Unimodal vs. Multimodal Prediction of Antenatal Depression from Smartphone-based Survey Data in a Longitudinal Study

Mengyu Zhong
Department of Information Technology and Women's Mental Health during the Reproductive Lifespan – Womher, Uppsala University
Uppsala, Sweden
mengyu.zhong@it.uu.se

Vera Van Zoest
Department of Information Technology, Uppsala University
Uppsala, Sweden
vera.van.zoest@it.uu.se

Ayesha Mae Bilal
Department of Medical Sciences, Psychiatry and Women's Mental Health during the Reproductive Lifespan – Womher, Uppsala University
Uppsala, Sweden
ayesha.bilal@neuro.uu.se

Fotios C Papadopoulos
Department of Medical Sciences, Psychiatry, Uppsala University
Uppsala, Sweden
fotis.papadopoulos@neuro.uu.se

Ginevra Castellano
Department of Information Technology, Uppsala University
Uppsala, Sweden
ginevra.castellano@it.uu.se

ABSTRACT

Antenatal depression impacts 7-20% of women globally, and can have serious consequences for both the mother and the infant. Preventative interventions are effective, but are cost-efficient only among those at high risk. As such, being able to predict and identify those at risk is invaluable for reducing the burden of care and adverse consequences, as well as improving treatment outcomes. While several approaches have been proposed in the literature for the automatic prediction of depressive states, there is a scarcity of research on automatic prediction of perinatal depression. Moreover, while there exist some works on the automatic prediction of *postpartum depression* using data collected in clinical settings and applied the model to a smartphone application, to the best of our knowledge, no previous work has investigated the automatic prediction of late *antenatal depression* using data collected via a smartphone app in the first and second trimesters of pregnancy. This study utilizes data measuring various aspects of self-reported psychological, physiological and behavioral information, collected from 915 women in the first and second trimester of pregnancy using a smartphone app designed for perinatal depression. By applying machine learning algorithms on these data, this paper explores the possibility of automatic early detection of antenatal depression (i.e., during week 36 to week 42 of pregnancy) in everyday life without the administration of healthcare professionals. We compare uni-modal and multi-modal models and identify predictive markers related to antenatal depression. With multi-modal approach the model reaches a BAC of 0.75, and an AUC of 0.82.

ICMI '22, November 7–11, 2022, Bengaluru, India

ACM ISBN 978-1-4503-9390-4/22/11.
https://doi.org/10.1145/3536221.3556605

CCS CONCEPTS

• **Human-centered computing** → *Ubiquitous and mobile computing*; • **Computing methodologies** → *Machine learning*; • **Applied computing** → *Life and medical science.*

KEYWORDS

Antenatal depression; multimodal markers; mobile surveys; longitudinal study; machine learning

ACM Reference Format:
Mengyu Zhong, Vera Van Zoest, Ayesha Mae Bilal, Fotios C Papadopoulos, and Ginevra Castellano. 2022. Unimodal vs. Multimodal Prediction of Antenatal Depression from Smartphone-based Survey Data in a Longitudinal Study. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '22), November 7–11, 2022, Bengaluru, India.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3536221.3556605

1 INTRODUCTION

Depression in the perinatal period can occur during pregnancy (antenatal depression) or within 12 months after birth (postpartum depression) [92]. Antenatal depression impacts 7-20% of women globally [11], and can have serious consequences for both the mother and the infant. It is associated with complications during pregnancy and birth, such as preterm birth, preeclampsia, and low birth weight [1, 34], as well as a poor perinatal quality of life, sexual dysfunction, and difficulties in relationship with partner [77]. Furthermore, antenatal depression is a key risk factor for postpartum depression [49], which extends the consequences to the child's social, emotional and cognitive development [32, 77]. Moreover, it is a leading cause of maternal mortality by suicide, both during pregnancy and in the postpartum period [63].

Despite the profound consequences and public health burden associated with antenatal depression, research has historically focused largely on postpartum depression [11]. Previous studies have identified biological, psychological and sociodemographic correlates for antenatal depression [48, 58], however, early detection of those at risk, particularly during pregnancy, continues to be a

challenge, with approximately 30-70% of cases being undetected and only 15% receiving adequate treatment [14, 23]. An important reason for this is that universal screening is carried out in the postpartum period [10]. As such, detection in the antenatal period largely relies on women reporting symptoms; however, women often hesitate doing so because of stigma or shame arising from their perinatal experience not meeting their expectations of how they perceive pregnancy and motherhood *should* be. Furthermore, it can be difficult to separate symptoms of depression from common emotional and somatic changes experienced in the perinatal period [11].

Preventative and early interventions for antenatal depression are effective [88], but are cost-efficient only among those at high risk. As such, being able to predict and detect those at risk is invaluable for reducing the burden of care and adverse consequences, as well as improving treatment outcomes [64]. However, the development of mental diseases can be subtle. It is a continuous but fluctuating progress. In traditional screening, women only have brief contact with clinicians in a controlled healthcare setting with limited time. Clinicians will not have access to long-term data collected outside of the clinics, and will often rely on women's retrospective reporting to assess the risk. This process is susceptible to recall biases, which can affect the integrity of information reported by women and the subsequent likelihood of missed cases [76].

At the same time, machine learning-based methods have proven to be successful in predicting depression using a variety of features, from audio-visual behavioural data [44, 73] to mobility patterns captured by GPS in smartphones [55]. However, previous work on the automatic prediction of perinatal depression is limited [19]. Moreover, while there exists some work on the automatic prediction of *postpartum* depression using data collected by interviews in clinical settings and applied on smartphone application [42], to the best of our knowledge, no previous work investigated the automatic prediction of late antenatal depression using data collected via a smartphone app in the first and second trimesters of pregnancy.

We utilized self-reported data, measuring psychological, physiological and behavioral factors, collected in the first and second trimesters of pregnancy using a smartphone app designed for perinatal depression research Mom2b [12]. By applying machine learning algorithms on these data, this paper explores the possibility of automatic early detection of antenatal depression (during week 36 to week 42 of pregnancy) in everyday life without the administration of healthcare professionals. Furthermore, we identified predictive markers of antenatal depression. We used data collected from 1505 (915 after selection) women in the first and second trimesters of pregnancy as part of a longitudinal study to compare the application of five machine learning algorithms: Support Vector Machine (SVM), Logistic Regression (LR), K-Nearest Neighbour (KNN), eXtreme Gradient Boosting (XGBoost), and Multi-layer Perception (MLP). For each of these machine learning algorithms, we compared the performance of single modality models to multi-modal approaches. We also investigated which features were most important to predict antenatal depression. The results of this work provide insights on the feasibility of different modelling approaches, as well as on which types of data are important to collect at an early stage during the pregnancy.

2 RELATED WORK

2.1 Smartphone Applications in Mental Healthcare

The emergence of ubiquitous smart devices provides potential new methods in healthcare. Mobile health (mHealth) apps based on mobile smart devices, such as smartphones, tablets or wearables, can be beneficial for symptom monitoring, community support, screening and assessment, education, and management [5, 61]. The availability of built-in sensors facilitates the collection of large amounts of data, and the prevalence of smartphones enables access to large populations of users. The advantage of such technologies in monitoring has been shown for data collection and transmission, but there are still critical issues on data processing, in aspects of efficacy, reliability, data security and privacy [5, 84].

Unsurprisingly, mHealth technology has also received much attention in mental healthcare. By employing multiple tools, e.g. surveys, sensors, and voice recordings, clinicians can collect large amounts of multi-modal, real-time data regarding a patient's behavioral, psychological, and physiological patterns [85]. The access to such data provides an opportunity to address a longstanding difficulty in psychiatry, i.e., quantifying a disease's phenotype accurately and reliably. In 2016, the term "digital phenotyping" has been introduced as the "moment-by-moment quantification of the individual-level human phenotype in-situ using data from smartphones and other personal digital devices" [83, p.1]. The authors also offered an open-source research application, Beiwe, for multimodal data collection purposes, which largely accelerated the development of the field [83].

Although the number of studies in the area is growing, we see a discrepancy between research and actual smartphone apps available in the market. There is still an urgent need of clinical investigation on mental health apps. A recent review reported that from 2011 to 2020, there were seventeen articles identified as qualified studies on smartphone apps targeting mental health [20]. Among these apps, only seven targeted depression or bipolar disorder screening, monitoring or intervention [16, 18, 36, 37, 43, 52, 59], and only one for perinatal depression [36]. Considering the paucity of studies in the field, the reality is troublesome: although the use of smartphone apps in mental health is still an on-going research topic, more than 10,000 mental health-related apps are available to download in commercial marketplaces, but most do not conform to clinical guidelines, are not supported by research data, and are not evidence-based [8, 84, 86]. In another survey conducted in 2019 on existing depression support apps, 553 apps in app stores were identified out of which 216 passed the selection criteria. Among these apps, 52 were assessment apps providing digital versions of screening questionnaires, 50 were mood trackers empowered by machine learning, and 3 apps were designed specifically for perinatal depression [81]. The efficacy of such AI-based applications and type of data to be collected for monitoring or prediction purposes still need further investigation, especially for the prediction of perinatal depression.

2.2 Automatic Prediction of Depression and Perinatal Depression

An increasingly high number of studies have investigated the automatic prediction of depression and depression-related states using machine learning-based methods. Different from traditional paper-based self-reported questionnaire assessments, data-driven approaches provide new opportunities to refine patient screening and may introduce new paradigms for mental healthcare by providing novel ways to identify individuals at risk of developing depression [19].

