Depressive Behavior Detection Using Sensor Signal Data: An Attention-based Privacy-Preserving Approach

Aijia Yuan
Indiana University
yuana@iu.edu

Edlin Garcia
Indiana University
eggarcia@iu.edu

Hongyi Zhu
University of Texas at San Antonio
hongyi.zhu@utsa.edu

Sagar Samtani
Indiana University
ssamtani@iu.edu

Abstract

Security concerns around using personally identifiable information (PII) introduces notable privacy concerns in sensor signal-based depression detection. In this study, we propose a novel attention-based privacy-preserving model that mitigates these concerns. It assigns greater weights to non-PII-releasing sensors and lesser to high-privacy risk sensors, leveraging the principles of differential privacy (DP). We compare the performance of machine learning and deep learning benchmark models with and without PII-releasing sensors. Our results underline a significant performance discrepancy, suggesting potential instability in prediction performance without these sensors. Our proposed model, with a recall, precision, F1 of 0.889, and an AUC of 0.9, illustrates that high-quality results are achievable while considering privacy. This privacy-conscious model has implications for promoting a more unobtrusive approach to mental healthcare. Furthermore, the model's potential for secure deployment in wide-reaching digital health applications and collaborative settings enhances its relevance for large-scale mental monitoring while preserving privacy.

Keywords: mental health, depression, privacy, sensor signal, machine learning.

1. Introduction

Mental health disorders have become a leading cause of disability worldwide, causing significant impacts on global human health, social and economic consequences (WHO, 2019). Depression is one of the most common and serious mental health disorders that can negatively affect our emotions, thoughts, and behaviors (APA, 2020). Approximately 280 million people globally have depression, and this group typically has higher suicide rates than those with other disorders (CDC, 2022). Thus, depression has become a salient societal concern with significant public health expenditures, such as medical costs. Standard methods for assessing depression include the 9-item Patient Health Questionnaire (PHQ-9), which is directly drawn from the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) (APA, 2020). Although widely used by clinicians to diagnose potential patients with depression, such approaches usually require interviews and self-reports, which are time-consuming, expensive, and labor-intensive. Self-reports such as the PHQ-9 survey may be susceptible to human recall biases and potentially miss out on moment-by-moment human behavioral patterns (Opoku Asare et al., 2021).

Mobile technology, such as smartphones and smartwatches, can provide cost-effective, ubiquitous, and objective ways to capture multiple dimensions of human behaviors (Miller et al., 2021). Previous studies have shown that sensor signal data can be an alternative method to assess people's mental health status (Burns et al., 2011; Bardhan et al., 2020). Given the high-dimensional and high-velocity nature of sensor signals, many past researchers have successfully leveraged machine learning (ML) and deep learning (DL)-based methods to analyze these data to identify depressive behaviors (Hussain et al., 2021; Srikanthan et al., 2021). Despite the numerous benefits of sensor signal-based analytics for mental health applications, concerns have been raised related to the privacy aspects of using sensor signal data (Al Ameen et al., 2010). In particular, using sensors that release personally identifiable information (PII), or data that can be used to identify an individual on its own or when combined with other information, could cause significant issues related to surveillance and privacy breaches (Krishnamurthy & Wills, 2010). For example, service providers might track patients' GPS locations and record sensitive voice messages without permission. Misuse of sensor data that leads to privacy breaches would limit or even damage such systems' potential benefits and adoptions.

Security and privacy are critical in AI-enabled digital health tools, especially in sensor-based mental health analysis (Shajari et al., 2023). Protecting AI models and the data they process is essential to prevent

URI: https://hdl.handle.net/10125/108885
978-0-9981331-8-8

misuse, breaches, and other cybersecurity threats (Ibrahim et al., 2020). Ensuring that sensitive health data is secure and privacy-preserving builds trust among users and facilitates wider adoption of these technologies in mental health care. Given these significant ramifications, this study aims to develop a novel attention-based privacy-preserving approach that integrates differential privacy principles directly into the scoring function of the attention mechanism by quantifying the privacy cost associated with each sensor for depression detection. This model has the potential to facilitate the trust of DL-based sensor signal analysis systems and could help promote their adoption for mental health assessments. Additionally, the potential for the model to be used collaboratively across different organizations and settings is noteworthy. This approach can support collaborative efforts in mental health monitoring by ensuring secure and privacy-preserving data sharing.

The paper is organized as follows. We first review literature on AI in secure digital health tools, sensor-based studies for depression, and general principles of attention mechanisms and differential privacy. Based on our literature review, we identify research gaps and pose research questions for study. Subsequently, we introduce the dataset used for our analysis, our research framework, and experimental design. We then present a set of benchmarking results and assess our proposed attention-based model. Finally, we discuss implications and propose future research directions.

2. Literature Review

We reviewed several areas of literature to set the foundation for this study. First, we reviewed the relevance of AI in secure digital health tools, particularly focusing on sensor-based mental health analysis. Second, we reviewed previous studies on sensor-based mental health analysis, specifically focusing on efforts made toward privacy preservation and identifying potential methodologies for depression detection. Third, we reviewed the general principle of the attention mechanism to identify its advantages and limitations for identifying the most important yet least privacy-sensitive sensors. Finally, we reviewed the concept of differential privacy to explore how it could be used to adjust the traditional attention mechanism.

