Emotional State Inference Using Mobile Sensing

Diogo Mota
200013237
School of Electronic Engineering and Computer Science
Queen Mary University of London
MSc Big Data Science
Supervisor: Usman Naeem

***Abstract*—Due to the heterogeneous characteristics of depression, diagnosis is often hard to perform and the time interval between first showing depression symptoms and the first treatment can be as big as 8 years. Given the ubiquitous nature of smartphones, it is not surprising that our emotions have a direct impact on our interactions with them. Furthermore with the continuous smartphone advancements, scientists have been able to collect important mobile sensing data that has already shown to link back to depression and its symptom severity. Therefore, the question arises: Can mobile sensing data be used to infer and predict daily aspects interconnected with depression to help its diagnosis? To answer this question we investigate the data produced from StudentLife, a study that monitored the mental state of a collection of students of Dartmouth University for a period of 10 weeks. We show that mobile sensing data provides reliable enough features to accurately predict stress signs, sleep quality and depression signs.**

***Index Terms*—Digital mental health, Depression, Machine Learning, Mobile Sensing**

I. Introduction

According to the World Health Organization, in 2015 alone, an estimated 322 million people suffered from depression, corresponding to an 18.4% increase compared to 2005 (Organization et al. 2017). The latest GDB study shows that in 2019, depression was the sixth most common cause of years lived with disability for people aged between 25-49 years old and only forth for 10-24 years old (Vos et al. 2020). Due to the heterogeneity of the disease, diagnosis is often hard to perform. According to Wang et al. (2005), in the U.S, the median time to treatment is actually 8 years. This means that the time before a person first showing symptoms of depression and their first treatment session can reach as long as 8 years. During this time the symptoms only worsen, making remission harder to achieve. Even though there are a plethora of effective treatments for depression, ranging from psychotherapy to antidepressants, a failure to detect depression is a major cause of a population-level disability, furthermore people often do not receive adequate care (Boenisch et al. 2012, Wittchen et al. 2011). In fact, more than two thirds of diagnosed patients do not actually receive treatment (Kessler et al. 2005).

Smartphones have become an essential, everyday object for humans. These devices allow users to connect with the world. It even has been shown that people create emotional attachments to their mobile phones (Vincent 2006). It is therefore not surprising that researchers have been investigating important aspects of our interaction with them. Several studies have already investigated the correlation between users' mood and their interactions with their smartphones (Alvarez-Lozano et al. 2014, LiKamWa et al. 2013, Vermeulen et al. n.d.). More interestingly however, was Mehrotra et al. (2017) investigation of the causal links between users emotional states and their interaction with mobile phones, which found that users emotions have a direct causal impact on different aspects of their mobile phone interactions.

Mobile phones can also be used to help clinicians detect users' emotional states on the fly. As smartphones evolved we have been able to capture important data that has already shown to link back to depression and it's symptom severity. One example is the mobile phone location sensor data, that Saeb et al. (2016) showed to have important features that provide reliable predictors of depression symptoms.

Smartphones can also be used to infer and alter users' emotional states, as demonstrated by Bendtsen et al. (2020), that developed a mobile app to increase the positive mental health of students of a Swedish University. The mHealth app proved to be more effective than usual care in increasing students' positive mental health, proving that mHealth solutions are more than an ideology in public mental health promotion.

Another key study that displayed the monitoring capabilities of mobile phones using their sensing data was that of Wang et al. (2014), when they assessed users' mental health in the StudentLife study. Using a mobile app, they monitored the mental state of a collection of students of the Dartmouth University for a period of 10 weeks, using mobile sensus and smartphones sensing capabilities. The resulting dataset, comprising a variety of features, such as audio inference, activity inference, call logs, app usage, EMAs (Ecological Momentary Assessment), and survey data was made publicly available online for anyone to use.

Even though a number of studies have shown that smartphones allow for a continuous monitorisation of users' mental states, these are limited to exactly this, monitoring. More often than not, the data extracted from these studies is only used for analysis after the experiment has finished, not during. This is due to the fact that the usual goal of these studies is to unobstructedly collect data to infer on the participants day-to-day mental state. In contrast, this project, using mobile sensing data collected during the StudenLife study, aims to show that smartphones can not only be used to monitor user's

mental state, but also predict user behaviour and provide them feedback in order to increase user awareness and help them cope with their mental struggles. We show that mobile sensing provides sufficiently reliable features that allow for the construction of machine learning algorithms capable of effectively predicting users' emotional states. This opens the way to the creation of more attentive mobile apps capable of delivering personalised mental health interventions for its users.

II. Contribution

In light of the difficulty in detecting depression and the challenges current mobile sensing studies present, this project, inspired by the work of Saeb et al. (2015), uses data produced in the StudentLife dataset to investigate potential, reliable features that allow machine learning models to infer users' mental state. This project is divided into two main tasks. Firstly, we focus on the prediction of aspects interconnected with depression, such as signs of stress and sleep quality by using and comparing the XGBoost and TabNet models. Secondly, through the implementation of a Fully Convolutional Network (FCN), we attempt to predict signs of depression in students after the completion of the StudentLife study.

III. Related Work

A. Previous mobile sensing studies

Mobile phones are an everyday item in everyone's lives. On account of their increasingly large complement of sensors, their potential to monitor behavioural patterns that provide insight into user's mental state has been proven in several studies. A review of the literature was performed to identify key studies that display the capabilities of mobile sensing to both monitor and alter user's emotional state.

The StudentLife study, developed by Wang et al. (2014), assessed the mental health of a collection of students during the 10 week term of the Dartmouth College. Influenced by a number of previous experiments, such as the friends and family study (Aharony et al. 2011) and the reality mining project (Eagle & Pentland 2006), they developed a mobile app to collect a range of sensing data types, had participants answer several EMAs and fill a set of surveys with the intent to assess students' mental health. With the purpose of implementing classifiers to infer some users given states, they built upon previous work. For instance, they implemented a sleep classifier developed in Chen et al. (2013), Lane et al. (2011) to unobtrusively infer sleep duration. Another example was the activity classifier Lane et al. (2011), Lu et al. (2010), implemented to infer stationary, walking, running, driving and cycling states. Other types of data were collected using mobile sensing such as app usage, bluetooth, wifi, EMA and survey data. Results from this study showed multiple significant correlations between mobile sensing data and mental health and educational outcomes. Furthermore, they identified a term life cycle in the University of Darmouth. The StudentLife experiment worked as a basis for several new studies that attempted to show the applicability of mobile sensors to monitor mental health (Sano et al. 2015). The data produced was also used to predict the students GPA in another study (Wang et al. 2015).

As previously stated, smartphones can also be used to alter users' mental state. This was the case of Bendtsen et al. (2020) study, that consisted of a 2-arm, single-blind, parallel groups randomised controlled trial that aimed to test the feasibility of an mHealth mobile app to increase the positive mental health of a group of students of a Swedish University. The primary outcome of the experiment was positive mental health and the secondary was depression and anxiety symptomatology. Participants were randomly assigned to either an intervention group or a control group. For the duration of the study (10 weeks), the intervention group had access to a mHealth positive psychology app, that aimed to enhance users' positive mental health and was developed based on empirical evidence. The application was comprised of information regarding well-being, self-help exercises, brief tips, self-monitoring and personalised feedback. Participants allocated to the control group were notified of their position through text message, which also included all essential contact details to their local health center, primary care center, and governmental national health website. At the 3 month follow-up, data from both groups was collected and showed that positive mental health, measured by the Mental Health Continuum Short Form (MHC-SF), was substantially higher in the intervention group compared to the control group. On the one hand, the study proved the viability of using a fully automated mHealth app to enhance users' mental state. On the other hand, it is important to note that the interventions given were not tailored for different individuals. All participants were given the same interventions and no adjustments were made in order to attend to users' preferences. If feedback from the participants had been collected, a machine learning algorithm could have been deployed to inform what given intervention was better suited for a given participant.

