a more varied sample, participants were recruited from different kinds of educational institutions, namely schools, colleges and universities. The cohort includes school students, undergraduates,

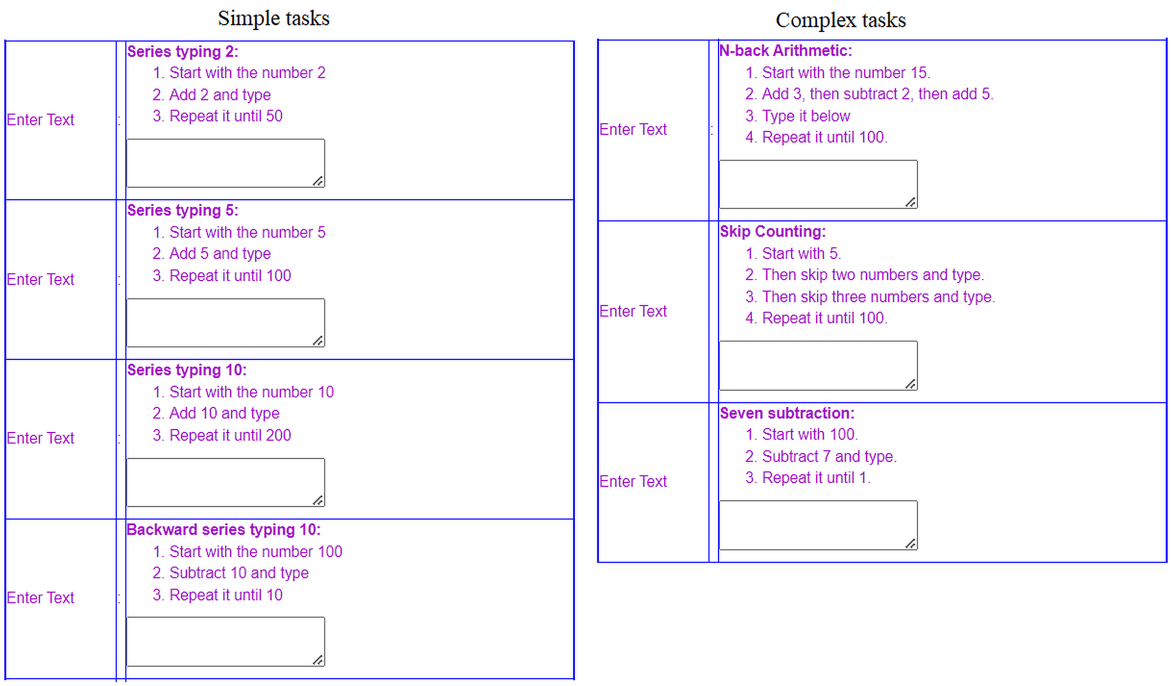


Fig 3. Experimental protocol adopted for behavioural data acquisition under baseline and acute stress conditions.

postgraduates, research scholars, teachers and other adults, and therefore covers different levels of cognitive maturity, smartphone experience and typing skill. Students were well represented, a group in which stress is common and increasingly studied with machine learning[22].

Demographic data were recorded on the same platform before each session: age, gender, educational qualification, handedness and typing style. This allowed us to examine behavioural differences between user groups. Participants belong to four age groups (below 18, 18–25, 26–45 and 46–60 years), and their qualifications range from school level to doctoral degree. Handedness and typing style were recorded for every participant because both can influence typing. Figure 4 shows the demographic composition of the dataset.

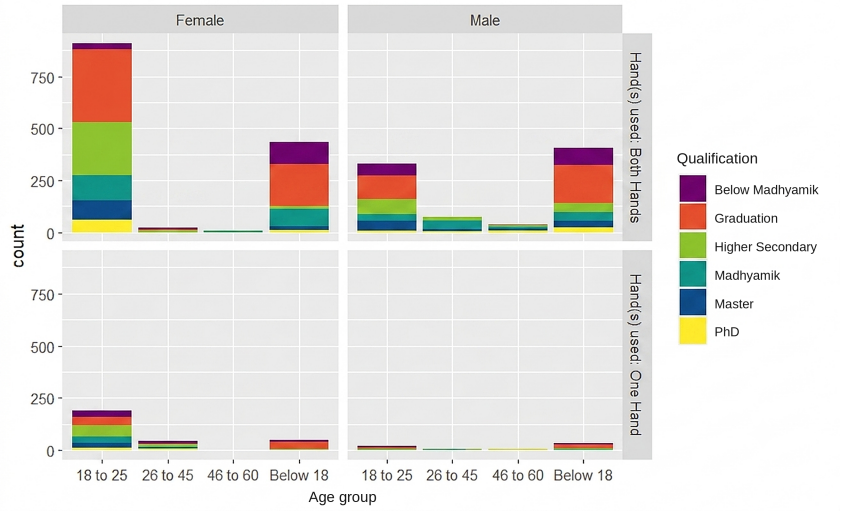


Fig 4. Demographic characteristics of the participants considered for behavioural data acquisition.

Every participant completed the full protocol, baseline and stress, in the same web application and under the same settings, which kept environmental and software differences small. Throughout each session the application recorded keystroke events and accelerometer and gyroscope readings while the participant worked through the tasks, so each sample is again tied to a known cognitive state.