The data used in the models available in the literature range from bio-markers [28], electronic health records (EHRs) [60], and magnetic resonance imaging (MRI), to audio-visual recordings [68], social media data [51], and smartphone-based digital phenotyping [67]. The latter may include passive data (e.g., sensor data, phone usage patterns) [9] and/or active data (e.g., self-reported medical history, sociodemographic data) [94].

In the multimodal interaction and ubiquitous computing communities, previous work has shown that it is possible to automatically predict clinical depression from audio-visual recording and smartphone data. As far as audio-visual data are concerned, previous work has proposed computational methods to predict depression using non-verbal (i.e., facial, eye and body features) [26, 44, 79] and verbal behaviour (i.e., acoustic features) [73, 78]. When it comes to the automatic prediction of depression using smartphones data, previous work has shown that machine learning methods applied to smartphone data can be used to monitor subjects affected by depressive states and automatically predict the latter by analyzing their mobility patterns from GPS traces [17, 35, 55] and other smartphone interaction features [54].

Contrary to the large variety of methods proposed for the automatic detection of depression, few studies report on machine learning approaches used for the purpose of automatically predicting perinatal depression. A recent review surveyed 11 studies on postpartum depression prediction [19] and found that sociodemographic and clinical variables (i.e., psychiatric and gynecological factors) seem to be the most reliable. Other studies investigated biological variables (i.e., blood, genetic and epigenetic features), while one study included in the survey discussed the use of smartphone apps without actually collecting smartphone data [42]. Another review on mHealth for perinatal depression and anxiety found 22 unique studies including protocols [41], where only four studies for screening or treatment included the antenatal period, and none of those four used machine learning approaches.

This work builds on findings from Andersson et al.'s study, which applied machine learning methods to clinical, demographic, and psychometric data collected via an online survey four times during pregnancy and after birth for the automatic prediction of *postpartum depression*. The authors found that antenatal depression and anxiety, as well as factors associated with resilience and personality, were important risk factors for postpartum depression. In this paper we apply machine learning methods to clinical, demographic, and psychometric data collected via a smartphone app for the purpose of automatically predicting *antenatal depression* during week 36 to week 42 of pregnancy. Compared to Andersson et al.'s use of online surveys, the use of a smartphone app allowed us to collect data more frequently during pregnancy[4].

3 METHODS AND MATERIALS

3.1 Dataset

Data for the development of prediction models in this study were obtained from the Mom2b cohort study. Mom2b is an ongoing population-based prospective cohort study based in Sweden. Data are collected through the Mom2b smartphone app that was launched in 2019 to App Store and Google Play[12]. All Swedish-speaking women who are pregnant or within three months postpartum, above the age of 18, and own a smartphone, are eligible to participate by downloading the Mom2b app, where they can register to the study and provide consent. 1505 women, who are part of the existing Mom2b cohort and had completed, at minimum, the Edinburgh Postnatal Depression Scale (EPDS) [24] in between week 36 to week 42 of pregnancy were considered for analysis in this study. Data with greater than 80% missing value rate were removed, and subsequently, a total of 915 participants were included in our analysis. Details regarding how missing values were handled are described in Section 3.2.2.

Survey data collected from the Mom2b app were used to develop our models. A diverse combination of validated and self-developed instruments are delivered within the Mom2b app at baseline (whenever the participant joins the study), and periodically throughout the pregnancy and postpartum period. Surveys remain available on the app for specified time windows, and disappear once completed. Figure 1 gives an overview of the surveys included in our analysis, as well as a timeline for when, how long, and how often each survey is available for participants to complete. Surveys were grouped into 6 categories: sociodemographic information, psychological health, general health, behavioral, social, and personality. Surveys completed by participants in the first and second pregnancy trimesters (up till week 28) were included in our analysis. Additionally, 16 surveys that we considered time-independent (such as those that collected information on sociodemographics, preconception health history, or personality traits) were also included from the period after week 28.

The outcome, risk of depression, was assessed using the EPDS, available for completion from pregnancy week 36 to week 42 of pregnancy. The EPDS is 10-item self-report instrument that has been validated as a tool for screening for depression in the perinatal period. We considered a cut-off score of 12 as indicative of depression, based on the Swedish validation study [91].

3.2 Preprocessing

To handle the diverse data types in our dataset, we apply several preprocessing procedures, including encoding, standardization, and imputation according to their unique characteristics.

3.2.1 *Encoding and Standardizing.*

- **Categorical variables**: variables in likert scales, single choices and multiple choices are considered as categorical variables. For most of the categorical data, we keep the numerical labels. Only for multiple choice variables, we apply one-hot encoding for each option.

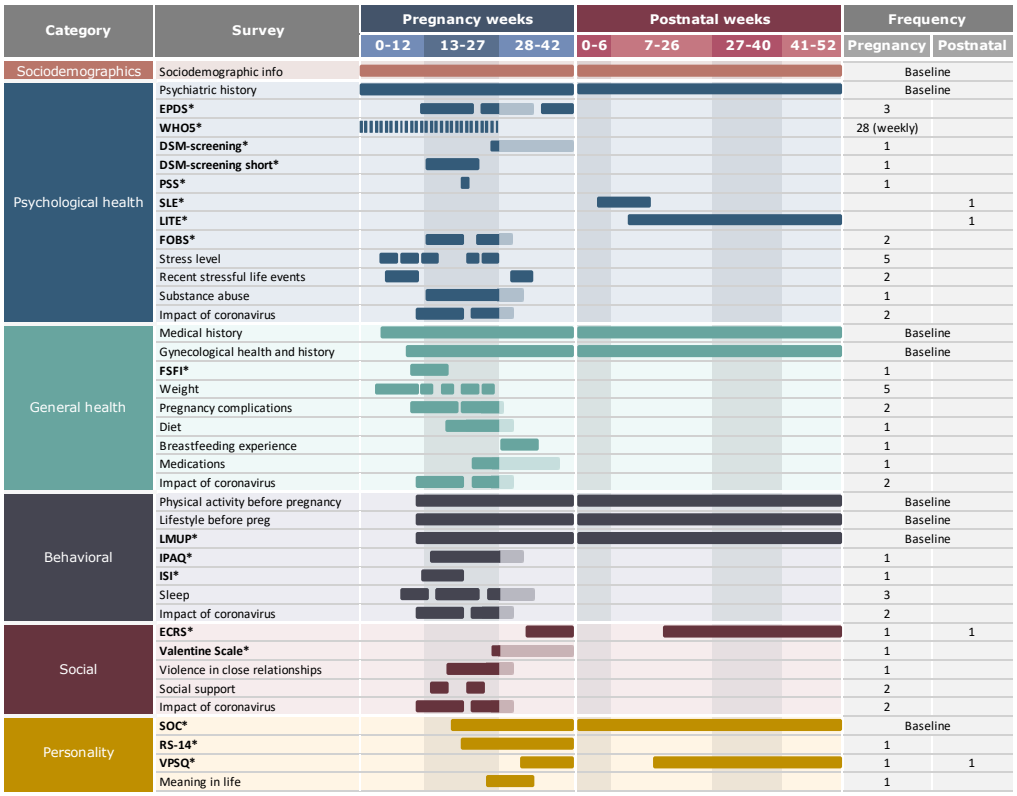


Figure 1: Timeline, description and frequency of surveys included in the analysis. *Validated instruments. EPDS, Edinburgh Postnatal Depression Scale; WHO5, WHO-5 Well-Being Index [82]; DSM-screening, Diagnostic and Statistical Manual of Mental Disorders, 5th Edition, criterion for depression; DSM-screening short (shortened version of the DSM-screening); PSS, Perceived Stress Scale [29]; SLE, Stressful Life Events [70]; LITE, Lifetime Influence of Traumatic Experiences [33]; FOBS, Fear of Birth Scale [39]; FSFI, Female Sexual Function Index [69]; LMUP, London Measure of Unplanned Pregnancy [6]; IPAQ, International Physical Activity Questionnaire [25]; ISI, Insomnia Severity Index [7]; ECRS, Experience in Close Relationships Scale [15]; Valentine Scale (relationship with your partner) [3]; SOC, Sense of Coherence [30]; RS-14, Resilience Scale [90]; VPSQ, Vulnerability Personality Style Questionnaire [27].

