Less is More: Leveraging Digital Behavioral Markers for Real-Time Identification of Loneliness in Resource-Limited Settings

Md. Sabbir Ahmed(✉) and Nova Ahmed

Design Inclusion and Access Lab (DIAL), North South University, Dhaka, Bangladesh
msg2sabbir@gmail.com, nova.ahmed@northsouth.edu

Abstract. The resource-constrained nature of developing regions and also the positive impact of early intervention show the need for a minimal and faster system to identify loneliness. However, existing pervasive device-based promising systems' requirement to run in the background for prolonged periods can be costly in terms of resources and also may not be effective for early intervention. Thus, we conducted a study (N = 105) in Bangladesh by developing a minimal system that can retrieve the past 7 days' app usage behavioral data within a second (Mean = 0.31 s, SD = 1.1 s). Leveraging only the instantly accessed data, we developed models through features selected by 3 different methods and exploration of 14 diverse machine learning (ML) algorithms including 8-tree-based algorithms. We found that the Gaussian Naïve Bayes model, developed by filter method Information Gain selected features, can identify 90.7% of lonely participants correctly with an F1 score of 82.4%. Through SHapley Additive exPlanations (SHAP), we explained the ML models showing how the features impacted the model's outcome. Due to being minimal, faster, and explainable, our system can play a role in resource-limited settings for early identification of loneliness which may create a positive impact by mitigating the loneliness rate.

Keywords: Loneliness · Smartphone · Resource-limited settings · Real-time · Explainable ML

1 Introduction

Mental health of people in Low- and Middle-Income Countries (LMIC) is much neglected than in high-income countries [1]. For example, compared to low-income countries, high-income countries have more than 35 times mental health workers for every 100,000 people [1]. The large divide persists in mental health research also where only 6% of literature emanated from the LMIC [2]; where, over 80% of people having a mental disorder, and also 80% of the world's total population reside in LMIC [3], highlighting the necessity of prioritizing research in that context. Loneliness, a perceived feeling of being separated from others [12], is linked with poor sleep, dementia, depression, and suicidal ideation [5] which may deteriorate if it remains for a prolonged

A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 460–476, 2023.
https://doi.org/10.1007/978-3-031-34586-9_31

period. In Bangladesh, around 43% of university students feel high loneliness [7]. However, there are only 270 psychiatrists and 565 psychologists in the country having over 160 million people [6] and 1.2 million students of universities [34]. For the early identification of loneliness in resource-limited settings like Bangladesh, the smartphone can be incorporated due to its high availability among the youth where 86.62% of university students in Bangladesh own smartphones [14].

Given the usability of the digital behavioral markers, researchers explored pervasive devices for a wide range of problems including loneliness [8–11, 21]. Existing studies leveraged phone usage [8–10, 21], smartphone sensed [9, 11], Fitbit sensed [9] and other systems [21] retrieved data to develop machine learning (ML) models for loneliness identification. While Austin et al. [21] focused on older adults, other studies [8–11] used youth as the samples who have higher loneliness [22]. ML models of the existing studies show promising performance. For example, Doryab et al. [9] found an accuracy of over 80% in identifying lonely participants. The positive impact of early intervention of psychological problems [20] shows the need for a faster system. However, the main limitation of the existing pervasive device-based system is the requirement for a prolonged data collection period (e.g., 2 weeks [10], 10 weeks [8], 16 weeks [9], and several months [21]) which may not be effective for early intervention. Also, existing systems' [8–11, 21] requirement for running a tool in the background throughout the whole study period may introduce research reactivity problems (e.g., Hawthorne effect). In addition, due to the consumption of much battery power of the background services [23], there may be significant barriers to having quality data in low-resource settings where electricity and internet services are limited [25].

To overcome these limitations, we present a minimal and unobtrusive system that can identify loneliness in real-time. Our study makes 3-fold contributions:

- Leveraging our tool's instantly (Mean = 0.31 s, SD = 1.1 s) retrieved app usage behavioral markers only, our ML model can identify 90.7% of lonely participants correctly (F1 = 82.4%). As far as we know, compared to any other existing pervasive device-based systems, to identify lonely students, our system is faster and minimal which can enable it to play a significant role in low-resource settings.
- We developed ML models by 14 classification algorithms including the linear, nonlinear, Stacking and Weighted Voting algorithms. We selected important features by 1 feature selection (FS) algorithm from each of the 3 main FS methods [19]. With comprehensive exploration, we presented a parsimonious ML model developed with around 6 features (Mean = 5.8, SD = 1.3) which has a sensitivity and specificity of over 70%. Due to having a lower number of features, this finding can be useful to have a more resource-insensitive system.
- Through SHapley Additive exPlanations (SHAP), we present how different behavioral markers impacted the predicted class. We discuss the findings of explainable ML which can facilitate mental healthcare professionals to understand lonely students more and take steps in intervention accordingly.

2 Related Work

2.1 Relation of App Usage Behavioral Markers with Loneliness

Smartphone usage behavior has a relation with loneliness [9]. In a smartphone, there remains a diverse set of apps and each app's unique features keep it in a distinct category [27]. Apps such as Facebook, Instagram, and Snapchat are in the Social Media category and their higher usage has a relation to lower loneliness [26]. However, depending on the category, there is a variation in users' behavior [27] and also the relation between loneliness and app usage [8, 10, 28]. For instance, while the number of messages has a negative relation, browsing searchers at late night has a positive relation with loneliness [10]. Even within the same category, there can be variation in relation depending on the behavioral markers. For instance, though loneliness has a negative association with the number of incoming calls, the number of missed calls does not have any significant association [10]. This reflects the importance of leveraging different behavioral markers while developing computational models to identify loneliness.

2.2 Identifying Loneliness Through Pervasive Devices

Comparatively, a significantly higher number of research about the identification of particular mental problem through sensing devices has focused on depression as presented by a recent systematic review [30]. However, loneliness has an effect on depression [31] and depression can be mitigated by mitigating loneliness [32] which presents the importance of prioritizing research on loneliness identification.