Participation was voluntary. Before data collection, each participant was told the purpose of the study, which behavioural data would be recorded and how the dataset would be used, and written informed consent was obtained. All data were anonymised before analysis to protect privacy. Figure 5 shows the consent and data acquisition procedure.

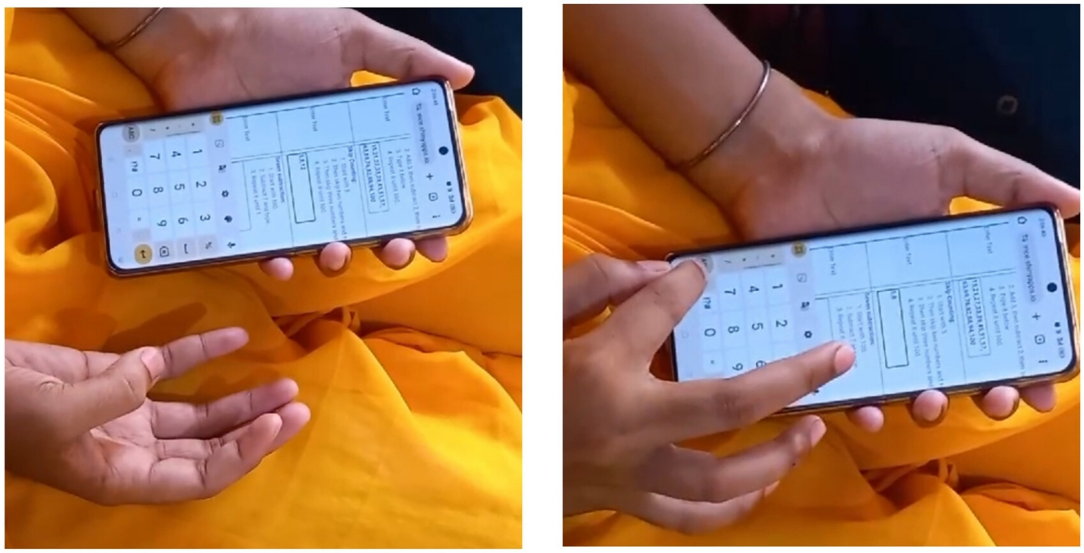


Fig 5. Participant consent and behavioural data acquisition process adopted for dataset construction.

The result is a set of synchronised multimodal behavioural recordings from a participant group that is varied in age and education and that was collected under a single protocol. This variety reduces demographic bias and gives a more representative basis for evaluating stress detection models on smartphones.

2.4 Behavioural feature representation

The synchronised signals from the web application were converted into features describing typing behaviour and phone motion. Acute stress affects cognitive processing, motor coordination and hand movement together, so features were taken from two complementary modalities, keystroke dynamics and inertial measurements, and joined into one multimodal representation for the machine learning stage.

Keystroke dynamics was described by five timing features, each reflecting a different aspect of typing. *Key Hold Time* is the time a key stays pressed. *Press–Press Latency* is the interval between two successive key presses, and *Release–Release Latency* the interval between two successive releases. *Release–Press Latency* is the time from releasing one key to pressing the next, and *Digraph Latency* is the time from pressing one key to releasing the next. Together these features describe typing rhythm, response latency and motor consistency, all of which change with cognitive workload and stress.

Motion features were taken from the phone's inertial sensors: linear acceleration along three axes from the accelerometer and the corresponding angular velocities from the gyroscope, both describing how the device moved during typing. Orientation was recorded as roll, pitch and yaw, which adds information on hand movement and device handling over the session. With both groups of features, the framework captures the fine timing of individual keystrokes as well as the coarser motor behaviour that accompanies acute stress.

Figure 6 summarizes the resulting multimodal behavioural features.

These feature vectors are the input to the machine learning stage. Before model development, the synchronised features were preprocessed and normalised so that they were consistent across participants and sessions, and the resulting feature matrix was used to train and evaluate the supervised classifiers.

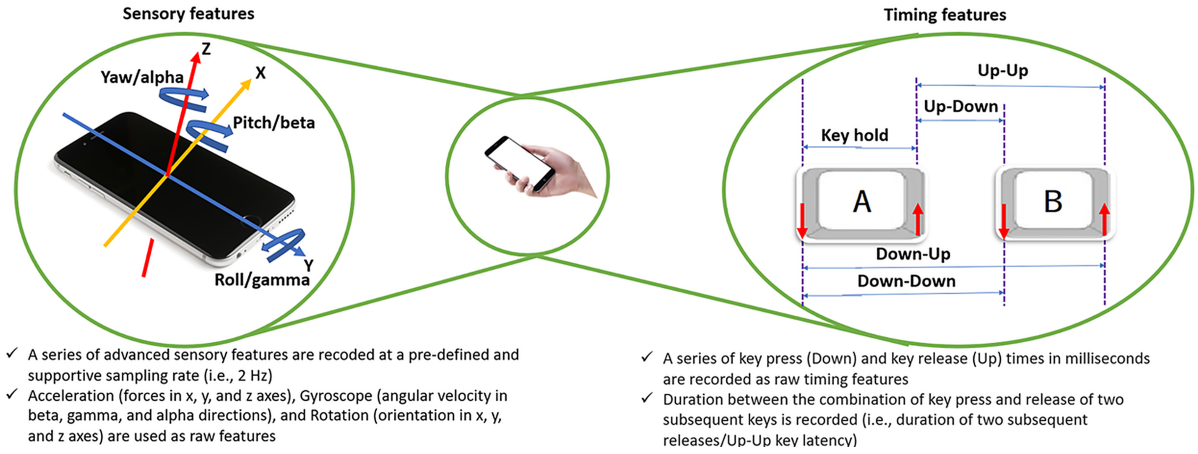


Fig 6. Behavioural features extracted from keystroke dynamics and smartphone inertial sensors for acute stress detection.