- **Continuous variables**: some questions are in continuous scale, such as weight and height. We applied a standard scale in which the score of a sample x is calculated as: $z = (x-u)/s$, where u is the mean of the training samples, and s is the standard deviation of the training samples.
- **Text-based variables**: a few free response questions exist in our data. We dropped the free response questions that generally have long answers, and label encoded the short text variables answering questions like, *What other medicines do you use? Which region / county do you live in?*

3.2.2 Imputing Missing Values. Missing rates of our dataset are generally high, which is another common challenge in mHealth and especially when data are collected longitudinally [13, 31]. Our dataset likely contains data missing at random (MAR) and missing not at random (MNAR) [72]:

- **Missing at random**: the missing values before participants join the study is the majority missing. With our inclusion criteria, women can join the study during pregnancy or within three months postpartum. Surveys that expired before they joined the study would never be delivered to them. Meanwhile, since the app is deployed in the wild to collect a massive amount of features in a long-term and relative loose setting, it is easy for participants to miss some surveys due to forgotten or other unseen factors.
- **Missing not at random**: self-administrated surveys delivered by the smartphone app give participants great freedom to choose to stop answering the surveys anytime. Missing data can sometimes be indicative of participants' mental health status. That is, it is tenable that participants may stop answering surveys due to symptoms of depressed mood. The occurrence of missing values could, therefore, be an interesting predictor in itself when operationalized as a variables [50]. Among included participants of the study, 17.5% were depressed, while in excluded participants, 26.5% were depressed. The average EPDS score of included participants is 6.58 and of excluded participants is 8.22. Participants who answered less than 20% of mobile surveys during the first two trimesters had significantly higher EPDS scores ($p < 0.001$ in t-test). However, we didn't do further analysis on

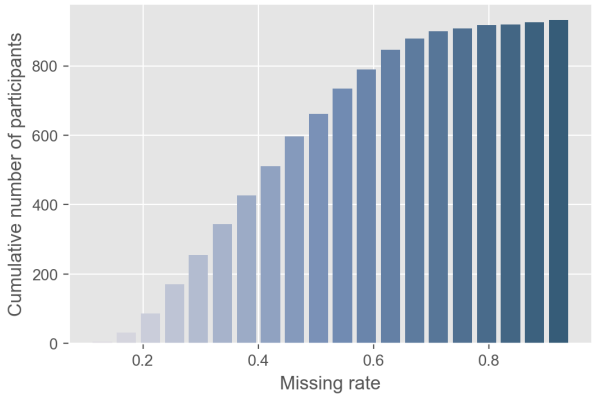


Figure 2: Cumulative number of participants when apply different inclusion thresholds of missing rate

the data with high missing rate, as it is out of the scope of this current paper.

Complete case analysis is widely used to combat missing value issues, which however may introduce bias and omit some useful information [97]. Nevertheless, in this particular study with no rigidly controlled experimental setting, complete case analysis is not suitable because of the high missing rate of data. As shown in Figure 2, we barely have any complete case in the dataset. Therefore, we can either use algorithms which can handle missing values, such as decision tree ensembles [87], or use imputation to complete the dataset. In order to be able to compare different algorithms, we have imputed missing values. We included *most frequent imputation* for categorical variables; *mean imputation* and *most frequent imputation* for continuous variables in grid search.

3.3 Machine Learning Models

We consider the task as a binary classification task for the automatic prediction of depression, and compared five classifiers, namely Logistic Regression (LR), Support Vector Machine Classifier (SVC), K-Nearest Neighbours (KNN), eXtreme Gradient Boosting Classifier (XGBoost), and Multi-layer Perceptron (MLP) in both uni-modal and multi-modal experiments. The efficacy of these algorithms are proven in previous studies on depression prediction [19, 74]. All the models are trained after the same preprocessing procedures mentioned in 3.2, where the data go through encoding, scaling, and imputation.

K-Nearest Neighbours, as a non-parametric supervised learning, is used as baseline of the study. K-NN will take k nearest neighbors from the known dataset to predict the unknown label of a new case, thus it is not sensitive to missing values. With grid search, we tested a range of $k = 3, 5, 7, ...13$.

Support Vector Machine projects the data into a higher-dimensional space by a kernel function and use a separating hyperplane to separate two groups of samples [62]. Different kernels were tested, namely linear, poly, sigmoid, and rbf. We also adjusted margin of the separating hyperplane and class weight to maximize the performance of the model by tuning the parameters C and $class\_weight$.

Logistic Regression is commonly used for a binary outcome. There are a wealth of studies using LR for clinical prediction models. In the area of clinical predictive models, Christodoulou et al. have done a systemic review on the applications of LR and other machine learning algorithms and there was no significant performance difference shown. We applied a multivariate logistic regression model with the implementation provided in Scikit-learn. Optimization algorithms, *solver*, is searched in (*'sag', 'saga', 'liblinear'*), *L1*, *L2* and *elastic net* regularization were also tested when applicable. Different $class\_weight$ and C were also tested for better performance.

Gradient boosting trees is essentially an ensemble of weak prediction models. As a popular algorithm among the decision tree ensembles, the use of XGBoost (eXtreme Gradient Boosting [21]) in depression prediction was explored by [74]. Like previous algorithms, to reach better performance, we tuned some parameters including *booster, learning_rate, max_depth, subsample, reg_lambda, n_estimators, scale_pos_weight, min_child_weight.* When using DART [46] algorithm as *booster*, we also tuned *rate_drop* and *skip_rate.*

Multi-layer Perceptron is also known as a type of artificial neural network. It has at least one hidden layer between input and output layers. We applied the MLP classifier with stochastic gradient descent and Adam optimization algorithm proposed by [45]. The size of the hidden layers is determined following [75, 80], with size $N_i - 1$, $\sqrt{(N_i \times N_i)}$ for three-layer models, and $\frac{N_i}{2} + 3$ for a four-layer model, where N_i is the number of input neurons and N_o is the number of output neurons. Two activation functions, tanh and relu, are examined. Learning rates were set to be constant and adaptive separately. As we have a relative big number of features considering the number of samples, over-fitting could very likely be an issue. Beside a value of 0.0001, we tried a bigger alpha for the *L2* penalty, 0.05.

3.4 Class Imbalance

Clinical predictive models are often dealing with imbalanced classification tasks. In practice, the number of positive cases is generally less than that of negative cases, which can bias the performance of the models towards the negative prediction. Although in our dataset we have 17% positive cases (21% before selection), which is higher than normal report of perinatal depression ratio, it is still problematic. We adjusted class weight parameters for our models to reduce the influence of class imbalance, except for MLP as it is not yet supported by Scikit-learn.

3.5 Uni-modal and Multimodal Experiments

We conducted uni-modal and multimodal experiments. In both sets of experiments, the dataset is randomly divided into a *train set* and a *test set* with a ratio of 75%:25%. The same random seed is used throughout the experiments. Positive cases take 17.28% and 17.98% in *trian set* and *test set* respectively. The models are first developed and selected within *train set* using grid search with stratified 5-fold cross validation, then the selected models are retrained with the whole *train set* and tested with *test set* for a final unbiased evaluation. During the design of the methodology, we use the PROBAST

(Prediction model Risk Of Bias Assessment Tool) to avoid bias of the developed prediction model [93].

In the uni-modal experiments, we used the five classifiers described above to train models for each individual modality, i.e. *Psychological Health, General Health, Behavioral, Social, and Personality.*

In multimodal experiments, considering the distinctive content of the five input streams of our dataset, we applied different fusion operations to combine the information on both feature-level and decision-level for comparison.

Feature-level fusion: after preprocessing procedures, we employed the concatenation to combine the five modalities of data to input vectors, as concatenation has been used to combine low-level input features in many studies [96]. Then we passed the fused input in grid search with cross validation to find the better model.

Decision-level fusion: for decision fusion we trained uni-modal models separately for the five modalities, i.e. *Psychological Health, General Health, Behavioral, Social, and Personality. Sociodemographics* are added into all modalities to account for potential confounding. Next, we applied prevailing decision level fusions on the predictions of the models, including fixed fusion rules [47] and trainable fusion rules [56, 66], i.e., max fusion, average fusion, and median fusion (the same as majority vote in our case), and fusion based on supervised learning. The uni-modal models were selected using grid search with the same set of parameters as feature level fusion.

Similar to Kuncheva's definition [47], let $C = \{C_1, \ldots, C_L\}$ be the a set of uni-modal classifiers. With the fusion algorithms, we aim to a better performance than classifiers C. As our classifiers can produce soft class labels, we assume that $d_{j,i}(\mathbf{x}) \in [0, 1]$ is an estimate of the probability $P(\omega_i \mid \mathbf{x})$ offered by classifier C_j for an input $\mathbf{x} \in \Re^n, i = 1, 2,\ j = 1, \ldots, L$. There are two possible classes $\Omega = \{\omega_1, \omega_2\}$, and for any $\mathbf{x}$, $d_{j,1}(\mathbf{x}) + d_{j,2}(\mathbf{x}) = 1, j = 1, \ldots, L$. For each input point $\mathbf{x}$, label ω_1 will be assigned if $P(\omega_1 \mid \mathbf{x}) = p > 0.5$.

Where for the *Average Fusion, Maximum Fusion and Median Fusion*:

$$\hat{P}(\omega_i \mid \mathbf{x}) = \mathcal{F}(P(\omega_i \mid \mathbf{x})_1, \ldots, P(\omega_i \mid \mathbf{x})_L) \tag{1}$$

Where $\mathcal{F}$ stands for fixed fusion rules used, i.e. average, maximum, and median.

While for the *Supervised Learning Based Fusion:*

$$\hat{P}(\omega_i \mid \mathbf{x}) = \mathcal{S}(P(\omega_i \mid \mathbf{x})_1, \ldots, P(\omega_i \mid \mathbf{x})_L) \tag{2}$$

Where $\mathcal{S}$ represents trainable fusion algorithms. In this work, we used several supervised learning algorithms, i.e., Gaussian Naive Bayes (GaussianNB), Logistic Regression, and Multi-layer Perceptron.

3.6 Evaluation metrics

Considering the class imbalance of our case, we take *balanced accuracy (BAC) = (sensitivity + specificity) / 2* as our main evaluation metric. Area under the curve (AUC), weighted-averaged F1 score, sensitivity (SENS), specificity (SPEC), accuracy (ACC), negative predictive values (NPV), and positive predictive value (PPV) are reported as the supplementary matrix.