Nevertheless, little research has been conducted in identifying loneliness through unobtrusive ways. Some of these studies [8, 11] leveraged smartphone data. In a study [8] on 46 participants, researchers correctly identified 67.89% of the lonely participants by their smartphone usage and sensed data. Another study [10] used 2 weeks' smartphone usage, WIFI, and Bluetooth sensed data and their model correctly identified loneliness with an accuracy of 90%. However, due to having only 9 participants and due to unavailable details of their ML model (e.g., building and evaluation details of the classifier) [10], their findings may not be generalizable. In a previous study [11], utilizing the smartphone sensed geospatial data, an ML model distinguished the lonely from the non-lonely with an AUC value of .74. But their study has some limitations. For instance, the researchers [11] used the response of a single question as ground truth to group the lonely and non-lonely which may not be a standard method such as the validated UCLA Loneliness Scale-8 (ULS-8) [12]. In addition, to get stable performance, they needed several days' data which was collected by running their tool in the background for an equal number of days. In other previous studies also, researchers needed to use the data for several weeks (e.g., 10 weeks [8], 16 weeks [9]) running the tool for the whole study period. However, loneliness is associated with physical and psychological problems [5] where early identification and intervention can play a role in mitigating the severe consequences as early intervention has a positive impact [20].

3 Behavioral Data Collection Tool

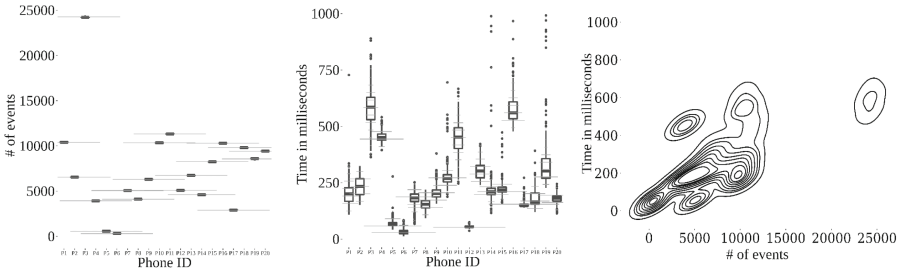
3.1 Development and Validation of the Tool

To retrieve the actual app usage behavioral markers (e.g., launch) unobtrusively, we developed an Android app. We chose the Android platform as it is used by 95.68% [41] phone users of Bangladesh. To retrieve the foreground and background events' data, we used several functions of the Java Class *UsageStatsManager* [33]. However, since app usage events are kept only for a few days [33], our tool can retrieve instantly (Sect. 3.2.) the app usage data of the past 7 days.

There were 3 steps in the app testing phase. We compared our tool's retrieved app usage data with the manually calculated data and retrieved data of such tools (e.g., [13]) available in Google Play Store which needs to run in the background. Considering the variations of phones, we also checked our app's retrieved data in 9 different phones.

3.2 Required Time to Retrieve Data

To estimate the generalizable time required to retrieve the foreground and background events from smartphones, we tested our tool on 20 smartphones of 19 different models and 8 different operating system (OS) versions of Android. From each phone, we collected the past 7 days' app usage data 500 times, and in total, we calculated the required time 10,000 times. The average number of retrieved foreground and background events was 7,447.61 (Min. = 306, Max. = 24,297, Median = 6,641, SD = 4986.62) (Fig. 1(a)) and on average, it took 307.94 ms (ms) (Min. = 13 ms, Max. = 61,087 ms, Median = 211 ms, SD = 1103.91 ms) (Fig. 1(b)).



(a) Number of retrieved events (b) Required time (c) KDE plot showing relation between time and number of events

Fig. 1. Performance of the data collection tool. KDE: Kernel Density Estimation. In figure b and figure c, 97 instances where the required time was more than 1000 ms were truncated to make the smaller values visible and we did not scale the required time to present the actual data.

To understand the factors that can impact the time in data retrieval, we explored the correlation of required time with 20 phones' API level and the number of retrieved events. We calculated the Spearman correlation coefficient (r_s) as the data did not satisfy the assumptions of the parametric test. We found no significant relation between the Android API level and required time ($r_s = .18$, $p = .44$). In exploring the relation with

the number of events, we used the average number of events and the required average time for each phone as within each phone, there was almost no variation in the number of retrieved events (Fig. 1(a)). We found that the events' number has a significant positive relation with the required time to retrieve data ($r_s = .56$, $p = .0096$) (Fig. 1(c)). After that, to estimate the plausible number of events on the students' phones, we used our constructed dataset for this study having 105 students (please, see Sect. 4.1 and Sect. 4.3 for details). In the dataset, on average, there were 8,174.04 events (SD = 4972.5). Among our data retrieval for 10,000 times, there were 4,500 instances for which the number of retrieved events was more than 8,000, and to retrieve this large number of events, our app needed 430.31 ms (SD = 1596.455 ms) which reveals that on average, our app can retrieve past 7 days' app usage data within a second.

4 Methodology

4.1 Research Ethics and Participants

Our study was approved by the Center for Research and Development of a university. All participants provided their consent before voluntary participation and data were stored in secure storage where only the researchers of this project can access. We collected data during the COVID-19 pandemic from January to June 2021. From 8 educational institutes, 105 students participated whose mean age was 22.3 years (SD = 1.57) and they were from 33 districts among the 64 districts of Bangladesh.

4.2 Categorization of Lonely and Non-lonely Participants

To understand loneliness and to use as the ground truth labels for ML models, we used the score of ULS-8 [12] which has been used for identification of loneliness in different countries (e.g., USA [12], Bangladesh [15]) showing the validity. There are 8 items and participants responded to the items through our developed app while donating app usage data. The options to respond for each item is Never (score 1), Rarely (score 2), Sometimes (score 3), and Always (score 4). Having a score of more than 16 means there is at least one loneliness measuring item that was bothered sometimes. Following the previous studies [9, 15], we categorized the participants having ULS-8 score of more than 16 into the lonely, and others were kept in the non-lonely group.

4.3 Feature Extraction and App Usage Behavioral Markers

In 7 days, 105 students used 867 apps and there were 868,636 foreground and background events' data from which we extracted the features.