2.5 Proposed smartphone behavioural sensing framework

This subsection describes the complete behavioural sensing framework. It links web-based data acquisition, controlled stress induction, multimodal feature extraction and machine learning classification in one pipeline. Conventional physiological monitoring needs dedicated sensors; this framework uses only what a smartphone already provides during normal use, namely its inertial sensors and the keyboard. Figure 7 shows the full workflow.

The pipeline in Figure 7 has four stages. Behavioural data are first recorded by the web application while participants perform the cognitive tasks in the baseline and stress conditions. The keystroke events and the accelerometer and gyroscope readings are then synchronised so that their timing matches. Next, behavioural features are extracted from the synchronised signals. Finally, the features are passed to the supervised classifiers, which are all evaluated under the same protocol.

3 Results and Discussion

The evaluation had two aims: to test whether multimodal behavioural features from phone use are useful for detecting acute stress, and to compare several supervised learning algorithms under one protocol. Many earlier studies test only a few classifiers or a single behavioural modality. Here, classifiers from several learning families were applied to one multimodal feature set built from synchronised keystroke, accelerometer and gyroscope data. Comparisons of this kind are increasingly used in smartphone stress research to set benchmarks and to identify models suitable for deployment [8], [9], [10], [16].

3.1 Experimental evaluation protocol

Ten supervised algorithms from linear, probabilistic, distance-based, neural, tree-based and ensemble families were evaluated: Logistic Regression (LR), Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Naïve Bayes (NB), Multilayer Perceptron (MLP), C4.5 Decision Tree, AdaBoost, Random Forest (RF) and Extreme Gradient Boosting (XGBoost). All of them were trained on the same multimodal feature set, so the comparison is fair.

The dataset was divided into 80% for training and 20% for testing. Each classifier was assessed by Accuracy, Area Under the ROC Curve (AUC), weighted Precision, Recall and F-score, and macro Precision, Recall and F-score. Accuracy measures overall correctness, AUC measures discrimination independently of the threshold, and the weighted and macro scores show performance per class; this set of metrics is widely used in machine learning and behavioural classification [24].

The classifiers separated into clear groups. Ensemble methods beat the linear, probabilistic, distance-based and shallow neural models on almost every metric. XGBoost and Random Forest were well ahead, AdaBoost followed at some distance, and Logistic Regression, LDA, SVM, KNN, MLP and Naïve Bayes were clearly behind. The detailed figures are given below. The overall pattern suggests that ensemble methods handle the non-linear relations in multimodal smartphone features better than the other model families.

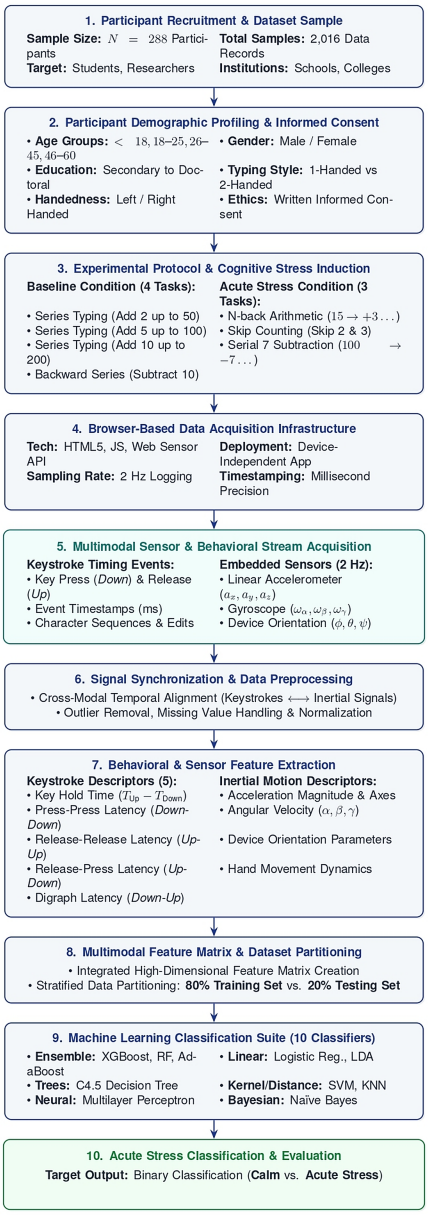


Fig 7. Overall workflow of the proposed smartphone behavioural sensing framework for acute stress detection.

3.2 Comparative analysis of classification models

Table 2 gives the complete results of the ten algorithms on the combined feature set. XGBoost ranks first and Random Forest a close second; AdaBoost and C4.5 form the next group, and the remaining classifiers are further behind. The combined keystroke and inertial features therefore separate Calm from Stress well. More interesting than the first place of XGBoost is the small distance between the two best models. It suggests that tree ensembles in general suit this representation, rather than XGBoost having a particular advantage over Random Forest.