With the goal of discovering the reliability of mobile phone GPS and app usage sensing for the prediction of both signs and levels of depression, Saeb et al. (2015) work consisted of an experiment comprised of 40 participants, that in a span of 2 weeks provided GPS and phone usage data from a data acquisition app named Purple Robot (Schueller et al. 2014). Based on the collected data, they defined a given number of features and built classification and regression models to investigate their relationship with depressive symptom severity. In order to determine if the obtained features would provide useful insight regarding users' depressed symptoms, correlation analysis between the different features and the PHQ-9 scores was performed. Additionally, participants were divided into those with depressive symptoms (PHQ-9 > 5) and those without (PHQ-9 ≤ 5). With this they tested whether it was possible to distinguish people with depressive symptoms from those without. Two types of models were created for 2 different tasks: in order to estimate the PHQ-9 score of the participants, a regression model was built, using the features extracted from the data. A logistic regression classifier was built to identify participants with depressive symptoms. The

regression model predicted the scores with an average error of 23.5% and the logistic classifier performed with an accuracy of 86.5%.

B. The XGBoost and TabNet Models

All the data produced in the previously discussed studies, more importantly the StudentLife dateset, is of one type: tabular. Up until recently Gradient Boosting Machines (GBMs) were the heavy favorite for classification problems involving tabular data since, when compared to neural networks, they require less coding, perform better in smaller datasets (neural networks often require very large datasets), are faster to train and easily interpretable (whereas neural networks are often considered black boxes). At the top of GBMs stands the XGBoost. Introduced in Chen & Guestrin (2016), the XGBoost is a speed and performance oriented implementation of gradient boosted decision trees. At its core, XGBoost is an ensemble tree method where each tree boosts the attributes that led the previous tree to a misclassification. However, XGBoost improves upon standard GBMs with it's novel system optimisations: parallelisation (the process of sequential tree building is parallelised by interchanging the order of the loops used for building base learners), tree pruning (uses the max_depth parameter and starts pruning trees backwards improving performance), hardware optimisation (by introducing cache awareness and out-of-core computing) and algorithmic enhancements, such as regularisation (uses both L1 and L2 regularisation methods to prevent overfitting), sparsity awareness (to handle sparsity patterns in the data more efficiently), weighted quantile sketch (to be able to handle weighted data) and cross validation (incorporated at each iteration). With these novel changes, XGBoost currently outperforms state-of-the-art decision tree based algorithms.

However, after the introduction of the TabNet, the paradigm has since shifted. TabNet, a deep neural model tackled all the previous disadvantages of neural networks compared to GBMs. Developed in Arik & Pfister (2021), the TabNet, although complex, is a highly interpretable model. Its design allows it to inherit the explainability of tree based methods, while still providing the key benefits of deep learning models. It possesses a set of unique features that allowed it to stand out from other architectures: it is able to input raw data, without any need for preprocessing, which currently constitutes a major step for data scientists when developing machine learning models; it's trained using gradient descent; uses sequential attention to choose features at each decision step (with a built-in explainability in the form of masks); employs single deep learning architectures and enables two types of interpretability: global through the quantification of feature contributions and local by visualising the importance of features and how they are combined for a single row.

C. Fully Convolutional Networks for multivariate time-series classification

Besides tabular, large amounts of data are stored in the form of time series, such as, stock market data, weather data, electroencephalography results, etc. The tabular data produced in the StudentLife dataset can also be viewed as several time series, one per data type, per student.

When it comes to the task of time series classification, more specifically, multivariate time series classification, Fully Convolutional Networks (FCNs) have shown to be a very effective architecture. Usually, image classification is the common association when talking about Convolutional Neural Networks (CNNs). However, the less famous 1D convolutions allow for a more generalised use of CNNs.

According to Wang et al. (2017), a study that explored 3 different types of networks for time series classification: Multilayer Perceptron (MLP), Fully Convolutional Network (FCN) and the Residual Network (ResNet), the FCN achieved premium performance. Its basic block is a convolutional layer (the convolution is performed by 3 1D kernels) followed by a batch normalisation (to increase convergence speed and improve generalisation) and a ReLU activation layer. After the convolution blocks, the features are passed to a global average pooling layer, instead of a fully connected layer, to greatly reduce the number of weights. Finally, the prediction is produced by the softmax layer (Figure 1).

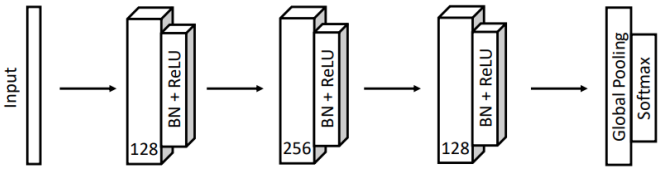

Fig. 1. **Example of a FCN architecture. (From Wang et al. (2017))**

IV. Methodology

This project comprises 2 main tasks, yielding a total of 3 classification tasks. The first task was focused on predicting key aspects that influence users' emotional states, such as: if a student, on a given day, will show signs of stress and if a user, on a given day, will be able to sleep the recommended amount of hours. The second task aimed at using the data in the form of time-series to predict signs of depression in students.

We therefore propose the methodology presented in Figure 2. The first step consists of preprocessing the data: all missing values are imputed using a K-Nearest-Neighbours approach. Secondly, through feature engineering we extract statistical features from the chosen data types. Thirdly, we aggregate the data depending on the given task. For both tasks, the features are organized by data type and then aggregated by day. Specific to each task is how the data is organised. In the first task, we joined the features from all participants and given the existence of restrictive data types that shortened the final size of the dataset we created two datasets: dataset A, which excluded the restrictive data types and dataset B that included them. For the second Task, each data instance is a set of time series that corresponds to a given student. Fourthly, the features are standardised by removing the mean and scaling to unit variance. Moreover, in order to optimise our models'

performances, we removed unnecessary or conflicting features using Recursive Feature Elimination (RFE) for the first task and selected the set of features that allowed the largest dataset possible for the second.

A. KNN Imputer

The sklearn KNNImputer method provides imputation for filling in missing values using the k-Nearest-Neighbors approach. Using the euclidean distance metric, the nearest neighbours of the missing values are found. Each missing value is imputed by averaging (either uniformly or wheighted) the available features of the corresponding k nearest neighbours.

B. Feature Engineering

From the plethora of available data types in the StudenLife dataset, we selected the following: Call log, EMA, Sensing, SMS, Survey. Details regarding the types of data and their collection can be found in Wang et al. (2014). The reasons for this particular subset of data types vary: many data types had a large amount of missing values, details regarding the meaning of the data or how it was collected were not given or the data was not related to a user's emotional state.

From the Call log data we had access to all the calls durations in a day. During the 10 week experiment, students responded to psychological and behavioural EMAs on their smartphones. These EMAs ranged from stress, mood, sleep, social aspects, physical exercise, time spent on different activities, short personality item and PAM (Photographic Affect Meter), where users identified from a set of images which they found best described how they felt. Due to the vast amount of missing values and being related with specific events, such as the Boston Bombing, several EMAs had to be excluded from being used to extract features. Out of the 27 EMAs collected during the experiment we used only 4. For the sensing data there are a total of 10 different sensor data, 8 of which were used to extract features: physical activity, audio inferences, conversation inferences, light sensor, GPS, phone charge, phone lock and wifi location. From the SMS data we were only interested in knowing the amount of SMS exchanged per day. Survey data was only used for the purpose of predicting depression signs in the students, since these were only collected twice: once before the study began and again when it ended.

For the majority of the collected data types, the standard procedure for feature engineering was aggregating the available values by day and extracting statistical features such as the mean, median, standard deviation, minimum and maximum values, the skew and the variance. Any other process will be mentioned below:

- **Call log data -** with the duration of the calls received/taken in a given day, the standard procedure was applied to obtain features;
- **Sensing data:**
  - Activity: there were 4 possible values for the activity inference: 0 (stationary), 1 (walking), 2 (running) and 3 (unknown). With these values collected on a daily basis, often collected more than once a day the standard procedure was applied to get the features. As an additional features, the sum of theses values per day was calculated;
  - Audio: the same procedure as for the activity inference data was used. There were also 4 possible values for the audio inference: 0 (silence), 1 (voice), 2 (noise) and 3 (unknown);
  - Conversation: calculated the conversation duration for each day this data was available. Afterwards the standard procedure was applied;
  - Light sensor: calculated the amount of time the phone was at a dark environment and executed the standard procedure;
  - GPS: using the latitude, longitude, altitude, bearing, speed and travelstate values the standard procedure was performed. Note that the travelstate values were encoded as 0 if the value was "stationary" and 1 if it was "moving";
  - Phone Charge: once again we calculated the amount of time the phone was charging and used the standard procedure to extract features;
  - Phone Lock: an analogous procedure to the phone charge data was performed;
  - Wifi Location: calculated the amount of different locations a user visited during a day;
- **EMA data:**
  - PAM: with the indexes of the images students chose, we followed the mentioned standard procedure to obtain features;
  - Sleep: this EMA comprised of 3 questions whose id were: hour, rate and social. For each of these questions the standard procedure was applied;
  - Social: comprised of only one question, we applied the standard procedure for feature extraction;
  - Stress: the same process as for the social EMAs was applied.
- **SMS data -** for SMS data we calculated the amount of messages exchanged in a given day and used this as the feature.