App categories. For app categorization, at first, we retrieved the developers' preferred category available in Play Store using a Java HTML parser. Students used several apps which were not available there, and thus, using the package name, we explored those apps' features in online app stores (e.g., apkmonk.com) and developers' websites. Finally, we did categorization by understanding the process of previous studies (e.g., [17]) and through a discussion with 2 graduate students of engineering faculty. We found 105 students

App Category	Example Apps	# of Apps	% of Apps	App Category	Example Apps	# of Apps	% of Apps
Tools	Wi-Fi, VPN Private	282	32.45	Launcher	Home, Launcher3	18	2.07
Photo & Video	1Gallery, Album	121	13.92	Finance	bKash, GPay	14	1.61
Communication	Duo, Discord	58	6.67	Shopping	Daraz, Evaly	12	1.38
Games	TotM, Sudoku	55	6.33	Weather	Weather	9	1.04
Productivity	Calendar, Notebook	49	5.64	News & Magazines	Briefing, Jobs BD	8	0.92
Books & Reference	Booknet, Al Hadith	36	4.14	Health & Fitness	GloryFit, Mi Health	7	0.81
Music & Audio	Music Party, Radio	33	3.8	Lifestyle	Athan, My Galaxy	7	0.81
Entertainment	Flixoid, Netflix	30	3.45	Travel & Local	Uber, Earth	7	0.81
Browser & Search	Aloha, Chrome	28	3.22	Sports	Cricbuzz, SofaScore	5	0.58
Social	LIKE, Facebook	25	2.88	Food & Drink	eFood, foodpanda	2	0.23
Personalization	Wallpapers, Themes	21	2.42	Medical	Surokkha, Maya	2	0.23
Business	Fiverr, Kormo	19	2.19	Auto & Vehicles	Android Auto	1	0.12
Education	Arabits, Englishplz	19	2.19	Unknown	Not Applicable	1	0.12

Fig. 2. Example apps along with the number (#) and percentage (%) of apps of each category.

using 867 apps of 26 categories (Fig. 2). Most apps were from Tools (32.45%) and the least apps were from the Auto & Vehicles (0.12%) category (Fig. 2).

Ratio of hamming distance. As uniqueness in terms of app usage varies among smartphone users (e.g., between the depressed and non-depressed [17]), in the case of each student, we calculated the ratio of the hamming distance from the nearest lonely to the nearest non-lonely student. To get an unbiased ML model, we did not consider the group (e.g., non-lonely) of that student while calculating the distance: $DL_{ij} = (A_i \cup A_j) - (A_i \cap A_j)$ where DL_{ij} denotes the distance of the i^{th} student from the j^{th} student of the lonely group; A_i and A_j denote the set of apps used by the i^{th} and j^{th} students respectively. In this way, we calculated the minimum distance DL_i of the i^{th} student in the lonely group. We followed the same process to calculate the minimum distance DNL_i of the i^{th} student in the non-lonely group. Finally, we calculated the ratio of hamming distance for the student: $\frac{DL_i}{DNL_i}$. The main motivation behind using ratio, instead of global distance (minimum distance among all participants regardless group) is that it tells us how much or less a participant is unique compared to the lonely and non-lonely participants which are intuitively more informative.

Other behavioral data: For total smartphone usage (i.e., regardless of app category) and each of the 26 app categories, we also extracted duration, frequency of launch, duration per launch, launch per app, duration per app, and session data to extract the participants' app usage patterns. In addition, as app usage behavior varies by days [17], we extracted features regarding the difference in app usage between weekdays and weekends dividing the days (Weekdays: Sunday to Thursday, Weekends: Friday and Saturday) based on the working schedule of Bangladesh.

Characteristics of the features: Apart from the features using the 24 h data, for each of the above-mentioned behavioral markers, we also extracted diurnal usage data dividing days into 4 equal time periods: Night: 00:01–6:00; Morning: 6:01–12:00; Afternoon: 12:01–18:00; Evening: 18:01–00:00. After that, we calculated 8 different data from the diurnal usage to be used as features: minimum, maximum, range, mean, standard deviation, entropy, skewness, and kurtosis. To understand the app usage patterns more, we extracted features in two ways using the entropy formula. Firstly, for the diurnal usage

data, we calculated the entropy $E_t(j) = -\sum_{t=1}^{4} P_d(i) log P_d(i)$ where for the student j who has an unequal spending duration in each of the 4 time periods, E_t will be lower compared to the student who has an equal spending duration in each time period. Let's say, a student has 1, 1, 3, and 3 min whereas another student has 2, 2, 2, and 2 min per app spending duration on apps of a category in the night, morning, afternoon, and evening periods respectively. Then, the 1st student will have an entropy E_t of 1.81 which is lower than the entropy E_t of 2.0 of the 2nd student presenting that the 1st student is more focused on the phone during certain time periods and also has variation in app usage over the day compared to the 2nd student. Secondly, we calculated entropy $E_{dl}(j)$ for the student j on the basis of spending duration and frequency of launching of each app of an app category.

$E_{dl}(j) = -\frac{1}{2}(\sum_{i=1}^{n} P_d(i) log P_d(i) + \sum_{i=1}^{n} P_l(i) log P_l(i))$; here, P_d and P_l denote the probability to use i^{th} app based on spending duration and launch respectively.

4.4 Feature Selection

As there is no feature selection (FS) method that can find the best set of features ensuring the maximum performance of the ML models, we explored 1 FS approach from each of the 3 main FS categories [19]: wrapper, filter, and embedded method. As a wrapper method, we used the Boruta where, unlike the minimal-optimal methods, all-relevant features are selected [18]. In Boruta, Random Forest (RF) is used as the base estimator [18] and it is suggested to use 3 to 7 as the base estimator's maximum depth [35]. As the filter and embedded methods, we used the Information Gain (IG) and RF respectively. However, unlike Boruta, these two methods do not inform a fixed set of features that can have the best performance. Hence, we set the lower boundary of features using the 1 in 10 rule [38] where 5 features are to be selected due to having 54 lonely participants in our study. We increased the number of features gradually up to 20 and did not increase further to prevent the possibility of having overfitted models.

4.5 Model Development, Validation, and Explanations

We preferred machine learning (ML) to deep learning since ML models have higher transparency. We extracted the features (Fig. 3(a)) ourselves and also explained the ML models which can be insightful, particularly for mental healthcare professionals. As there is no ML model which can fit for all solutions, we developed models by a set of classification algorithms where both linear and non-linear including 8-tree based algorithms were explored: AdaBoost, CatBoost, Decision Trees, Extra Tree, Extreme Gradient Boosting (XGBoost), Gaussian Naïve Bayes (NB), Gradient Boosting (GB), K-Nearest Neighbor (KNN), Light GBM, Logistic Regression (Logit), RF, C-Support Vector Classifier (SVC). In addition, as a baseline, we used a Dummy classifier.