XGBoost gave the best overall result, 89.30% Accuracy and 95.42% AUC, so it assigned the correct state to about nine in ten test samples. The AUC of 95.42% shows that this separation does not depend on one particular decision threshold. The other metrics agree. Weighted Precision, Recall and F1 lie within one point of each other (89.28%, 89.30%, 89.28%), and the macro values are almost the same (89.17%, 88.93%, 89.04%). In a binary task this agreement between weighted and macro scores

Table 2. Performance comparison of classification algorithms.

Metric	XGBoost	Random Forest	AdaBoost	C4.5	Logistic Regression	KNN	LDA	SVM	MLP	Naive Bayes
Accuracy	89.30	88.72	84.82	81.13	72.96	71.01	69.46	68.48	65.56	57.98
AUC	95.42	95.32	91.71	80.59	77.73	79.03	76.44	79.18	71.27	62.84
Precision (Weighted)	89.28	88.71	84.90	81.09	73.78	71.09	71.54	70.79	66.95	68.66
Recall (Weighted)	89.30	88.72	84.82	81.13	72.96	71.01	69.46	68.48	65.56	57.98
F1 (Weighted)	89.28	88.68	84.70	81.10	71.81	70.18	66.98	65.62	62.29	43.69
Precision (Macro)	89.17	88.69	85.02	80.77	74.21	71.14	72.28	71.55	67.39	70.51
Recall (Macro)	88.93	88.19	83.87	80.59	70.35	68.83	65.81	64.61	61.55	50.97
F1 (Macro)	89.04	88.41	84.28	80.67	70.62	69.06	65.24	63.73	60.18	38.75

Note: The target variable represents the behavioural state, with two classes: *Calm* and *Stress*. All algorithms were evaluated using an 80%/20% train/test split.

matters, because it shows that the high overall score is not produced by one class at the cost of the other. XGBoost is thus both the most accurate and a well-balanced classifier in this evaluation.

Random Forest was only slightly behind, with 88.72% Accuracy and 95.32% AUC, weighted Precision, Recall and F1 of 88.71%, 88.72% and 88.68%, and macro values of 88.69%, 88.19% and 88.41%. Here too the weighted and macro scores are close, so performance is balanced across the two classes. The difference between the two leading models is 0.58 percentage points in Accuracy and 0.10 in AUC. The AUC difference is the more telling figure: it shows that both models discriminate almost equally well, whatever threshold is chosen. The fairer reading is therefore that both ensembles extract about the same information from the multimodal features. We treat XGBoost as the leading classifier because its scores are higher, with Random Forest as a close second.

AdaBoost follows the same ensemble pattern, with 84.82% Accuracy and 91.71% AUC, weighted Precision, Recall and F1 of 84.90%, 84.82% and 84.70%, and macro values of 85.02%, 83.87% and 84.28%. These values are close together, so AdaBoost also treats both classes evenly, although it is below the two leading models. C4.5 reaches 81.13% Accuracy and 80.59% AUC (weighted F1 81.10%, macro F1 80.67%). The gap is informative. C4.5 is a single decision tree, whereas AdaBoost, Random Forest and XGBoost combine many trees, and performance rises in this order. Combining trees evidently lets the model use more of the heterogeneous information in the features than one tree can. The AUC shows this most clearly, rising from 80.59% for C4.5 to 91.71%, 95.32% and 95.42% for the ensembles. These results cannot show which features or interactions cause the improvement, but they do show that ensembles outperform a single tree under the conditions tested.

The conventional classifiers perform clearly worse. Logistic Regression reaches 72.96% Accuracy and 77.73% AUC and LDA 69.46% and 76.44%; their linear decision boundaries do not separate Calm from Stress well in this feature space. KNN does slightly better with 71.01% Accuracy and 79.03% AUC, and SVM reaches 68.48% Accuracy with a similar AUC of 79.18%. These AUC values show that KNN and SVM can still partly distinguish the classes, although well below the ensembles. Their AUC is also about ten points above their Accuracy, which means that the separation they achieve across thresholds is not turned into correct decisions and good F1 scores at the default threshold. Overall, the conventional classifiers make less use of the combined keystroke and inertial information, which fits the view that the heterogeneous multimodal features need the more flexible boundaries that ensembles can form.

MLP and Naïve Bayes are far below the leading models, for different reasons. MLP reaches 65.56% Accuracy and 71.27% AUC, with weighted and macro F1 of 62.29% and 60.18%. This modest AUC reflects the specific architecture and training settings used here and should not be read as evidence that neural networks are unsuitable for behavioural stress recognition. Naïve Bayes has the lowest Accuracy (57.98%) and AUC (62.84%) of all classifiers. Its F1 scores are more revealing: weighted F1 drops to 43.69% and macro F1 to 38.75%, with a macro Recall of 50.97%. The gap between Accuracy and F1 shows that the share of correct predictions hides a strong imbalance between the classes, and the low macro Recall confirms that the model cannot discriminate consistently. This agrees with the conditional independence assumption of Naïve Bayes, which does not hold for multimodal features derived from the same key events and movements. Accuracy alone can therefore mislead, and AUC and class-based metrics are needed, especially when a model behaves unevenly across classes.