Specific to the second task is the preprocessing of the survey data. Students were required to fill a set of surveys at the beginning of the study and again at the end (10 weeks after). One of the surveys filled was the PHQ-9 (Kroenke et al. 2001), a self administered questionnaire, scoring each of the nine DSM-IV criteria as "0" (not at all) to "3" (nearly every day). Scores lower than 5 indicate no signs of depression, whereas scores equal or above include mild (5), moderate (10), moderately severe (15) and severe depression (20).

C. Data Standardisation and Feature Selection

Feature standardisation is an essential step of data preprocessing, it allows for different features to have similar weights between each other. Without it, different features that have larger values carry more weight when compared to smaller features.

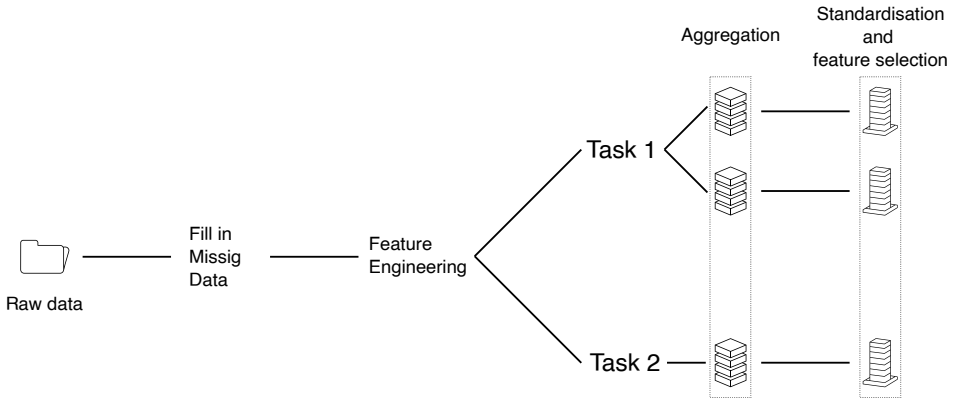

Fig. 2. Schematic representation of the adopted methodology. (Image vectors from https://www.vectorstock.com/royalty-free-vector/49-data-icons-vector-36842999)

The final step used to build the datasets for the first task is called feature selection. In order to optimise the models' performances we removed unnecessary or conflicting features using Recursive Feature Elimination (RFE). This automated feature selection algorithm trains a model on a gradually smaller set of features. In this project, the RFE trained a Random Forest Classifier. Since this method requires a specific number of features to keep, which usually is not known, we added cross-validation. This way, the algorithm selects the best subset of features for the given estimator and then selects the best subset based on the cross-validation score of the model.

D. Models

Both the XGBoost and the TabNet models have many hyperparameters that need to be tuned so we can get the best possible performance out of them. A schematic representation of the set of hyperparameters tuned, per model, is presented in Figure 3.

All of the mentioned hyperparameters were tuned and only the best set was used for the classification tasks.

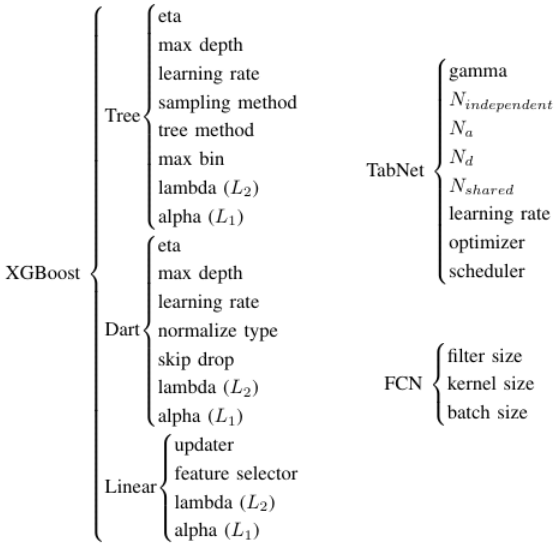

Fig. 3. **Set of hyperparameters tuned per model.** For the XGBoost we experimented with the 3 available boosters and tuned their respective hyperparameters. Both Dart and Tree booster are tree based methods and the linear applies linear functions. The most important hyperparameters were the ones focused on tackling overfitting such as eta, alpha, lambda and max_depth. Being a deep learning model, the TabNet has some familiar parameters: optimisers (methods to change the attributes of the NN) and schedulers (to change the learning rate during training). Others are unique to TabNet: N_d (width of the decision prediction layer), N_a (width of the attention embedding for each mask), N_{steps}, the number of independent ($N_{independent}$) and shared (N_{shared}) Gated Linear Unites layers at each step, the momentum (for batch normalization) and an extra sparsity loss coefficient. For the final task, based on the implementation of Ismail Fawaz et al. (2019), a custom FCN was created and it's standard CNN hyperparameters tuned.

V. Results and Discussion

This section regards the results and discussion from each experiment. The first two sections correspond to predicting signs of stress and whether a given student will sleep the recommended hours in a day. As previously mentioned, the data used for this task was organised by day. We noticed that including both the SMS and Call log features reduces the number of data instances from 397 (dataset A) to 117 (dataset B), therefore one question arose: Are the restricting features important or reliable enough to compensate for the difference in dataset sizes? To answer this question we trained and tested the XGBoost and TabNet models with and without the SMS and Call log data.

The final section corresponds to the task of multivariate time-series classification where we used an FCN to predict signs of depression in students after the study was completed.

Note that since the analysis in the first two subsections are very similar, only the first subsection presents the produced plots. The remaining (analogous graphics) were placed and explained in the Appendix. Similarly, the TabNet Loss, Accuracy and Learning rates plots were also placed in the Appendix, since their analysis is meant to simply show adequate training.

A. Stress Prediction

To predict signs of stress we turned to the responses of the stress EMA. Students had to fill in the following sentence: "Right now, I am..." with one of the available answers:

- A little stressed [1];
- Definitely stressed [2];
- Stressed out [3];
- Feeling good [4];
- Feeling great [5];

With these answers we obtained the mean stress level of each student per day. Afterwards, we encoded these results to obtain the labels: based on the meaning of the available responses of the stress EMA, we defined as showing signs of stress if the mean level of stress $\in [1, 4[$ Levels above meant no signs of stress.

For the rest of the section we will be presenting and discussing the obtained results: feature correlation, model performance, feature importance, model explainability.

When it comes to feature correlation, both datasets yielded small values of correlations between the features and the mean level of stress a student experiences in a day. Based on the work of Saeb et al. (2016), we hypothesised that features related to users movement, such as the number of distinct locations, activity inferences and GPS features would yield the highest correlation values. Even though higher values were expected, our hypothesis is supported by the findings of Tables I and II. When using dataset B, we observed a general increase in feature correlation. Furthermore, features related to Call logs appear in the top 5 correlated features, indicating good predictor reliability.

Unexpected, was the lack of correlation between stress related features and sleep and vice-versa. This of course can be due to a plethora of reasons: the feature engineering process yielded only statistical features which could have destroyed any correlation or not explored an existing one, it could additionally be due to the way the data was collected. Interestingly, the features with the highest correlation values were not the ones the models found to be the most important (Figure 4). In fact, the importance distribution links the association between sleep and stress we were looking for, which is in accordance with the literature. When trained with dataset A, the TabNet model attributed the most importance to the mean hours of sleep. According to Hall et al. (2000), the tendency to experience stressful thoughts is associated with sleep complaints. Furthermore, Carlson & Garland (2005) found that using a Mindfulness-Based Stress Reduction program not only significantly reduced stress, but also improved sleep quality in patients. In addition, the second most important feature for both models supports the previously stated link between GPS data and stress since it corresponds to the maximum value of the latitude, for the TabNet, and the median value of the latitude, for the XGBoost.

Both models performed exceptionally well, being able to predict signs of stress with an accuracy significantly above chance (Table III). When trained on dataset A, the TabNet, even though very slightly, was able to outperform the XGBoost model. However, this difference is practically negligible.