To develop the models, we used the nested cross-validation approach which shows a generalizable and unbiased performance [16]. In the outer loop, we used the Leave-One-Out-Cross-Validation (LOOCV) method which has a lower variance [36] and where we divided the dataset into n equal portions presenting each participant's data. In the inner loop, $n-1$ participants' data were used for 2 purposes: to select a set of important

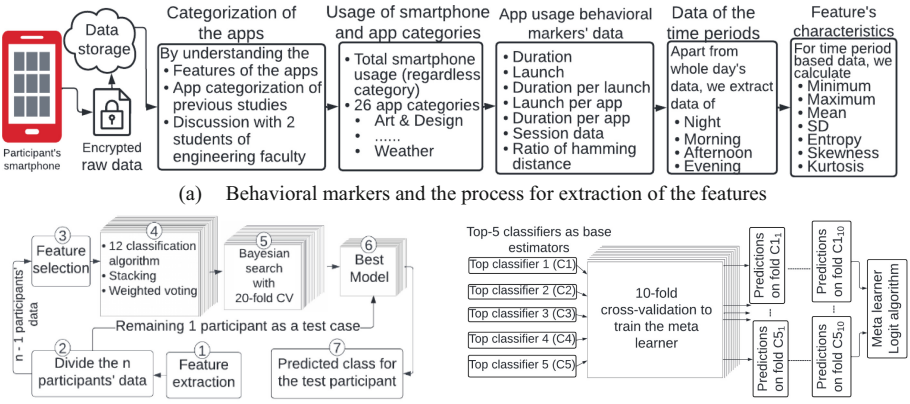


Fig. 3. Overview of the ML model development process.

features and to tune the hyper-parameters by the Bayesian search optimization technique using 20-fold CV instead of the LOOCV method to reduce the time complexity. In the Bayesian optimization technique, the informed search technique is used where the next step is taken based on the performance of the previous step and unlike Grid Search, this method does not need to explore every combination which makes it faster. After finding the best estimator, we predict the remaining 1 participant's class who was not included in FS and hyper-parameter tuning steps (Fig. 3(b)). To predict each of the 105 student's class, we repeat the same process. It can be noted that to build the ML models, we used open-sourced Python libraries.

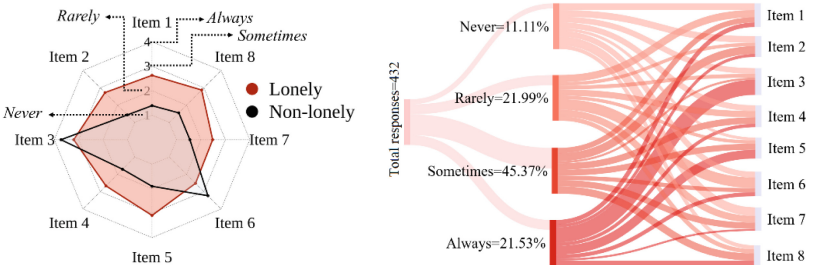
After developing the individual ML models based on the aforementioned 12 classification algorithms, using the best 5 classification algorithms, we developed an ensemble model Stacking and Weighted Voting. In the Stacking classifier, the predictions of the individual estimators were stacked to train the meta-learner Logit (Fig. 3(c)). On the other hand, in the Weighted Voting classifier, we calculated the weights by dividing the training data into 10-folds. The weight was multiplied by the final predicted probability of each of the top-5 classifiers for the test participant. After that, the final class for the participant is decided based on the soft-voting.

We evaluated the models' performance by comparing the models' predicted class with the ground truth class based on ULS-8 score. We calculated sensitivity and specificity which present how many of the lonely and non-lonely were identified correctly respectively. In addition, we presented precision which informs us the percentage of predicted lonely participants was truly lonely. We focused on maximizing the F1 score which is the harmonic mean of precision and sensitivity. To explain the ML models, we used the SHAP [4] which works based on the concept of cooperative game theory.

5 Findings

5.1 Participants' Loneliness

Among 105 participants, 54 participants (51.4%) were lonely and 51 participants (48.6%) were non-lonely (please, see Sect. 4.2 for categorization process). Except for the reverse items (3rd item: *I am an outgoing person*; 6th item: *I can find companionship when I want it*), lonely participants had a much higher frequency of bothering with loneliness (Fig. 4(a)). For instance, lonely participants' average frequency of feeling isolated from others (5th item) was more than sometimes which was rarely in the case of non-lonely participants (Fig. 4(a)). Also, lonely participants' most (45.37%) responses for the items were sometimes and 21.53% responses were always (Fig. 4(b)) presenting around 67% responses containing feeling of loneliness at least sometimes.



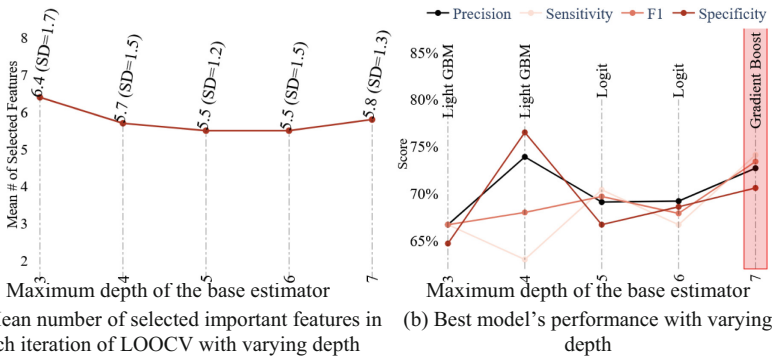
(a) (b)

Fig. 4. (a) The difference between the lonely and non-lonely students in frequency of the ULS-8 scale's items' appearance. (b) The link between the items and the frequency of appearance, based on the 432 (54 lonely students * responses of the 8 items) responses of the lonely students.

5.2 ML Models' Performance

Exploring the Boruta selected features, we found that the mean number of selected important features varied with the variation of the maximum depth of the base estimator Random Forest (RF) algorithm. In each depth, the average number of selected features in each iteration of LOOCV was around 6 (Fig. 5(a)). However, the performance of the models varied largely where we found the minimum F1 score of 66.7% at depth 3 and a maximum of 73.4% at depth 7 (Fig. 5(b)). At depth 7, the average number of selected features in LOOCV was 5.8 (SD = 1.3) and the best performing model among the explored classification algorithms was Gradient Boosting (GB) which had sensitivity, specificity, and precision of 74.1%, 70.6%, and 72.7% respectively (Fig. 5(b)). These present that the GB model identified 74.1% lonely and 70.6% non-lonely correctly. Also, the predicted lonely class was correct in 72.7% of cases.