Comparing weighted and macro averages gives one more view of classifier behaviour. Weighted averages weight each class by its frequency in the test set, while macro averages give both classes equal weight. For the best models the two hardly differ: weighted and macro F1 are 89.28% and 89.04% for XGBoost, 88.68% and 88.41% for Random Forest, and 84.70% and 84.28%

for AdaBoost. These small differences, together with the similar weighted and macro Precision and Recall values, show that the high scores of the ensembles are not driven by one class. The weaker classifiers, especially Naïve Bayes, show a much larger difference, so an acceptable overall score can hide real problems with one class. Read together, the two averages give a fuller picture than Accuracy and confirm that the ensembles perform consistently on both behavioural states.

The metrics in Table 2 all depend on a decision threshold, so ROC analysis was added. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) as the threshold changes, and the AUC summarises how well a classifier separates the two states over all thresholds[24]. This is relevant in practice because a deployed system may need a different threshold depending on whether missed stress episodes or false alarms are more costly. The ROC analysis therefore checks whether the results in Table 2 hold over a range of thresholds and not only at the one used for Accuracy, Precision, Recall and F1.

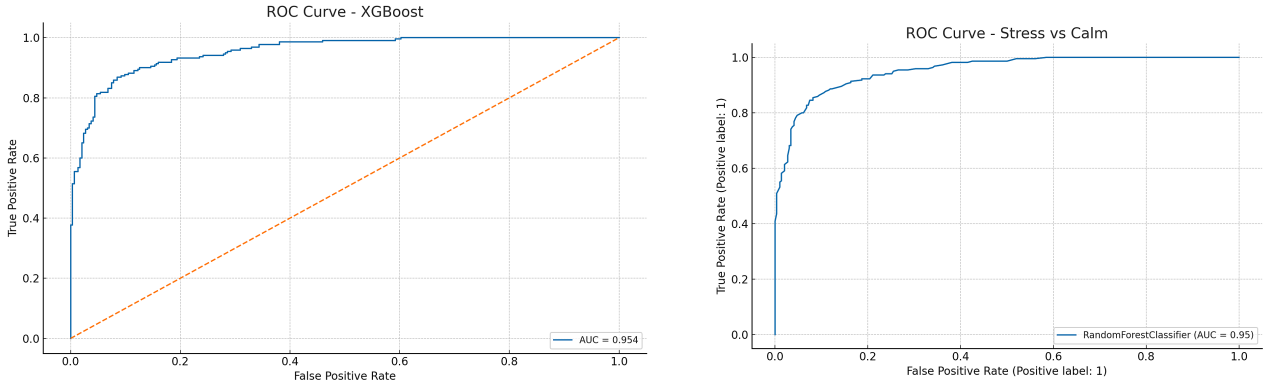


Fig 8. ROC curves of XGBoost and Random Forest for distinguishing Calm and Stress behavioural states.

Figure 8 shows the ROC curves of XGBoost and Random Forest, the two best classifiers. Both curves stay close to the upper-left corner, indicating a good TPR/FPR trade-off over the tested thresholds, in line with their AUC values of 95.42% and 95.32%. The two curves are almost identical, which confirms the result of Table 2 that the two models discriminate equally well. The AUC advantage of XGBoost, 0.10 percentage points, is real but too small to be given much weight. The ROC analysis thus confirms that both ensembles separate Calm from Stress well, with XGBoost marginally ahead.

The comparison as a whole favours ensemble classifiers for this multimodal framework. XGBoost comes first, Random Forest a close second and AdaBoost third, while the conventional classifiers and Naïve Bayes are well behind. The same two models lead on Accuracy, AUC, weighted and macro measures and in the ROC analysis. Because the gap between them is small, XGBoost is taken as the leading classifier and Random Forest as an equally strong alternative.

3.3 Comparison with prior behavioural and multimodal stress-detection studies

Table 2 shows which learning family suits the proposed features best, but not how the framework compares with published behavioural stress detection systems. To answer that question, our results are compared with four studies that, like ours, infer stress or a related affective state from unobtrusive behavioural signals (keystroke dynamics, mouse dynamics, smartphone inertial sensing or a combination) instead of dedicated physiological hardware, and with one recent wearable-sensor study that used a similar classifier comparison. Table 3 lists the modality, sample size, best model and reported performance of each study, and states whether its findings support or contradict ours. The discussion below also draws on other studies cited earlier in this paper, so that both agreeing and disagreeing evidence is considered.

3.3.1. Findings consistent with earlier studies

Several findings of earlier work agree with ours. First, ensemble and boosting models perform best whenever they are included in a comparison. Pepa et al.[19] obtained their best results with a Random Forest variant, Naegelin et al.[20] found Gradient Boosting strongest, and da Silva et al.[21] used Random Forest and SVM as their best participant-independent models. Roy et al.[8] likewise reported good results with ensemble classification of keystrokes on modern touch devices. XGBoost and Random Forest leading our ten-classifier comparison, well ahead of Logistic Regression, LDA and Naïve Bayes, is therefore consistent with the literature and not an isolated result. Second, all of these studies, including ours, find that everyday interaction with a computer or phone contains a stress signal that can be extracted without physiological sensors. Hoeijmakers et al.[14] support

Table 3. Comparison of the proposed framework with prior behavioural stress-detection studies.

