Deep CNN-Based Continuous Monitoring of Depression Using Smartphone Usage Patterns: Algorithm Development and Validation

Girish Srinivasan, Artur Trzesiok, Soumya Chowdhary, Roy Cohen, Janine Ellenberger

Submitted to: Journal of Medical Internet Research
on: December 09, 2024

Disclaimer: **© The authors. All rights reserved.** This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Deep CNN-Based Continuous Monitoring of Depression Using Smartphone Usage Patterns: Algorithm Development and Validation

Girish Srinivasan[1] BE, MS, PhD; Artur Trzesiok[1]; Soumya Chowdhary[1]; Roy Cohen[1]; Janine Ellenberger[1]

[1]Behavidence New York US

Corresponding Author:
Janine Ellenberger
Behavidence
99 Wall Street #4004
New York
US

Abstract

Background: Depression is a global mental health challenge, with traditional assessment methods like the Patient Health Questionnaire-9 (PHQ-9) limited by infrequent data collection and susceptibility to recall bias. Recent advances in digital phenotyping offer the potential to capture real-time behavioral data through smartphones. However, existing models primarily focus on binary classification, often overlooking the need for continuous and granular symptom-specific monitoring.

Objective: This study aims to address limitations in traditional depression assessments by developing and validating a convolutional neural network (CNN)-based framework. The objectives are to predict continuous PHQ-9 scores, enable symptom-specific analysis, and provide confident classifications of depression severity using passive smartphone data.

Methods: A novel digital phenotyping approach was employed, leveraging raster plots to encode smartphone usage patterns in 48-hour windows. Data from 491 participants were collected via a smartphone application, which tracked app usage, screen time, and interaction frequency. Participants also completed periodic PHQ-9 assessments. The CNN model was trained using five-fold cross-validation, optimized through grid search, and benchmarked against a random forest model using metrics such as precision, recall, mean absolute error (MAE), and fraction classified (f).

Results: The CNN model demonstrated superior performance over the random forest baseline, achieving an overall accuracy of 83.1%, a precision of 90.3% for positive cases, and a low MAE of 0.81 for motor activity predictions. The fraction classified (f) metric indicated 95% of cases were confidently categorized as either negative or positive, with only 5% falling into the uncertain range (PHQ-9 scores 10–15). Continuous tracking of PHQ-9 scores illustrated the model’s ability to monitor stable and dynamic depressive trajectories. For instance, User ID 38867 showed strong alignment between predictions and self-reports, while User ID 56084 highlighted the model's sensitivity to symptom variability.

Conclusions: This study introduces a robust framework for passive, continuous depression monitoring, advancing the application of digital phenotyping in mental health care. By leveraging CNNs and raster plots, the approach bridges gaps in traditional assessments, providing actionable, symptom-specific insights. The results emphasize the potential for personalized, scalable mental health monitoring and support future integration of multimodal data and real-time feedback mechanisms for enhanced clinical applicability.

(JMIR Preprints 09/12/2024:69836)
DOI: https://doi.org/10.2196/preprints.69836

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Deep CNN-Based Continuous Monitoring of Depression Using Smartphone Usage Patterns: Algorithm Development and Validation

Abstract

Background

Depression is a global mental health challenge, with traditional assessment methods like the Patient Health Questionnaire-9 (PHQ-9) limited by infrequent data collection and susceptibility to recall bias. Recent advances in digital phenotyping offer the potential to capture real-time behavioral data through smartphones. However, existing models primarily focus on binary classification, often overlooking the need for continuous and granular symptom-specific monitoring.

Objective

This study aims to address limitations in traditional depression assessments by developing and validating a convolutional neural network (CNN)-based framework. The objectives are to predict continuous PHQ-9 scores, enable symptom-specific analysis, and provide confident classifications of depression severity using passive smartphone data.

Methods

A novel digital phenotyping approach was employed, leveraging raster plots to encode smartphone usage patterns in 48-hour windows. Data from 491 participants were collected via a smartphone application, which tracked app usage, screen time, and interaction frequency. Participants also completed periodic PHQ-9 assessments. The CNN model was trained using five-fold cross-validation, optimized through grid search, and benchmarked against a random forest model using metrics such as precision, recall, mean absolute error (MAE), and fraction classified (f).

Results

The CNN model demonstrated superior performance over the random forest baseline, achieving an overall accuracy of 83.1%, a precision of 90.3% for positive cases, and a low MAE of 0.81 for motor activity predictions. The fraction classified (f) metric indicated 95% of cases were confidently categorized as either negative or positive, with only 5% falling into the uncertain range (PHQ-9 scores 10–15). Continuous tracking of PHQ-9 scores illustrated the model’s ability to monitor stable and dynamic depressive trajectories. For instance, User ID 38867 showed strong alignment between predictions and self-reports, while User ID 56084 highlighted the model's sensitivity to symptom variability.

Conclusions

This study introduces a robust framework for passive, continuous depression monitoring, advancing the application of digital phenotyping in mental health care. By leveraging CNNs and raster plots, the approach bridges gaps in traditional assessments, providing actionable, symptom-specific insights. The results emphasize the potential for personalized, scalable mental health monitoring and support future integration of multimodal data and real-time feedback mechanisms for enhanced clinical applicability.

Keywords

Digital Phenotyping; Depression; Deep Learning; Convolutional Neural Networks; Remote Monitoring; Mental Health; PHQ-9 Score Prediction

Introduction

Depression remains a leading cause of disability worldwide, affecting over 280 million people, as estimated by the World Health Organization (WHO)(Malhi & Mann, 2018). The chronic and fluctuating nature of depression makes timely and accurate monitoring critical for effective diagnosis and intervention(McAllister-Williams et al., 2020) . Traditionally, depression is assessed through clinical interviews or self-reported questionnaires, such as the Patient Health Questionnaire-9 (PHQ-9), which provides a validated measure of depression severity(Kroenke et al., 2001). However, these tools are limited by recall bias, underreporting, and infrequent monitoring, which can result in delayed interventions.

In recent years, there has been growing interest in leveraging digital health technologies, particularly digital phenotyping, for continuous and passive monitoring of mental health(Torous et al., 2021). Digital phenotyping involves the collection of behavioral data from personal devices, such as smartphones, to infer mental health status in real-time(Bufano et al., 2023). The potential of such approaches lies in their ability to provide objective, ongoing monitoring without requiring active input from the user, making them less burdensome for patients and more reflective of daily life behaviors. By using smartphone data such as screen time, app usage, and sleep patterns, researchers aim to develop models that can detect early signs of mental health deterioration, offering a more scalable and cost-effective solution for monitoring large populations(Torous et al., 2014).

This paper builds upon existing research in digital phenotyping and explores the application of deep learning, particularly convolutional neural networks (CNNs), to move beyond simple binary classifications of depression. By predicting continuous PHQ-9 scores and enabling symptom-specific analysis, this approach offers a more nuanced understanding of depression severity, potentially improving personalized mental health care and early intervention.

Background

Digital Phenotyping for Mental Health Monitoring

Digital phenotyping has emerged as a promising approach for passive and continuous monitoring of mental health, leveraging the ubiquity of smartphones and other personal devices. The term was first introduced by Onnela and Rauch (2016), who proposed using smartphone data to quantify behaviors associated with mental health conditions in real-time. This approach is particularly well-suited to mental health disorders like depression, which manifests in behavioral changes that can be detected passively, such as shifts in daily routines, social interaction, and sleep patterns.

Several studies have validated the potential of smartphone-based digital phenotyping for detecting mental health conditions. A paper, for example, demonstrated that features like reduced phone usage and irregular sleep patterns were significantly correlated with depression severity(Saeb et al., 2015). Similarly, Torous et al. emphasized the value of passive smartphone data in predicting early warning signs of mental health deterioration, noting that these tools could potentially enhance accessibility to mental health care in underserved populations(Torous et al., 2015). These studies, along with others, highlight the value of continuous data collection in capturing subtle behavioral changes that might go unnoticed through traditional assessments.

Machine Learning Models in Depression Detection

In addition to digital phenotyping, machine learning (ML) models have been increasingly applied to predict mental health outcomes based on smartphone data (Choudhary et al., 2022; Grzenda et al., 2021). Early models, such as random forests and support vector machines, have been used to classify users as either "depressed" or "not depressed" based on features such as screen time, app usage patterns, frequency of app switching, and digital sleep time (i.e., time spent away from the phone(Patel et al., 2025). These models typically rely on binary classification, which simplifies the complex spectrum of depression into two categories, limiting their ability to provide detailed insights into symptom severity or specific behavioral patterns associated with different levels of depression.

A recent paper used random forest models to predict depression based on smartphone usage data, achieving accuracies as high as 87% for binary classification of PHQ-9-based depression scores(Choudhary et al., 2022). However, while these models showed promise, they were limited by their inability to predict continuous PHQ-9 scores or offer symptom-specific insights, which are critical for personalized mental health care. Similar studies by Jacobson et al. (2021) and Meyerhoff et al. (2021) have also demonstrated the use of machine learning for detecting mental health conditions, although their focus remained on binary outcomes and often used intrusive data such as GPS location.

Advancements with Deep Learning and CNNs

The use of deep learning, particularly convolutional neural networks (CNNs), offers new possibilities for more detailed mental health assessments(Su et al., 2020). CNNs have been widely adopted in image recognition and medical diagnostics for their ability to capture complex patterns in high-dimensional data(Mall et al., 2023). By converting smartphone usage data into raster-like images, which visually represent the time spent in various app categories, CNNs can analyze temporal shifts in user behavior that are more challenging for traditional models to detect.

While deep learning models have been applied in adjacent fields, few studies have explored their potential for predicting continuous PHQ-9 scores or performing symptom-specific analysis in mental health(Gadzama et al., 2024). Previous work has shown that CNNs can effectively capture subtle patterns in health-related data, as demonstrated by Wang et al. (2021) in their study on mood disorder detection using smartphone data. This study expands on such approaches by introducing CNNs as a tool for predicting depression severity on a continuous scale, moving beyond the limitations of binary classification.

Challenges in Predicting PHQ-9 scores

Accurately predicting PHQ-9 scores, which measure depression severity across nine specific symptoms, remains a challenge in the field of digital phenotyping. Traditional machine learning models have struggled to predict these scores, often resulting in oversimplified binary classifications(Choudhary & Srinivasan, 2022). This limitation prompted researchers to explore alternative methods, such as symptom-specific modeling, where each PHQ-9 item is individually predicted based on corresponding behavioral data. Such an approach allows for a more granular understanding of how specific behaviors—such as irregular sleep or reduced social interaction—correlate with particular symptoms of depression, thereby facilitating more personalized interventions.