In the models based on the filter method Information Gain (IG) selected 5 to 11 important features, the best models' sensitivity score varied from 53.7% to 68.5%, and the specificity score varied from 56.9% to 70.6% (Fig. 6(a)). At the number of features



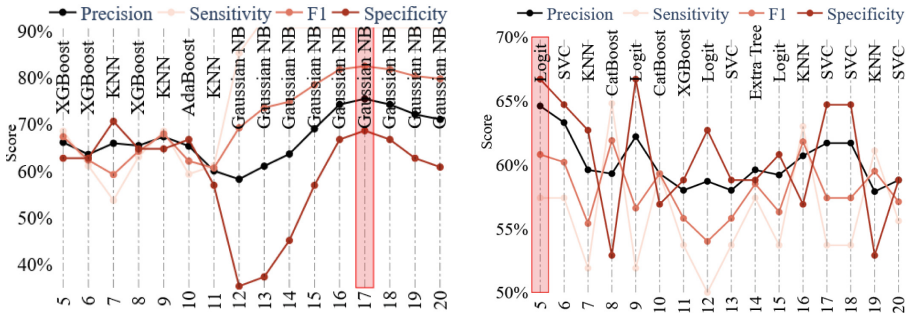
(a) Mean number of selected important features in each iteration of LOOCV with varying depth

(b) Best model's performance with varying depth

Fig. 5. (a) Wrapper method Boruta selected important features and (b) performance of the best models. The rectangle box presents the model which had optimal performance.

12, though the specificity (35.3%) reached a minimum, sensitivity (85.2%) increased. Gradually, increasing the number of selected important features, we found the best performance when the Gaussian Naïve Bayes (NB) was built on 17 important features selected in each iteration of LOOCV (Fig. 6(a)). The NB model identified 90.7% of lonely students correctly (sensitivity = 90.7%) and also predicted lonely students were truly lonely in 75.4% of cases (precision = 75.4%).

On the other hand, in the embedded method RF selected 17 features, the C-Support Vector Classifier (SVC) model performed best which had a sensitivity and specificity of 53.7% and 64.7% respectively (Fig. 6(b)). Among the RF selected important features from 5 to 20 (range setting process is available in Sect. 4.4), we found the best model while the Logit model was developed based on 5 features. The Logit model had a sensitivity, specificity, and precision score of 57.4%, 66.7%, and 64.6% respectively.



(a) Filter method Information Gain (IG)

(b) Embedded method Random Forest (RF)

Fig. 6. Best models' performance in a varying number of selected important features by the (a) IG and (b) RF FS methods. The rectangle box presents the model which had optimal performance.

Interestingly, it appears that the Boruta selected important features-based ML models have a higher performance with a lower number of features. For instance, with 6 features

selected by the IG, the XGBoost performed best having a sensitivity of 61.1% (Fig. 6(a)) while the RF selected 6 features based best performing model SVC had a sensitivity of 57.4% (Fig. 6(b)). However, in all of the explored maximum depths of the Boruta, the number of selected features was around 6 (Fig. 5(a)) where the minimum and maximum sensitivity were 63% and 74.1% respectively and the maximum performing model was GB (Fig. 5(b)) as aforementioned. This presents GB as a parsimonious model having higher predictability with a lower number of features.

Among all models of all feature selection (FS) methods, we found the best performance in terms of F1 score, sensitivity, precision, and accuracy from NB model (sensitivity = 90.7%, F1 score = 82.4%) which was developed based on 17 important features selected by the IG (Fig. 7). However, the specificity of the model was 68.6% whereas the GB model based on the Boruta selected features had a specificity score of 70.6% (Fig. 7) presenting relatively higher ability to identify the non-lonely participants.

Filter (Information Gain (IG), # of features=17)						Wrapper (Boruta, maximum depth=7, # of features: Mean=5.75, SD=1.32)						Embedded (Random Forest (RF), # of features=5)					
Model Name	Precision	Sensitivity	F1 Score	Specificity	Accuracy	Model Name	Precision	Sensitivity	F1	Specificity	Accuracy	Model Name	Precision	Sensitivity	F1	Specificity	Accuracy
N. Bayes	75.4	90.7	82.4	68.6	80	G. Boost	72.7	74.1	73.4	70.6	72.4	Logit	64.6	57.4	60.8	66.7	61.9
E. Tree	56.8	77.8	65.6	37.3	58.1	Logit	72.5	68.5	70.5	72.5	70.5	E. Tree	61.4	64.8	63.1	56.9	61
Logit	60.3	64.8	62.5	54.9	60	SVC	71.2	68.5	69.8	70.6	69.5	SVC	60.4	59.3	59.8	58.8	59
SVC	58.3	64.8	61.4	51	58.1	XGBoost	68.4	72.2	70.3	64.7	68.6	CatBoost	59.3	64.8	61.9	52.9	59
D. Tree	50.6	75.9	60.7	21.6	49.5	N. Bayes	67.3	68.5	67.9	64.7	66.7	AdaBoost	58.7	68.5	63.2	49	59
Stacking	59.1	72.2	65	47.1	60	Stacking	66	64.8	65.4	64.7	64.8	Stacking	60.3	64.8	62.5	54.9	60
W. Voting	65.3	87	74.6	51	69.5	W. Voting	67.3	64.8	66	66.7	65.7	W. Voting	58.3	64.8	61.4	50.98	58.1
Baseline	51.4	100	67.9	0	51.4	Baseline	51.4	100	67.9	0	51.4	Baseline	51.4	100	67.9	0	51.4

Fig. 7. Performance of the top-5 classifiers, based on the best (in terms of ML models' performance) set of features in each FS method. "# of features" present the number of features used in each iteration of LOOCV. D: Decision, E: Extra, G: Gradient, N: Naïve, W: Weighted.

Based on each FS method's top-5 classifiers, we developed Stacking and Weighted Voting models. Among these, IG selected features based Weighted Voting model had a higher performance which identified 87% lonely and 51% non-lonely students correctly (Fig. 7) having a balanced accuracy ($\frac{Sensitivity+Specificity}{2}$) of 69%. Though all the models had a higher performance than the baseline Dummy classifier's balanced accuracy 50%, and specificity of 0%, the Stacking, and Weighted Voting classifiers' performance were not higher than the best classifier of each FS method (Fig. 7).

5.3 Explanation of the ML Models

Exploring the top-30 important features which were used for the ML model development, we did not find a common feature that appeared in each of the 3 FS methods (Fig. 8). This is reflected in the performance of the ML models also where we found a higher variation across the FS methods (Fig. 7) which presents FS methods' different mechanisms of selecting the important features and also explains the rationale behind the exploration of 3 different FS methods. Among the top-30 features, 3 (10%) features were regarding the whole day whereas the remaining 27 (90%) important features were based on the four time periods of a day (night, morning, afternoon, and evening) (Fig. 8). This presents

that the diurnal usage data contains more information in identifying the differences between the lonely and non-lonely students. Similarly, compared to the features on total smartphone usage regardless of the app category (6.67%), there were a higher number of important features in the case of the app categories (93.3%).

