Journal of Information Systems Engineering and Management
2025, 10(57s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

LifeGuardAI : A Privacy-Preserving Multimodal Framework for Suicide and Self-Harm Risk Detection on Social Media Using Blockchain and Hybrid Deep Learning.

Hadj Ahmed Bouarara[1], Kadda Benyahia[2]

[1] *Department of computer science, GeCoDe laboratory university of saida algeria*

[2] *Department of computer science, LTC laboratory university of saida algeria*

ARTICLE INFO

Received: 20 Oct 2024

Revised: 10 Feb 2025

Accepted: 25 Feb 2025

ABSTRACT

Introduction: Suicide and self-harm risks are increasingly prevalent on social media, making early detection and intervention a critical societal need. However, analyzing sensitive user-generated content for such risks raises significant challenges, especially in terms of privacy protection, data heterogeneity, and the need for effective multimodal and multilingual understanding.

Objectives: This work aims to develop a collaborative and privacy-preserving system **LifeGuardAI** capable of accurately detecting and preventing suicide and self-harm risks across multiple social media platforms, while strictly maintaining user confidentiality.

Methods: LifeGuardAI leverages personalized federated learning to handle heterogeneous and non-IID user data distributed over various platforms. Multimodal and multilingual content is processed using advanced deep learning models, including BLIP-2 and BERT. To ensure privacy, all training and model updates are encrypted via homomorphic encryption, and model aggregation is performed through blockchain-based mechanisms for added security and data integrity. The system performs continuous, day-by-day temporal analysis of each user's posts across all connected social networks, providing individualized and comprehensive risk assessments. Upon identifying potential risks, LifeGuardAI automatically alerts trusted contacts to enable timely intervention..

Results: In extensive evaluations, LifeGuardAI achieved a detection accuracy of **98.8%**, outperforming conventional and state-of-the-art models such as CNN (80.3%), Decision Tree (90.78%), SVM (88.96%), Transformer (97.4%), BiLSTM-CNN (69.83%), ResNet50 2D (97.59%), and VGG16 2D (98.73%). These results highlight the system's superior performance and its robustness across heterogeneous, real-world data.

Conclusions: LifeGuardAI demonstrates that it is possible to deliver highly accurate, personalized, and privacy-preserving detection of suicide and self-harm risks across diverse social media platforms. Its collaborative architecture and advanced privacy mechanisms make it a promising solution

Journal of Information Systems Engineering and Management
2025, 10(57s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

for proactive mental health intervention while upholding the highest standards of user security and ethical AI deployment.

Keywords: Suicidal slf-harm detection, Personalised Federated learning; social media; deep learning.

1. INTRODUCTION

Human beings today face unprecedented levels of competition and pressure. Teenagers, in particular, inevitably experience various psychological challenges related to academics, social interactions, emotional development, and self-identity. Rapid changes in society and the economy often leave many young people feeling confused and overwhelmed, as their self-awareness and coping skills are still developing. Today's youth are experiencing even greater psychological pressure due to the rise of social media: according to the Digital 2024 report, 95% of 16-24 year olds use Instagram, TikTok, Snapchat, Facebook, or Reddit daily. The WHO reported in 2023 that anxiety and depression among adolescents have surged by 80% in the past decade, mainly because of cyberbullying and constant social comparison. A French study from 2024 found that 60% of young people suffer from anxiety linked to intensive social media use, and nearly one in four show significant symptoms of depression. In some tragic cases, a young person has posted depressive and self-harm content online before dying by suicide, emphasizing the urgent need to detect and respond to such warning signs. When these signals such as depressive posts, suicidal thoughts, or images of self-harm are identified early, friends, family, or professionals can intervene and potentially prevent suicide or self-harm.