Study	Behavioural Modality	N	Best Model	Reported Performance	Relation to present findings
Sağbaş et al. [9]	Smartphone accelerometer + gyroscope during touchscreen typing	46	k-NN	87.56% Accuracy (binary: stress vs. calm)	Supports feasibility of phone sensing; contradicts on best model (k-NN reached 71.01% here)
Pepa et al. [19]	Keyboard + mouse dynamics, in-the-wild web application	62	Random Forest (MIL)	76% Accuracy (keyboard), 63% (mouse); 3-class stress level	Supports ensemble superiority; lower figures reflect a harder 3-class, in-the-wild task
Naegelin et al. [20]	Mouse + keyboard + heart-rate variability, simulated office task	90	Gradient Boosting	F1 = 0.625 (3-class perceived stress)	Supports boosting as strongest model; task not directly comparable
da Silva et al. [21]	Pressure-instrumented keyboard and mouse (keystroke + mouse trajectory + grip/typing pressure)	–	SVM/RF (participant-independent)	Pressure features improved keystroke- and mouse-based detection by 6% and 3% respectively over dynamics alone	Partly supports (RF among best); shows gain from extra hardware and need for participant-independent tests
Gedam et al. [25]	ECG + GSR + skin temperature (wearable sensors), four stressors	200	XGBoost	96.17% Accuracy (binary: stress vs. no stress)	Supports ensemble ranking (XGBoost > RF); requires body-worn sensors
Present study	**Keystroke dynamics + smartphone accelerometer/gyroscope (synchronized, multimodal)**	**288**	**XGBoost**	**89.30% Accuracy, 95.42% AUC (binary: Calm vs. Stress)**	–

Note: Reported figures are taken directly from the cited papers and reflect each study's own task formulation (binary or multi-class), sample, and evaluation protocol; they are not obtained on a common benchmark and so should be read as an indication of relative standing rather than a strictly like-for-like comparison.

the same premise from a clinical angle, showing that passively collected smartphone keystrokes carry information about disease severity in multiple sclerosis. Third, our accuracy (89.30%) is higher than that of Sağbaş et al. [9] (87.56%), the study closest to ours in sensing modality. Their result was obtained from accelerometer and gyroscope signals during typing, without a separate keystroke timing channel, and on 46 participants, whereas our cohort is more than six times larger. This suggests that keystroke timing adds information to device motion, but because population and protocol differ, it is an indication and not a controlled ablation.

3.3.2. Findings that differ from earlier studies

Other findings contradict ours or limit how far our results can be generalised. First, Sağbaş et al. [9] found k-NN to be the best classifier, while k-NN reached only 71.01% here. Distance-based classifiers are sensitive to features on different scales and to heterogeneous populations; our cohort ranges from school students to adults over 45 and includes both one-handed and two-handed typists, which increases the spread within each class. Second, SVM performed strongly on statistical EMG features for driver stress [12] but was among the weaker models here (68.48%). Smooth physiological features may suit kernel boundaries better than irregular behavioural timings combined with motion sampled at 2 Hz. Third, deep learning achieved good results on camera-based rPPG signals [15] and on smartphone motion data [10], whereas our MLP was one of the weakest models (65.56%). Deep models benefit from long, densely sampled signals and careful tuning, which our compact feature vectors and standard configuration did not offer. Fourth, the lower results of Pepa et al. [19] (76%/63% accuracy) and Naegelin et al. [20] (F1 of 0.625) do not directly contradict ours: both solved a three-class problem on in-the-wild or office-simulation data, which is harder than our controlled binary Calm/Stress task. Finally, da Silva et al. [21] showed that pressure sensors in the keyboard and mouse add 3–6 percentage points to detection and stressed participant-independent evaluation. Pressure sensing is a possible route to higher accuracy, but it needs extra hardware, which our framework deliberately avoids. Their emphasis on participant-independent testing also indicates that our results, obtained with a single 80:20 split, should be confirmed with subject-wise validation.

3.3.3. Comparison with recent studies (2025–2026)

Recent studies published in 2025 and 2026 add to this picture. Gedam et al. compared nine classifiers on ECG, galvanic skin response and skin-temperature signals recorded from 200 Indian students and found XGBoost best (96.17%), closely followed by Random Forest (96.03%), with ensemble methods ahead of single classifiers [25]. The ranking matches ours, XGBoost first and Random Forest a close second, although their data came from dedicated body-worn sensors rather than an unmodified phone. Their higher accuracy reflects the richer physiological signal and shows the price paid for the unobtrusiveness of our set-up. Mac Donald et al. showed that stress conditions such as time pressure, a knowledge constraint and typing with the non-dominant hand can change keystroke patterns significantly [26], which supports the premise that typing behaviour carries a stress signal. Kallio et al. reviewed sensor-based stress monitoring in knowledge work and identified the effort required from the user and the limited information available from single sensors as the main barriers to continuous use [27]; combining keystroke and inertial data without extra hardware, as done here, responds to both barriers. Not every recent result favours classical ensembles. Ahmad and Khan applied a vision transformer to ECG spectrograms for multilevel stress classification [28], showing that deep architectures can work well when dense physiological signals are available. This contrasts with the weak MLP result in our study and suggests that the advantage of tree ensembles observed here is linked to compact behavioural features sampled at 2 Hz, and is not universal.