It is also important to note that the performance of both models increased when we included the SMS and Call log data (dataset B), meaning that their reliability, as per Table II, allowed for a better model performance (Table III). The fact that the XGBoost performs better than the TabNet is expected given that the size of the dataset decreased drastically and deep neural networks tend to perform worse with small datasets due to their high number of parameters. This requires a large number of iterations to find all optimal values. Consequently, using a high number of iterations with small datasets often results in overfitting. We are unable to compare feature importance when using dataset B since the best performing model of the XGBoost is obtained using the Linear booster which implements linear functions and the concept of feature importance is specific to tree based methods.

TABLE I
DATASET A TOP 5 CORRELATED FEATURES FOR THE STRESS PREDICTION TASK.

Feature	Correlation
N of distinct locations	**0.136**
Sum of audio inference	0.133
Median of activ. inference	0.119
Sum of activ. inference	0.116
Median number social	0.105

TABLE II
DATASET B TOP 5 CORRELATED FEATURES FOR THE STRESS PREDICTION TASK.

Feature	Correlation
Sum of audio inference	**0.183**
N of distinct locations	0.181
Median of call duration	0.156
Skew conv. duration	0.147
Mean of call duration	0.141

Even though GBMs have a good level of explainability, TabNet provides a deeper look at feature importance through masks. From Figure 5 we can observe 3 matrices, each corresponding to one decision step of the TabNet when trained on dataset A. Each row of these matrices corresponds to a test sample and a column to a feature. By looking at these matrices we can investigate which features were given more importance per prediction. For instance, if we look at the first decision step (Mask 0) we see that for the majority of predictions, the highest importance was given to the last 2 features which are sleep related features (median of the hours sleep and the minimum of hours of sleep respectively).

B. Sleep Prediction

From the sleep EMA we used the results of the question "How many hours did you sleep last night?" and calculated

TABLE III
MODEL PERFORMANCE WHEN PREDICTING WHETHER A USER SHOWED SIGNS OF STRESS.

	Dataset A	Dataset B
XGBoost	0.830	**0.890**
TabNet	**0.833**	0.850

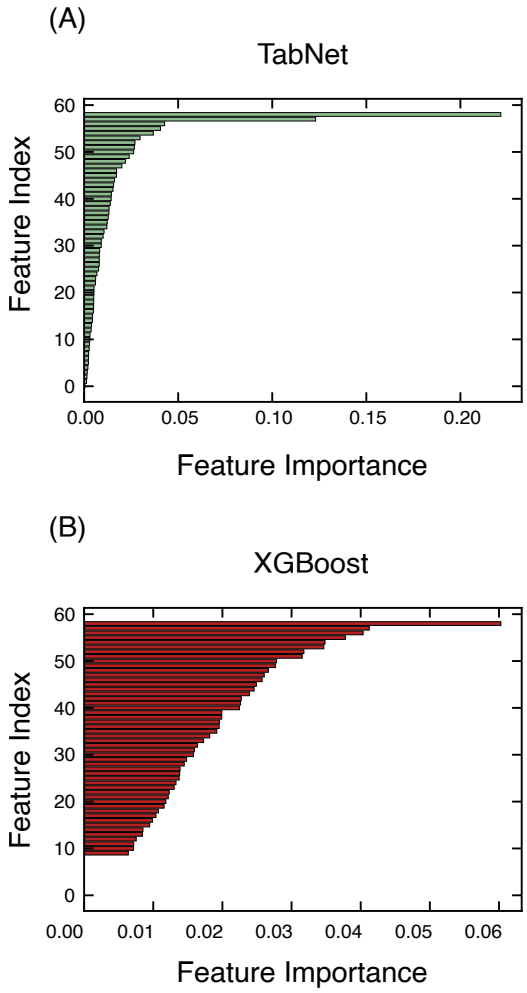

Fig. 4. **Feature Importance histogram for dataset A**. An advantage of using a tree ensemble and TabNet is their explainability. These models allow for feature exploration so we can better understand how and why a model makes a given prediction. As previously stated, TabNet was designed to learn a "decision-tree-like" mapping in order to benefit from the explainability of tree based models. (A)Feature Importance histogram using the TabNet model. (B) Feature Importance histogram using the XGBoost model.

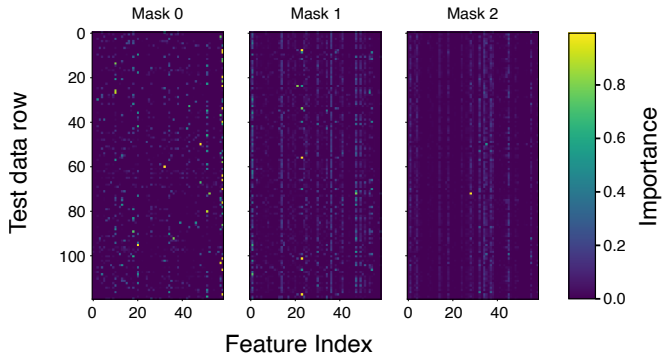

Fig. 5. **TabNet Feature Importancs Masks for dataset A.** Each mask corresponds to a decision step during the testing stage of the TabNet. These masks allow us to understand which features were the most important on a individual prediction level.

the average score of the answer which corresponds to a given number of hours slept, therefore calculating the mean number of hours a person sleeps in a day.

According to a significant body of research (Hirshkowitz, Whiton, Albert, Alessi, Bruni, DonCarlos, Hazen, Herman, Hillard, Katz et al. 2015, Hirshkowitz, Whiton, Albert, Alessi, Bruni, DonCarlos, Hazen, Herman, Katz, Kheirandish-Gozal et al. 2015, Panel et al. 2015), the recommended sleep duration for adults is between 7 to 8 hours, therefore we encoded the mean hours of sleep the following way: if the mean hours of sleep were higher than 7, the given student slept the recommended amount of time that day, otherwise he/she did not.

Analogously to what was done in the previous section, we performed the same set of experiments during this task.

TABLE IV
DATASET A TOP 5 CORRELATED FEATURES FOR THE SLEEP PREDICTION TASK.

Feature	Correlation
Variance of phone lock duration	**0.265**
Maximum duration of phone lock	0.262
Standard deviation of the phone lock duration	0.247
Variance of light duration	0.171
Maximum of light duration	0.134

TABLE V
DATASET B TOP 5 CORRELATED FEATURES FOR THE SLEEP PREDICTION TASK.

Feature	Correlation
Maximum duration of phone charge	**0.372**
Maximum duration of phone lock	0.364
Standard deviation of the phone lock duration	0.285
Variance of the phone charge duration	0.279
Variance of the phone lock duration	0.275

Tables IV and V show that the features with the highest correlation values are all related to phone usage. This finding is expected since several studies have already concluded that phone usage impacts both quality and sleep duration. Such is the case of the work of Gupta et al. (2016), who found that night time usage of mobile phones was significantly associated with difficulty in waking up, waking up tiredness and other related effects.

Given that we were able to associate sleep with signs of stress, we looked to replicate the finding in this experiment. However we were unsuccessful. Similarly to what occurred in the previous section, most of the features awarded with the most importance were not the highest correlated ones. This time around, for dataset A, both the TabNet and XGBoost attributed more importance to features related with motion, audio inference and some phone usage features (not coincident with features with high correlation values). However the XGBoost gave most importance to the PAM (mood inferences) related features, which according to Finan et al. (2015) are connected with sleep quality.

TABLE VI
MODEL PERFORMANCE WHEN PREDICTING WHETHER A USER SLEPT THE RECOMMENDED AMOUNT OF HOURS.

	Dataset A	Dataset B
XGBoost	**0.860**	**0.810**
TabNet	0.842	0.723

Similar conclusions about model performance can be made as in the previous section: both models predicted if a student slept the recommended number of hours significantly above chance. When trained on datset B, the XGBoost outperformed the TabNet model.

Even though the feature importance distributions (Appendix, Figure 9) behave similarly as in Figure 4, the XGBoost achieves a higher performance, leading us to believe that distributing importance more equally, despite being a TabNet characteristic, does not translate to better performance. Table II also supports this finding since the difference between model performances is practically negligible.

C. Depression Prediction

Similarly to what was done in Saeb et al. (2015), the post PHQ-9 scores were encoded as 0 if the score was less than 5, meaning no signs of depression, and 1 otherwise.