Because of the worrying rise in suicide and depression rates around the world, finding suicidal self-harm using machine learning (ML) and deep learning (DL) techniques has become an important area of research. These technologies make it possible to find people at risk early and help them quickly, which could save lives by giving them proactive support. Researchers have looked into a lot of different ML and DL models in the last few years. They often use data from social media and other digital platforms to find signs of suicidal thoughts with more and more accuracy and speed. Support Vector Machines (SVM) in[10] and Logistic Regression are two of the classical ML methods that have worked very well. Studies have shown that SVM can be up to 95% accurate and Logistic Regression can be up to 93.96% accurate, showing that they can make very good predictions on certain datasets. Also, models like the Random Forest and Passive Aggressive Classifier have both been shown to be useful for finding people who are likely to commit suicide, with an accuracy of about 93% [10]. XGBoost [11], a popular ensemble method, has also worked well with text features, achieving 91.5% accuracy, though more complex deep learning architectures have sometimes done better.

Journal of Information Systems Engineering and Management
2025, 10(57s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Deep learning models have shown even more promise in picking up on the nuances of suicidal thoughts because they can process and understand language and context. Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks have done very well. CNN got 92.73% correct, and LSTM got 92.84% correct, which is a little better. Researchers have also looked into hybrid models that use both CNN and LSTM. These models have been able to reach accuracies of about 92.26% by taking advantage of the strengths of both architectures. More complex structures, like the Bidirectional Gated Recurrent Unit (Bi-GRU) combined with CNN [9], have shown even more promise, with accuracies of 93.07% and 92.47% across a range of datasets. The CNN-BiLSTM hybrid model [11] is one of the best setups, with an amazing 95% accuracy. It beats traditional ML methods like XGBoost and shows how useful it is to combine spatial and sequential data processing into one framework [5]. The type and quality of the data used have a big effect on these results [1]. Most studies use content from social media sites like Reddit, Twitter, Facebook, and Instagram, where people are more likely to share their thoughts and feelings. We use natural language processing (NLP) and sentiment analysis to find emotional signals and language patterns that might show someone is in trouble. Word embedding methods like Word2Vec, GloVe, and Skip-Gram have been used to improve textual representation, especially when used with optimization algorithms like Adam [2]. This makes the model better at picking up on the subtle differences in meaning in user posts. Some studies have used multimodal methods that combine facial gestures, voice patterns, and written language with textual data to get a more complete and accurate picture of suicidal behavior [15]. Even though there have been some promising developments, there are still a number of important issues that need to be resolved to make sure that suicidal ideation detection technologies are effective, fair, and ethical:

- **Data Privacy**: Protecting the privacy of users is still a big worry. Because mental health data is so sensitive, it needs to be properly stored, anonymized, and handled to keep it from being misused or accidentally shared.
- **Cross-Platform User Activity**: A lot of people are active on more than one platform, like Facebook, Twitter, Instagram, and Reddit. This makes it hard to make unified user profiles or keep track of behavioral patterns across all platforms.
- **Risks to security and inversion attacks**: Adversarial threats, like inversion attacks that try to get sensitive input data back from model outputs, are big problems for privacy and security that need to be fixed.
- **Different and inconsistent input data**: The content that users create can be very different in terms of language, tone, format, and subject matter, both within and between platforms. Because of this, models need to be very strong and flexible so they can handle a wide range of input that isn't always consistent.
- **Language and Resource Gaps**: Some platforms don't work well with languages that aren't spoken as much. For instance, Facebook might not have enough resources

Journal of Information Systems Engineering and Management
2025, 10(57s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

for Chinese-language content, which makes detection tools less useful in some areas and for some groups of people.

- **Problems with dataset imbalance and generalization**: Many datasets that are available are either unbalanced or too small, which makes it hard to trust the results across different user groups, settings, and cultural backgrounds [9].
- **No Longitudinal Analysis**: Most current methods don't take into account the historical context of user activity, like past tweets or posts. Risk assessments may not be as accurate if you don't take a user's history into account, and early intervention efforts may be less effective [16].
- **Problems with processing multimodal data**: It's still hard to analyze and combine different types of content, like text, images, and videos. You need special tools to get useful information out of different types of multimodal data and make sure that everyone understands it the same way.
- **Posts from users who speak more than one language**: Users often switch between languages in and between posts. This variety of languages makes it harder to find the right ones, so we need models that can handle multilingual content and code-switching [14].