In this context, our study introduces raster plots as a novel representation of smartphone usage data and applies CNNs to predict both continuous PHQ-9 scores and individual symptoms. This approach aims to provide a more detailed and accurate understanding of depression severity, with potential applications in decentralized clinical trials and remote mental health monitoring.

Challenges of Traditional Models and the Role of Innovations

Traditional approaches to depression monitoring using passive smartphone data, such as those employing random forest classifiers, rely heavily on manually engineered features. These features include metrics such as average phone usage across app categories, frequency of app switches, and estimated digital sleep time(Nepal et al., 2024; Price et al., 2023). While such models have demonstrated moderate success in binary classifications of depression (e.g., depressed vs. not depressed), they face significant limitations:

1. *Simplification of Depression*: Binary classifications fail to capture the nuanced spectrum of depression severity that tools like the Patient Health Questionnaire-9 (PHQ-9) aim to measure. This oversimplification reduces the clinical utility of these models for personalized care.
2. *Loss of Temporal Patterns*: Manually engineered features often fail to preserve the temporal dynamics of user behavior, such as how app usage patterns vary across time. This temporal information is critical for understanding symptoms like sleep disturbances or inconsistent energy levels, which are hallmark signs of depression.
3. *Limited Symptom-Specific Insights*: Traditional models are typically designed to predict a single outcome (e.g., depression presence), offering little insight into how specific symptoms manifest in behavioral data.

To address these challenges, this study introduces two key innovations:

1. *Raster Plot Representation*:
   - By converting smartphone usage data into raster plots, this method transforms behavioral patterns into a visual, image-like format that captures both spatial (across app categories) and temporal (over time) dimensions of user activity.
   - Raster plots allow for the preservation of fine-grained temporal details, such as sleep-wake cycles and shifts in app engagement patterns, enabling the model to detect subtle behavioral changes associated with different levels of depression severity.
2. *Convolutional Neural Networks (CNNs)*:
   - CNNs, widely used in medical image analysis, excel at identifying complex patterns in high-dimensional data. By applying CNNs to raster plots, the model learns to automatically extract features that reflect temporal and spatial behavioral patterns without relying on manual engineering.
   - This approach moves beyond binary classification to predict continuous PHQ-9 scores, enabling a more precise assessment of depression severity. Additionally, by training symptom-specific models, the CNN provides granular insights into how individual depressive symptoms, such as sleep disturbances (PHQ-9 Item 3) or low energy (PHQ-9 Item 4), manifest through smartphone behavior.

These innovations not only enhance the accuracy and interpretability of depression predictions but also open new possibilities for continuous, passive monitoring of mental health. By leveraging the rich, image-like structure of raster plots and the pattern-recognition capabilities of CNNs, this study aims to bridge the gap between passive data collection and clinically meaningful mental health insights.

Objective

The objective of this study is to develop and evaluate a novel approach to depression monitoring that leverages passive smartphone usage data, focusing on addressing the limitations of traditional models. Specifically, this study aims to:

1. *Introduce Raster Plot Representation:*
   - Transform smartphone usage data into raster plots, a visual representation that preserves both spatial (app category usage) and temporal (time-dependent behavior) patterns of user activity. This approach captures subtle behavioral changes critical for understanding depression severity.
2. *Leverage Convolutional Neural Networks (CNNs):*
   - Apply CNNs to analyze raster plots, enabling the model to automatically extract complex temporal and spatial features without relying on manual feature engineering.
   - Develop a CNN-based framework capable of predicting continuous PHQ-9 scores, providing a nuanced assessment of depression severity that goes beyond binary classification.
3. *Enable Symptom-Specific Analysis:*
   - Train separate CNN models to predict individual PHQ-9 items, offering detailed insights into specific depressive symptoms (e.g., sleep disturbances, low energy) and their behavioral correlates.
4. *Compare Against Traditional Models:*
   - Benchmark the performance of the CNN model against a traditional random forest classifier trained on manually engineered features. Evaluate improvements in accuracy, precision, recall, and other performance metrics.
5. *Ensure Generalizability Across Demographics:*
   - Assess the model’s prediction accuracy across diverse demographic groups, including variations in age and gender, to evaluate its applicability in real-world, heterogeneous populations.

Through this work, the study aims to advance the field of digital phenotyping by demonstrating how raster plots and CNNs can bridge the gap between passive smartphone data collection and clinically meaningful mental health assessments. This approach holds potential for continuous, low-burden depression monitoring in decentralized clinical trials and personalized mental health care.

Methods

Participants and Data Collection

This study involved 491 participants from the United States and the United Kingdom. All participants provided informed consent for their data to be used for scientific purposes, in line with ethical guidelines. Data was collected through the Behavidence mobile application(*Mental Health Application*, 2024.), which passively monitors smartphone usage without the need for active user input.

Each participant completed the Patient Health Questionnaire-9 (PHQ-9), a widely used instrument for measuring depression severity(Kroenke et al., 2001). The PHQ-9 assesses nine depressive symptoms with total scores ranging from 0 to 27, where higher scores indicate greater severity. Participants completed the PHQ-9 at various intervals during the study, and passive smartphone data was collected continuously throughout the study period.

The primary data collected included:

- Total screen time: The cumulative amount of time participants used their smartphones.
- App launches: The frequency with which participants opened specific apps.
- App category usage: The duration of use across 22 predefined app categories, such as communication, social media, and productivity.

Smartphone usage data was recorded in 30-minute intervals to ensure a high-resolution behavioral record.

Data Preprocessing

The raw smartphone usage data, consisting of time-stamped logs of app interactions, was processed to create structured daily datasets. For each 30-minute interval, two key metrics were computed:

1. Total time spent on each app category.
2. Frequency of app launches within each category.

To better capture behavioral patterns across sleep and wake cycles, we aggregated data over a 48-hour period, combining two consecutive days to form a single data window. This helped account for behavioral shifts between day and night and ensured that relevant sleep patterns were included.

Missing Data Handling

Missing data, which occurred due to connectivity issues or device power constraints, were addressed using mean imputation. Adjacent time intervals were used to fill gaps, ensuring the continuity of the dataset.

Raster Plot Representation of Smartphone Usage

To enable the application of deep learning techniques, smartphone usage data was transformed into raster-like images. These images visually represent the time spent and frequency of app launches across the 22 categories, using two color channels:

- Red: Represents the total time spent in each app category.
- Teal: Represents the frequency of app launches.

Each raster plot provides a detailed, visual snapshot of a participant's smartphone behavior over a 48-hour period. This method captures temporal dynamics and subtle changes in behavior that are often indicative of depression, such as changes in app usage, sleep disturbances, and patterns of social interaction.

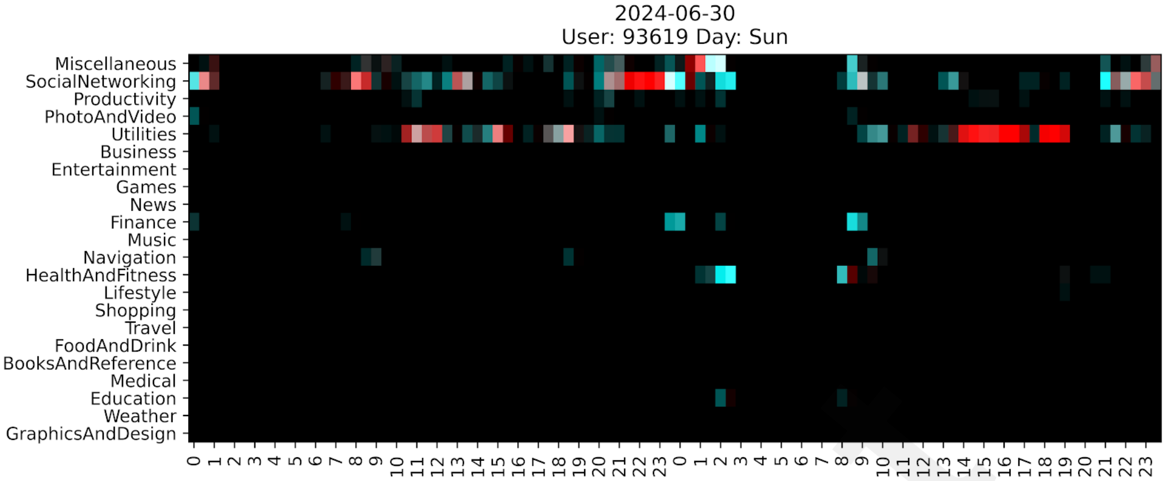


Figure 1: Raster plot for user 93619 showing a 48-hour period ending June 30th 2024 with time spent and frequency of app launches across categories.

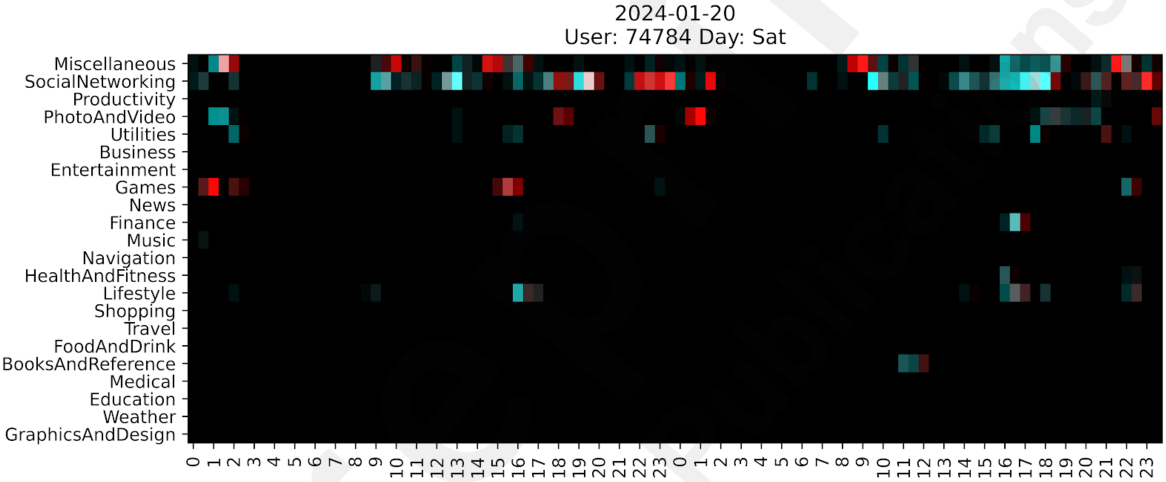


Figure 2: Raster plot for user 74784 showing a 48-hour period ending Jan 20th 2024 with time spent and frequency of app launches across categories.