3.3.4. Distinctive contributions of the present work

Two features distinguish the present framework from these studies. The first is scale and diversity: with 288 participants and 2,016 samples, ranging from school students to doctoral-level adults, our dataset is larger and more varied than any of the behavioural comparator studies, all of which used fewer than 100 participants, and it also exceeds the 200-participant wearable cohort of Gedam et al. [25]. The second is the breadth of the model comparison. The cited behavioural studies each evaluated at most three to five classifiers, whereas ten algorithms from linear, probabilistic, distance-based, neural and ensemble families were compared here, which shows more completely which learning methods do and do not suit this kind of multimodal behavioural data. Together with the higher accuracy and AUC, the comparison supports the validity of the proposed framework. It also shows that a strict like-for-like comparison with earlier work is not yet possible, because no shared public dataset of keystroke and inertial data under stress exists. Releasing the present dataset in future work could help close this gap.

4 Conclusion

This study presented a web-based framework that combines keystroke dynamics with smartphone inertial sensing for unobtrusive acute stress detection. The framework records keystrokes and accelerometer and gyroscope data at the same time through an ordinary web application, without physiological sensors or a platform-specific app. Using one protocol, we collected a multimodal dataset of 2,016 samples from 288 participants under calm and experimentally induced stress conditions and evaluated the features with ten supervised machine learning algorithms.

The results show that behavioural data from routine phone use are a reliable basis for acute stress detection. Ensemble methods outperformed linear, probabilistic, distance-based and shallow neural models throughout. XGBoost performed best, with 89.30% accuracy and 95.42% AUC, and Random Forest was close behind. The clear lead of the ensembles suggests that they capture the non-linear relations in the synchronised keystroke and inertial features better than the other models.

The work also provides a complete behavioural sensing pipeline, from web-based acquisition through synchronised feature extraction to the comparison of classifiers. It offers a practical and reproducible way to study stress with ordinary smartphones and shows that continuous stress assessment is possible without wearable physiological sensors.

The study has limitations. The data come from controlled stress-induction tasks and were evaluated on one dataset, so performance during unconstrained daily phone use may be lower. Differences in hardware, typing style and individual interaction patterns may also affect how well the models transfer to other populations.

Future work should validate the framework in real-world, longitudinal studies with larger and more varied groups of participants. Personalised stress models, more advanced fusion of the two modalities, explainable machine learning and lightweight deep learning models may improve both accuracy and interpretability. These steps would move the approach towards scalable and privacy-preserving stress monitoring on smartphones for digital healthcare and human-centred intelligent systems.

5 Acknowledgements

The authors thank all the volunteers from schools, colleges and universities who took part in the data collection. The APC defrayed partly by Indian Society for Education and Environment (iSee) is greatly appreciated.