Since this task is user specific, meaning that each student has a set of time-series (one per feature) that are unique to them, the size of the dataset was particularly small, furthermore we are using an FCN as our machine learning model, which presents a problem already mentioned: neural networks tend to perform poorly on small datasets. An additional constraint was the amount of days worth of data each participant had. The post PHQ-9 scores were obtained at the end of the study, 10 weeks after the study began. This means that, in order for our predictions and analysis to be credible, we need at least 50 days worth of data for each student. In order to tackle the first problem we tuned the hyperparameters as per the Models subsection in the Methodology. As for the latter constraint, the following approach was taken: in order to maximise the number of features used as well as the number of days worth of data (for every user) we determined what was the lowest amount of data types we could use where, after feature engineering, the majority of the students would have more than 50 days worth of data. The result were 3 data types: activity inference, audio inference and conversation duration, which yielded a total of 23 features per student, for a total of 33 students. Moreover, the model was trained using 4 folds of cross-validation to test the model's ability to generalise and predict new instances of data and also flag issues such as overfitting.

The FCN was able to predict depression signs with an accuracy slightly above chance of 53.5%. However with an average False Negative rate of 54.2% (False positive rate was 33.3%). This underwhelming performance is mainly due to the size of the dataset. Nontheless, it shows that mobile sensing features are reliable enough to predict depression signs in students, as per Saeb et al. (2015). With a larger dataset, we firmly believe the model would provide a satisfactory performance.

VI. LIMITATIONS

During the feature engineering process it became clear that, although extremely rich, the StudentLife dataset lacked quality. This is not due to the design of the experiment, but with the participants. As it is common in "in the wild" experiments, as the study progresses participant adherence decreases. This results in several gaps in the daily collection of data that resulted in a heavy size constriction of the datasets built in this study. This was a problem during the prediction of signs of depression, since the more features we used, fewer students had sufficient days worth of data. This resulted in a model with an underwhelming performance and displaying a high False Negative Rate (a common issue as stated in Saeb et al. (2016)). Furthermore, the feature importance analysis performed is restricted by the feature selection process. Even though we implemented a 5 step cross validation, different uses of the REFCV resulted in different selected features, which consequently influences which features the models find the most important.

VII. CONCLUSION AND FUTURE WORK

Regardless of these limitations, the present project displayed the potential and reliability of mobile sensing features to predict both interconnected aspects of depression (stress and sleep), extending the work of Saeb et al. (2015), and depression signs.

We show that it is possible to accurately predict both signs of stress (83.3% and 89.0%) and if a given person will sleep the recommended amount of hours (86.0% and 81.0%).

Given the richness of the StudentLife dataset, future work could study further associations between mobile sensing and a user's emotional state, for instance, predicting the amount of people a given user will meet the following day, based on the previous day's set of features. Given the unsatisfactory performance of the FCN, additional architectures should be tested, namely the ResNet and LSTMs. Furthermore, given the size constriction of the dataset, either a new, refined version of the StudentLife should be performed to increase the dataset quality, or merging similar datasets to increase the final dataset size.

All in all this project shows that mobile sensing features can and should be used in either a clinical environment to help track patient progress, or in mobile applications to increase users awareness of their mental state and mitigate the time between first showing depression symptoms and the first treatment session and to tailor user interventions in order to optimise their effectiveness.

REFERENCES

Aharony, N., Pan, W., Ip, C., Khayal, I. & Pentland, A. (2011), 'Social fmri: Investigating and shaping social mechanisms in the real world', *Pervasive and Mobile Computing* **7**(6), 643–659.

Alvarez-Lozano, J., Osmani, V., Mayora, O., Frost, M., Bardram, J., Faurholt-Jepsen, M. & Kessing, L. V. (2014), Tell me your apps and i will tell you your mood: correlation of apps usage with bipolar disorder state, *in* 'Proceedings of the 7th International Conference on Pervasive Technologies Related to Assistive Environments', pp. 1–7.

Arik, S. & Pfister, T. (2021), Tabnet: Attentive interpretable tabular learning.

Bendtsen, M., Müssener, U., Linderoth, C. & Thomas, K. (2020), 'A mobile health intervention for mental health promotion among university students: randomized controlled trial', *JMIR mHealth and uHealth* **8**(3), e17208.

Boenisch, S., Kocalevent, R.-D., Matschinger, H., Mergl, R., Wimmer-Brunauer, C., Tauscher, M., Kramer, D., Hegerl, U. & Bramesfeld, A. (2012), 'Who receives depression-specific treatment? a secondary data-based analysis of outpatient care received by over 780,000 statutory health-insured individuals diagnosed with depression', *Social psychiatry and psychiatric epidemiology* **47**(3), 475–486.

Carlson, L. E. & Garland, S. N. (2005), 'Impact of mindfulness-based stress reduction (mbsr) on sleep, mood, stress and fatigue symptoms in cancer outpatients', *International journal of behavioral medicine* **12**(4), 278–285.

Chen, T. & Guestrin, C. (2016), Xgboost: A scalable tree boosting system, *in* 'Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining', pp. 785–794.

Chen, Z., Lin, M., Chen, F., Lane, N. D., Cardone, G., Wang, R., Li, T., Chen, Y., Choudhury, T. & Campbell, A. T. (2013), Unobtrusive sleep monitoring using smartphones, *in* '2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops', IEEE, pp. 145–152.

Eagle, N. & Pentland, A. S. (2006), 'Reality mining: sensing complex social systems', *Personal and ubiquitous computing* **10**(4), 255–268.

Finan, P. H., Quartana, P. J. & Smith, M. T. (2015), 'The effects of sleep continuity disruption on positive mood and sleep architecture in healthy adults', *Sleep* **38**(11), 1735–1742.

Gupta, N., Garg, S. & Arora, K. (2016), 'Pattern of mobile phone usage and its effects on psychological health, sleep, and academic performance in students of a medical university', *National Journal of Physiology, Pharmacy and Pharmacology* **6**(2), 132–139.

Hall, M., Buysse, D. J., Nowell, P. D., Nofzinger, E. A., Houck, P., Reynolds III, C. F. & Kupfer, D. J. (2000), 'Symptoms of stress and depression as correlates of sleep in primary insomnia', *Psychosomatic medicine* **62**(2), 227–230.

Hirshkowitz, M., Whiton, K., Albert, S. M., Alessi, C., Bruni, O., DonCarlos, L., Hazen, N., Herman, J., Hillard, P. J. A., Katz, E. S. et al. (2015), 'National sleep foundation's updated sleep duration recommendations', *Sleep health* **1**(4), 233–243.

Hirshkowitz, M., Whiton, K., Albert, S. M., Alessi, C., Bruni, O., DonCarlos, L., Hazen, N., Herman, J., Katz, E. S., Kheirandish-Gozal, L. et al. (2015), 'National sleep foundation's sleep time duration recommendations: methodology and results summary', *Sleep health* **1**(1), 40–43.

Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L. & Muller, P.-A. (2019), 'Deep learning for time series classification: a review', *Data Mining and Knowledge Discovery* **33**(4), 917–963.

Kessler, R. C., Chiu, W. T., Demler, O. & Walters, E. E. (2005), 'Prevalence, severity, and comorbidity of 12-month dsm-iv disorders in the national comorbidity survey replication', *Archives of general psychiatry* **62**(6), 617–627.

Kroenke, K., Spitzer, R. L. & Williams, J. B. (2001), 'The phq-9: validity of a brief depression severity measure', *Journal of general internal medicine* **16**(9), 606–613.

Lane, N. D., Mohammod, M., Lin, M., Yang, X., Lu, H., Ali, S., Doryab, A., Berke, E., Choudhury, T. & Campbell, A. (2011), Bewell: A smartphone application to monitor, model and promote wellbeing, *in* '5th international ICST conference on pervasive computing technologies for healthcare', pp. 23–26.

LiKamWa, R., Liu, Y., Lane, N. D. & Zhong, L. (2013), Moodscope: Building a mood sensor from smartphone usage patterns, *in* 'Proceeding of the 11th annual international conference on Mobile systems, applications, and services', pp. 389–402.

Lu, H., Yang, J., Liu, Z., Lane, N. D., Choudhury, T. & Campbell, A. T. (2010), The jigsaw continuous sensing engine for mobile phone applications, *in* 'Proceedings of the 8th ACM conference on embedded networked sensor systems', pp. 71–84.