4 RESULTS

We investigated the performance of machine learning models on unimodal and multimodal data. For uni-modal models, we used feature importance ranking provided by XGBoost to further identify the predictive items.

4.1 Uni-modal Models

Table 1 shows the performance of selected classifiers on the test set, which reached the highest mean BAC during the cross-validation process. In General, XGBoost has the best performance on unimodal data, while SVM and LR also show some predictive ability, reaching a mean BAC of 0.70 and mean AUC of 0.75 and 0.77, respectively. KNN is not suitable for the task, getting a mean BAC a little greater than 0.5.

Psychological health, which is proven to be an important predictor in previous works [4, 48], outperformed other kinds of data, reaching the highest mean BAC (at 0.74) and mean AUC (at 0.81) on the validation set during the cross-validation and highest AUC (at 0.81) and F1 score (0.79) on the test set.

Fig. 3 shows the feature importance of XGBoost in five modalities. Although sociodemographic information is added to all modalities as a confounder, only age, weight, and height appear at the top of the ranks in most uni-modal models.

In the most predictive modality, psychological health, previous EPDS scores, WHO5 scores, stress, anxiety, and fear of birth (as measured by FOBS) seem to be the most important.

Personality is also a relatively strong predictor. In the feature importance rank, we see the RS14 total score and certain questions, the SOC scores and certain questions, and one 'meaning of life' question: *'to what extent do you have inner peace?'* rank in the top fifteen. Age, weight before pregnancy, height, children in the household, and place of residence also seem to be informative in this aspect.

In general health features, we find that time of trying to get pregnancy, weights during pregnancy, diet (fiber-rich foods, seafood, and sugary drinks), and previous pregnancy times can be strong predictors. Particularly, as the study started in 2019, right before the COVID-19 pandemic, we see an influence of COVID-19 related factors, e.g. having symptoms similar to description of COVID-19. Place of residence, age, weight before pregnancy, and height are also used by the model to make decisions.

Among behavioral features, sleep (i.e., feeling rested and hours asleep), and physical activity (IPAQ) before pregnancy and during pregnancy are the main predictors. At the same time, the impact of the COVID-19 pandemic is not obviously predictive in the behavioral aspects. Many sociodemographic details, namely age, height, weight and place of residence are also important bricks constructing the model.

Although social factors are less predictive than others, we still see a feature importance ranking compliant with the clinical studies [48, 57]. The sociodemographic features, i.e., age, weight, height, place of residence, and education, and support from relatives and partner rank top on the list. Maternity care and the social influence of the pandemic also contribute to the prediction. Violence in close relationships, Experience in Close Relationships Scale (ECSR), and

Figure 3: Top 15 feature importance rank of uni-modal models trained with (a) behavioral data; (b) general health data; (c) personality data; (d) psychological health data; (e) social-related data

Table 1: Performance of uni-modal models

Modal	Classifier	Cross Validation		Test Set							
		BAC	AUC	BAC	AUC	F1	ACC	SENS	SPEC	PPV	NPV
Behavioral	KNN	0.51	0.56	0.55	0.57	0.74	0.78	0.17	0.93	0.40	0.81
Behavioral	LR	0.63	0.66	0.58	0.67	0.66	0.63	0.51	0.66	0.28	0.84
Behavioral	MLP	0.57	0.59	0.56	0.62	0.73	0.76	0.21	0.90	0.36	0.81
Behavioral	SVM	0.62	0.65	0.60	0.62	0.56	0.52	0.72	0.47	0.26	0.87
Behavioral	**XGB***	**0.66**	0.66	0.61	0.65	0.62	0.58	0.66	0.56	0.28	0.86
General health	KNN	0.52	0.55	0.50	0.62	0.70	0.78	0.02	0.98	0.25	0.79
General health	LR	0.67	0.69	0.61	0.65	0.65	0.61	0.62	0.61	0.29	0.86
General health	MLP	0.59	0.66	0.52	0.59	0.68	0.67	0.28	0.77	0.24	0.80
General health	SVM	0.66	0.71	0.62	0.61	0.63	0.66	0.59	0.57	0.29	0.87
General health	**XGB***	**0.68**	0.69	0.61	0.64	0.66	0.57	0.63	0.64	0.30	0.85
Personality	KNN	0.52	0.61	0.55	0.58	0.74	0.79	0.13	0.97	0.50	0.81
Personality	LR	0.69	0.74	**0.76**	0.77	0.78	0.77	**0.74**	0.77	0.46	**0.92**
Personality	MLP	0.61	0.70	0.58	0.73	0.76	0.80	0.21	0.95	0.53	0.82
Personality	SVM	0.68	0.74	0.69	0.76	0.74	0.72	0.64	0.74	0.39	0.89
Personality	**XGB***	**0.69**	0.73	0.74	0.76	0.75	0.73	**0.74**	0.73	0.42	**0.92**
Psychological health	KNN	0.57	0.64	0.55	0.64	0.74	0.79	0.15	0.96	0.47	0.81
Psychological health	LR	0.70	0.77	0.71	0.79	**0.79**	0.78	0.60	0.83	0.48	0.89
Psychological health	MLP	0.67	0.75	0.68	0.76	0.78	0.78	0.49	0.86	0.48	0.87
Psychological health	SVM	0.70	0.75	0.66	0.75	0.64	0.60	**0.74**	0.57	0.31	0.89
Psychological health	**XGB***	**0.74**	**0.81**	0.73	**0.81**	0.78	0.77	0.68	0.79	0.46	0.90
Social	KNN	0.52	0.56	0.57	0.53	0.76	**0.81**	0.17	0.98	**0.67**	0.82
Social	LR	0.60	0.63	0.61	0.68	0.71	0.69	0.47	0.75	0.33	0.84
Social	MLP	0.56	0.60	0.50	0.65	0.70	0.79	0.00	**1.00**	0.00	0.79
Social	**SVM***	**0.61**	0.64	0.62	0.69	0.65	0.62	0.64	0.61	0.30	0.87
Social	XGB	0.61	0.62	0.55	0.60	0.68	0.67	0.34	0.76	0.27	0.81

* selected for decision-level fusion

relationship with your partner (Valentine Scale) do not appear in the rank, which could be because of the similarity of the instruments.

4.2 Multi-modal Models

The comparison of fusion methods is shown in Table 2. The model reaches the highest performance with the Logistic Regression fusion rule, getting an AUC of 0.82 and a BAC of 0.75. With Multi-layer Perception fusion rule the model gets a highest AUC of 0.83.

Fig. 4 shows the XGBoost feature importance ranking of feature-level fusion. Most of the predictive features from the psychological health modality rank high, especially previous EPDS scores. Personality, behavioral patterns, and weights also contribute to the prediction. However, the feature-level fusion approach doesn't show a performance improvement. With psychological health features only, XGBoost can reach similar BAC and AUC.

The trainable rule decision fusion have better performance on BAC and AUC compared to fixed rule decision fusion, but with the maximum fusion rule, we achieved the highest SENS (at 0.91) without a dramatic drop in BAC (at 0.63).

Fig. 5 are the receiver operating characteristic (ROC) curves of three best performed multi-modal models. Although the three models all get AUC higher than 0.81, the balance of SENS and SPEC is different at various threshold settings, e.g. when we require at least 0.70 of SENS, MLP decision fusion and LR decision fusion get better SPEC at around 0.80, while feature-level fusion with XGBoost only get SPEC of around 0.70.

5 CONCLUSION AND DISCUSSION

Antenatal depression has been linked with poor obstetric outcomes and maternal well-being, and problems in the cognitive, social and psychological development of the child, and is the biggest predictor of postpartum mental ill-health [95]. However, the focus of research largely remains on depression after birth. The results of this study highlight the utility of machine learning methods in early prediction of symptoms of antenatal depression using survey data collected via a smartphone app throughout the first two trimesters of pregnancy. Even with a missing rate as high as 80%, machine learning algorithms can still predict future antenatal depression with a BAC of 0.75 and an AUC of 0.82, higher than the AUC of 0.81 reported by Andersson et al. [4]. The best performing model, LR fusion, also achieved an overall good balance between between sensitivity (0.74) and specificity (0.78), although the highest sensitivity was achieved by max fusion method (0.91). Achieving a high sensitivity may be important for clinicians to avoid missed cases, which is one of the most significant advantages sought through

Table 2: Performance of multi-modal models

Fusion	Model/ Fusion Rule	Cross Validation		Test Set							
		BAC	AUC	BAC	AUC	F1	ACC	SENS	SPEC	PPV	NPV
Feature-Level Fusion	LR	0.68	0.74	0.67	0.78	**0.81**	0.82	0.43	0.92	0.59	0.86
	KNN	0.51	0.54	0.50	0.52	0.70	0.77	0.04	0.96	0.20	0.79
	SVM	0.63	0.69	0.68	0.75	0.69	0.66	**0.72**	0.64	0.35	0.90
	XGB	**0.73**	**0.81**	**0.74**	**0.80**	0.79	0.77	0.68	0.79	0.46	**0.91**
	MLP	0.61	0.72	0.67	0.73	**0.81**	**0.82**	0.40	**0.93**	**0.59**	0.86
Decision-Level Fusion	Max Fuison	0.61	0.72	0.57	0.71	0.37	0.37	**0.91**	0.23	0.24	**0.91**
	Mean Fusion	0.69	0.77	0.68	0.75	0.74	0.72	0.60	0.76	0.39	0.88
	Median Fusion	0.71	0.77	0.75	0.78	0.77	0.74	0.77	0.74	0.43	**0.92**
	MLP Fusion	0.61	**0.83**	0.65	**0.80**	0.81	**0.83**	0.34	0.96	0.70	0.85
	GaussianNB Fusion	0.70	0.81	0.72	0.80	**0.81**	0.81	0.57	0.87	0.54	0.89
	LR Fusion	**0.75**	0.82	**0.76**	0.80	0.79	0.78	0.74	0.78	0.47	0.92