2.1. AI in Secure Digital Health Tools

AI is increasingly integral to digital health tools, offering advanced capabilities for monitoring and analyzing health data. In sensor-based mental health analysis, AI can facilitate real-time, continuous assessments that complement traditional methods (Shajari et al., 2023). However, this integration raises significant security and privacy concerns, necessitating the protection of AI models and the data they process to prevent personal information misuse and breaches (Rieke et al., 2020). The deployment of AI in healthcare settings has shown vulnerabilities to adversarial attacks, data poisoning, privacy leaks, and the unauthorized use of personal health data, leading to user distrust and negative sentiments (Khalid et al., 2023; Mitchell & El-Gayar, 2023; Silva et al., 2020).

Despite these concerns, limited research focuses on the security or privacy of AI models in mental health monitoring. Past studies on passive sensing data have identified privacy concerns such as loss of confidentiality and data misuse (Rogan et al., 2024). Existing literature often neglects the balance between preserving privacy and maintaining AI models' accuracy and reliability in mental health. More importantly, such concerns could negatively impact therapeutic relationships between patients and therapists (Byrne et al., 2022). Thus, this gap highlights the need for research on integrating privacy-preserving techniques within AI models, especially for sensor-based mental health assessments such as depression detection.

2.2. ML and DL Approaches for Sensor Signal-based Depression Detection

Prior research has focused on utilizing sensor signal data for depression analysis, providing more objective and ubiquitous measures for mental health assessments (Tigga & Garg, 2023; Jiao et al., 2023; Yuan et al., 2023; Morshed et al., 2019). Most studies have primarily concentrated on analyzing various aspects of human behavior, including physical activity, social activity, and sleep patterns, to detect depression (Pfaff et al., 2022; Ding et al., 2021). These studies typically collected or analyzed passive sensor data reflecting the physical activity aspects of human behavior. For instance, mobile phones were used to record patient locations, which were then analyzed to identify mood disorders like depression or bipolar disorder (Delgado-Santos et al., 2022; Trifan et al., 2019). However, location data such as GPS coordinates are considered as PII and are thus sensitive (Ren et al., 2016). As a result, some studies proposed using self-reported survey data or non-PII-releasing sensors as alternatives. Despite these adaptations, these studies primarily focused on the data collection aspect rather than data analysis. Therefore, we place specific emphasis on studies that have analyzed sensor signal data for mental health.

Within the existing literature, classical ML, DL, and statistical analysis emerged as the most frequently used methods for sensor data analysis. Specifically, Support Vector Machines (SVM) and K-Nearest Neighbors (KNN) algorithms were widely employed due to their proficiency in generating accurate predictions on existing mental health datasets (Jiao et al., 2023; Kim et al., 2021). Nevertheless, these methods may struggle to capture sensor data's complex and non-linear relationships. To address such limitations, Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and their variants were commonly used DL methods due to their inherent ability to process high-dimensional sensor signal data without the need for manual feature engineering (Tigga & Garg, 2023; Khan et al., 2022; Anjum et al., 2022). While most studies focused on supervised learning tasks such as depression detection, only a few addressed unsupervised learning tasks like clustering analysis (Nguyen et al., 2021).

Despite the effectiveness of these methods, a significant gap persists in the literature as the privacy concerns associated with PII have not been adequately addressed. Thus, there is a need to develop a privacy-preserving approach that maintains depression detection accuracy. This challenge necessitates techniques that could help identify and prioritize less privacy-sensitive but highly predictive sensor data. One promising method is the attention mechanism, a technique in neural networks that focuses on the input data's most important features, which could be adapted for our privacy-oriented context.

2.3. Attention Mechanisms

Attention mechanisms are a technique in neural networks that assigns trainable weights to source data, helping extract the most important features of the task (Du et al., 2019). Attention mechanisms can help model performance by focusing on relevant parts of the input data. Attention mechanisms comprise a query (Q), the key (K), and the value (V) (Brauwers and Frasincar, 2021). The query represents the expected outcome or intermediate result of a neural network, the key is the source data, and the value is the internal representation extracted by the neural network (Niu et al., 2021). Attention mechanisms typically use a scoring function to compute the similarity between the Q and each key K (Samtani et al., 2021). This function assigns weights to the input features, reflecting their importance or contribution to the task. In our context, the scoring function can evaluate the importance of each sensor for depression detection.

However, this traditional attention approach does not account for privacy. Intuitively, sensors with higher privacy concerns should be seen as more sensitive or risky and, thus, should receive lower weights. This is a significant limitation of standard attention mechanisms. Hence, our objective should be to develop a modified scoring function that balances the importance of sensor data for depression detection and its associated privacy risk. This way, we can identify the most relevant sensors with the least privacy sensitivity. By incorporating differential privacy principles into the scoring function of the attention mechanism, we can potentially address the limitation of traditional attention mechanisms in addressing privacy concerns.

2.4. Differential Privacy

Differential privacy (DP) is a privacy framework and mathematical concept that provides a rigorous and quantifiable measure of privacy preservation in data analysis (Dwork, 2006). Its objective is to protect individuals' privacy by ensuring that including or excluding their data does not significantly affect the outcome or conclusions of an analysis or computation. Differential privacy applications have been explored in various domains, including healthcare, finance, social sciences, and more (Ficek et al., 2021; Byrd & Polychroniadou, 2020; Kenny et al., 2021). These studies underscore the increasing recognition of differential privacy as an effective approach for safeguarding privacy across different fields. Differential privacy also holds tremendous potential, for developing a privacy-preserving model for sensor signal depression detection.