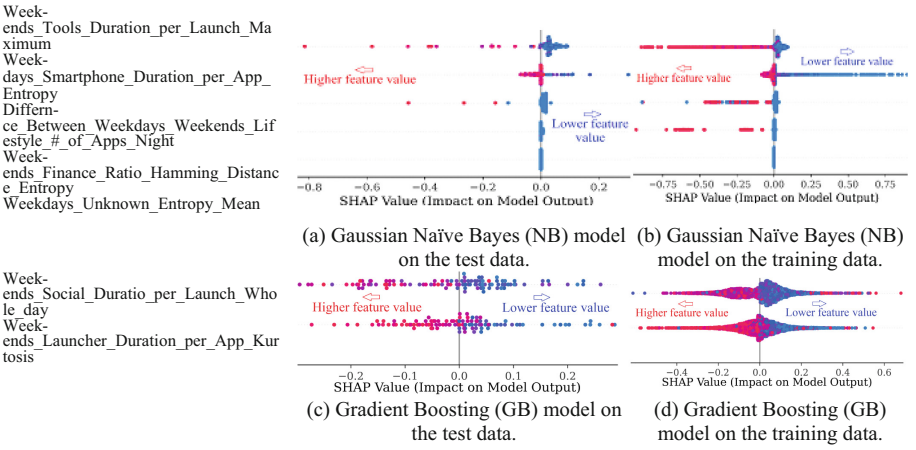
Feature	IG	Boruta	RF	Feature	IG	Boruta	RF
Weekend_Launcher__Duration_per_App_Kurtosis	0	100	80	Weekend_Music_Launch_Skew	56.2	0	0
Weekend_Social_Duration_per_Launch_whole_day	0	100	42	Weekend_Entertainment_Ratio_of_Hamming_SD	53.3	0	0
Dif_bet_weekdays_ends_Communication_Entropy_Evening	0	90.5	13	Weekend_Unknown_Ratio_of_Hamming_SD	51.4	0	0
Weekday_Smartphone_Duration_per_App_Entropy	100	0	0	Weekday_Sports_Ratio_of_Hamming_Range	48.6	0	0
Dif_bet_weekdays_ends_Lifestyle_#_of_Apps_Night	100	0	0	Weekday_Medical_Ratio_of_Hamming_Kurtosis	47.6	0	0
Weekend_Tools_Duration_per_Launch_Max	100	0	0	Weekday_Education_#_of_Apps_Range	46.7	0	0
Weekday_Unknown_Entropy_Mean	100	0	0	Weekday_Browser_Entropy_Max	0	0	46
Weekend_Finance_Ratio_of_Hamming_Entropy	100	0	0	Weekend_Auto_&_Vehicles_Duration_per_Launch_Max	44.8	0	0
Weekday_Browser_Entropy_whole_day	0	92.4	5.7	Dif_bet_weekdays_ends_Communication_#_of_Apps_Evening	0	41	3.8
Weekday_Productivity_Entropy_whole_day	0	84.8	13	Weekend_Finance_Ratio_of_Hamming_Skew	41	0	0
Weekday_Lifestyle_Launch_per_#_of_Apps_Kurtosis	93.3	0	0	Dif_bet_weekdays_ends_Books_Ratio_of_Hamming_Evening	37.1	0	0
Weekend_Shopping_Entropy_Mean	93.3	0	0	Weekday_Art_Duration_per_Launch_Mean	33.3	0	0
Weekend_Sports_#_of_Apps_SD	67.6	0	0	Weekend_Lifestyle_Duration_per_Launch_Min	33.3	0	0
Dif_bet_weekdays_ends_Health_Ratio_of_Hamming_Afternoon	61.9	0	0	Weekday_Health_Duration_per_App_Kurtosis	29.5	0	0
Weekday_Smartphone_Entropy_Skew	58.1	0	0	Weekend_Photo_Video_#_of_Apps_Skew	23.8	0	0

Fig. 8. Top-30 (average presence in each FS method) important features among the features used for the top-5 classifiers of each FS method. The values present the percentage of iterations a feature appeared as important among all iterations of LOOCV. Dif_bet_weekdays_ends: Difference between weekends and weekends, SD: Standard Deviation.

Exploring the features' data characteristics, we found more than half of the features (53.3%) among the top-30 contain the ratio of hamming distance and entropy data (Fig. 8) presenting these features' higher ability to differentiate the lonely and non-lonely students through the ML models. On the other hand, we found almost the same number of solely weekdays (12 features, 40%) and weekends' (13 features, 43.33%) data-based features (Fig. 8) which presents equal importance of weekdays and weekends' data.

Summary plot of the SHAP analysis showed that on the weekend when the Tools app category's (e.g., *Assistant*) maximum duration per launch among the usage in 4 time periods became lower, students were more likely to be in the lonely group (Fig. 9 (a, b)). In addition, when the entropy regarding duration per app became lower on weekdays, the students were also likely to be lonely. Having a lower entropy means there is a higher variation in spending duration per app over the 4 time periods of weekdays (a detailed discussion on entropy is available in Sect. 4.3).

Apart from these, the SHAP analysis demonstrated that when the difference in the number of used Lifestyle apps (e.g., *Athan*) in night time periods of weekdays and weekends became lower, students were more likely to be in the lonely group (Fig. 9(a, b)) which indicates that over the whole week's night, lonely students were likely to use Lifestyle category's apps. We found that among the students' used 7 apps of this category, 4 apps *(Athan, App Of Ramadan, Prayer Time Quran Qibla Dua Tasbih, Muslim Pro*) were related to prayer. Our findings also showed that the lonely students were more likely to have a lower duration per launch of Social Media apps (e.g., *Facebook*) on the weekend (Fig. 9(c, d)). In addition, lonely students' kurtosis values regarding the



(a) Gaussian Naïve Bayes (NB) model on the test data.

(b) Gaussian Naïve Bayes (NB) model on the training data.

(c) Gradient Boosting (GB) model on the test data.

(d) Gradient Boosting (GB) model on the training data.