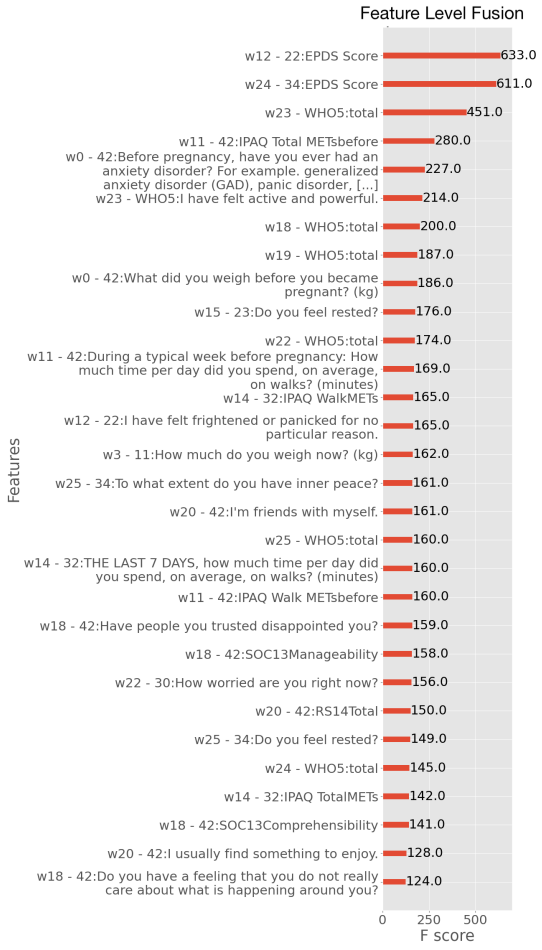


Figure 4: Top 30 feature importance rank of feature-level fusion

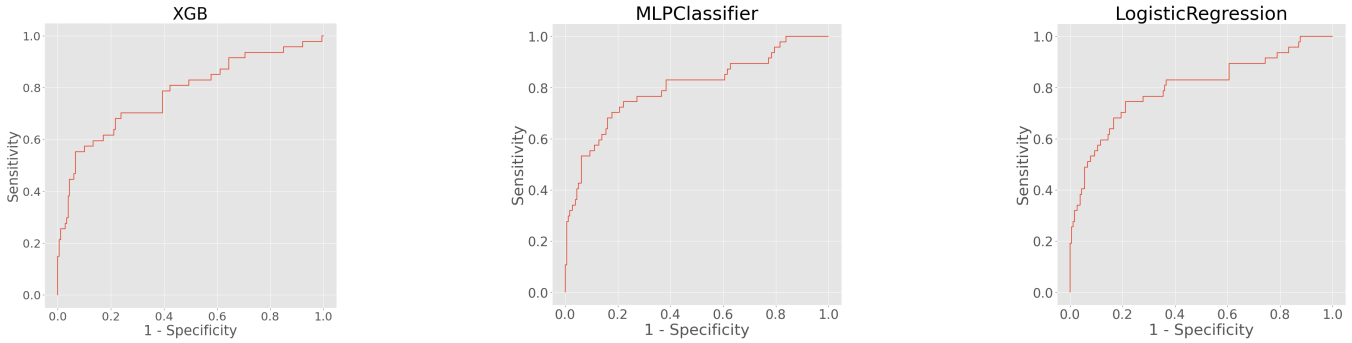
using machine learning algorithms. However, it is also equally important to achieve a specificity level that ensures the model will be cost-effective when applied in clinical settings. Using the ROC curves in Fig. 5, we can have a better understanding of what is the trade off when adjusting threshold for different sensitivities.

Considering the high missing rate of our data, this is a fairly good outcome. It shows the efficacy of machine learning-based mHealth early prediction of antenatal depression. The explainability of the models should also be taken into account. Compared with unimodal and feature-level fusion models, the hierarchical structure of the decision-level fusion model has better explainability of the prediction. The importance of Explainable Artificial Intelligence (XAI) and its practical value in healthcare is well-recognized [2, 53, 65, 89]. In this work, the necessity of XAI was particularly shown in aspects of justifying the predictions, controlling the performance and discovering novel predictors.

The feature importance analysis and uni-modal/multimodal approaches provide a possible solution for selecting useful surveys that can be collected on predictive mHealth apps. By comparing model performance and important features of five modalities in our dataset, we investigated the contributions of different factors to antenatal depression prediction.

Globally, including in Sweden, antenatal screening for depression is done in the early postpartum period [14], and this study is part of a growing body of evidence supporting the development and implementation of early detection and intervention protocols. However, before the implementation of these models in clinical settings can be done, it is important to analyze predictors in an explainable approach to build an unbiased and trustworthy tool for clinicians.

In the study, we use feature importance ranking on both XGBoost trained with unimodal features and concatenated multi-modal features, providing a better understanding of the contributions of different features. Psychometric scores during pregnancy, i.e., EPDS scores, WHO5 scores, anxiety history, physical activity, sleep, and weights came out as important predictive factors for antenatal depression. This is mostly consistent with previous studies [57]. Physical activity and sleep quality are more complex to interpret as it can be argued that they are also symptoms of depressive episodes. They may be acting as confounders, or might be mediated by other risk factors. We did not explore this relationship further, however, it would be fruitful for future studies to investigate the link between



(a) XGBoost classifier trained with Feature-level fusion rule

(b) Multi-layer perceptron decision fusion of five selected uni-modal models

(c) Logistic regression decision fusion of five selected uni-modal models

Figure 5: Receiver operating characteristic curves

sleep disturbances and changes in physical activity in early pregnancy, with depressive symptoms in the late antenatal period. A major advantage of this study was the inclusion of a wide range of variables over a longitudinal period that have been shown by previous studies to be important risk factors, as well as variables that lack research focus such as personality traits, such as life events [49]. The majority of previous studies have been cross-sectional in nature, which limits our ability to infer the direction of relationships [48].

This study uses the EPDS as the main outcome measure to quantify the risk of antenatal depression. The EPDS is considered a validated tool, with high sensitivity and specificity for detecting depression in the perinatal period [71]. We included several validated surveys with good psychometric properties to measure variables. Furthermore, we also included several self-developed surveys, usually with the intention of keeping surveys focused on the variable of choice, and short in length. While these surveys have not been validated, the questions were usually inspired from existing validated surveys. The results proved the feasibility of using shorter surveys instead of long surveys that may be validated, as many single questions get high importance ranking instead of the total score of a survey.

There are a few limitations in our work to consider. *Selection bias* is an important issue in this study. Data used in our analysis were acquired from a domestic cohort study, which excluded women who did not speak Swedish. The resulting cohort consists predominantly of women who were born in Sweden and have received post-secondary education. This may limit the generalizability of our findings. Another issue to consider in terms of generalizability is the use of a smartphone app to collect data. As the smartphone app also contains additional features that allow women to track their well-being, behavioral activity and receive health-related information about their pregnancy frequently, the potential influence of being able to track one's health and well-being must be kept in mind when interpreting our results and considering the implementation of such models.

Prevalence of depression was found to be 21% within the study population, which is higher than the Swedish average level reported in previous studies [38]. This discrepancy could be attributed to COVID-19 pandemic having worsened mental well-being [40], or the various cut-off values for the EPDS and sample characteristics of studies exploring prevalence. Selection bias was also considered as a reason for greater prevalence, such that women with depressive symptoms may simply be more inclined to participate in research exploring the issue. Furthermore, the ease of participating in research on a smartphone app may have allowed more women with depressive symptoms to respond.

ACKNOWLEDGMENTS

This work is founded by Women's Mental Health during the Reproductive Lifespan – WOMHER, Uppsala University.

REFERENCES

[1] Judith Alder, Nadine Fink, Johannes Bitzer, Irene Hösli, and Wolfgang Holzgreve. 2007. Depression and anxiety during pregnancy: a risk factor for obstetric, fetal and neonatal outcome? A critical review of the literature. *The Journal of Maternal-Fetal & Neonatal Medicine* 20, 3 (2007), 189–209.

[2] Julia Amann, Alessandro Blasimme, Effy Vayena, Dietmar Frey, and Vince I Madai. 2020. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making* 20, 1 (2020), 1–9.

[3] Gerhard Andersson, Maria Burman, Per Carlbring, and Anna-Karin Norlander. 2018. *Närmare varandra: Nio veckor till en starkare parrelation.* Natur & Kultur.

[4] Sam Andersson, Deepti R Bathula, Stavros I Iliadis, Martin Walter, and Alkistis Skalkidou. 2021. Predicting women with depressive symptoms postpartum with machine learning methods. *Scientific reports* 11, 1 (2021), 1–15.