In this context, we present LifeGuardAI, a new system that solves the important problems of privacy, diversity, and quick action in finding people who are at risk of suicide or self-harm on social media. Our method aims to give accurate and personalized risk assessments while strictly protecting user privacy. We do this by using advanced federated learning methods, multimodal analysis, and strong privacy-preserving mechanisms. This paper goes into great detail about how LifeGuardAI was created, put into use, and tested, making it a complete and useful solution in the larger field of mental health monitoring technologies.

The rest of this paper is set up like this. In Section 2, we talk about the datasets we used in our study, including where they came from, what they are like, and how they relate to finding people at risk of suicide and self-harm. In Section 3, we talk about the proposed LifeGuardAI solution, including its structure, main parts, and ways to protect privacy. Section 4 talks about the experimental setup and compares our method to a number of the best baseline models that are currently available. Finally, Section 5 wraps up the paper and talks about what needs to be done next.

2. THE SUICIDAL DETECTION DATASET (SDD[1]) :

The SDD is a meticulously curated collection of tweets designed to facilitate the creation and evaluation of automated systems capable of detecting suicidal ideation on social media platforms. This dataset can assist researchers and developers in constructing machine learning models capable of distinguishing between tweets expressing suicidal ideation and those that do not. The primary objective is to enhance

[1] https://www.kaggle.com/datasets/aunanya875/suicidal-tweet-detection-dataset/data

Journal of Information Systems Engineering and Management
2025, 10(57s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

natural language processing (NLP) and sentiment analysis, particularly in the context of mental health.

The collection comprises 1,778 tweets, each meticulously annotated to indicate the presence or absence of suicidal ideation. There exist two categories of tweets: "Not Suicide post," including 1,120 tweets devoid of indications of suicidal intent, and "Potential Suicide post," consisting of 658 tweets that exhibit emotional distress or potential suicidal ideation. This dataset is not fully balanced, as there are more tweets unrelated to suicide than those that are. This distribution resembles the frequency with which such tweets appear on social media in reality. Each entry in the dataset comprises two major columns. The initial column, "Tweet," contains the content of the tweet. These texts talk about a number of various things, moods, and ways of saying things. This illustrates the variability of internet communication. The second column, "Suicide," tells you what kind of suicide it is. Tweets labeled "Not Suicide post" do not express suicide ideation, whereas those labeled "Potential Suicide post" indicate suicidal thoughts, mental turmoil, or plans that may require assistance.

This dataset is highly beneficial for numerous NLP and machine learning applications, such as sentiment analysis, text classification, and early detection of mental health issues. Exhibiting both suicidal and non-suicidal content in reality facilitates the development of improved support networks and intervention tools for individuals at risk. Researchers and developers utilizing this dataset must use caution and adhere to ethical standards in its application. They must consider the sensitivity of the content, its impact on individual privacy, and the implications of their work for vulnerable populations collectively. Below are samples from the dataset that illustrate the functionality of the annotation scheme:

- “that crap took me forever to put together…” (Potential Suicide post)
- “kiwitweets hey jer since when did you start tw…” (Not Suicide post)
- “oh that’s good to hear but is it over already o…” (Not Suicide post)