Incorporating differential privacy principles into the scoring function allows us to assign appropriate sensor weights based on their privacy risk. This ensures that sensors with higher privacy concerns receive lower weights, reducing their impact on the model's decision-making process. Furthermore, integrating differential privacy principles can quantify the privacy cost associated with each sensor. Considering the privacy implications of the sensor data and accounting for the sensitivity of each sensor can help strike a balance between accurate depression detection and preserving user privacy.

3. Research Gaps and Questions

We identified several research gaps based on our literature review. First, while previous research in sensor-based mental health analysis has made strides in privacy preservation, the focus has primarily been on data collection rather than data analysis. This highlights a gap in understanding the privacy implications during the analysis phase. Second, although sensor data has been effectively utilized for depression detection, privacy concerns related to PII have often been overlooked. Consequently, there is a critical need to develop a privacy-preserving approach that ensures accurate depression detection while safeguarding individual privacy. Lastly, attention mechanisms, despite their potential to prioritize relevant sensor data, often fail to consider privacy risks when scoring the importance of sensors. This necessitates the development of a modified scoring function that balances sensor importance and privacy risks. Building upon these research gaps, we propose the following research questions:

- *What is the performance of ML and DL models when all available sensors are considered, and how does their performance change when privacy is taken into account by utilizing only non-PII-releasing sensors for depression detection?*
- *How can we develop an attention-based model that assigns higher weights to sensors that have the most significant impact on model performance while being the least sensitive in terms of privacy?*

4. Research Framework

To address the proposed research questions, we propose a research framework with three major components: (1) Data Collection & Data Preprocessing, (2) Proposed Attention-based Privacy-Preserving Model, and (3) Model Evaluations. We detail each component in the following sub-sections.

4.1. Data Collection and Preprocessing

We used the publicly available StudentLife dataset, which contains comprehensive phone sensor recordings from 48 students over 10 weeks (Wang et al., 2014). The dataset includes multiple categories of sensor signal data with timestamps that reflect various aspects of the student's physical activities, social activities, and sleep patterns (Table 1).

Table 1. StudentLife Dataset Overview

Sensor	**Description**	**PII**	**Feature**	**Total data points**
GPS	GPS coordinates collected every 10 minutes	Yes	timestamp	231,851
			latitude	231,851
			longitude	231,851
			altitude	231,851
Bluetooth	Surrounding Bluetooth signal every 10 minutes	Yes	timestamp	1,288,526
			signal level	1,288,526
WiFi	Surrounding Wifi signal strength	Yes	timetamp	18,429,544
			signal level	18,429,544
Conver-sation	Records when participants were around conversations	No	start_timestamp	79,023
			start_timestamp	79,023
Light	Records when at a dark environment for > an hour	No	start_time	7,254
			end_time	7,254
Phone Lock	Records when the phone was locked for > an hour	No	start_time	9,275
			end_time	9,275
Phone Charge	Records when the phone was charged for > an hour	No	start_time	3,318
			end_time	3,318

Past literature has indicated that GPS location coordinates (latitude, longitude, and altitude), Bluetooth signals, and Wi-Fi are PII-releasing sensors (Fanourakis, 2020). The non-PII-releasing sensors include surrounding sound, dark/light environment, phone lock frequency, and phone charge information of users. Besides the sensors with timestamps, the StudentLife dataset contains responses to the PHQ-9 survey collected from the students, which we used as labels in our classification tasks. 38 of the 48 were classified as having a low risk of depression, and 10 were identified as having a high risk.

To preprocess the data, we addressed the missing data, as the sensors are not always active. In cases where timestamps were included, we extracted the time of day and day of the week from the timestamp data, providing our model with potentially useful temporal information. Since we are working with DL models, we also normalized our data. This process ensures that all features have a similar scale and prevents certain features from dominating the learning process (Singh and Singh, 2020).

4.2. Proposed Attention-based Privacy-Preserving Model

Given the limitations of prevailing DL-based models and traditional attention mechanisms, we propose a novel Attention-based Privacy-Preserving Model (Figure 1) for depression detection.

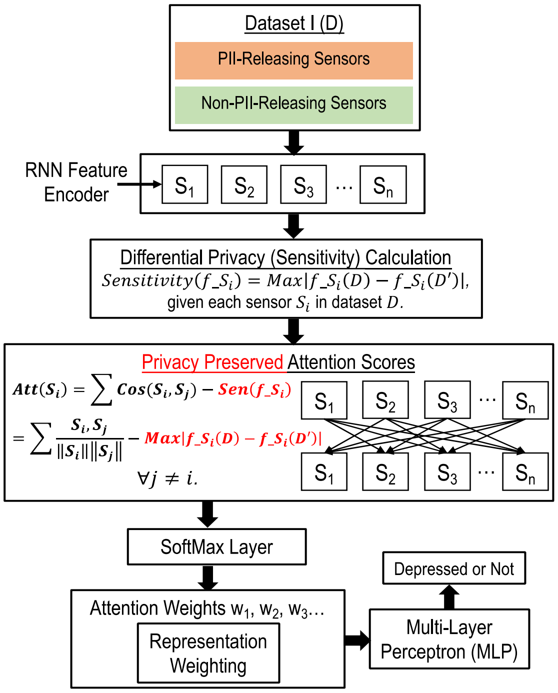
Dataset I (D)
PII-Releasing Sensors
Non-PII-Releasing Sensors
RNN Feature Encoder
S1
S2
S3
···
Sn
Differential Privacy (Sensitivity) Calculation
$Sensitivity(f\_S_i) = Max|f\_S_i(D) - f\_S_i(D')|$,
given each sensor S_i in dataset D.
Privacy Preserved Attention Scores
$Att(S_i) = \sum Cos(S_i, S_j) - Sen(f\_S_i)$
$= \sum \frac{S_i, S_j}{\|S_i\|\|S_j\|} - Max|f\_S_i(D) - f\_S_i(D')|$
$\forall j \neq i.$
SoftMax Layer
Attention Weights w1, w2, w3…
Representation Weighting
Multi-Layer Perceptron (MLP)
Depressed or Not