[5] Mirza Mansoor Baig, Hamid Gholamhosseini, and Martin J. Connolly. 2015. Mobile healthcare applications: system design review, critical issues and challenges. *Australasian Physical & Engineering Sciences in Medicine* 38, 1 (2015), 23–38. https://doi.org/10.1007/s13246-014-0315-4

[6] Geraldine Barrett, Sarah C Smith, and Kaye Wellings. 2004. Conceptualisation, development, and evaluation of a measure of unplanned pregnancy. *Journal of Epidemiology & Community Health* 58, 5 (2004), 426–433.

[7] Célyne H Bastien, Annie Vallières, and Charles M Morin. 2001. Validation of the Insomnia Severity Index as an outcome measure for insomnia research. *Sleep medicine* 2, 4 (2001), 297–307.

[8] Amit Baumel, John Torous, Stav Edan, and John M. Kane. 2020. There is a non-evidence-based app for that: A systematic review and mixed methods analysis of depression- and anxiety-related apps that incorporate unrecognized techniques. *Journal of Affective Disorders* 273 (2020), 410–421. https://doi.org/10.1016/j.jad.2020.05.011

[9] James Benoit, Henry Onyeaka, Matcheri Keshavan, and John Torous. 2020. Systematic review of digital phenotyping and machine learning in psychosis spectrum illnesses. *Harvard Review of Psychiatry* 28, 5 (2020), 296–304.

[10] Swedish Child Preventative Health Registry BHVQ. 2020. .

[11] Alessandra Biaggi, Susan Conroy, Susan Pawlby, and Carmine M Pariante. 2016. Identifying the women at risk of antenatal anxiety and depression: a systematic review. *Journal of affective disorders* 191 (2016), 62–77.

[12] Ayesha M Bilal, Emma Fransson, Emma Bränn, Allison Eriksson, Mengyu Zhong, Karin Gidén, Ulf Elofsson, Cathrine Axfors, Alkistis Skalkidou, and Fotios C

Papadopoulos. 2022. Predicting perinatal health outcomes using smartphone-based digital phenotyping and machine learning in a prospective Swedish cohort (Mom2B): study protocol. *BMJ open* 12, 4 (2022), e059033.

[13] Matthijs Blankers, Maarten W J Koeter, and Gerard M Schippers. 2010. Missing Data Approaches in eHealth Research: Simulation Study and a Tutorial for Non-mathematically Inclined Researchers. *Journal of Medical Internet Research* 12, 5 (2010), e54. https://doi.org/10.2196/jmir.1448

[14] Emma Bränn, Emma Fransson, Anna Wikman, Natasa Kollia, Diem Nguyen, Caroline Lilliecreutz, and Alkistis Skalkidou. 2021. Who do we miss when screening for postpartum depression? A population-based study in a Swedish region. *Journal of Affective Disorders* 287 (2021), 165–173.

[15] Kelly A Brennan, Catherine L Clark, and Phillip R Shaver. 1998. Self-report measurement of adult attachment: An integrative overview. (1998).

[16] Michelle Nicole Burns, Mark Begale, Jennifer Duffecy, Darren Gergle, Chris J Karr, Emily Giangrande, and David C Mohr. 2011. Harnessing Context Sensing to Develop a Mobile Intervention for Depression. *Journal of Medical Internet Research* 13, 3 (2011), e55. https://doi.org/10.2196/jmir.1838

[17] Luca Canzian and Mirco Musolesi. 2015. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. 1293–1304.

[18] Susan Caplan, Angelina Sosa Lovera, and Patricia Reyna Liberato. 2018. A feasibility study of a mental health mobile app in the Dominican Republic: The untold story. *International Journal of Mental Health* 47, 4 (2018), 311–345. https://doi.org/10.1080/00207411.2018.1553486

[19] Paolo Cellini, Alessandro Pigoni, Giuseppe Delvecchio, Chiara Moltrasio, and Paolo Brambilla. 2022. Machine learning in the prediction of postpartum depression: A review. *Journal of Affective Disorders* 309 (2022), 350–357. https://doi.org/10.1016/j.jad.2022.04.093

[20] Amy Hai Yan Chan and Michelle L. L. Honey. 2022. User perceptions of mobile digital apps for mental health: Acceptability and usability - An integrative review. *Journal of Psychiatric and Mental Health Nursing* 29, 1 (2022), 147–168. https://doi.org/10.1111/jpm.12744

[21] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

[22] Evangelia Christodoulou, Jie Ma, Gary S. Collins, Ewout W. Steyerberg, Jan Y. Verbakel, and Ben Van Calster. 2019. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology* 110 (2019), 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004

[23] Elizabeth Q Cox, Nathaniel A Sowa, Samantha E Meltzer-Brody, and Bradley N Gaynes. 2016. The perinatal depression treatment cascade: baby steps toward improving outcomes. *The Journal of clinical psychiatry* 77, 9 (2016), 20901.

[24] John Cox and Jeni Holden. 2003. *Perinatal mental health: A guide to the Edinburgh Postnatal Depression Scale (EPDS)*. Royal College of Psychiatrists.

[25] Cora L Craig, Alison L Marshall, Michael Sjöström, Adrian E Bauman, Michael L Booth, Barbara E Ainsworth, Michael Pratt, ULF Ekelund, Agneta Yngve, James F Sallis, et al. 2003. International physical activity questionnaire: 12-country reliability and validity. *Medicine & science in sports & exercise* 35, 8 (2003), 1381–1395.

[26] Mohamed Daoudi, Zakia Hammal, Anis Kacem, and Jeffrey F Cohn. 2019. Gram matrices formulation of body shape motion: an application for depression severity assessment. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, 258–263.

[27] CL Dennis and P Boyce. 2004. Further psychometric testing of a brief personality scale to measure vulnerability to postpartum depression. *Journal of Psychosomatic Obstetrics & Gynecology* 25, 3-4 (2004), 305–311.

[28] Antora Dev, Nipa Roy, Md. Kafiul Islam, Chiranjeeb Biswas, Helal Uddin Ahmed, Md. Ashraful Amin, Farhana Sarker, Ravi Vaidyanathan, and Khondaker A. Mamun. 2022. Exploration of EEG-Based Depression Biomarkers Identification Techniques and Their Applications: A Systematic Review. *IEEE Access* 10 (2022), 16756–16781. https://doi.org/10.1109/ACCESS.2022.3146711

[29] Mona Eklund, Martin Bäckström, and Hanna Tuvesson. 2014. Psychometric properties and factor structure of the Swedish version of the Perceived Stress Scale. *Nordic Journal of Psychiatry* 68, 7 (2014), 494–499.

[30] Monica Eriksson and Bengt Lindström. 2005. Validity of Antonovsky's sense of coherence scale: a systematic review. *Journal of Epidemiology & Community Health* 59, 6 (2005), 460–466.

[31] Hua Fang, Kimberly Andrews Espy, Maria L. Rizzo, Christian Stopp, Sandra A. Wiebe, and Walter W. Stroup. 2009. Pattern recognition of longitudinal trial data with nonignorable missingness: An empirical case study. *International Journal of Information Technology & Decision Making* 08, 03 (2009), 491–513. https://doi.org/10.1142/s0219622009003508

[32] Tiffany Field, Miguel Diego, and Maria Hernandez-Reif. 2006. Prenatal depression effects on the fetus and newborn: a review. *Infant Behavior and Development* 29, 3 (2006), 445–455.

[33] Ricky Greenwald and Allen Rubin. 1999. Assessment of posttraumatic symptoms in children: Development and preliminary validation of parent and child scales. *Research on Social Work Practice* 9, 1 (1999), 61–75.

[34] Nancy K Grote, Jeffrey A Bridge, Amelia R Gavin, Jennifer L Melville, Satish Iyengar, and Wayne J Katon. 2010. A meta-analysis of depression during pregnancy and the risk of preterm birth, low birth weight, and intrauterine growth restriction. *Archives of general psychiatry* 67, 10 (2010), 1012–1024.

[35] Agnes Gruenerbl, Venet Osmani, Gernot Bahle, Jose C Carrasco, Stefan Oehler, Oscar Mayora, Christian Haring, and Paul Lukowicz. 2014. Using smart phone mobility traces for the diagnosis of depressive and manic episodes in bipolar patients. In *Proceedings of the 5th augmented human international conference*. 1–8.

[36] Liisa Hantsoo, Stephanie Criniti, Annum Khan, Marian Moseley, Naomi Kincler, Laura J Faherty, C Neill Epperson, and Ian M Bennett. 2018. A mobile application for monitoring and management of depressed mood in a vulnerable pregnant population. *Psychiatric Services* 69, 1 (2018), 104–107.

[37] Diego Hidalgo-Mazzei, Ainoa Mateu, María Reinares, Andrea Murru, Caterina del Mar Bonnín, Cristina Varo, Marc Valentí, Juan Undurraga, Sergio Strejilevich, José Sánchez-Moreno, et al. 2016. Psychoeducation in bipolar disorder with a SIMPLe smartphone application: feasibility, acceptability and satisfaction. *Journal of Affective Disorders* 200 (2016), 58–66.

[38] Ingegerd Hildingsson and Christine Rubertsson. 2022. Depressive symptoms during pregnancy and after birth in women living in Sweden who received treatments for fear of birth. *Archives of women's mental health* 25, 2 (2022), 473–484.

[39] Ingegerd Hildingsson, Christine Rubertsson, Annika Karlström, and Helen Haines. 2018. Exploring the Fear of Birth Scale in a mixed population of women of childbearing age—A Swedish pilot study. *Women and Birth* 31, 5 (2018), 407–413.