3. PROPOSED SOLUTION

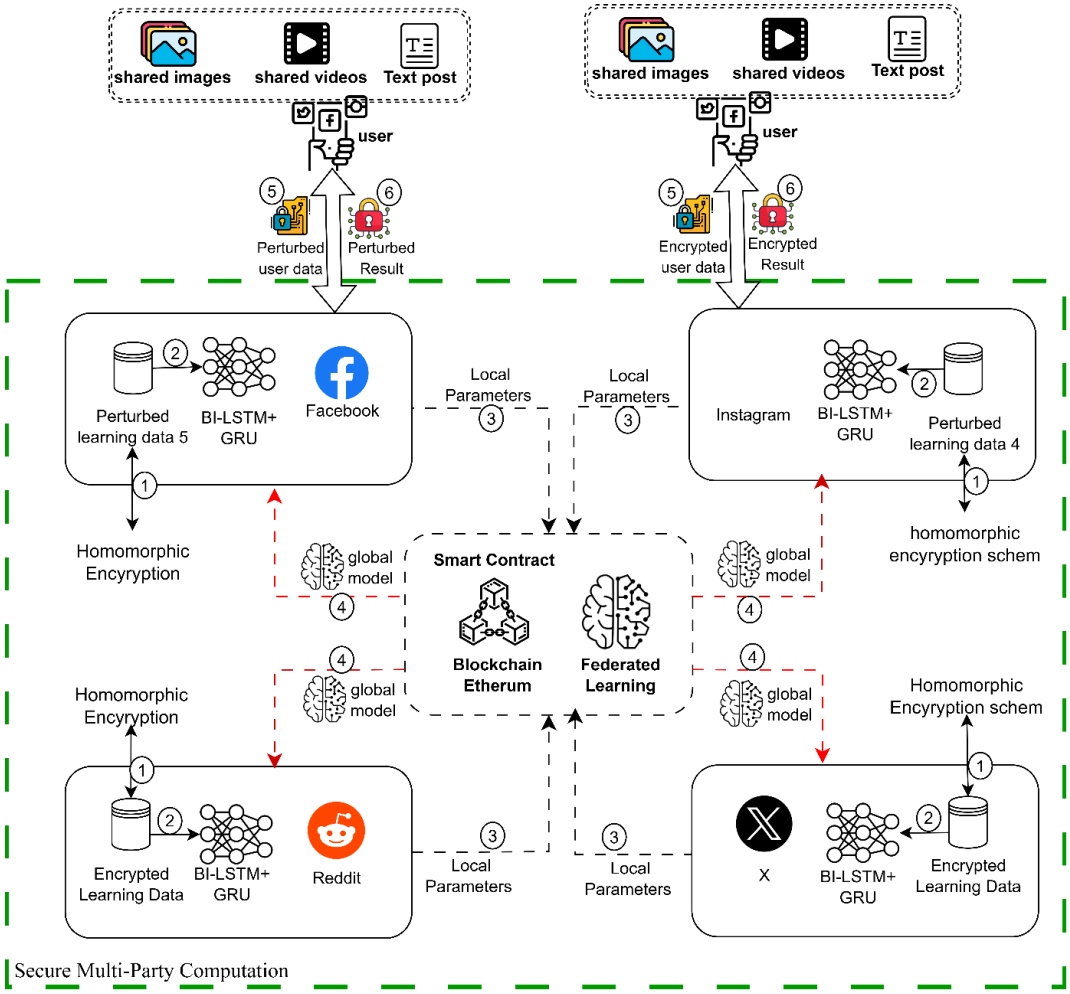
Using collaborative and privacy-preserving analysis of user-generated content from multiple social media platforms (e.g., Reddit, Facebook, Twitter), our system presents a novel framework for the early detection of psychological distress and suicidal behaviors. By using its own proprietary data to train a local model, each platform engages in federated learning, which guarantees data sovereignty and privacy while also allowing the system to manage heterogeneous and multimodal data sources, such as text, images, and videos. In order to further protect user privacy, we use fully homomorphic encryption (FHE), which makes it possible to train and infer models directly on encrypted data while guaranteeing that even participating servers cannot access sensitive user data.

Journal of Information Systems Engineering and Management
2025, 10(57s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

We incorporate extra security measures like differential privacy and gradient obfuscation to mitigate the risks of model inversion attacks, in which adversaries try to reconstruct private user information from shared model parameters. By taking these steps, the possibility of information leakage during cooperative model updates is greatly decreased. Additionally, we use smart contracts implemented on the Ethereum blockchain to ensure decentralization, transparency, and trust in the aggregation of model updates. By removing single points of failure and improving the integrity of the federated learning process, the integration of blockchain technology allows for the decentralized, immutable, and auditable aggregation of model parameters.

Our architecture also utilizes personalized and adaptive models capable of effectively processing the diverse and heterogeneous data from multiple cultural and contextual backgrounds, ensuring highly accurate and contextualized risk assessments. When a high risk of suicidal ideation or self-harm is detected, the system issues clear and actionable alerts to the affected user and, in an ethical and secure manner, to their trusted network, thereby facilitating timely intervention and proactive prevention.