References

1) Agyapong B, Brett-MacLean P, Burback L, Agyapong VIO, Wei Y. Interventions to reduce stress and burnout among teachers: a scoping review. *International Journal of Environmental Research and Public Health*. 2023;20(9):5625. https://doi.org/10.3390/ijerph20095625.
2) Banerjee P, Roy S, Roy U. Analysing and predicting acute stress in smartphone-based online activities using keystroke dynamics and advanced sensory features. *International Journal of Computer Applications*. 2025;186(71):29–38. https://doi.org/10.5120/ijca2025924564.
3) Taskasaplidis G, Fotiadis DA, Bamidis PD. Review of stress detection methods using wearable sensors. *IEEE Access*. 2024;12:38219–38246. https://doi.org/10.1109/ACCESS.2024.3373010.
4) Pataca AO, Zdravevski E, Coelho PJ, Garcia NM, Deryuck M, Albuquerque C, et al. Use of machine learning for predicting stress episodes based on wearable sensor data: a systematic review. *Computers in Biology and Medicine*. 2025;198:111166. https://doi.org/10.1016/j.compbiomed.2025.111166.
5) Ladakis I, Fotopoulos D, Chouvarda I. Integrative analysis of open datasets for stress prediction. *Journal of Medical and Biological Engineering*. 2025;45(3):385–399. https://doi.org/10.1007/s40846-025-00958-z.
6) Nguyen TM, Leow AD, Ajilore O. A review on smartphone keystroke dynamics as a digital biomarker for understanding neurocognitive functioning. *Brain Sciences*. 2023;13(6):959. https://doi.org/10.3390/brainsci13060959.
7) Rupp LH, Kumar A, Sadeghi M, Schindler-Gmelch L, Keinert M, Eskofier BM, et al. Stress can be detected during emotion-evoking smartphone use: a pilot study using machine learning. *Frontiers in Digital Health*. 2025;7:1578917. https://doi.org/10.3389/fdgth.2025.1578917.
8) Roy S, Roy U, Sinha D, Pal RK. A novel approach to identify Parkinson's disease and other similar neural stress by analysing keystrokes on modern active devices with ensemble classification. *Multimedia Tools and Applications*. 2025;84(29):36163–36202. https://doi.org/10.1007/s11042-024-20380-y.
9) Sağbaş EA, Korukoglu S, Ballı S. Stress detection via keyboard typing behaviors by using smartphone sensors and machine learning techniques. *Journal of Medical Systems*. 2020;44(4):68. https://doi.org/10.1007/s10916-020-1530-z.
10) Khan NS, Qadir S, Anjum G, Uddin N. StresSense: real-time detection of stress-displaying behaviors. *International Journal of Medical Informatics*. 2024;185:105401. https://doi.org/10.1016/j.ijmedinf.2024.105401.
11) Shahin A, Deshpande D. Stress detection and prediction using wearable sensors and machine learning: a data-driven approach. In: Ganokratanaa T, Londhe ND, Joshi A, Kitsing M, editors. Information Systems for Intelligent Systems (ISBM 2025);vol. 468 of Smart Innovation, Systems and Technologies. Cham. Springer. 2026. https://doi.org/10.1007/978-3-032-12999-4_19.
12) Vargas-Lopez O, Perez-Ramirez CA, Valtierra-Rodriguez M, Yanez-Borjas JJ, Amezquita-Sanchez JP. An explainable machine learning approach based on statistical indexes and SVM for stress detection in automobile drivers using electromyographic signals. *Sensors*. 2021;21(9):3155. https://doi.org/10.3390/s21093155.
13) Finnegan OL, White JW, Armstrong B, Adams EL, Burkart S, Beets MW, et al. The utility of behavioral biometrics in user authentication and demographic characteristic detection: a scoping review. *Systematic Reviews*. 2024;13(1):61. https://doi.org/10.1186/s13643-024-02451-1.
14) Hoeijmakers A, Licitra G, Meijer K, Lam KH, Molenaar P, Strijbis E, et al. Disease severity classification using passively collected smartphone-based keystroke dynamics within multiple sclerosis. *Scientific Reports*. 2023;13(1):1871. https://doi.org/10.1038/s41598-023-28990-6.
15) Fontes L, Machado P, Vinkemeier D, Yahaya S, Bird JJ, Ihianle IK. Enhancing stress detection: a comprehensive approach through rPPG analysis and deep learning techniques. *Sensors*. 2024;24(4):1096. https://doi.org/10.3390/s24041096.
16) Haque Y, Zawad RS, Rony CSA, Banna HA, Ghosh T, Kaiser MS, et al. State-of-the-art of stress prediction from heart rate variability using artificial intelligence. *Cognitive Computation*. 2024;16(2):455–481. https://doi.org/10.1007/s12559-023-10200-0.
17) Roy S, Roy U, Sinha D, Pal RK. Predicting useful information from typing patterns using a bootstrapped-based homogeneous ensemble approach;vol. 555. Singapore. Springer. 2023. https://doi.org/10.1007/978-981-19-6791-7_1.
18) Yaacob MN, Idrus SZS, Ali WNAW, Mustafa WA, Jamlos MA, Wahab MHA. Decision making process in keystroke dynamics. *Journal of Physics: Conference Series*. 2020;1529(2):022087. https://doi.org/10.1088/1742-6596/1529/2/022087.
19) Pepa L, Sabatelli A, Ciabattoni L, Monteriù A, Lamberti F, Morra L. Stress detection in computer users from keyboard and mouse dynamics. *IEEE Transactions on Consumer Electronics*. 2021;67(1):12–19. https://doi.org/10.1109/TCE.2020.3045228.
20) Naegelin M, Weibel RP, Kerr JI, Schinazi VR, Marca RL, von Wangenheim F, et al. An interpretable machine learning approach to multimodal stress detection in a simulated office environment. *Journal of Biomedical Informatics*. 2023;139:104299. https://doi.org/10.1016/j.jbi.2023.104299.
21) da Cunha Silva DR, Wang Z, Gutierrez-Osuna R. Towards participant-independent stress detection using instrumented peripherals. *IEEE Transactions on Affective Computing*. 2023;14(1):773–785. https://doi.org/10.1109/TAFFC.2021.3061417.
22) Shahapur SS, Chitti P, Patil S, Nerurkar CA, Shivannagol VS, Rayanaikar VC, et al. Decoding minds: estimation of stress level in students using machine learning. *Indian Journal of Science and Technology*. 2024;17(19):2002–2012. https://doi.org/10.17485/IJST/v17i19.2951.
23) Muhammad S, Akand MR, Tamim FF, Talapatra S, Biswas SK, Rahman MS, et al. An assessment of researchers' mental workload and fatigue while engaged in research using machine learning and NASA-TLX. *Indian Journal of Science and Technology*. 2025;18(5):347–356. https://doi.org/10.17485/IJST/v18i5.3583.
24) Rainio O, Teuho J, Klén R. Evaluation metrics and statistical tests for machine learning. *Scientific Reports*. 2024;14:6086. https://doi.org/10.1038/s41598-024-56706-x.
25) Gedam S, Dutta S, Jha R. Analyzing mental stress in Indian students through advanced machine learning and wearable technologies. *Scientific Reports*. 2025;15(1):20610. https://doi.org/10.1038/s41598-025-06918-6.
26) Donald MM, Drevin L, Kruger H. Biometric authentication: stress as a factor in keystroke dynamics. In: Venter H, Eloff M, Eloff J, Botha R, Loock M, Mushtaq U, editors. Information and Cyber Security. ISSA 2024;vol. 2661 of Communications in Computer and Information Science. Cham. Springer. 2026. https://doi.org/10.1007/978-3-032-09660-9_6.
27) Kallio J, Vildjiounaite E, Tervonen J, López MB. A survey on sensor-based techniques for continuous stress monitoring in knowledge work environments. *ACM Transactions on Computing for Healthcare*. 2025;6(3):1–31. https://doi.org/10.1145/3720551.
28) Ahmad Z, Khan N. Stress classification from ECG signals using vision transformer. *IEEE Journal of Selected Areas in Sensors*. 2026;3:47–56. https://doi.org/10.1109/JSAS.2025.3648684.