[40] Chung Ho-Fung, Ewa Andersson, Huang Hsuan-Ying, Ganesh Acharya, and Simone Schwank. 2022. Self-reported mental health status of pregnant women in Sweden during the COVID-19 pandemic: a cross-sectional survey. *BMC pregnancy and childbirth* 22, 1 (2022), 1–12.

[41] Neesha Hussain-Shamsy, Amika Shah, Simone N Vigod, Juveria Zaheer, and Emily Seto. 2020. Mobile Health for Perinatal Depression and Anxiety: Scoping Review. *Journal of Medical Internet Research* 22, 4 (2020), e17011. https://doi.org/10.2196/17011

[42] Santiago Jiménez-Serrano, Salvador Tortajada, and Juan Miguel García-Gómez. 2015. A mobile health application to predict postpartum depression based on machine learning. *Telemedicine and e-Health* 21, 7 (2015), 567–574.

[43] Geneva Jonathan, Elizabeth A Carpenter-Song, Rachel M Brian, and Dror Ben-Zeev. 2019. Life with FOCUS: A qualitative evaluation of the impact of a smartphone intervention on people with serious mental illness. *Psychiatric rehabilitation journal* 42, 2 (2019), 182.

[44] Heysem Kaya and Albert Ali Salah. 2014. Eyes whisper depression: A cca based multimodal approach. In *Proceedings of the 22nd ACM international conference on Multimedia*. 961–964.

[45] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[46] Rashmi Korlakai Vinayak and Ran Gilad-Bachrach. 2015. DART: Dropouts meet Multiple Additive Regression Trees. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 38)*, Guy Lebanon and S. V. N. Vishwanathan (Eds.). PMLR, San Diego, California, USA, 489–497. https://proceedings.mlr.press/v38/korlakaivinayak15.html

[47] L.I. Kuncheva. 2002. A theoretical study on six classifier fusion strategies. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24, 2 (2002), 281–286. https://doi.org/10.1109/34.982906

[48] Christie A Lancaster, Katherine J Gold, Heather A Flynn, Harim Yoo, Sheila M Marcus, and Matthew M Davis. 2010. Risk factors for depressive symptoms during pregnancy: a systematic review. *American journal of obstetrics and gynecology* 202, 1 (2010), 5–14.

[49] Bronwyn Leigh and Jeannette Milgrom. 2008. Risk factors for antenatal depression, postnatal depression and parenting stress. *BMC psychiatry* 8, 1 (2008), 1–11.

[50] Jau-Huei Lin and Peter J. Haug. 2008. Exploiting missing clinical data in Bayesian network modeling for predicting medical problems. *Journal of Biomedical Informatics* 41, 1 (2008), 1–14. https://doi.org/10.1016/j.jbi.2007.06.001

[51] Danxia Liu, Xing Lin Feng, Farooq Ahmed, Muhammad Shahid, and Jing Guo. 2022. Detecting and Measuring Depression on Social Media Using a Machine Learning Approach: Systematic Review. *JMIR Mental Health* 9, 3 (2022), e27244. https://doi.org/10.2196/27244

[52] Kien Hoa Ly, Elsa Janni, Richard Wrede, Mina Sedem, Tara Donker, Per Carlbring, and Gerhard Andersson. 2015. Experiences of a guided smartphone-based behavioral activation therapy for depression: a qualitative study. *Internet Interventions* 2, 1 (2015), 60–68.

[53] Sherin Mary Mathews. 2019. Explainable artificial intelligence applications in NLP, biomedical, and malware classification: a literature review. In *Intelligent computing-proceedings of the computing conference*. Springer, 1269–1292.

[54] Abhinav Mehrotra, Robert Hendley, and Mirco Musolesi. 2016. Towards Multimodal Anticipatory Monitoring of Depressive States through the Analysis of Human-smartphone Interaction. In *Proceedings of 1st Mental Health: Sensing and Intervention Workshop. Colocated with ACM UbiComp'16* (Heidelberg, Germany). ACM, 1132–1138.

[55] Abhinav Mehrotra and Mirco Musolesi. 2018. Using autoencoders to automatically extract mobility features for predicting depressive states. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–20.

[56] Tong Meng, Xuyang Jing, Zheng Yan, and Witold Pedrycz. 2020. A survey on machine learning for data fusion. *Information Fusion* 57 (2020), 115–129. https://doi.org/10.1016/j.inffus.2019.12.001

[57] M Carmen Míguez and M Belén Vázquez. 2021. Risk factors for antenatal depression: A review. *World Journal of Psychiatry* 11, 7 (2021), 325.

[58] Jeannette Milgrom, Alan W Gemmill, Justin L Bilszta, Barbara Hayes, Bryanne Barnett, Janette Brooks, Jennifer Ericksen, David Ellwood, and Anne Buist. 2008. Antenatal risk factors for postnatal depression: a large prospective study. *Journal of affective disorders* 108, 1-2 (2008), 147–157.

[59] Lisa A Mistler, Dror Ben-Zeev, Elizabeth Carpenter-Song, Mary F Brunette, and Matthew J Friedman. 2017. Mobile mindfulness intervention on an acute psychiatric unit: feasibility and acceptability study. *JMIR Mental Health* 4, 3 (2017), e7717.

[60] Matthew D. Nemesure, Michael V. Heinz, Raphael Huang, and Nicholas C. Jacobson. 2021. Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence. *Scientific Reports* 11, 1 (2021). https://doi.org/10.1038/s41598-021-81368-4

[61] Jennifer Nicholas, Mark Erik Larsen, Judith Proudfoot, and Helen Christensen. 2015. Mobile Apps for Bipolar Disorder: A Systematic Review of Features and Content Quality. *Journal of Medical Internet Research* 17, 8 (2015), e198. https://doi.org/10.2196/jmir.4581

[62] William S Noble. 2006. What is a support vector machine? *Nature Biotechnology* 24, 12 (Dec 2006), 1565–1567. https://doi.org/10.1038/nbt1206-1565

[63] Laura Orsolini, Alessandro Valchera, Roberta Vecchiotti, Carmine Tomasetti, Felice Iasevoli, Michele Fornaro, Domenico De Berardis, Giampaolo Perna, Maurizio Pompili, and Cesario Bellantuono. 2016. Suicide during perinatal period: epidemiology, risk factors, and clinical correlates. *Frontiers in psychiatry* 7 (2016), 138.

[64] Elizabeth O'Connor, Caitlyn A Senger, Michelle L Henninger, Erin Coppola, and Bradley N Gaynes. 2019. Interventions to prevent perinatal depression: evidence report and systematic review for the US Preventive Services Task Force. *Jama* 321, 6 (2019), 588–601.

[65] Urja Pawar, Donna O'Shea, Susan Rea, and Ruairi O'Reilly. 2020. *Explainable AI in Healthcare.* https://doi.org/10.1109/cybersa49311.2020.9139655

[66] Šarūnas Raudys. 2006. Trainable fusion rules. I. Large sample size case. *Neural Networks* 19, 10 (2006), 1506–1516. https://doi.org/10.1016/j.neunet.2006.01.018

[67] Rouzbeh Razavi, Amin Gharipour, and Mojgan Gharipour. 2020. Depression screening using mobile phone usage metadata: a machine learning approach. *Journal of the American Medical Informatics Association* 27, 4 (2020), 522–530. https://doi.org/10.1093/jamia/ocz221

[68] Fabien Ringeval, Bjorn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Messner Eva-Maria Messner, Siyang Song, Shuo Liu, Ziping Zhao, Adria Mallol-Ragolta, Zhao Ren, Mohammad Soleymani, and Maja Pantic. 2019. AVEC 2019 workshop and challenge: State-of-mind, detecting depression with ai, and cross-cultural affect recognition. In *AVEC 2019 - Proceedings of the 9th International Audio/Visual Emotion Challenge and Workshop, co-located with MM 2019.* Association for Computing Machinery (ACM), United States, 3–12. https://doi.org/10.1145/3347320.3357688 Publisher Copyright: © 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.; 9th Audio/Visual Emotion Challenge and Workshop, AVEC 2019, AVEC 2019 ; Conference date: 21-10-2019 Through 21-10-2019.

[69] R. Rosen, C. Brown, J. Heiman, S. Leiblum, C. Meston, R. Shabsigh, D. Ferguson, and R. D'Agostino. 2000. The Female Sexual Function Index (FSFI): A Multidimensional Self-Report Instrument for the Assessment of Female Sexual Function. *Journal of Sex & Marital Therapy* 26, 2 (2000), 191–208. https://doi.org/10.1080/009262300278597

[70] Annika Rosengren, Kristina Orth-Gomer, Hans Wedel, and Lars Wilhelmsen. 1993. Stressful life events, social support, and mortality in men born in 1933. *British medical journal* 307, 6912 (1993), 1102–1105.

[71] Christine Rubertsson, Karin Börjesson, Anna Berglund, Ann Josefsson, and Gunilla Sydsjö. 2011. The Swedish validation of Edinburgh postnatal depression scale (EPDS) during pregnancy. *Nordic journal of psychiatry* 65, 6 (2011), 414–418.

[72] Donald B Rubin. 1976. Inference and missing data. *Biometrika* 63, 3 (1976), 581–592.