Figure 1. Proposed Attention-based Privacy-Preserving Model

The proposed model seeks to understand which non-PII-releasing sensors have the highest impact on model performance through a novel attention mechanism and calculate each sensor's privacy cost (sensitivity). This quantitative privacy measure helps create a balance between achieving predictive accuracy and preserving data privacy. The proposed model contains four components:

a. **Feature encoding:** Due to the complexity and high-dimensionality of sensor signal data, we initially process each feature (both PII-releasing and non-PII-releasing) with an RNN/LSTM encoder. This operation condenses the features into more compact embeddings and lowers their dimensions, helping to simplify further computations.

b. **Differential privacy term calculation:** Next, we define a function f for each sensor that captures the essence of the sensor data. For instance, f could be a function that calculates the mean sensor value. Then we calculate the sensitivity of f for each sensor, representing its privacy cost. This design was motivated by the concept of differential privacy, a paradigm stating that if we remove the information of a single user in the dataset, the output of the algorithm should not change significantly (Zhao and Chen, 2022). In our case, for each student m, we create a new version of the dataset D'_m with that student's data removed. We call these datasets D'_1, D'_2,..., D'_n for n students. Next, we calculate the absolute difference between $f(D)$ and each of $f(D')$, which gives a set of absolute changes: $|f(D) - f(D'_1)|$, $|f(D) - f(D'_2)|$, ..., $|f(D) - f(D'_n)|$.

Thus, for each sensor, $Sensitivity(f) = Max|f(D) - f(D')|$, the maximum value in this set of absolute changes. The idea here lies in understanding the sensitivity of a sensor to the absence of a student's data. Essentially, a sensor is deemed highly influential and potentially privacy-releasing/violating if its function's output greatly changes due to the removal of a single person's data. This change is what we call as "sensitivity". By capturing the worst-case change in the function's output, we obtain a robust measure of privacy risk.

c. **Attention score calculation:** For each sensor S_i (both non-PII and PII-releasing), we first calculate its similarities with all other sensors S_j using the cosine similarity measure. We then sum the similarity scores for each sensor S_i across all other sensors S_j to represent its total information contribution, as shown in (1). Also in the previous step, we calculated each sensor's sensitivity term, also called the DP term, as shown in (2).

$$\sum cos(S_i, S_j) = \sum \frac{S_i, S_j}{\|S_i\|\|S_j\|}, \quad \forall j \neq i \quad (1)$$

$$Sensitivity(f\_S_i) = Max|f\_S_i(D) - f\_S_i(D')| \quad (2)$$

Next, an attention score for each sensor S_i is calculated as the sum of its similarity scores with all other sensors S_j minus its sensitivity term. In this case, the intuition is that the more privacy-releasing a certain sensor is, the higher sensitivity it is likely to have, thus the calculated attention weight will be lower. This represents the balance between the sensor's information contribution and its privacy cost. Thus, the attention score for S_i would be $(1) - (2)$. We then normalize these attention scores using the SoftMax function to get a weight w_i for each sensor. This ensures the weights across all sensors sum to 1.

d. **Depression detection:** Finally, we assign w_i to the initial features and pass the weighted representations to three layers of multi-layer perceptron (MLP) to perform the depression detection task. We chose MLP due to its effectiveness in handling high-dimensional data and its demonstrated success in similar contexts (Yu et al., 2024). However, our approach is flexible and allows for the use of any classifier based on specific requirements and constraints.

The novelty of our proposed attention-based privacy-preserving framework lies in its integration of differential privacy principles directly into the scoring function of the attention mechanism. This integration is achieved by quantifying the privacy cost or sensitivity associated with each sensor and incorporating it into the computation of attention scores. This approach effectively balances each sensor's information contribution and its potential privacy risks.

4.3. Model Evaluations

Our model evaluation consists of two experiments designed to assess the performance and effectiveness of our proposed approach. In experiment I, we conduct a comparison between different benchmark models by considering the inclusion or exclusion of personally identifiable information (PII)-releasing sensors. Specifically, we run the models with both the inclusion and exclusion of PII-releasing sensors such as GPS, Bluetooth, and Wi-Fi. By analyzing the variations in model performance, we can understand the significance and indispensability of PII-releasing sensors in the detection of depression. In experiment II, we proceed to evaluate the full model that we have designed. In this evaluation, we assign weights to distinct sensors based on their importance and privacy risk. By incorporating the principles of differential privacy into the attention mechanism, we can effectively balance the contribution of each sensor while considering privacy concerns. This experiment allows us to examine the performance of the model when attention weights are applied to different sensors, providing insights into the effectiveness of our privacy-preserving approach.

Our benchmark model selections include classical ML-based models frequently employed in prior sensor signal analysis literature for depression detection. These models include SVM, KNN, Logistic Regression (LR), and Gradient Boosting (GB) classifier. Additionally, we incorporate prevailing DL-based techniques, such as CNN, RNN, Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (BiLSTM), and Gated Recurrent Unit (GRU) identified in related literature, as the second set of our benchmarks.