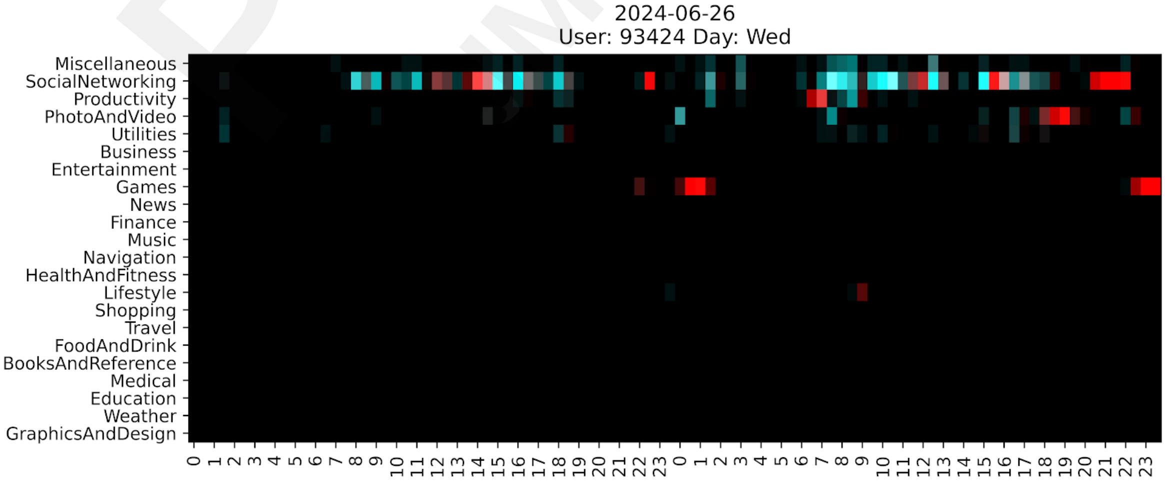


Figure 3: Raster plot for user 93424 showing a 48-hour period ending June 26th 2024 with time spent and frequency of app launches across categories.

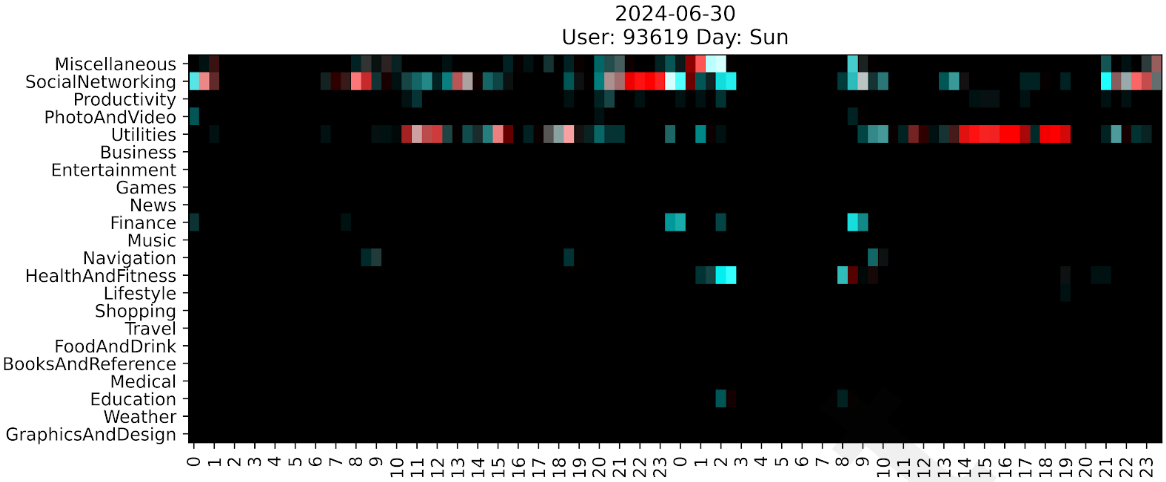


Figure 4: Raster plot for user 93619 showing a 48-hour period ending June 30th 2024 with time spent and frequency of app launches across categories.

Convolutional Neural Network (CNN) Architecture

To analyze the rasterized smartphone data, we designed a 3D convolutional neural network (CNN) capable of detecting patterns associated with depressive behavior. The CNN architecture was optimized for image-like data and was designed to capture both spatial and temporal relationships in the smartphone usage patterns.

The key components of the CNN architecture are as follows:

1. Input Layer: The input consists of raster images representing the two-dimensional (time vs. app category) smartphone usage data for each 48-hour period.
2. Category Squash Layer: This layer processes data from all 22 app categories over a 3.5-hour window, shifted by 30 minutes, to capture immediate behavioral patterns.
3. High-Level Feature Extraction Layer: This layer extends the temporal window to 7.5 hours, allowing the model to extract more complex, high-level features from the raster data that represent behavioral trends over longer time periods.
4. Fully Connected Layers: The extracted features are flattened into a 1024-dimensional vector, followed by two fully connected dense layers that map the feature space to a final output prediction.

Dropout regularization was applied to the fully connected layers to prevent overfitting, and batch normalization was used after each convolutional layer to standardize the input to each layer and accelerate training.

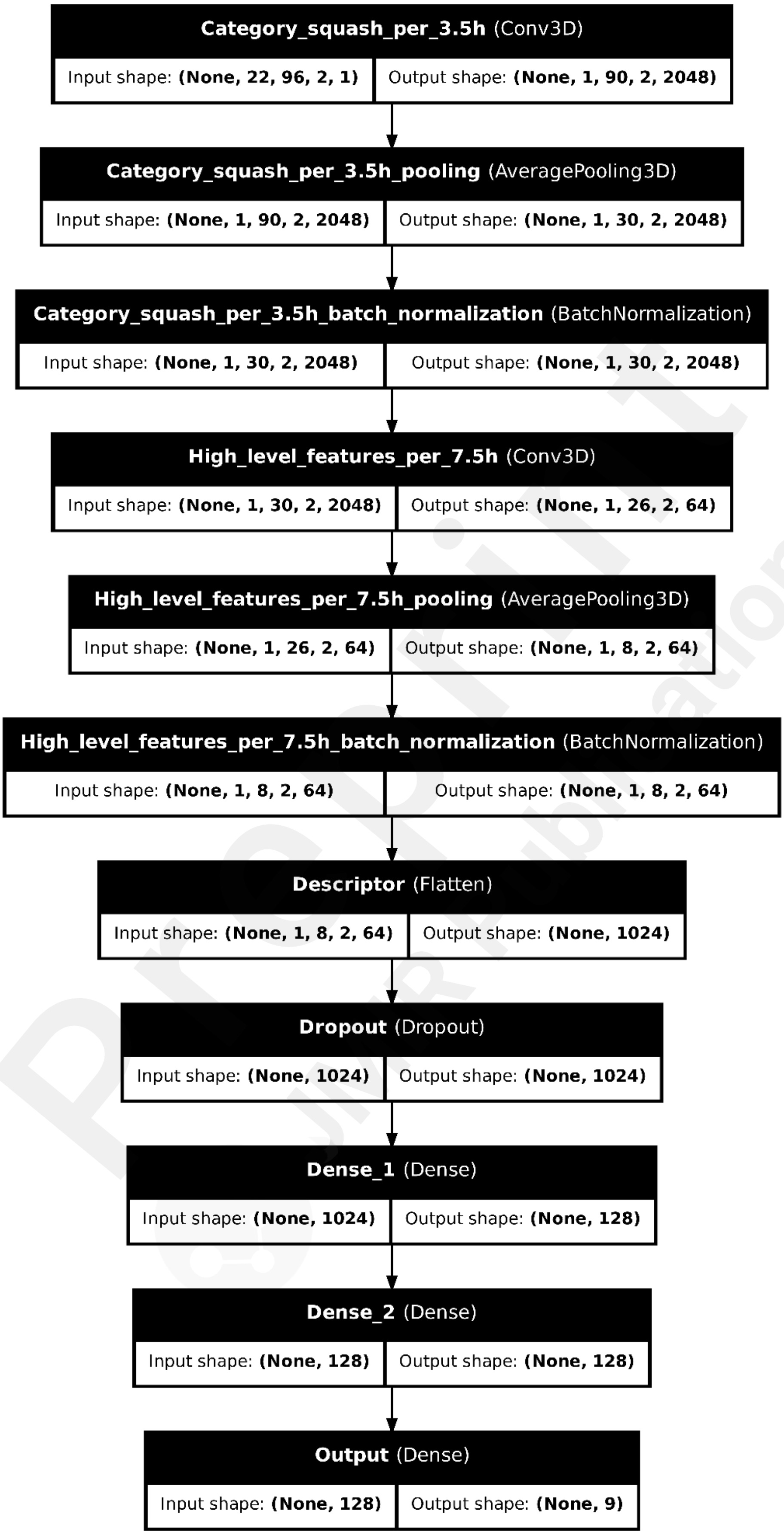

Category_squash_per_3.5h (Conv3D)
Input shape: (None, 22, 96, 2, 1)
Output shape: (None, 1, 90, 2, 2048)
Category_squash_per_3.5h_pooling (AveragePooling3D)
Input shape: (None, 1, 90, 2, 2048)
Output shape: (None, 1, 30, 2, 2048)
Category_squash_per_3.5h_batch_normalization (BatchNormalization)
Input shape: (None, 1, 30, 2, 2048)
Output shape: (None, 1, 30, 2, 2048)
High_level_features_per_7.5h (Conv3D)
Input shape: (None, 1, 30, 2, 2048)
Output shape: (None, 1, 26, 2, 64)
High_level_features_per_7.5h_pooling (AveragePooling3D)
Input shape: (None, 1, 26, 2, 64)
Output shape: (None, 1, 8, 2, 64)
High_level_features_per_7.5h_batch_normalization (BatchNormalization)
Input shape: (None, 1, 8, 2, 64)
Output shape: (None, 1, 8, 2, 64)
Descriptor (Flatten)
Input shape: (None, 1, 8, 2, 64)
Output shape: (None, 1024)
Dropout (Dropout)
Input shape: (None, 1024)
Output shape: (None, 1024)
Dense_1 (Dense)
Input shape: (None, 1024)
Output shape: (None, 128)
Dense_2 (Dense)
Input shape: (None, 128)
Output shape: (None, 128)
Output (Dense)
Input shape: (None, 128)
Output shape: (None, 9)

Figure 5: CNN architecture diagram showing the flow from raster plot input to PHQ-9 score prediction.