Mehrotra, A., Tsapeli, F., Hendley, R. & Musolesi, M. (2017), 'Mytraces: Investigating correlation and causation between users' emotional states and mobile phone interaction', *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* **1**(3), 1–21.

Organization, W. H. et al. (2017), Depression and other common mental disorders: global health estimates, Technical report, World Health Organization.

Panel, C. C., Watson, N. F., Badr, M. S., Belenky, G., Bliwise, D. L., Buxton, O. M., Buysse, D., Dinges, D. F., Gangwisch, J., Grandner, M. A. et al. (2015), 'Recommended amount of sleep for a healthy adult: a joint consensus statement of the american academy of sleep medicine and sleep research society', *Journal of Clinical Sleep Medicine* **11**(6), 591–592.

Saeb, S., Lattie, E. G., Schueller, S. M., Kording, K. P. & Mohr, D. C. (2016), 'The relationship between mobile phone location sensor data and depressive symptom severity', *PeerJ* **4**, e2537.

Saeb, S., Zhang, M., Karr, C. J., Schueller, S. M., Corden, M. E., Kording, K. P. & Mohr, D. C. (2015), 'Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study', *Journal of medical Internet research* **17**(7), e175.

Sano, A., Phillips, A. J., Amy, Z. Y., McHill, A. W., Taylor, S., Jaques, N., Czeisler, C. A., Klerman, E. B. & Picard,

R. W. (2015), Recognizing academic performance, sleep quality, stress level, and mental health using personality traits, wearable sensors and mobile phones, *in* '2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN)', IEEE, pp. 1–6.

Schueller, S. M., Begale, M., Penedo, F. J. & Mohr, D. C. (2014), 'Purple: a modular system for developing and deploying behavioral intervention technologies', *Journal of medical Internet research* **16**(7), e3376.

Vermeulen, A. M. V. P. J., Hendley, R. & Musolesi, M. (n.d.), 'My phone and me: Understanding people's receptivity to mobile notifications'.

Vincent, J. (2006), 'Emotional attachment and mobile phones', *Knowledge, Technology & Policy* **19**(1), 39–44.

Vos, T., Lim, S. S., Abbafati, C., Abbas, K. M., Abbasi, M., Abbasifard, M., Abbasi-Kangevari, M., Abbastabar, H., Abd-Allah, F., Abdelalim, A. et al. (2020), 'Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019', *The Lancet* **396**(10258), 1204–1222.

Wang, P. S., Berglund, P., Olfson, M., Pincus, H. A., Wells, K. B. & Kessler, R. C. (2005), 'Failure and delay in initial treatment contact after first onset of mental disorders in the national comorbidity survey replication', *Archives of general psychiatry* **62**(6), 603–613.

Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., Zhou, X., Ben-Zeev, D. & Campbell, A. T. (2014), Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones, *in* 'Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing', pp. 3–14.

Wang, R., Harari, G., Hao, P., Zhou, X. & Campbell, A. T. (2015), Smartgpa: how smartphones can assess and predict academic performance of college students, *in* 'Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing', pp. 295–306.

Wang, Z., Yan, W. & Oates, T. (2017), Time series classification from scratch with deep neural networks: A strong baseline, *in* '2017 International joint conference on neural networks (IJCNN)', IEEE, pp. 1578–1585.

Wittchen, H.-U., Jacobi, F., Rehm, J., Gustavsson, A., Svensson, M., Jönsson, B., Olesen, J., Allgulander, C., Alonso, J., Faravelli, C. et al. (2011), 'The size and burden of mental disorders and other disorders of the brain in europe 2010', *European neuropsychopharmacology* **21**(9), 655–679.

APPENDIX

As mentioned in the Results and Discussion section, in order to avoid repetitiveness, analogous plots to the ones analysed during the stress prediction section are shown and explained below.

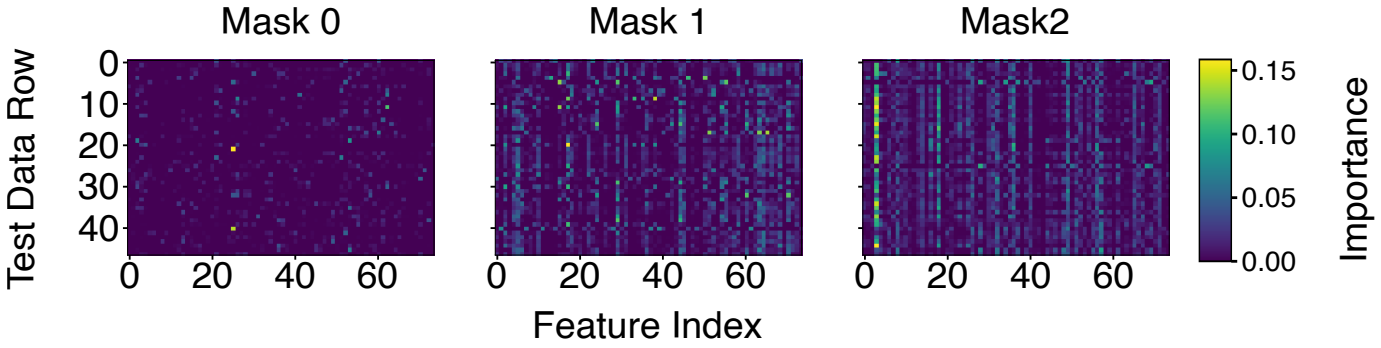

Fig. 6. TabNet Feature Importance Masks for the task of stress signs predictions when using dataset B.

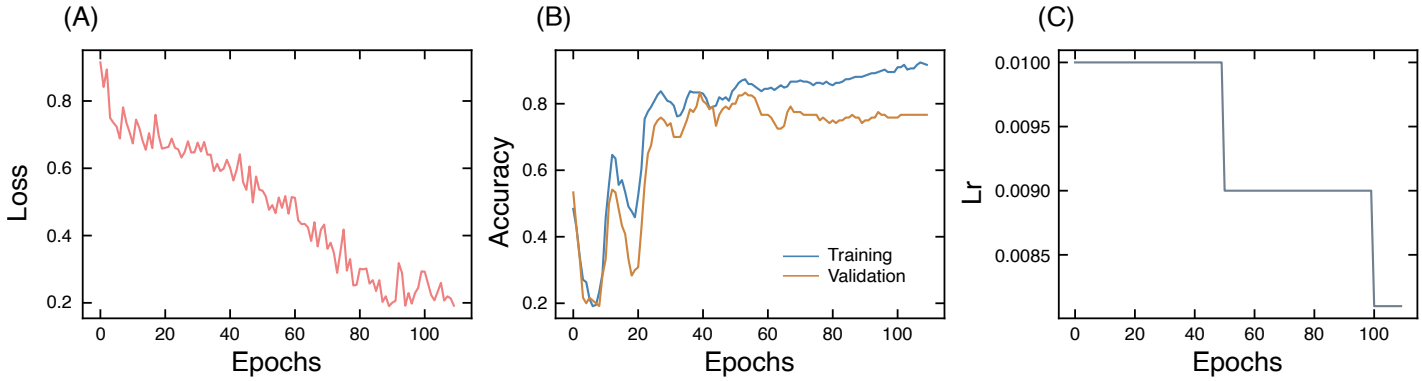

Fig. 7. **Standard TabNet training plots for stress signs prediction, when using dataset A.** (A) Training Loss through the training epochs shows an adequate loss decay, indicating a correct choice of learning rate. (B) Accuracy plot that shows that optimal performance is achieved around the 40th epoch, afterwards the validation accuracy decays, whereas the training accuracy increases indicating overfitting. (C) Learning rate decay when using the SetpLR scheduler.