We select recall, precision, and F1-score as our evaluation metrics since these were the most commonly used in prior relevant literature. These metrics are calculated using the following formulas where True Positives (TP) refer to correctly predicted positive values, True Negatives (TN) to correctly predicted negative values, False Positives (FP) to instances where the actual class is no but the predicted class is yes, and False Negatives (FN) to instances where the actual class is yes but the predicted class is no: (1) *Precision = TPs / (TPs + FPs)*, (2) *Recall = TPs / (TPs + FNs)*, (3) *F1 = 2 ((Precision * Recall) / (Precision + Recall))*.

5. Results and Discussions

5.1. Experiment I

In experiment I, we ran ML and DL-based benchmark models with and without including PII-releasing sensors, aiming to discern differences in model performance to address our first proposed research question. Table 2 summarizes the model performance with all PII-releasing sensors vs. without PII-releasing sensors.

Table 2. ML/DL-based Benchmark Performance

ML/DL Model	All Sensors (PII & non-PII)			Non-PII Sensors Only		
	Rec.	Pre.	F1	Rec.	Pre.	F1
KNN	0.875	0.917	0.883	0.400	0.250	0.308
LR	0.750	0.857	0.750	0.875	0.917	0.883
SVM	0.500	0.278	0.357	0.500	0.222	0.308
GB	1.000	1.000	1.000	1.000	1.000	1.000
RNN	0.889	0.889	0.889	0.889	0.853	0.870
LSTM	0.889	0.860	0.874	0.889	0.821	0.853
GRU	0.889	0.889	0.889	0.889	0.889	0.889
BiLSTM	0.889	0.889	0.889	0.889	0.869	0.879
CNN	0.889	0.889	0.889	0.889	0.888	0.888

Our findings reveal that the elimination of PII sensors led to different detection performances in 7 out of 9 evaluated ML/DL models (appearing in boldface). The exclusion of PII-releasing sensors significantly affects certain ML models. For example, the KNN model exhibits a high dependence on these sensors, which sees its recall drop from 0.875 to 0.4, precision drop from 0.917 to 0.25, and F1-score drop from 0.883 to 0.308 when excluding PII-releasing sensors. Conversely, few models, such as GB, show resilience against excluding PII-releasing sensors, maintaining comparably stable performance levels. This indicates their potential as privacy-friendly predictive models. Nevertheless, this apparent disparity underlines the impact of sensor type selection on model performance. It's also important to note that traditional ML models inherently depend on more aggregated level data as they struggle with high-dimensional sensor data. This trait might contribute to the extreme values and instability we observe in the results.

In addition, our initial analysis demonstrates that DL benchmarks consistently outperform ML benchmarks on average, highlighting their superior capability and stability in sensor-based depression detection. However, the presence or absence of PII-releasing sensors also impacts the performance of DL models. Excluding PII-releasing sensors negatively affects most DL models, including RNN, LSTM, BiLSTM, and CNN, leading to average declining performance metrics. For example, the LSTM model experiences a reduction in its precision from 0.86 to 0.821 and in its F1 score from 0.874 to 0.853 when PII-releasing sensors are excluded. This suggests that including PII-releasing sensors contributes to the improved performance of these models in detecting depression. On the other hand, only GRU demonstrates higher resilience against the exclusion of PII-releasing sensors, maintaining stable performance without heavy reliance on PII-related information.

Conducting further testing and evaluation on multiple datasets is crucial to enhancing the reliability and generalizability of our findings. However, the observed performance discrepancies in both our benchmark settings highlight the instability of model performance when PII-releasing sensors are removed. These findings underscore the importance of developing a more robust privacy-preserving approach for depression detection.

5.2. Experiment II

Our proposed attention-based privacy-preserving model achieved a recall of 0.889, precision of 0.889, F1 of 0.889, and AUC of 0.9. These scores outperform most DL-based benchmarks and are on par with the best-performing DL benchmark, GRU, which also achieved 0.889 for recall, precision, and F1. These performances indicate that standard DL models' reliance on high-dimensional data, some of which may be private and sensitive, which increases the risk of privacy invasion. Our proposed model, however, is designed to strike a balance between privacy and performance by leveraging an attention mechanism that weighs sensors based on their informational contribution and privacy risk. Our model's performance proves that it's possible to maintain high-quality results while protecting privacy. Further, using an attention mechanism presents an interesting avenue for model interpretability, making our model more transparent. By assigning weights to each sensor feature, our model also provides a clear explanation of the role each sensor plays in the depression detection task, which might be beneficial for future research and model improvements.

6. Theoretical Contributions and Practical Implications

6.1. Theoretical Contributions

This study introduces a novel integration of differential privacy principles into the attention mechanisms of DL models, addressing a key limitation in traditional approaches that focus on enhancing model performance but often overlook privacy risks. By refining feature selection while proactively mitigating privacy concerns, our approach represents a theoretical advancement that bridges the gap between maximizing data utility and ensuring robust privacy protections in ML/DL models.

Our proposed attention mechanism represents a shift in how privacy is treated in AI model design for healthcare. Traditionally, privacy considerations are often addressed after a model has been developed, through external safeguards or post-hoc adjustments. In contrast, our approach embeds privacy-preserving techniques directly into the core of the model's decision-making process, making privacy an integral aspect of the model's architecture and throughout the model's lifecycle, from initial design to deployment.