Model Training and Hyperparameter Optimization

The convolutional neural network (CNN) was developed and optimized to predict continuous PHQ-9 scores from raster plots generated from smartphone usage data. This section details the training process and the systematic hyperparameter optimization performed to achieve robust and accurate predictions while minimizing overfitting.

Training and Validation Split

The dataset was divided into:

- Training set (80%): Used to allow the model to learn patterns in smartphone usage behavior associated with different levels of depression severity.
- Validation set (20%): Used to monitor the model's performance on unseen data during training and guide hyperparameter tuning.

A 5-fold cross-validation strategy was employed to validate the model's generalizability. In this approach, the dataset was split into five subsets, with the model trained on four subsets and validated on the fifth. The results from all five folds were averaged to provide a comprehensive evaluation.

Optimization and Loss Function

The CNN was optimized using the Adam optimizer, an adaptive learning algorithm that dynamically adjusts the learning rate to improve convergence. The mean squared error (MSE) loss function was used, aligning with the study's objective to minimize the difference between predicted and true PHQ-9 scores.

Hyperparameter Optimization

To determine the optimal configuration for the CNN, a systematic grid search was conducted, varying two key hyperparameters:

1. Dropout Rate: Regularization rates ranging from 0.5 to 0.98 were evaluated to reduce overfitting and enhance the model's ability to generalize.
2. Model Complexity: The number of trainable parameters in the CNN, ranging from 11,171 to 3,857,545, was varied by adjusting the number of layers and neurons in the fully connected layers.

The evaluation focused on three primary metrics:

1. Fraction Classified (f):
    - Predicted scores were categorized into:
        - Negative cases: PHQ-9 scores <= 10.
        - Positive cases: PHQ-9 scores >= 15.
        - Uncertain cases: Scores between 10 and 15, reflecting cases where the model's confidence was low.
    - f represents the percentage of cases confidently classified as either negative or positive, excluding uncertain cases. For example, an f=75% indicates that 25% of cases were classified as uncertain.
2. Mean Absolute Error (MAE):
    - Measures the average magnitude of prediction errors across all cases, providing an overall assessment of accuracy.
3. Balanced Classification Metric (TNR + TPR)/2:
    - The True Negative Rate (TNR) and True Positive Rate (TPR) were averaged to evaluate the model's performance in detecting both negative and positive cases.

Grid Search Results:

- Dropout Rate: A dropout rate of 0.95 yielded the best balance between underfitting and overfitting, achieving high f, low MAE, and strong balanced performance.

- Model Complexity: Models with 52,945 trainable parameters consistently performed best, balancing sufficient capacity for learning complex patterns with the risk of overfitting.

Visualizations:

- Heatmaps were generated to illustrate the effects of dropout rates and model complexity on f, MAE, and balanced metrics. These results guided the selection of the final model configuration.

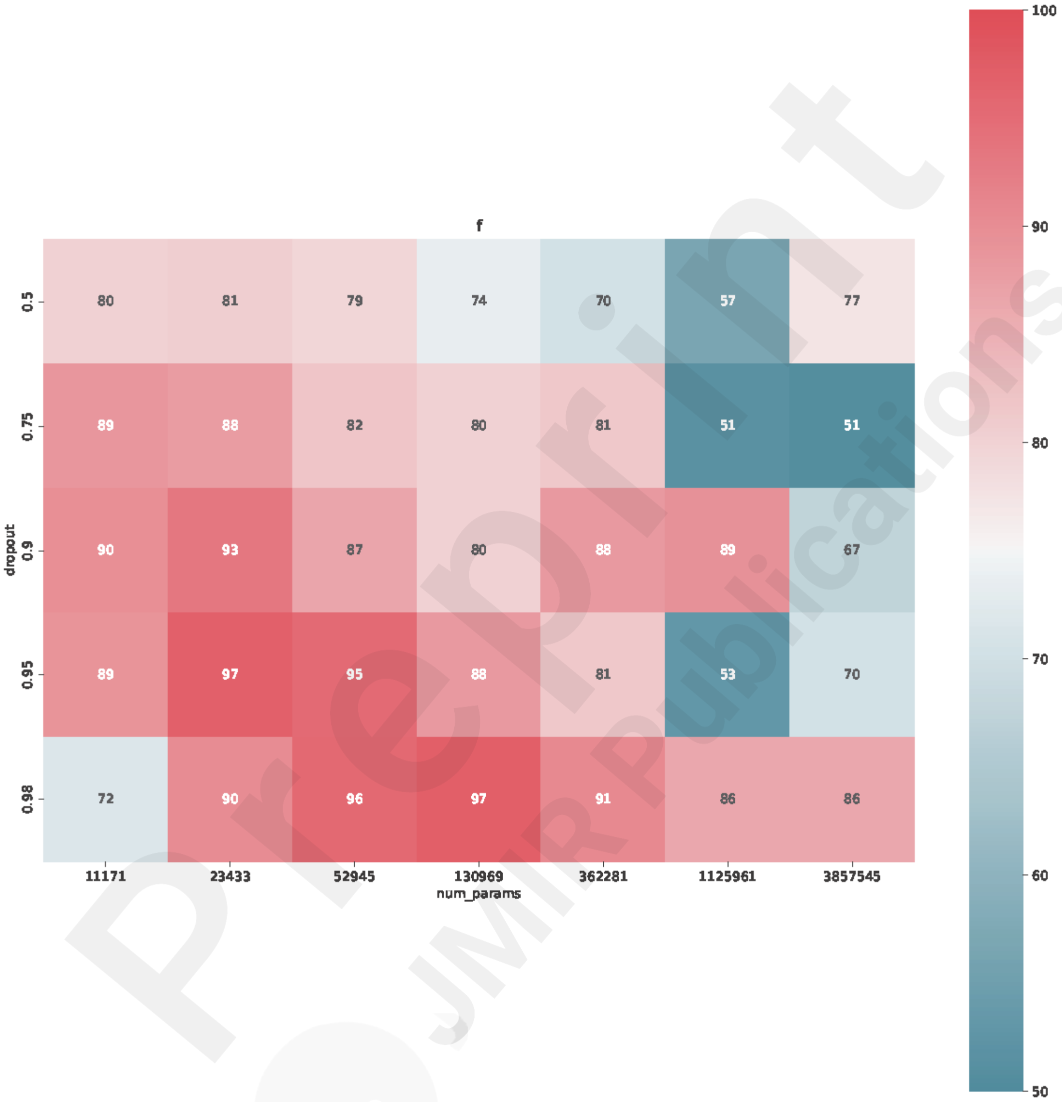


Figure 6: Heatmap of Fraction Classified (f) Across Dropout Rates and Model Complexity

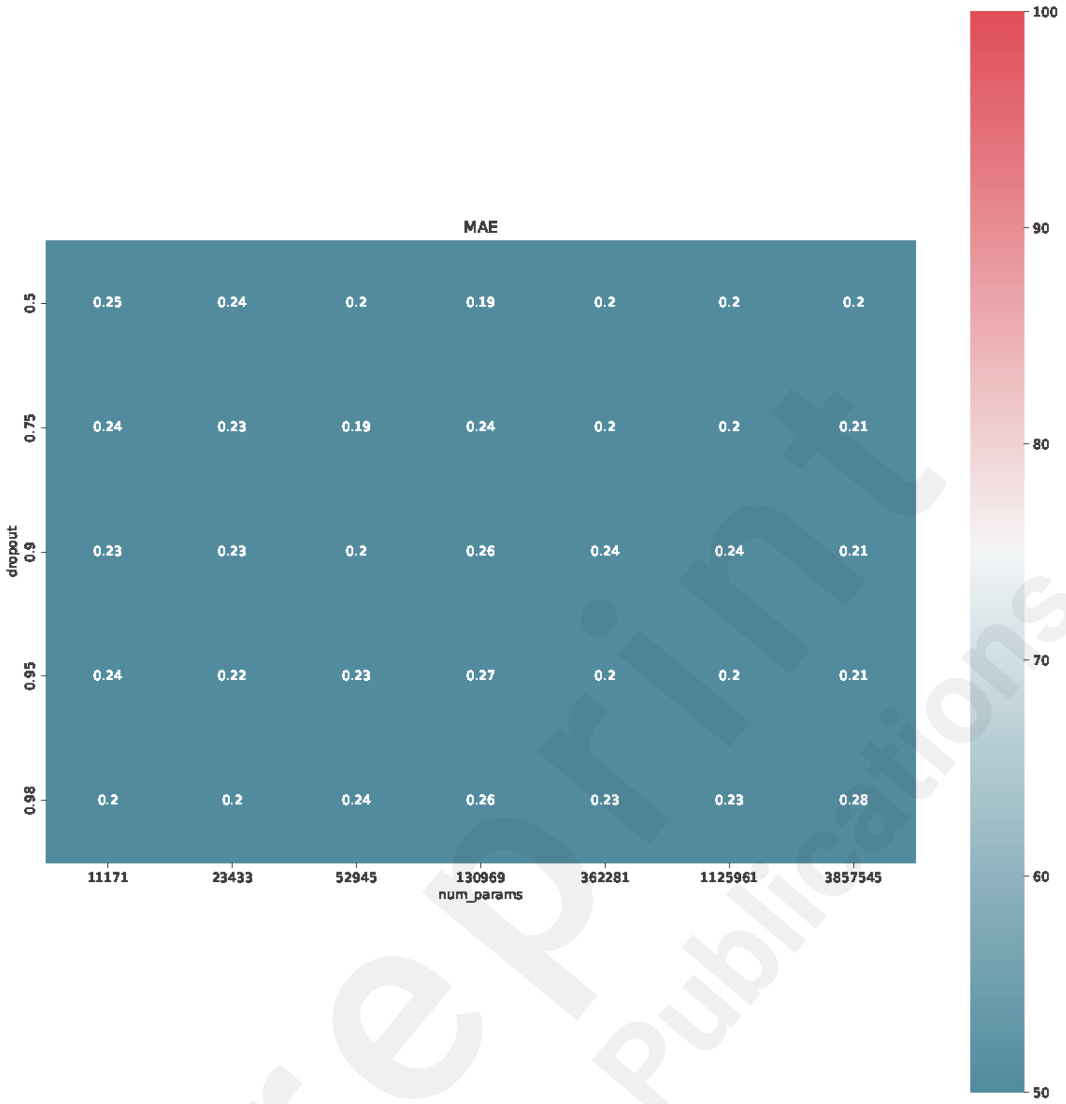
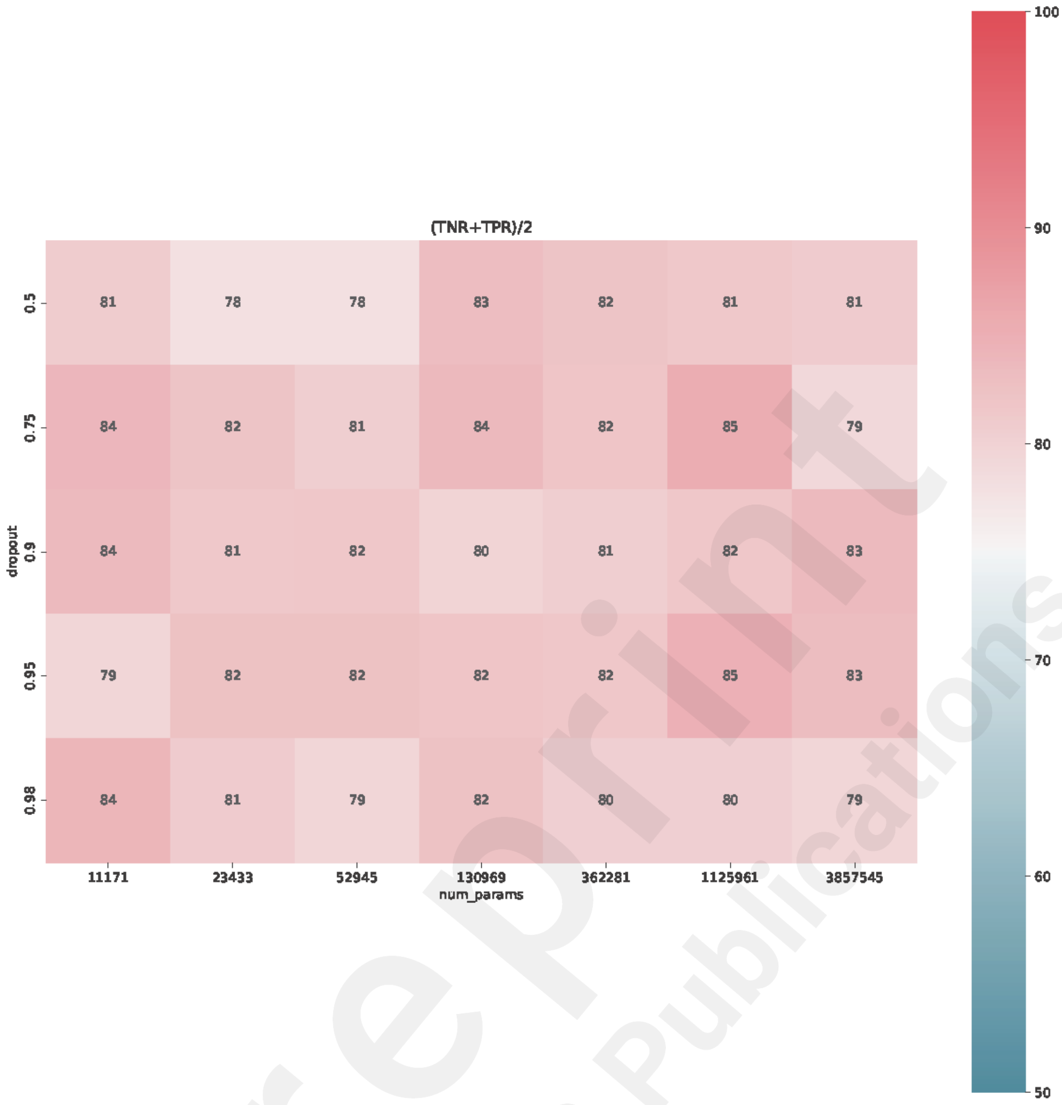