[73] Stefan Scherer, Zakia Hammal, Ying Yang, Louis-Philippe Morency, and Jeffrey F Cohn. 2014. Dyadic behavior analysis in depression severity assessment interviews. In *Proceedings of the 16th International Conference on Multimodal Interaction.* 112–119.

[74] Amita Sharma and Willem JMI Verbeke. 2020. Improving diagnosis of depression with XGBOOST machine learning model and a large biomarkers Dutch dataset (n= 11,081). *Frontiers in big Data* 3 (2020), 15.

[75] Katsunari Shibata and Yusuke Ikeda. 2009. Effect of number of hidden neurons on learning in large-scale layered neural networks. In *2009 ICCAS-SICE.* 5008–5013.

[76] Saul Shiffman, Arthur A Stone, and Michael R Hufford. 2008. Ecological momentary assessment. *Annu. Rev. Clin. Psychol.* 4 (2008), 1–32.

[77] Justine Slomian, Germain Honvo, Patrick Emonts, Jean-Yves Reginster, and Olivier Bruyère. 2019. Consequences of maternal postpartum depression: A systematic review of maternal and infant outcomes. *Women's Health* 15 (2019), 1745506519844044.

[78] Cynthia Solomon, Michel F Valstar, Richard K Morriss, and John Crowe. 2015. Objective methods for reliable detection of concealed depression. *Frontiers in ICT* 2 (2015), 5.

[79] Giota Stratou, Stefan Scherer, Jonathan Gratch, and Louis-Philippe Morency. 2015. Automatic nonverbal behavior indicators of depression and ptsd: the effect of gender. *Journal on Multimodal User Interfaces* 9, 1 (2015), 17–29.

[80] S. Tamura and M. Tateishi. 1997. Capabilities of a four-layered feedforward neural network: four layers versus three. *IEEE Transactions on Neural Networks* 8, 2 (1997), 251–255. https://doi.org/10.1109/72.557662

[81] Ariel Teles, Ivan Rodrigues, Davi Viana, Francisco Silva, Luciano Coutinho, Markus Endler, and Ricardo Rabêlo. 2019. Mobile mental health: A review of applications for depression assistance. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS).* IEEE, 708–713.

[82] Christian Winther Topp, Søren Dinesen Østergaard, Susan Søndergaard, and Per Bech. 2015. The WHO-5 Well-Being Index: a systematic review of the literature. *Psychotherapy and psychosomatics* 84, 3 (2015), 167–176.

[83] John Torous, Mathew V Kiang, Jeanette Lorme, and Jukka-Pekka Onnela. 2016. New Tools for New Research in Psychiatry: A Scalable and Customizable Platform to Empower Data Driven Smartphone Research. *JMIR Mental Health* 3, 2 (2016), e16. https://doi.org/10.2196/mental.5165

[84] John Torous and Laura Weiss Roberts. 2017. Needed Innovation in Digital Health and Smartphone Applications for Mental Health. *JAMA Psychiatry* 74, 5 (2017), 437. https://doi.org/10.1001/jamapsychiatry.2017.0262

[85] John Torous, Patrick Staples, and Jukka-Pekka Onnela. 2015. Realizing the potential of mobile mental health: new methods for new data in psychiatry. *Current psychiatry reports* 17, 8 (2015), 1–7.

[86] John Torous, Hannah Wisniewski, Gang Liu, and Matcheri Keshavan. 2018. Mental Health Mobile Phone App Usage, Concerns, and Benefits Among Psychiatric Outpatients: Comparative Survey Study. *JMIR Mental Health* 5, 4 (2018), e11715. https://doi.org/10.2196/11715

[87] Bhekisipho Twala. 2009. An empirical comparison of techniques for handling incomplete data using decision trees. *Applied Artificial Intelligence* 23, 5 (2009), 373–405. https://doi.org/10.1080/08839510902872223

[88] Kim van Zoonen, Claudia Buntrock, David Daniel Ebert, Filip Smit, Charles F Reynolds III, Aartjan TF Beekman, and Pim Cuijpers. 2014. Preventing the onset of major depressive disorder: a meta-analytic review of psychological interventions. *International journal of epidemiology* 43, 2 (2014), 318–329.

[89] Giulia Vilone and Luca Longo. 2020. Explainable Artificial Intelligence: a Systematic Review. *ArXiv* abs/2006.00093 (2020).

[90] Gail M Wagnild and Heather M Young. 1993. Development and psychometric. *Journal of nursing measurement* 1, 2 (1993), 165–17847.

[91] Birgitta Wickberg and CP Hwang. 1996. The Edinburgh postnatal depression scale: validation on a Swedish community sample. *Acta Psychiatrica Scandinavica* 94, 3 (1996), 181–184.

[92] Katherine L Wisner, Eydie L Moses-Kolko, and Dorothy KY Sit. 2010. Postpartum depression: a disorder in search of a definition. *Archives of women's mental health* 13, 1 (2010), 37–40.

[93] Robert F. Wolff, Karel G.M. Moons, Richard D. Riley, Penny F. Whiting, Marie Westwood, Gary S. Collins, Johannes B. Reitsma, Jos Kleijnen, and Sue Mallett. 2019. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Annals of Internal Medicine* 170, 1 (2019), 51. https://doi.org/10.7326/m18-1376

[94] Shiqi Yang, Ping Zhou, Kui Duan, M Shamim Hossain, and Mohammed F Alhamid. 2018. emHealth: towards emotion health through depression prediction and intelligent health recommender system. *Mobile Networks and Applications* 23, 2 (2018), 216–226.

[95] Xiaoxv Yin, Na Sun, Nan Jiang, Xing Xu, Yong Gan, Jia Zhang, Lei Qiu, Chenhui Yang, Xinwei Shi, Jun Chang, et al. 2021. Prevalence and associated factors of antenatal depression: systematic reviews and meta-analyses. *Clinical psychology review* 83 (2021), 101932.

[96] Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. 2020. Multimodal Intelligence: Representation Learning, Information Fusion, and Applications. *IEEE Journal of Selected Topics in Signal Processing* 14, 3 (2020), 478–493. https://doi.org/10.1109/jstsp.2020.2987728

[97] Zhongheng Zhang. 2016. Missing data imputation: focusing on single imputation. *Annals of translational medicine* 4, 1 (2016).

6 APPENDICES

Table 3: Parameters setting for grid-search

LR	penalty	['elasticnet'],
LR	solver	['saga'],
LR	C	[0.1,1,100,1000,2000,3000,4000],
LR	class_weight	[{1:6},{1:10},{1:14},'balanced'],
LR	l1_ratio	[0.4,0.5,0.6,0.7],
LR	penalty	['l2','none'],
LR	solver	['newton-cg','lbfgs','sag'],
LR	C	[0.1,1,100,1000,2000,3000,4000],
LR	class_weight	[{1:6},{1:10},{1:14},'balanced'],
LR	penalty	['l2','l1','none'],
LR	solver	['saga'],
LR	C	[0.1,1,100,1000,2000,3000,4000],
LR	class_weight	[{1:6},{1:10},{1:14},'balanced'],
LR	penalty	['l2','l1'],
LR	solver	['liblinear'],
LR	C	[0.1,1,100,1000,2000,3000,4000],
LR	class_weight	[{1:6},{1:10},{1:14},'balanced'],
SVM	kernel	["linear",'poly','sigmoid','rbf'],
SVM	gamma	[1000,100,10,1,0.1,0.01,'auto','scale'],
SVM	C	[0.1,1,10,100,1000,2000,3000,4000],
SVM	class_weight	[{1:6},{1:10},{1:14}],
SVM	degree	[2,3,4],
SVM	probability	[True],
XGB	booster	['gbtree'],
XGB	objective	['binary',logistic'],
XGB	learning_rate	[0.001,0.002,0.0001,0.0002,0.01,0.02],
XGB	max_depth	[9,12,15],
XGB	subsample	[0.7,0.6],
XGB	colsample_bytree	[0.7],
XGB	reg_lambda	[0.001,0.0001,0.00001],
XGB	n_estimators	[300,900],
XGB	seed	[42],
XGB	use_label_encoder	[False],
XGB	eval_metric	['logloss'],
XGB	scale_pos_weight	[10,12,14],
XGB	min_child_weight	[3,9,12],
XGB	booster	['dart'],
XGB	objective	['binary',logistic'],
XGB	learning_rate	[0.001,0.002,0.0001,0.0002,0.01,0.02],
XGB	max_depth	[9,12,15],
XGB	subsample	[0.7,0.6],
XGB	colsample_bytree	[0.7],
XGB	reg_lambda	[0.001,0.0001,0.00001],
XGB	n_estimators	[300,900],
XGB	seed	[42],
XGB	use_label_encoder	[False],
XGB	eval_metric	['logloss'],
XGB	scale_pos_weight	[10,12,14],
XGB	min_child_weight	[3,9,12],
XGB	rate_drop	[0.1,0.2],
XGB	skip_drop	[0.1,0.5],
XGB	sample_type	['weighted'],

Table 4: Parameters setting for grid-search – continued

MLP	hidden_layer_sizes	[(round(N/2+3),2,),(N-1,), (round(np.sqrt(N*2)),),(round (np.power(1,2)/N)+1,),(100,300,)],
MLP	activation	['tanh','relu'],
MLP	solver	['sgd','adam'],
MLP	alpha	[0.0001,0.05],
MLP	learning_rate	['constant','adaptive'],
KNN	n_neighbors	[3,5,7,9,11,13],