By embedding differential privacy directly into AI models, our study contributes to the broader theoretical discourse on the convergence of AI and cybersecurity. It offers a new perspective on designing AI systems that are secure by design where privacy and security are foundational elements, not secondary concerns. This perspective is critical in fields like healthcare, where the sensitivity of data demands rigorous privacy safeguards. Our approach demonstrates that it is possible to build AI models that are both high-performing and privacy-conscious, opening new avenues for research that balances these often-competing objectives.

6.2. Practical Implications

The practical implications of our study extend to various stakeholders involved in mental health care, with a strong emphasis on enhancing security and privacy in mental health assessments. Our model ensures the protection of sensitive data while providing a respectful, non-intrusive tool for monitoring mental health.

Privacy-focused healthcare entities: Integrating our privacy-preserving model into healthcare systems can potentially benefit stakeholders involved in regulatory compliance, such as healthcare providers and administrative bodies. This approach supports compliance with key regulatory standards, such as the

Health Insurance Portability and Accountability Act (HIPAA) in the United States, thereby reducing the risks associated with data breaches. By bolstering user confidentiality and fostering trust, hospitals and mental health clinics can elevate their digital health solutions, creating a safer and more reliable healthcare environment. These advancements benefit healthcare providers and regulatory authorities by ensuring that privacy standards are consistently met and maintained, thereby strengthening trust in healthcare systems.

Patients: Our privacy-preserving model reassures patients by prioritizing the safety of their personal data. It balances accurate depression detection with privacy preservation, fostering trust in AI-assisted mental health technology. Patients can engage confidently with our system, reassured by its robust performance and commitment to protecting sensitive information. This increased trust may lead to broader adoption of such technologies in mental health management, paving the way for earlier detection and intervention.

Organizations: Organizations can leverage our model for large-scale mental health monitoring, enabling them to safely support the mental well-being of their employees. By integrating our model into existing healthcare systems, organizations can enhance data security and privacy while providing management with valuable insights into overall mental health trends within their workforce. This approach allows for early identification of potential issues, enabling timely interventions that improve employee well-being, reduce absenteeism, and maintain high levels of productivity. Additionally, it can help foster an organizational culture that prioritizes mental health without compromising employee privacy.

7. Discussions

Our study has implications for collaborative systems and technologies in environments where multiple stakeholders need to securely share and analyze sensitive data. Our privacy-preserving model enables this by focusing on data patterns rather than individual data points. By analyzing aggregate patterns across datasets, our model can derive meaningful insights, such as feature importance, without exposing individual-level information.

This approach allows institutions to collaborate effectively, leveraging overall trends and relationships within the aggregated data to identify key factors affecting mental health, all while ensuring individual privacy is strictly maintained. For instance, universities could collaborate to monitor the mental health of student populations across different campuses using our model. By securely aggregating data from various institutions, these universities can identify common mental health trends, such as increased anxiety or depression during exam periods, and implement targeted interventions like counseling services or stress management workshops. Similarly, organizations could use our model to monitor the mental well-being of employees across different branches or departments. By sharing data securely, organizations can detect workplace stress patterns, such as those linked to tight deadlines or high workloads and introduce wellness programs or support systems to mitigate these issues.

One limitation of our study is that the data utilized was collected on an individual level. However, this does not diminish the model's relevance for collaborative use. The core privacy-preserving principles of our model allow it to focus on aggregate data patterns rather than individual data points, enabling secure aggregation and anonymization of individual data. This makes the model suitable for collaborative analysis across multiple entities, as it can derive meaningful insights from overall trends without exposing sensitive information.

8. Conclusion

Previous studies in sensor-based mental health analysis have made valuable contributions but have often placed more emphasis on data collection than analysis and have not adequately addressed privacy preservation concerns. This study contributes to addressing the critical issue of privacy in sensor-based depression detection by developing a novel attention-based model. Our model helps address these gaps by demonstrating the feasibility and potential of building privacy-preserving ML models without sacrificing performance. By incorporating differential privacy principles into the attention mechanism's scoring function, we effectively address the privacy concerns associated with PII-releasing sensors. Furthermore, the proposed model has significant potential for use in wide-reaching digital health applications for mental health, ensuring data security and privacy on a large scale.

There are several promising directions for future research. First, it is crucial to conduct further evaluation and validation of our attention-based model on larger and more diverse datasets. This will enable us to ensure its generalizability and effectiveness across different populations and settings, establishing its robustness and reliability in real-world applications. Second, extending our current model by exploring methods to avoid using any PII-releasing sensors while still maintaining high detection accuracy is essential. One example could include developing a knowledge distillation architecture that can effectively transfer the knowledge gained from PII-releasing sensors to future tasks, even in the absence of PII data. Finally, future research could explore the development of models for federated learning, where organizations train on decentralized data without sharing raw information. This can further validate the effectiveness of privacy-preserving methods in collaborative settings, enhancing joint analysis in multi-institutional studies and mental health initiatives while ensuring user privacy.

9. References

Al Ameen, M., Liu, J., & Kwak, K. (2012). Security and privacy issues in wireless sensor networks for healthcare applications. *Journal of medical systems*, *36*, 93-101.

American Psychiatric Association. Availablle at: https://www.psychiatry.org/news-room/apa-blogs/light-sleep-and-mental-health.

Anjum, F., Alam, S., Bahadur, E. H., Masum, A. K. M., & Rahman, M. Z. (2022, February). Deep learning for depression symptomatic activity recognition. In *2022 International Conference on Innovations in Science, Engineering and Technology (ICISET)* (pp. 510-515). IEEE.