Figure 7: Heatmap of Mean Absolute Error (MAE) Across Dropout Rates and Model Complexity



dropout	11171	23433	52945	130969	362281	1125961	3857545
0.5	81	78	78	83	82	81	81
0.75	84	82	81	84	82	85	79
0.9	84	81	82	80	81	82	83
0.95	79	82	82	82	82	85	83
0.98	84	81	79	82	80	80	79

Figure 8: Heatmap of Balanced Metric ((TNR + TPR)/2) Across Dropout Rates and Model Complexity

Training Process

The final CNN configuration was trained using the following settings:

- Dropout Rate: 0.95
- Trainable Parameters: 52,945
- Epochs: Training was conducted for up to 100 epochs, but early stopping was applied when validation loss did not improve for 10 consecutive epochs. This ensured optimal training duration without overfitting.
- Batch Size: A batch size of 32 was used, balancing computational efficiency and training stability.

Performance Monitoring

Throughout training, the following metrics were monitored:

1. Mean Absolute Error (MAE): Provided a measure of overall prediction accuracy for continuous PHQ-9 scores.
2. Fraction Classified (f): Tracked the percentage of cases confidently classified as either

negative or positive.

3. True Negative Rate (TNR) and True Positive Rate (TPR): Assessed the model’s ability to correctly identify non-depressed and depressed cases, respectively.

These metrics ensured that the CNN not only provided accurate predictions but also achieved high confidence in classifications, meeting the study’s objectives of robust PHQ-9 score prediction and reliable classification.

Summary of Final Configuration

- Dropout Rate: 0.95
- Trainable Parameters: 52,945
- MAE: Achieved a low error of 0.23
- Fraction Classified (f): Achieved f =95%, indicating that only 5% of cases fell into the uncertain range (10–15 PHQ-9 scores).

The selected configuration balances accurate predictions, confident classifications, and generalizability, aligning with the study’s goals of advancing depression monitoring through passive smartphone data.

Continuous Tracking Framework

To assess the ability of the behavioral phenotyping model to provide daily predictions of PHQ-9 scores, continuous tracking was implemented as part of the study. The model utilized passive smartphone usage data, generating daily PHQ-9 score predictions that were aligned with periodic self-reported PHQ-9 questionnaire responses. This dual approach allowed for validation of the model's predictions and evaluation of its ability to capture symptom dynamics.

Data from individual users were plotted to illustrate trends in model predictions over time, highlighting instances where:

1. Predictions aligned closely with self-reported PHQ-9 scores, demonstrating accuracy.
2. Fluctuations in predictions captured behavioral variability between static assessments.
3. Discrepancies between predictions and self-reports indicated areas for further investigation.

The data were analyzed to:

- Identify consistent users, where predictions closely matched self-reports.
- Highlight cases with substantial variability, emphasizing the model’s sensitivity to changes in depressive symptoms.
- Evaluate uncertain cases where predictions fell between ranges, providing insights into the model's robustness.

Results

Model Performance for PHQ-9 Prediction

The primary goal of this study was to predict continuous PHQ-9 scores and provide accurate classifications for depression severity. The CNN model outperformed the benchmark random forest model across all performance metrics, demonstrating its ability to leverage raster plots for nuanced predictions.

Table 1: Comparison of performance metrics between CNN and random forest models.

Method	*Metric*	*Negative Class (PHQ-9 <= 10)*	*Positive Class (PHQ-9 >= 15)*	*Overall Accuracy*
Random Forest	Precision	67.9%	73.4%	70.4%
	Recall	76.2%	62.8%	
	F1-Score	71.5%	66.9%	
CNN	Precision	75.6%	90.3%	83.1%
	Recall	88.1%	78.2%	
	F1-Score	81.2%	84.0%	

The CNN achieved an overall accuracy of 83.1%, significantly higher than the random forest's 70.4%. It also demonstrated superior precision and recall, particularly for positive cases (depression severity > 15), with a precision of 90.3% compared to 73.4% for the random forest.

Classification Confidence and Fraction Classified (*f*)

The CNN model was evaluated based on its ability to confidently classify cases as either positive or negative, with cases in the PHQ-9 range of 10–15 considered uncertain. The fraction classified (*f*) represents the percentage of cases confidently classified, excluding uncertain predictions.

- Fraction Classified (*f*): The model classified 95% of cases confidently, with only 5% falling into the uncertain range.
- The grid search optimization confirmed that a dropout rate of 0.95 and model complexity of 52,945 trainable parameters maximized *f*, enabling both high accuracy and confident classifications.

Demographic Analysis of Model Performance

The performance of the CNN model was further evaluated across different demographic groups to ensure generalizability. Metrics were assessed based on gender and age categories.

Table 2: Performance of PHQ-9 prediction across gender and age groups.

Gender	*N*	*True Negative Ratio (TNR)*	*True Positive Ratio (TPR)*
Female	101	81.2%	60.7%
Male	112	84.0%	89.7%
Other	57	84.6%	68.9%

Age Group	*N*	*TNR*	*TPR*
18–25	134	68.3%	82.8%
26–35	30	80.5%	40.2%
36–55	56	86.4%	75.0%
Other	46	89.5%	55.6%

The CNN model achieved the highest TPR (82.8%) for younger participants (18–25 years), indicating its effectiveness in detecting depression severity in this group. Male participants showed the strongest overall performance, with TPR reaching 89.7%.

Prediction of Individual PHQ-9 Items

A secondary objective was to predict individual PHQ-9 items, enabling symptom-specific analysis. The CNN model achieved the following mean absolute error (MAE) values for each PHQ-9 item.

Table 3: Mean Absolute Error (MAE) for individual PHQ-9 item predictions.

PHQ-9 Item	*MAE*
Q1 (Anhedonia)	0.87
Q2 (Depressed Mood)	0.88
Q3 (Sleep Disturbances)	0.88
Q4 (Low Energy)	0.93
Q5 (Appetite Changes)	0.89
Q6 (Feelings of Failure)	0.94
Q7 (Concentration Issues)	0.95
Q8 (Motor Activity)	0.81
Q9 (Suicidal Ideation)	0.85

The lowest MAE was observed for Q8 (Motor Activity) at 0.81, while the highest error was seen for Q7 (Concentration Issues) at 0.95. These results demonstrate the model's ability to provide granular insights into specific depressive symptoms.