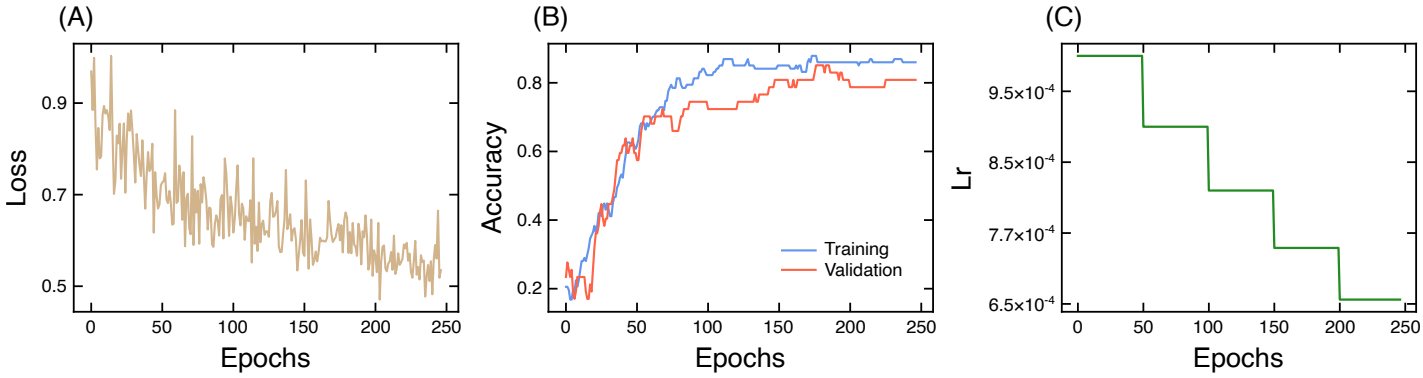

Fig. 8. **Standard TabNet training plots for stress signs prediction, when using dataset B.** (A) Training Loss through the training epochs shows an adequate loss decay, indicating a correct choice of learning rate. (B) Accuracy plot that shows no signs of overfitting since both the training and validation accuracy remain fairly constant after the 200th epoch. (C) Learning rate decay when using the SetpLR scheduler.

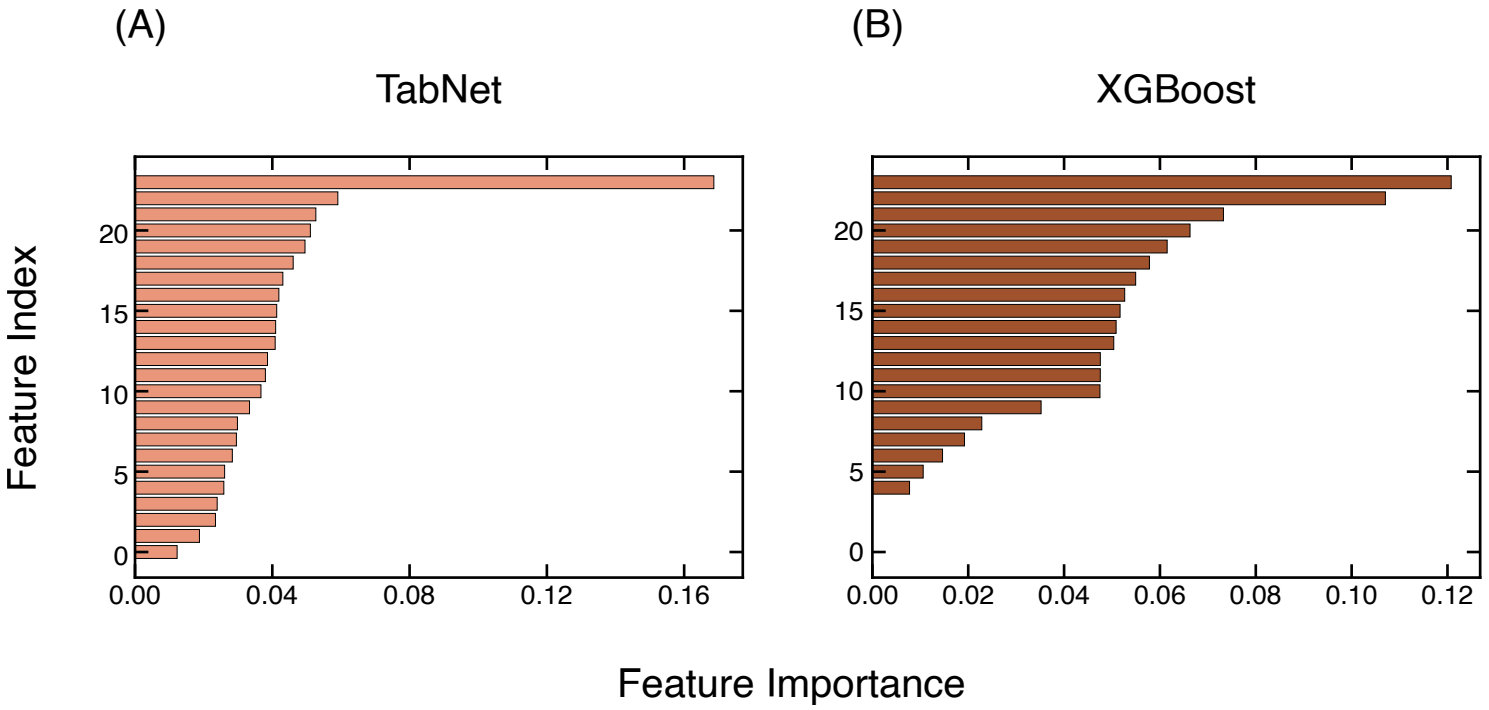

Fig. 9. **Feature importance for sleep prediction when training with dataset B.** The histograms show identical patterns to Figure 4, where the TabNet distributes weights more equally than the XGBoost. (Note that feature importance comparison is not possible for dataset A, since the best XGBoost model used a linear booster)

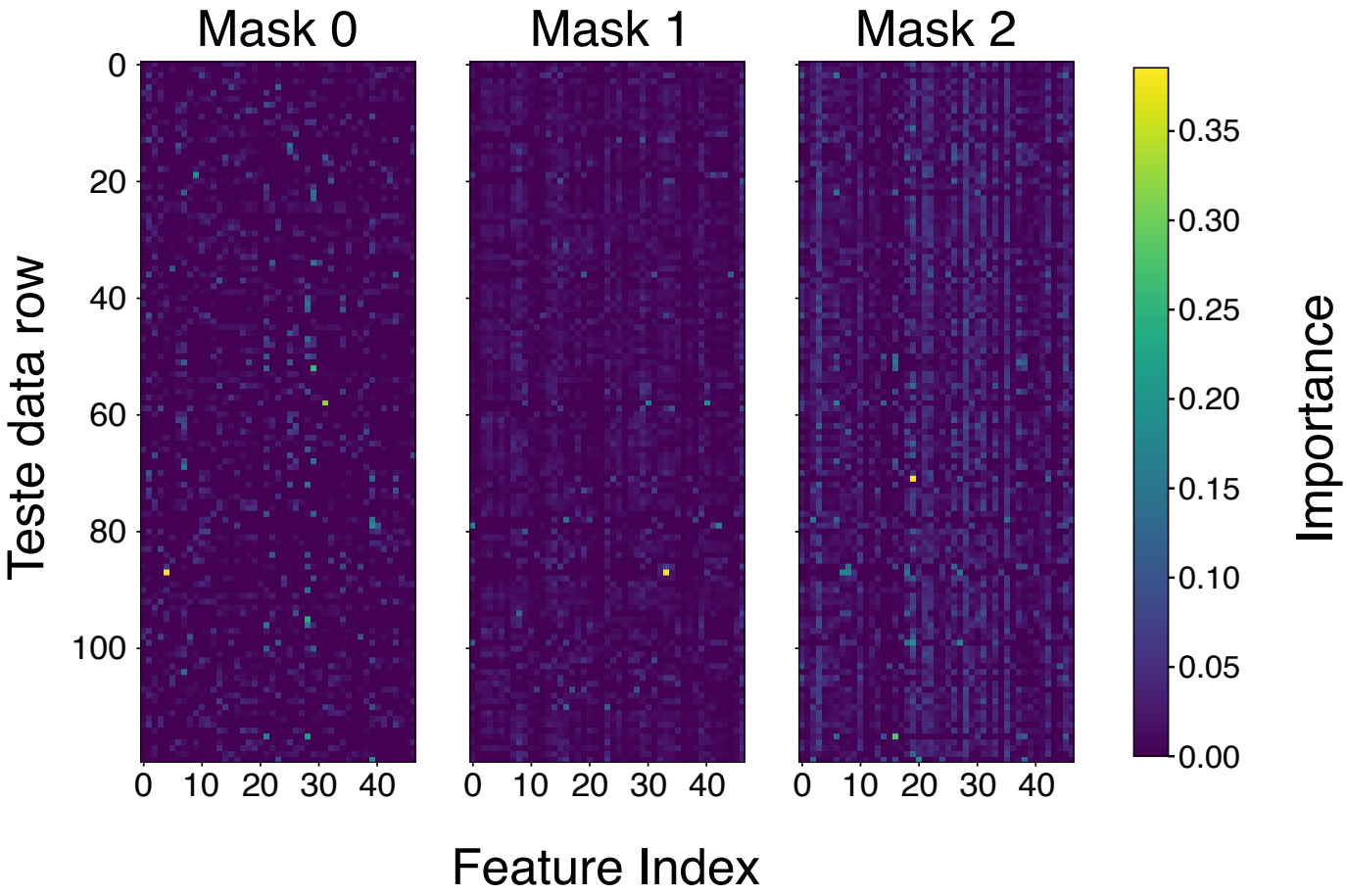

Fig. 10. TabNet Feature Importance Masks for the task of sleep predictions when using dataset A.