Bardhan, I., Chen, H., & Karahanna, E. (2020). Connecting systems, data, and people: A multidisciplinary research roadmap for chronic disease management. *MIS Quarterly*, *44*(1), 185-200.

Brauwers, G., & Frasincar, F. (2021). A general survey on attention mechanisms in deep learning. *IEEE Transactions on Knowledge and Data Engineering*, *35*(4), 3279-3298.

Burns, M. N., Begale, M., Duffecy, J., Gergle, D., Karr, C. J., Giangrande, E., & Mohr, D. C. (2011). Harnessing context sensing to develop a mobile intervention for depression. *Journal of medical Internet research*, *13*(3), e1838.

Byrd, D., & Polychroniadou, A. (2020, October). Differentially private secure multi-party computation for federated learning in financial applications. In *Proceedings of the First ACM International Conference on AI in Finance* (pp. 1-9).

Byrne, S., Tohamy, A., Kotze, B., Ramos, F., Starling, J., Karageorge, A., ... & Harris, A. (2022). Using a mobile health device to monitor physiological stress for serious mental illness: A qualitative analysis of patient and clinician-related acceptability. *Psychiatric Rehabilitation Journal*, *45*(3), 219.

Cai, H., Han, J., Chen, Y., Sha, X., Wang, Z., Hu, B., ... & Gutknecht, J. (2018). A pervasive approach to EEG-based depression detection. *complexity*, *2018*(1), 5238028.

Cao, J., Truong, A. L., Banu, S., Shah, A. A., Sabharwal, A., & Moukaddam, N. (2020). Tracking and predicting depressive symptoms of adolescents using smartphone-based self-reports, parental evaluations, and passive phone sensor data: development and usability study. *JMIR mental health*, *7*(1), e14045.

Centers for Disease Control and Prevention. Avaliable at: https://www.cdc.gov/suicide/facts/index.html.

Delgado-Santos, P., Stragapede, G., Tolosana, R., Guest, R., Deravi, F., & Vera-Rodriguez, R. (2022). A survey of privacy vulnerabilities of mobile device sensors. *ACM Computing Surveys (CSUR)*, *54*(11s), 1-30.

Ding, X., Clifton, D., Ji, N., Lovell, N. H., Bonato, P., Chen, W., ... & Zhang, Y. T. (2020). Wearable sensing and telehealth technology with potential applications in the coronavirus pandemic. *IEEE reviews in biomedical engineering*, *14*, 48-70.

Du, M., Liu, N., & Hu, X. (2019). Techniques for interpretable machine learning. *Communications of the ACM*, *63*(1), 68-77.

Dwork, C. (2006, July). Differential privacy. In *International colloquium on automata, languages, and programming* (pp. 1-12). Berlin, Heidelberg: Springer Berlin Heidelberg.

Fanourakis, M. (2020). A report on personally identifiable sensor data from smartphone devices. *arXiv preprint arXiv:2003.06159*.

Ficek, J., Wang, W., Chen, H., Dagne, G., & Daley, E. (2021). Differential privacy in health research: A scoping review. *Journal of the American Medical Informatics Association*, *28*(10), 2269-2276.

Gerych, W., Agu, E., & Rundensteiner, E. (2019, January). Classifying depression in imbalanced datasets using an autoencoder-based anomaly detection approach. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)* (pp. 124-127). IEEE.

Hussain, Z., Waterworth, D., Aldeer, M., Zhang, W. E., Sheng, Q. Z., & Ortiz, J. (2021, September). Do you brush your teeth properly? an off-body sensor-based approach for toothbrushing monitoring. In *2021 IEEE International Conference on Digital Health (ICDH)* (pp. 59-69). IEEE.

Ibrahim, A., Thiruvady, D., Schneider, J. G., & Abdelrazek, M. (2020). The challenges of leveraging threat intelligence to stop data breaches. *Frontiers in Computer Science*, *2*, 36.

Jiao, Y., Wang, X., Liu, C., Du, G., Zhao, L., Dong, H., ... & Liu, Y. (2023). Feasibility study for detection of mental stress and depression using pulse rate variability metrics via various durations. *Biomedical Signal Processing and Control*, *79*, 104145.

Kenny, C. T., Kuriwaki, S., McCartan, C., Rosenman, E. T., Simko, T., & Imai, K. (2021). The use of differential privacy for census data and its impact on redistricting: The case of the 2020 US Census. *Science advances*, *7*(41), eabk3283.

Khalid, N., Qayyum, A., Bilal, M., Al-Fuqaha, A., & Qadir, J. (2023). Privacy-preserving artificial intelligence in healthcare: Techniques and applications. *Computers in Biology and Medicine*, 106848.

Khan, D. M., Masroor, K., Jailani, M. F. M., Yahya, N., Yusoff, M. Z., & Khan, S. M. (2022). Development of wavelet coherence EEG as a biomarker for diagnosis of major depressive disorder. *IEEE Sensors Journal*, *22*(5), 4315-4325.

Kim, J., Hong, J., & Choi, Y. (2021, July). Automatic depression prediction using screen lock/unlock data on the smartphone. In *2021 18th International Conference on Ubiquitous Robots (UR)* (pp. 1-4). IEEE.

Krishnamurthy, B., & Wills, C. E. (2009, August). On the leakage of personally identifiable information via online social networks. In *Proceedings of the 2nd ACM workshop on Online social networks* (pp. 7-12).