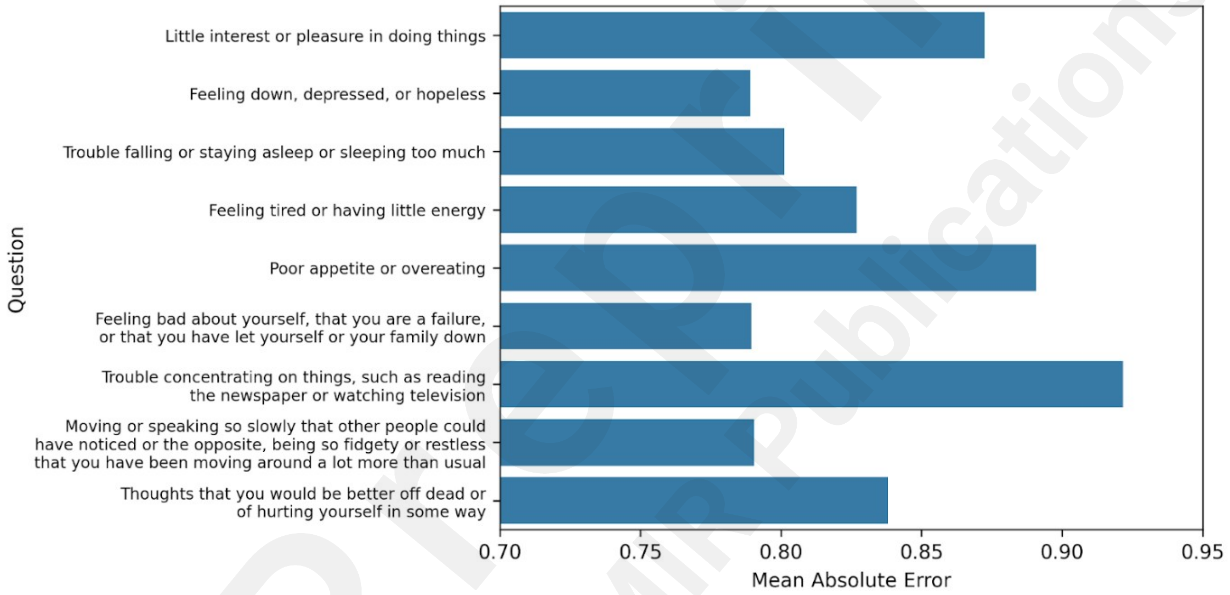


Figure 9: Graph showing the Mean Absolute Error (MAE) for each individual PHQ-9 item prediction by the CNN model. Lower MAE values indicate higher accuracy in predicting symptom-specific scores, with the lowest error observed for Q8 (Motor Activity) and the highest error for Q7 (Concentration Issues).

Comparison with Independent Symptom Models

In addition to predicting all PHQ-9 items together, independent models were trained for each item. The joint model consistently outperformed the independent models, achieving lower MAE values across most items.

Table 4: Comparison of MAE between the joint model and independent models for PHQ-9 item predictions.

PHQ-9 Item	*Joint Model MAE*	*Independent Model MAE*
Q1 (Anhedonia)	0.87	0.92
Q2 (Depressed Mood)	0.88	0.94
Q3 (Sleep Disturbances)	0.88	0.90

PHQ-9 Item	*Joint Model MAE*	*Independent Model MAE*
Q4 (Low Energy)	0.93	0.95
Q5 (Appetite Changes)	0.89	0.94
Q6 (Feelings of Failure)	0.94	0.95
Q7 (Concentration Issues)	0.95	0.97
Q8 (Motor Activity)	0.81	0.84
Q9 (Suicidal Ideation)	0.85	0.81

The joint model showed a clear advantage, particularly for items like Q1 (Anhedonia) and Q8 (Motor Activity), demonstrating the benefits of learning interdependencies among symptoms.

Continuous Tracking of PHQ-9 scores

The model's ability to continuously track depressive symptoms was evaluated by plotting daily PHQ-9 predictions for individual users alongside their periodic questionnaire responses. Selected examples demonstrate the model's strengths and highlight key insights:

- ***Consistent Tracking: User ID 50524:***
    - As shown in Figure 10, User ID 38867 demonstrates a consistent alignment between the model's predictions and self-reported PHQ-9 scores (denoted by "X" marks). Over an extended period, the predicted scores exhibit minimal fluctuations, closely mirroring self-reported scores across multiple time points. This user exemplifies the model's ability to maintain accuracy and reliability in stable depressive trajectories, reinforcing its suitability for longitudinal mental health monitoring.
    - Key Observations:
        - The model maintained strong consistency with self-reported scores over the study period.
        - Predicted scores aligned particularly well during periods of stable depression severity, as indicated by minimal variability between daily predictions and questionnaire responses.

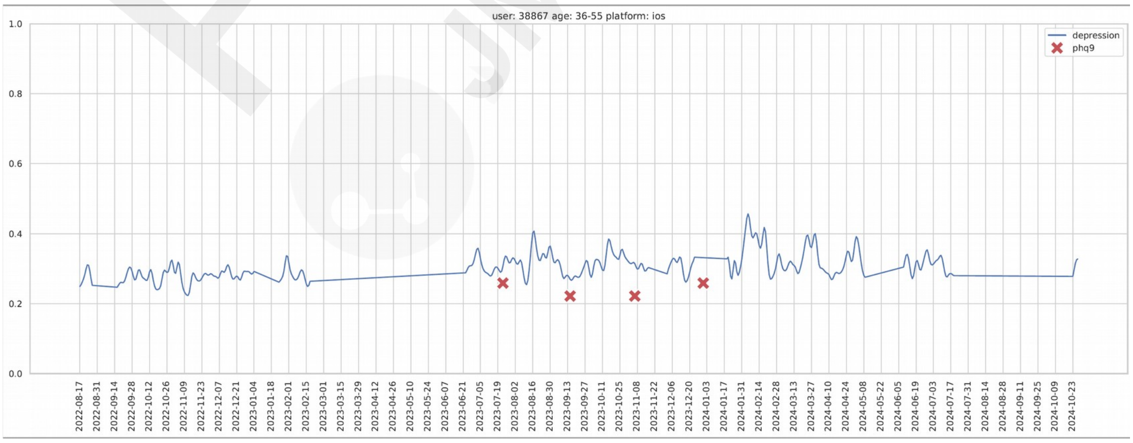


Figure 10: Continuous tracking of PHQ-9 scores for User ID 38867, demonstrating close alignment between daily predictions and self-reports over time

- ***Fluctuating Trends: User ID 44126:***
    - In contrast, Figure 11 showcases User ID 56084, whose predicted PHQ-9 scores

exhibit substantial fluctuations over time. The model captures rapid increases and decreases in predicted scores, reflecting dynamic changes in depressive symptom severity. These fluctuations align with self-reported PHQ-9 scores at critical time points, validating the model's sensitivity to variations in user behavior and symptom expression.

- Key Observations:
  - The model effectively tracked shifts in symptom severity, capturing temporal patterns of worsening and improvement.
  - Predicted scores showed alignment with self-reports during periods of symptom escalation and recovery, demonstrating the model's utility in capturing real-time variability.

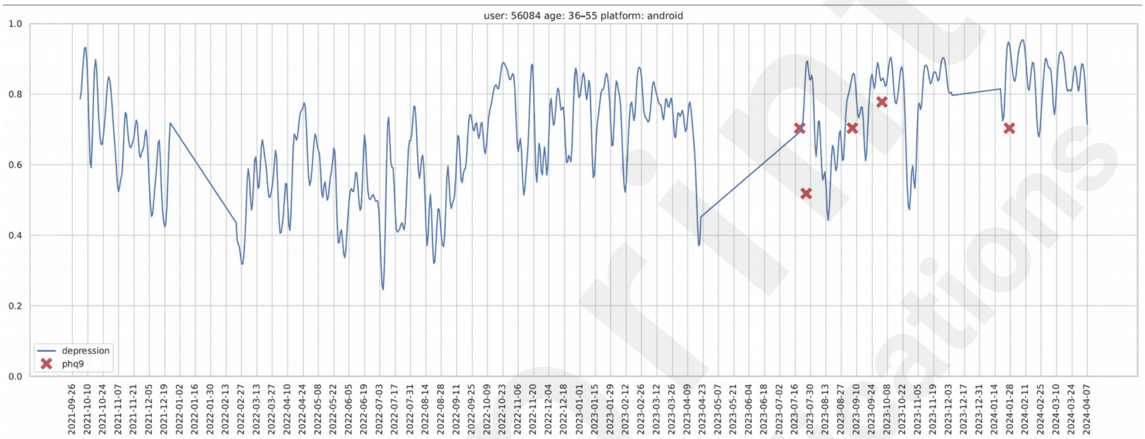


Figure 11: Dynamic tracking of PHQ-9 scores for User ID 56084, highlighting the model's sensitivity to temporal variations and its ability to reflect rapid changes in symptom severity

These results underscore the model's capability for continuous monitoring, capturing both consistent and dynamic symptom profiles across diverse cases.

Discussion

Key Findings and Their Alignment with Study Objectives

This study developed a CNN-based framework for continuous depression monitoring using smartphone-derived raster plots, with three primary goals: (1) to predict continuous PHQ-9 scores, (2) to enable symptom-specific analysis, and (3) to provide confident classifications. The results affirm that:

- The CNN model outperformed traditional random forests, achieving an accuracy of 83.1%, with a fraction classified (f) of 95%, ensuring confident classifications for most cases.
- Symptom-specific predictions were successful, with the lowest MAE observed for motor activity (Q8) and slightly higher errors for more abstract symptoms like concentration issues (Q7).
- Demographic analysis highlighted robust generalizability, particularly among younger participants and males.

These findings establish the feasibility of leveraging smartphone data for non-intrusive, continuous mental health monitoring, marking a significant step forward in digital phenotyping.

Interpretation of Results and Implications

Superior Performance of CNN and Raster Plots

The CNN model's superior performance highlights its ability to effectively model the complex, non-linear relationships inherent in behavioral data derived from smartphone usage. Unlike traditional methods such as random forests, which rely on manually engineered features, the CNN leverages raster plots to extract intricate temporal and spatial patterns. This innovation addresses a critical limitation of prior approaches and advances the capability of digital phenotyping systems in several key ways.

Capturing Temporal Variations

Raster plots provide a detailed visual representation of smartphone usage over time, encoding app interaction frequency and duration as distinct color channels. This temporal resolution enables the model to identify behavioral patterns that correlate strongly with depressive symptoms. For example:

- Sleep Disturbances (PHQ-9 Item 3):
    - Temporal irregularities, such as prolonged late-night phone usage or frequent app switching during typical sleeping hours, are preserved in the raster representation. These patterns, which might signal insomnia or disrupted circadian rhythms, are detectable because the CNN processes sequential data over 48-hour windows.
    - Prior models using manually averaged features (e.g., mean app usage) lose this temporal granularity, obscuring important behavioral fluctuations.
- Low Energy (PHQ-9 Item 4):
    - Decreased smartphone engagement during active hours, reflected as reduced intensity in raster plots, indicates reduced motivation or fatigue—a hallmark symptom of depression. The CNN captures these declines in engagement over time, providing a richer context than static features like total daily screen time.