Fig. 9. Shapley summary plot for the NB model (a, b) based on 17 features selected by the IG and for the GB model (c, d) based on the Boruta selected features when the maximum depth of the base estimator was 7. The plot shows the impact of the features (which appeared in all iterations of LOOCV) on the ML models' outcome. In figures b and d presenting models' outcome based on the training data, there is a higher number of feature values since for each iteration of LOOCV, $n - 1$ participants' data were in the training.

duration per Launcher (e.g., *Launcher3*) apps' usage on the weekends in the 4 time periods were lower which presents a lower tail in the distribution of their data.

6 Discussion

Our findings present that it is possible to identify loneliness unobtrusively and accurately (Sensitivity = 90.7%, F1 = 82.4%) within a second (Mean = 0.31 s, SD = 1.1 s). On the other hand, the existing robust systems have the need to run in the background for a prolonged period (Table 1) such as 10 weeks [8] and 16 weeks [9] which may not work for early intervention. As a consequence, lonely students' situations can deteriorate. Moreover, current systems' need of running in the background [8–11, 21], and also the need for access to the sensors such as GPS [8, 9, 11] which consumes significant battery power [23] may make the systems infeasible for the resource-limited settings, especially, where limited electricity and internet are two of the barriers to use technology [25]. However, our system does not need to run in the background. In addition, it relies solely on the instantly accessed computationally cheaper app usage data where simple mathematical formulas (please, see Sect. 4.3) are used to extract the features and does not have any dependency on additional computational models (e.g., usage of conversation detection classifiers to extract conversation related features as used in explored dataset of [8]) for feature extraction. All these make our system minimal which can have potential implications for loneliness identification in resource-limited settings such as in developing and underdeveloped countries.

We explored 14 different classification algorithms and the models were built by the features selected by a FS method from each of the 3 main categories [19]. We found

Table 1. Comparison of our system with existing pervasive device-based systems in classifying the lonely and non-lonely. Existing systems' "data retrieval time" is mentioned based on the systems' description available in the research article. NA: Not Available.

Reference	Sample size (N)	Example of the explored data	System's need to run in background	Duration of the collected data	Data retrieval time	Sensitivity	Accuracy	F1
[8]	46	App usage, location, conversation	Yes	10 weeks	10 weeks	67.89	68.67	66.54
[9]	160	Screen, sleep, steps, location	Yes	16 weeks	16 weeks	80.1	80.2	80.1
[10]	9	App usage, Bluetooth, Wi-Fi sensed	Yes	2 weeks	2 weeks	NA	98.0	NA
Our system	**105**	**Only app usage data**	**No**	**1 week**	**Mean: 307.94 ms**	**90.7**	**80**	**82.4**

that the GB model developed by Boruta selected around 6 features (Mean = 5.8, SD = 1.3) have maximum sensitivity and specificity of 74.1% and 70.6% where the sensitivity is more than 10% compared to the filter method Information Gain (IG) and embedded method RF selected 6 features-based models. One of the plausible reasons for having a higher performance is that the Boruta works by selecting all-relevant features to the target variable [18], unlike the minimal-optimal method which is followed by the filter and embedded methods. Due to having a higher performance with a lower number of features, the GB model appears as a parsimonious model. With the lower features, the requirement of computational resources to develop models decreases [40] and thus, the presented GB model can be used to develop a more resource-insensitive system.

SHAP analysis [4] on our best ML model NB based on the IG selected 17 features showed that the lonely students were more likely to have a higher variation in duration per app on weekdays over the 4 time periods. This finding aligns with the studies conducted through conventional statistical methods showing depressed students' variation of diurnal app usage patterns [17]. The variation can reflect students' mood swings while going through a negative experience [24]. In addition, with the variation of time periods, people stay at different places which is related with the variation in usage behavior of app categories [39]. That being said, aggregated data of the whole day may not contain such contextual information and as a result, that may not be informative enough to find subtle differences between the lonely and non-lonely students. This phenomenon is reflected in important feature analysis where compared to the whole day, we found a much higher number of diurnal usage data-based features of the app categories as important. However, the diurnal usage data of the app categories were unexplored in the previous studies [8–10] to develop models to assess loneliness. To improve performance, our findings

suggest incorporating the diurnal usage data-based app categories features also while developing computational models to assess loneliness.

With the goal to facilitate mental healthcare professionals, we explained the best ML models by SHAP. The analysis showed that the lonely students were more likely to have a lower spending duration per launch of Social Media apps on the weekends. This can be explained by the fact that negative feelings can lead to the launch of Social Media apps to seek social support or to be distracted [29]. We speculate that the lonely students' duration per launch can become lower due to facing negative experiences (e.g., negative content, comparison with others [29]) while using Social Media apps. In the case of the lonely students, we also found that they were likely to use a higher number of Lifestyle apps during the night time period. In that app category, students used apps mostly related to prayer which may also reflect their support-seeking behavior. Our findings through explainable ML which was based on the unobtrusively collected data, are in line with a qualitative study's findings [37] where prayer was found as a coping strategy while going through negative experiences amid the pandemic. Going beyond the study's main focus on loneliness identification, these findings can help mental healthcare professionals to take steps in the interventions.

7 Limitations

Mental health being a taboo topic in Bangladesh [6], the sample size was limited (N = 105). This research is expected to generate interest as mental health is a growing research field in developing countries and our findings can facilitate the researchers to develop a better system. Currently, we are conducting a country-wide study and are expecting to come up with a more robust system to more precisely predict loneliness.

8 Conclusion

We present a minimal and real-time system that can identify loneliness by leveraging the instantly (Mean = 0.31 s, SD = 1.1 s) accessed app usage behavioral data. In our study on 105 students of Bangladesh, our developed ML model correctly identified 90.7% lonely students with an F1 score of 82.4%. This shows the efficacy of our minimal system in faster identification of loneliness which can make a worthwhile contribution to minimizing the loneliness rate in low-resource settings.