Miller, K., Baugh, C. W., Chai, P. R., & Hasdianda, M. A. (2021, January 5). Deployment of a wearable biosensor system in the emergency department: A technical feasibility study. In *Hawaii International Conference on System Sciences (HICSS)* (pp. 3567-3572).

Mitchell, D., & El-Gayar, O. (2023). Discovering mHealth users' privacy and security concerns through social media mining. In *Hawaii International Conference on System Sciences (HICSS),* (pp. 3267-3276).

Morshed, M. B., Saha, K., Li, R., D'Mello, S. K., De Choudhury, M., Abowd, G. D., & Plötz, T. (2019). Prediction of mood instability with passive sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *3*(3), 1-21.

Nguyen, B., Kolappan, S., Bhat, V., & Krishnan, S. (2021, November). Clustering and feature analysis of smartphone data for depression monitoring. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (pp. 113-116). IEEE.

Niu, Z., Zhong, G., & Yu, H. (2021). A review on the attention mechanism of deep learning. *Neurocomputing*, *452*, 48-62.

Opoku Asare, K., Terhorst, Y., Vega, J., Peltonen, E., Lagerspetz, E., & Ferreira, D. (2021). Predicting depression from smartphone behavioral markers using machine learning methods, hyperparameter optimization, and feature importance analysis: exploratory study. *JMIR mHealth and uHealth*, *9*(7), e26540.

Pfaff, E. R., Haendel, M. A., Kostka, K., Lee, A., Niehaus, E., Palchuk, M. B., ... & Chute, C. G. (2022). Ensuring a safe (r) harbor: Excising personally identifiable information from structured electronic health record data. *Journal of Clinical and Translational Science*, *6*(1), e10.

Ren, J., Rao, A., Lindorfer, M., Legout, A., & Choffnes, D. (2016, June). Recon: Revealing and controlling pii leaks in mobile network traffic. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services* (pp. 361-374).

Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., ... & Cardoso, M. J. (2020). The future of digital health with federated learning. *NPJ digital medicine*, *3*(1), 1-7.

Rogan, J., Bucci, S., & Firth, J. (2024). Health Care Professionals' Views on the Use of Passive Sensing, AI, and Machine Learning in Mental Health Care: Systematic Review With Meta-Synthesis. *JMIR Mental Health*, *11*, e49577.

Samtani, S., Chai, Y., & Chen, H. (2022). Linking Exploits from the Dark Web to Known Vulnerabilities for Proactive Cyber Threat Intelligence: An Attention-Based Deep Structured Semantic Model. *MIS quarterly*, *46*(2).

Shajari, S., Kuruvinashetti, K., Komeili, A., & Sundararaj, U. (2023). The emergence of AI-based wearable sensors for digital health technology: a review. *Sensors*, *23*(23), 9498.

Silva, W., Sacramento, C., Silva, E., Garcia, A. C. B., & Ferreira, S. B. L. (2020, January). Health Information, Human Factors and Privacy Issues in Mobile Health Applications. In *Hawaii International Conference on System Sciences (HICSS)* (pp. 1-10).

Singh, D., & Singh, B. (2020). Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, *97*, 105524.

Srikanthan, S., Asani, F., Patel, B. K., & Agu, E. (2021, September). Smartphone TBI Sensing using Deep Embedded Clustering and Extreme Boosted Outlier Detection. In *2021 IEEE International Conference on Digital Health (ICDH)* (pp. 122-132). IEEE.

Tigga, N. P., & Garg, S. (2023). Efficacy of novel attention-based gated recurrent units transformer for depression detection using electroencephalogram signals. *Health Information Science and Systems*, *11*(1), 1.

Trifan, A., Oliveira, M., & Oliveira, J. L. (2019). Passive sensing of health outcomes through smartphones: systematic review of current solutions and possible limitations. *JMIR mHealth and uHealth*, *7*(8), e12649.

Wang, R., Wang, W., DaSilva, A., Huckins, J. F., Kelley, W. M., Heatherton, T. F., & Campbell, A. T. (2018). Tracking depression dynamics in college students using mobile phone and wearable sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(1), 1-26.

Wang, T., Li, C., Wu, C., Zhao, C., Sun, J., Peng, H., ... & Hu, B. (2020). A gait assessment framework for depression detection using kinect sensors. *IEEE Sensors Journal*, *21*(3), 3260-3270.

World Health Organization. Available at: https://www.who.int/news-room/fact-sheets/detail/depression.

Xu, X., Chikersal, P., Doryab, A., Villalba, D. K., Dutcher, J. M., Tumminia, M. J., ... & Dey, A. K. (2019). Leveraging routine behavior and contextually-filtered features for depression detection among college students. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *3*(3), 1-33.

Yu, S., Chai, Y., Samtani, S., Liu, H., & Chen, H. (2024). Motion Sensor–Based Fall Prevention for Senior Care: A Hidden Markov Model with Generative Adversarial Network Approach. *Information Systems Research*, *35*(1), 1-15.

Yuan, A., Xu, M., Zhu, H., Samtani, S., & Garcia, E. (2023, July). Towards Privacy-Preserving Depression Detection: Experiments on Passive Sensor Signal Data. In *2023 IEEE International Conference on Digital Health (ICDH)* (pp. 115-117). IEEE.

Zhao, Y., & Chen, J. (2022). A survey on differential privacy for unstructured data content. *ACM Computing Surveys (CSUR)*, *54*(10s), 1-28.