Modeling Spatial Relationships

Raster plots also encode spatial relationships across app categories, mapping usage into an interpretable structure. The CNN can identify inter-category behavioral dependencies, such as:

- Shifts from high social media usage to increased time in communication apps, which may reflect withdrawal or seeking reassurance.
- Reduced diversity in app usage, where a participant engages predominantly with a single app category (e.g., streaming) rather than distributing usage across multiple categories. Such behaviors can signify rumination or escapism, common in depression.

Traditional methods fail to capture these interdependencies because they treat app usage as isolated variables rather than part of a cohesive behavioral profile.

Overcoming Limitations of Manual Feature Engineering

Manually engineered features, such as average phone usage, frequency of app switching, or inferred sleep time, often oversimplify behavioral data, discarding subtle patterns that are critical for nuanced depression detection. Raster plots overcome this limitation by serving as a high-dimensional input that preserves:

- Temporal consistency: By representing behavior over consecutive intervals, the model avoids reliance on aggregated data that masks variability.
- Contextual richness: Raster plots integrate app category and temporal usage, providing a holistic view of behavioral changes.

Evidence from Model Performance

The results substantiate the advantages of this approach:

- The CNN achieved an accuracy of 83.1%, significantly higher than the random forest's 70.4%, validating the added value of raster plots in capturing predictive behavioral signals.
- For positive cases (PHQ-9 scores >15), the CNN demonstrated superior precision (90.3%) and recall (78.2%), reflecting its ability to reliably identify severe depression.
- The model's lower mean absolute error (MAE) across all PHQ-9 predictions further emphasizes the efficacy of using raster plots to model complex behaviors.

Advancing Digital Phenotyping

This study is among the first to integrate raster plots into digital phenotyping for mental health monitoring, offering a methodological innovation that bridges the gap between passive data collection and actionable clinical insights. By providing a structured, image-like representation of behavioral data:

- The CNN leverages techniques originally developed for medical imaging, adapting them to the unique challenges of behavioral signal processing.
- This approach extends the potential of digital phenotyping to address disorders with multifaceted behavioral manifestations, such as depression.

Scientific Implications

The ability of raster plots to preserve both temporal and spatial patterns in behavioral data opens new avenues for digital phenotyping:

- Scalability: Raster plots can be applied to diverse behavioral datasets beyond smartphone usage, such as wearables or smart home systems, broadening the scope of automated mental health monitoring.
- Generalizability: The success of raster plots in this study suggests that similar representations could enhance other domains of AI-driven healthcare, such as sleep tracking or substance use monitoring.

Granular Symptom-Specific Analysis

This study represents a significant advancement in depression monitoring by enabling predictions for individual PHQ-9 items, moving beyond traditional binary classifications to a granular, symptom-level analysis. By linking specific smartphone usage patterns to distinct depressive symptoms, the CNN model demonstrates its capacity to provide actionable insights for personalized care. The following results highlight key strengths and limitations of symptom-specific predictions.

High Accuracy for Observable Physical Symptoms

The model exhibited the lowest mean absolute error (MAE) for motor activity (Q8) and sleep disturbances (Q3):

- Motor Activity (Q8): The CNN achieved an MAE of 0.81 for motor activity, reflecting its strong ability to infer physical symptoms through passive data collection. Patterns such as reduced app switching or prolonged inactivity periods during active hours likely contributed to this high accuracy.
  - Scientific Context: Motor activity is a behavioral manifestation of psychomotor retardation or agitation, commonly observed in depression. These symptoms are often directly observable in smartphone usage, such as decreased frequency of app launches or prolonged engagement with passive apps (e.g., video streaming).
  - Clinical Implications: The ability to reliably monitor motor activity can aid in tracking treatment response for symptoms like lethargy or restlessness, providing clinicians

with quantifiable metrics for intervention adjustments.

- Sleep Disturbances (Q3): Temporal irregularities in phone usage, such as late-night app activity or reduced phone interaction during expected wake hours, were well-captured by the CNN. These patterns align with known disruptions in circadian rhythms associated with depression.
    - Clinical Implications: Sleep disturbances are both a symptom and a risk factor for worsening depression. Accurate monitoring of sleep-related behaviors can enable early detection of symptom escalation, particularly in longitudinal care settings.

Challenges in Predicting Cognitive and Emotional Symptoms

The model displayed higher MAE values for cognitive symptoms such as:

- Concentration Issues (Q7):
    - An MAE of 0.95 indicates greater difficulty in predicting concentration-related symptoms based on smartphone usage. Unlike physical symptoms, concentration issues may not directly manifest in measurable smartphone behaviors.
    - For example, reduced engagement with productivity apps or increased time on passive apps (e.g., streaming) might signal concentration problems, but these patterns can overlap with general usage trends unrelated to depression.
    - Scientific Context: Concentration issues are influenced by both internal states and external environmental factors, making them less reliably inferred from passive data alone.
- Feelings of Failure (Q6):
    - An MAE of 0.94 highlights a similar limitation in detecting emotional states that lack clear behavioral correlates. Unlike physical symptoms, feelings of failure may not translate into observable changes in phone interaction patterns.
    - Implications for Data Augmentation: The inclusion of multimodal data, such as text sentiment analysis from messaging apps or self-reported contextual inputs, could enhance predictions for these abstract symptoms.

Importance of Symptom-Specific Predictions

Symptom-specific analysis provides an unparalleled opportunity for precision mental health care, addressing limitations of traditional models that treat depression as a homogeneous condition. By quantifying symptom severity, the CNN enables:

- Dynamic Treatment Adjustments: Clinicians can monitor specific symptoms over time and tailor interventions accordingly. For instance:
    - Persistent sleep disturbances (Q3) might prompt pharmacological intervention targeting circadian rhythms.
    - Motor activity data (Q8) could help evaluate the effectiveness of physical activity interventions or monitor side effects of medications like sedatives.
- Symptom Targeting in Therapy: Behavioral therapies such as cognitive behavioral therapy (CBT) or interpersonal therapy could be adapted based on the specific symptoms a patient exhibit.

Contribution to Personalized Care

The ability to track individual symptoms addresses a critical gap in depression monitoring, where aggregate scores often mask underlying symptom heterogeneity. Personalized monitoring offers several advantages:

1. Granularity: Captures variations in symptom severity that may be missed when focusing on total PHQ-9 scores.

2. Scalability: Passive monitoring reduces the need for frequent in-clinic visits, enabling personalized care at scale.
3. Timeliness: Continuous symptom tracking facilitates early intervention, reducing the likelihood of symptom escalation or relapse.

Broader Scientific Implications

These findings contribute to the growing body of research emphasizing the multidimensional nature of depression. By linking behavioral patterns to specific symptoms, this study advances digital phenotyping toward:

- Improved Clinical Utility: Clinicians can gain deeper insights into symptom progression, enhancing the precision of diagnosis and treatment.
- Enhanced Research Applications: Symptom-specific predictions allow for more nuanced analyses in clinical trials, improving the evaluation of treatment efficacy at the symptom level.

Confidence in Classifications

The fraction classified (f) metric is a pivotal aspect of this study, demonstrating the model's ability to make confident and reliable predictions for the majority of cases while transparently identifying instances of uncertainty. With an f value of 95%, the CNN model confidently categorized cases as either negative (PHQ-9 scores < 10) or positive (PHQ-9 scores > 15), leaving only 5% in the uncertain range (PHQ-9 scores between 10 and 15). This confidence in classifications has significant implications for both the interpretability of the model and its practical application in clinical settings.

Handling Ambiguity in Predictions

The exclusion of uncertain cases from binary classification ensures that predictions falling within the ambiguous range (10–15) are not misrepresented as definitively positive or negative. This approach reflects a transparent handling of model outputs:

- Transparency in Predictions: By flagging uncertain cases, the model avoids overconfidence, a common pitfall in AI systems, particularly in sensitive applications like mental health monitoring.
- Clinical Relevance: In practice, ambiguous cases might warrant additional clinical assessment or follow-up rather than being treated as definitive predictions. This approach aligns with clinical workflows, where borderline cases are often subject to further evaluation.

Ethical AI Principles

The f metric supports the ethical deployment of AI in healthcare by adhering to principles of accountability and interpretability:

- Avoiding Misdiagnosis: Misclassifying cases in the uncertain range could lead to inappropriate interventions or missed opportunities for treatment. By explicitly flagging these cases, the system ensures that clinicians are aware of prediction limitations and can make informed decisions.
- Building Trust in AI: A transparent classification process builds trust among users and clinicians, who are more likely to adopt AI tools that acknowledge their limitations.

Balancing Confidence with Coverage

The high classification confidence (f = 95%) suggests that the model achieves a favorable balance between confidently classifying most cases and avoiding overconfidence in ambiguous predictions. This is particularly significant for:

- High-Stakes Scenarios: In mental health care, false positives (incorrectly labeling someone as depressed) and false negatives (failing to detect depression) can have serious consequences. A high f minimizes the likelihood of such errors by limiting classifications to cases where the model demonstrates clear confidence.

- Scalability: The fraction classified ensures that the system can handle large-scale deployments without overwhelming clinicians with false alarms or ambiguous results.

Benchmarking Against Traditional Models

Traditional models often lack explicit mechanisms for handling uncertain cases, forcing all predictions into binary categories. This can result in:

- Increased error rates in borderline cases, where predictions are inherently less reliable.
- Reduced trust in the model's outputs, as it fails to communicate the underlying uncertainty in its predictions.

In contrast, the CNN model's use of the f metric provides a more nuanced framework for classification:

- Higher Confidence Thresholds: By setting explicit boundaries for confident predictions, the model ensures that classifications align with clinically meaningful thresholds for PHQ-9 scores.
- Improved Interpretability: Clinicians can trust that predictions marked as confident are based on robust evidence from behavioral data, while flagged cases signal areas requiring further attention.

Implications for Clinical Decision-Making

The f metric directly supports clinical decision-making by:

- Reducing Cognitive Load: Clinicians can focus on confidently classified cases, streamlining workflows and improving efficiency in resource-constrained settings.
- Facilitating Tailored Interventions: Cases flagged as uncertain can be prioritized for more detailed assessments, such as self-reports or in-depth clinical evaluations.