References

1. World Health Organization (WHO): Mental health atlas 2017. WHO. (2018)
2. Saxena, S., Paraje, G., Sharan, P., Karam, G., Sadana, R.: The 10/90 divide in mental health research: trends over a 10-year period. Br. J. Psychiatry. **188**, 81–82 (2006)
3. Rathod, S., et al.: Mental health service provision in low- and middle-income countries. Health Serv. Insights. (2017)
4. Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st NeurIPS. Curran Associates Inc., Red Hook, NY, USA (2017)

5. Mushtaq, R., Shoib, S., Shah, T., Mushtaq, S.: Relationship between loneliness, psychiatric disorders and physical health ? A review on the psychological aspects of loneliness. J. Clin. Diagn. Res. 8, WE01–4 (2014). https://doi.org/10.7860/JCDR/2014/10077.4828
6. WHO: Bangladesh WHO special initiative for mental health situational assessment
7. Kundu, S., et al.: Depressive symptoms associated with loneliness and physical activities among graduate university students in Bangladesh: findings from a cross-sectional pilot study. Heliyon. **7**, e06401 (2021). https://doi.org/10.1016/j.heliyon.2021.e06401
8. Li, Z., Shi, D., Wang, F., Liu, F.: Loneliness recognition based on mobile phone data. In: Proceedings of the 2016 ISAEECE. Atlantis Press, Paris, France (2016)
9. Doryab, A., et al.: Identifying behavioral phenotypes of loneliness and social isolation with passive sensing: Statistical analysis, data mining and machine learning of smartphone and Fitbit data. JMIR MHealth UHealth. **7**, e13209 (2019)
10. Pulekar, G., Agu, E.: Autonomously sensing loneliness and its interactions with personality traits using smartphones. In: 2016 IEEE HI-POCT. IEEE (2016)
11. Wu, C., et al.: Improving prediction of real-time loneliness and companionship type using geosocial features of personal smartphone data. Smart Health (2021)
12. Hays, R.D., DiMatteo, M.R.: A short-form measure of loneliness. J. Pers. Assess. (1987)
13. YourHour - phone addiction tracker & controller. https://play.google.com/store/apps/details?id=com.mindefy.phoneaddiction.mobilepe. Accessed 28 March 2021
14. Ahmed, M.: 86pc university students own smartphones in Bangladesh: Survey. https://en.prothomalo.com/youth/education/86pc-university-students-own-smartphones-in-bangladesh-survey. Accessed 24 Aug 2021
15. Das, R., Hasan, M.R., Daria, S., Islam, M.R.: Impact of COVID-19 pandemic on mental health among general Bangladeshi population: a cross-sectional study. BMJ Open. (2021)
16. Vabalas, A., Gowen, E., Poliakoff, E., Casson, A.J.: Machine learning algorithm validation with a limited sample size. PLoS ONE **14**, e0224365 (2019)
17. Ahmed, M.S., Ahmed, N.: Exploring unique app signature of the depressed and non-depressed through their fingerprints on apps. In: Proceeding of the PervasiveHealth'21 (2022)
18. Kursa, M.B., Rudnicki, W.R.: Feature selection with the boruta package. J. Stat. Softw. 36, (2010). https://doi.org/10.18637/jss.v036.i11
19. Guyon, I., Nikravesh, M., Gunn, S., Zadeh, L.A. (eds.): Feature Extraction: Foundations and Applications. Springer, Berlin (2006)
20. Fusar-Poli, P., McGorry, P.D., Kane, J.M.: Improving outcomes of first-episode psychosis: an overview. World Psychiatry **16**, 251–265 (2017). https://doi.org/10.1002/wps.20446
21. Austin, J., et al.: A smart-home system to unobtrusively and continuously assess loneliness in older adults. IEEE J. Transl. Eng. Health Med. **4**, 2800311 (2016)
22. Coughlan, S.: Loneliness more likely to affect young people. https://www.bbc.com/news/education-43711606 (2018)
23. Guo, Y., Wang, C., Chen, X.: Understanding application-battery interactions on smartphones: a large-scale empirical study. IEEE Access. **5**, 13387–13400 (2017)
24. Rahiem, M.D.H., Krauss, S.E., Ersing, R.: Perceived consequences of extended social isolation on mental well-being: narratives from Indonesian university students during the COVID-19 pandemic. Int. J. Environ. Res. Public Health. **18**, 10489 (2021)
25. Owoyemi, A., Owoyemi, J., Osiyemi, A., Boyd, A.: Artificial intelligence for healthcare in Africa. Front Digit Health. **2**, 6 (2020). https://doi.org/10.3389/fdgth.2020.00006
26. Hunt, M.G., Marx, R., Lipson, C., Young, J.: No more FOMO: limiting social media decreases loneliness and depression. J. Soc. Clin. Psychol. **37**, 751–768 (2018)
27. Zhao, S., et al.: Discovering different kinds of smartphone users through their application usage behaviors. In: Proceedings of the ACM UbiComp'16 (2016)
28. Gao, Y., Li, A., Zhu, T., Liu, X., Liu, X.: How smartphone usage correlates with social anxiety and loneliness. PeerJ **4**, e2197 (2016). https://doi.org/10.7717/peerj.2197

29. Sarsenbayeva, Z., et al.: Does Smartphone Use Drive our Emotions or vice versa? A Causal Analysis. In: Proceedings of the ACM CHI'20 (2020)
30. Mendes, J.P.M., et al.: Sensing apps and public data sets for digital phenotyping of mental health: Systematic review. J. Med. Internet Res. **24**, e28735 (2022)
31. Erzen, E., Çikrikci, Ö.: The effect of loneliness on depression: a meta-analysis. Int. J. Soc. Psychiatry. **64**, 427–435 (2018). https://doi.org/10.1177/0020764018776349
32. Lee, S.L., et al.: The association between loneliness and depressive symptoms among adults aged 50 years and older: a 12-year population-based cohort study. Lancet Psychiatry (2021)
33. UsageStatsManager. https://developer.android.com/reference/android/app/usage/UsageStatsManager. Accessed 15 Sept 2022
34. BANBEIS: bangladesh education statistics 2021 (2022)
35. boruta_py: Python implementations of the Boruta all-relevant feature selection method
36. Zhang, Y., Yang, Y.: Cross-validation for selecting a model selection procedure. J. Econom. **187**, 95–112 (2015). https://doi.org/10.1016/j.jeconom.2015.02.006
37. Finlay, J.M., et al.: Coping during the COVID-19 pandemic: a qualitative study of older adults across the United States. Front. Public Health. **9**, 643807 (2021)
38. Peduzzi, P., Concato, J., Kemper, E., Holford, T.R., Feinstein, A.R.: A simulation study of the number of events per variable in logistic regression analysis. J. Clin. Epidemiol. (1996)
39. Mehrotra, A., et al.: Understanding the role of places and activities on mobile phone interaction and usage patterns. In: Proceedings of the ACM Interaction Mobile Wearable Ubiquitous Technology (2017)
40. Remeseiro, B., Bolon-Canedo, V.: A review of feature selection methods in medical applications. Comput. Biol. Med. **112**, 103375 (2019)
41. Mobile operating system market share Bangladesh. https://gs.statcounter.com/os-market-share/mobile/bangladesh. Accessed 15 Sept 2022