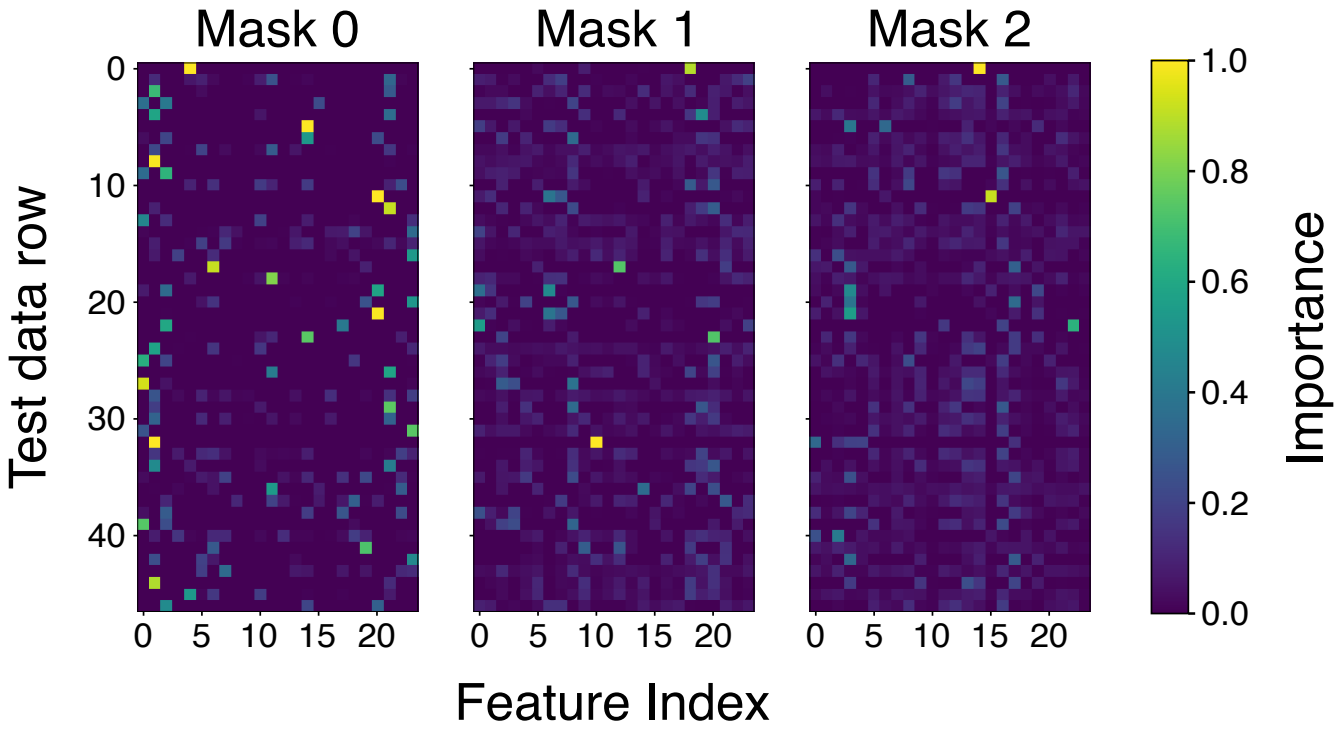

Fig. 11. TabNet Feature Importance Masks for the task of sleep predictions when using dataset B.

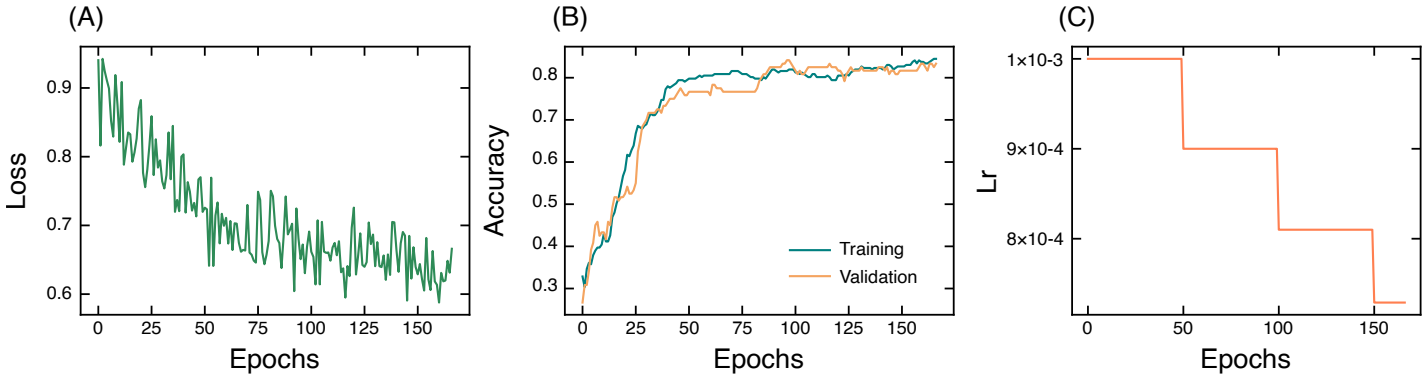

Fig. 12. **Standard TabNet training plots for sleep prediction, when using dataset A.** (A) Training Loss through the training epochs shows an adequate loss decay, indicating a correct choice of learning rate. (B) Accuracy plot that shows no signs of overfitting since both the training and validation accuracy remain fairly constant after the 100th epoch. (C) Learning rate decay when using the SetpLR scheduler.

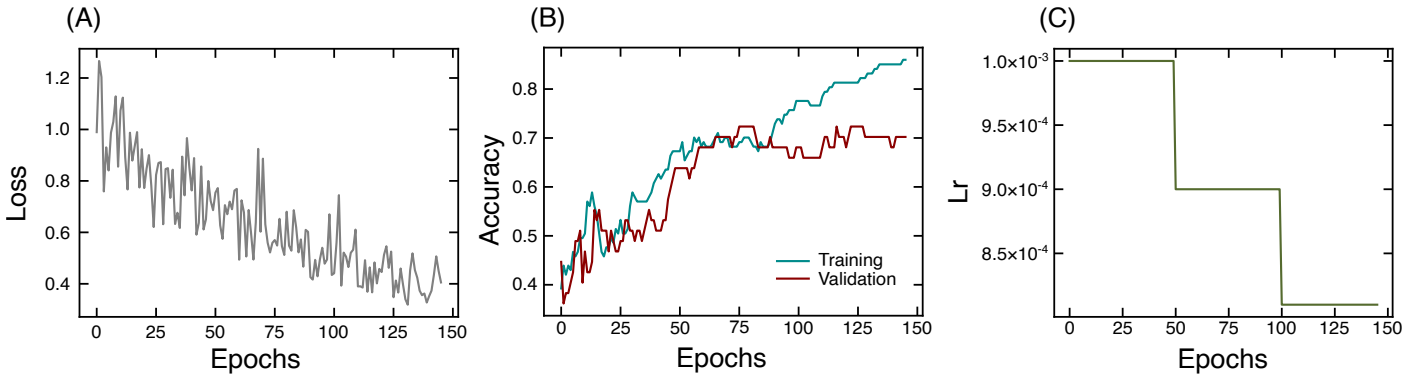

Fig. 13. **Standard TabNet training plots for sleep prediction, when using dataset B.** (A) Training Loss through the training epochs shows an adequate loss decay, indicating a correct choice of learning rate. (B) Accuracy plot that shows that the network requires training until the 75th epoch, however showing signs of overfitting in subsequent epochs as the validation accuracy will start decreasing while the training accuracy will continue on increasing. (C) Learning rate decay when using the SetpLR scheduler.