Implications of Continuous Tracking

Capturing Stable and Dynamic Symptom Trends

The examples of User ID 38867 and User ID 56084 underscore the model's dual ability to accurately track both stable and dynamic depressive trajectories:

- Stable Cases (User ID 38867): The close alignment between predictions and self-reports demonstrates the model's reliability in cases with minimal symptom variability. This is critical for long-term monitoring of patients with stable depression severity, allowing clinicians to assess treatment efficacy over time without frequent in-person evaluations.
- Dynamic Cases (User ID 56084): The ability to detect rapid fluctuations in depressive severity highlights the model's sensitivity to changes in user behavior. This capability is particularly valuable for identifying early signs of symptom escalation or improvement, enabling timely interventions in clinical practice.

Complementing Traditional PHQ-9 Assessments

The integration of continuous tracking with periodic self-reports bridges the gap between static assessments and real-time symptom monitoring:

- Self-reported PHQ-9 scores provide validated measures of depression severity but are limited by their infrequent administration.
- Continuous tracking offers a high-resolution view of symptom changes, capturing daily variability that static tools cannot detect. For example, User ID 56084's rapid symptom

changes between questionnaire points demonstrate the added value of daily predictions.

This complementary approach enhances the accuracy and responsiveness of depression monitoring systems, aligning with the goals of precision mental health care.

Implications for Personalized Care

The insights provided by continuous tracking enable clinicians to deliver more personalized and proactive mental health care:

- For stable cases (User ID 38867): Continuous monitoring can confirm treatment stability, reducing the need for frequent clinical visits while maintaining confidence in symptom management.
- For dynamic cases (User ID 56084): The model can alert clinicians to symptom changes in real-time, prompting early interventions such as therapy adjustments or medication reviews.

Novel Contributions to the Literature

This study advances the field in several critical ways:

1. Data Representation Innovation: The introduction of raster plots represents a methodological leap, transforming raw smartphone data into an interpretable, high-dimensional format that preserves temporal dynamics.
2. Continuous PHQ-9 Prediction: Unlike prior models focused on binary classifications, this framework provides nuanced predictions that reflect the full spectrum of depression severity.
3. Symptom-Specific Insights: The ability to predict individual PHQ-9 items addresses a major gap in the literature, enabling granular monitoring of depressive symptoms.
4. Confidence Metrics: The explicit handling of uncertain cases sets a new benchmark for transparency in AI-driven health monitoring.

Practical and Clinical Implications

Applications in Clinical Practice

The passive, non-intrusive nature of this system makes it particularly suited for:

- Personalized Mental Health Care: By tracking symptom-specific trends, clinicians can tailor interventions to target specific depressive symptoms dynamically.
- Early Detection: The ability to capture subtle behavioral changes enables early identification of depressive episodes, reducing delays in treatment.
- Longitudinal Monitoring: Continuous data collection allows for monitoring treatment progress over time without the need for frequent in-clinic visits.

Decentralized Clinical Trials

This approach aligns with the needs of decentralized clinical trials by:

- Reducing respondent burden through passive data collection.
- Providing high-resolution, longitudinal data for more accurate tracking of intervention effects.
- Enhancing participant retention by minimizing active reporting requirements.

Limitations and Future Directions

Despite its strengths, this study has several limitations:

1. Dependence on Smartphone Usage:
    - The reliance on smartphone behavior excludes populations with low or inconsistent device engagement, such as older adults or individuals in low-resource settings.
    - Future work should integrate multimodal data (e.g., wearables, environmental

sensors) to improve inclusivity.

2. Indirect Measurement of Depression:
   - Behavioral proxies, such as app usage, may not fully capture abstract depressive symptoms like suicidal ideation or concentration issues. Combining smartphone data with contextual inputs (e.g., self-reports) could enhance prediction accuracy.
3. Real-Time Feedback:
   - The current framework is retrospective, limiting its immediate clinical applicability. Future iterations should incorporate real-time alerts to enable proactive interventions.
4. Ethical and Privacy Concerns:
   - Passive data collection raises concerns about consent, transparency, and data security. Developing robust privacy-preserving mechanisms, such as federated learning, will be critical for adoption.
5. Demographic Variability:
   - While the model performed well for younger populations and males, further tuning is required to address variability in older adults and underrepresented groups.

Conclusion

This study presents a novel framework for passive depression monitoring, leveraging raster plots and CNNs to predict PHQ-9 scores with high accuracy and confidence. The results demonstrate:

- Superior performance in capturing temporal and spatial behavioral patterns.
- Granular insights into symptom-specific manifestations of depression.
- Strong potential for scalable and personalized mental health monitoring.

Future work should focus on incorporating multimodal data, enhancing real-time capabilities, and addressing demographic variability to ensure equitable and impactful deployment. This approach lays the foundation for scalable, accessible, and clinically relevant digital phenotyping solutions.

References

Bufano, P., Laurino, M., Said, S., Tognetti, A., & Menicucci, D. (2023). Digital Phenotyping for Monitoring Mental Disorders: Systematic Review. *Journal of Medical Internet Research*. https://doi.org/10.2196/46778

Choudhary, S., & Srinivasan, G. (2022). The Importance of Using Binary Classification Models in Predicting Depression from a Machine Learning Perspective. *Digital Medicine and Healthcare Technology*. https://doi.org/10.5772/dmht.12

Choudhary, S., Thomas, N., Ellenberger, J., Srinivasan, G., & Cohen, R. (2022). A Machine Learning Approach Detecting Digital Behavioural Patterns of Depression Using Non-intrusive Smartphone Data - A Complementary Path to PHQ-9 Assessment: A Prospective Observational Study. *JMIR Formative Research*. https://doi.org/10.2196/37736

Gadzama, W. A., Gabi, D., Argungu, M. S., & Suru, H. U. (2024). The use of machine learning and deep learning models in detecting depression on social media: A systematic literature review. *Personalized Medicine in Psychiatry*, *45*. https://doi.org/10.1016/j.pmip.2024.100125

Grzenda, A., Speier, W., Siddarth, P., Pant, A., Krause-Sorio, B., Narr, K., & Lavretsky, H. (2021). Machine Learning Prediction of Treatment Outcome in Late-Life Depression. *Frontiers in Psychiatry*, *12*, 738494. https://doi.org/10.3389/fpsyt.2021.738494

Jacobson, N. C. & Bhattacharya, S.* (2021). Digital Biomarkers of Anxiety Disorder Symptom Changes Personalized Deep Learning Models Using Smartphone Sensors Accurately Predict Anxiety Symptoms from Ecological Momentary Assessments. Behaviour Research and Therapy. https://doi.org/10.1016/j.brat.2021.104013

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. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Malhi, G. S., & Mann, J. J. (2018). Depression. *Lancet (London, England)*, *392*(10161), 2299–2312. https://doi.org/10.1016/S0140-6736(18)31948-2

Meyerhoff J, Liu T, Kording K, Ungar L, Kaiser S, Karr C, Mohr D (2021). Evaluation of Changes in Depression, Anxiety, and Social Anxiety Using Smartphone Sensor Features: Longitudinal Cohort Study. J Med Internet Res; 23(9):e22844. https://www.jmir.org/2021/9/e22844

Mall, P. K., Singh, P. K., Srivastav, S., Narayan, V., Paprzycki, M., Jaworska, T., & Ganzha, M. (2023). A comprehensive review of deep neural networks for medical image processing: Recent developments and future opportunities. *Healthcare Analytics*, *4*, 100216. https://doi.org/10.1016/j.health.2023.100216

McAllister-Williams, R. H., Arango, C., Blier, P., Demyttenaere, K., Falkai, P., Gorwood, P., Hopwood, M., Javed, A., Kasper, S., Malhi, G. S., Soares, J. C., Vieta, E., Young, A. H., Papadopoulos, A., & Rush, A. J. (2020). The identification, assessment and management of difficult-to-treat depression: An international consensus statement. *Journal of Affective Disorders*, *267*, 264–282. https://doi.org/10.1016/j.jad.2020.02.023

Mental Health Application. (n.d.). Behavidence. Retrieved January 31, 2022, from https://www.behavidence.com

Nepal, S., Pillai, A., Wang, W., Griffin, T., Collins, A. C., Heinz, M., Lekkas, D., Mirjafari, S., Nemesure, M., Price, G., Jacobson, N. C., & Campbell, A. T. (2024). MoodCapture: Depression Detection Using In-the-Wild Smartphone Images. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI Conference*, *2024*, 996. https://doi.org/10.1145/3613904.3642680

Patel, J., Hung, C., & Katapally, T. R. (2025). Evaluating predictive artificial intelligence approaches used in mobile health platforms to forecast mental health symptoms among youth: A systematic review. *Psychiatry Research*, *343*, 116277. https://doi.org/10.1016/j.psychres.2024.116277

Price, G. D., Heinz, M. V., Song, S. H., Nemesure, M. D., & Jacobson, N. C. (2023). Using digital phenotyping to capture depression symptom variability: Detecting naturalistic variability in depression symptoms across one year using passively collected wearable movement and sleep data. *Translational Psychiatry*, *13*(1), 1–10. https://doi.org/10.1038/s41398-023-02669-y

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), e4273. https://doi.org/10.2196/jmir.4273

Su, C., Xu, Z., Pathak, J., & Wang, F. (2020). Deep learning in mental health outcome research: A scoping review. *Translational Psychiatry*, *10*(1), 1–26. https://doi.org/10.1038/s41398-020-0780-3

Torous, J., Bucci, S., Bell, I. H., Kessing, L. V., Faurholt-Jepsen, M., Whelan, P., Carvalho, A. F., Keshavan, M., Linardon, J., & Firth, J. (2021). The growing field of digital psychiatry: Current evidence and the future of apps, social media, chatbots, and virtual reality. *World Psychiatry*, *20*(3), 318–335. https://doi.org/10.1002/wps.20883

Torous, J., Chan, S. R., Yee-Marie Tan, S., Behrens, J., Mathew, I., Conrad, E. J., Hinton, L., Yellowlees, P., & Keshavan, M. (2014). Patient Smartphone Ownership and Interest in Mobile Apps to Monitor Symptoms of Mental Health Conditions: A Survey in Four Geographically Distinct Psychiatric Clinics. *JMIR Mental Health*, *1*(1), e5. https://doi.org/10.2196/mental.4004

Torous, J., Staples, P., Shanahan, M., Lin, C., Peck, P., Keshavan, M., & Onnela, J.-P. (2015).

Utilizing a Personal Smartphone Custom App to Assess the Patient Health Questionnaire-9 (PHQ-9) Depressive Symptoms in Patients With Major Depressive Disorder. *JMIR Mental Health*, *2*(1), e8. https://doi.org/10.2196/mental.3889