Journal of Information Systems Engineering and Management
2025, 10(57s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Fig. 1. System overview of LifeGuardAI. **1)** Fully homomorphic encryption is applied on learning data. **2)** social media Model Development: Each social media platform trains a personalized recurrent neural network (RNN) on the encrypted text (Figure 2) and the obtained common layers (weights + biases) are submitted to the federated learning and blockchain. **3)** Global Model Aggregation: The system performs federated averaging (FedAvg) across all platforms common layers parameters and A consensus global model is generated through blockchain-verified aggregation **4)** Model Personalization: social media platform fine-tune the global model while maintaining the core architecture Custom input/output layers are added to adapt to the local user inputs of each platform **5)** Secure Data Collection Loop: new user data are encrypted and transmitted to the platform via secure channels. **6)** the hospital uses the fine tuned global model to predict the results (encrypted output), and sent them back to user and its friends, who can obtain cipher results using multiparty security computation protocol (see figure 2).

3.1. MULTIMODAL DATA PROCESSING

We suggest integrating the BLIP-2 framework (Bootstrapping Language-Image Pre-training) [17] to improve our solution for identifying people who are at risk of suicide on social media. An advanced model called BLIP-2 can convert user-shared visual content, like pictures or videos, into pertinent textual descriptions. This makes it possible to analyze posts that aren't just text. Additionally, we will use the BERT model, which allows for the automatic translation of Arabic and French texts into English, to handle content written in these languages. This step guarantees that our main model for identifying depression, self-harm, or suicide risk can effectively analyze all messages, regardless of their original language. Our system can offer a more thorough and accurate analysis of users' psychological well-being by taking into consideration the variety of formats and languages used across social media platforms because of this multimodal and multilingual approach.

3.2. DATA USER PRIVACY

Our system makes use of homomorphic encryption, a cryptographic technique that enables computations to be carried out directly on encrypted data without ever disclosing the underlying sensitive information, to guarantee maximum privacy and data security throughout the learning process. To prevent external parties or central servers from accessing the raw data or any intermediate computations, each social media platform uses homomorphic encryption to encrypt its local training data prior to the start of the training phase. Three cutting-edge fully homomorphic encryption (FHE) schemes—CKKS, BGV, and FHEE—have been tested to ensure reliable encryption of both user and training data, as shown in Figure 4. The Microsoft SEAL library, which offers dependable and effective tools for implementing homomorphic encryption in real-world scenarios, was used for benchmarking and practical implementation.

We usually utilized a polynomial modulus degree of 8192 and a coefficient modulus of 218 bits for the Cheon-Kim-Kim-Song (CKKS) scheme, which enables deep learning operations on encrypted data and facilitates effective approximate arithmetic on encrypted real or complex numbers. In order to facilitate precise computations on

Journal of Information Systems Engineering and Management
2025, 10(57s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

encrypted integers for a variety of machine learning tasks, the BGV (Brakerski-Gentry-Vaikuntanathan) scheme was set up with a polynomial modulus degree of 4096, a plaintext modulus of 786433, and a security level of 128 bits. A plaintext modulus of 65537 and a similar security level were used to evaluate the FHEE (Fully Homomorphic Encryption over the Integers) scheme, which allows arbitrary computations—including addition and multiplication—on encrypted data. By integrating these cutting-edge cryptographic techniques with well chosen parameters, our system guarantees safe and effective collaborative training and inference while maintaining the highest levels of user data privacy.

3.3. PERSONALISED LOCAL MODEL PREDICTION

The "Personalised Local Prediction Model" section is specifically designed for the detection of suicidal and self-harm risks across social media platforms such as Facebook, Twitter/X, Reddit, and others. In our system, we consider five distinct social media platforms, each represented by a separate local model. The original dataset was divided into five different training subsets, corresponding to these platforms, while 30% of the original dataset was set aside for the testing phase to ensure robust evaluation. Using data augmentation techniques, we generated five diverse training sets from 70% of the original dataset, allowing each platform to train on a unique but representative portion of the data. This approach helps produce diverse local model weights, which is essential for effectively evaluating the federated learning process.

The model architecture includes a customised input and output layer for each platform, enabling adaptation to the specific data formats and characteristics of each social network. The core of the model is a shared block, consisting of lightweight bidirectional LSTM-GRU layers followed by two dense layers, which remains identical across all platforms. We have intentionally chosen this simple and efficient structure to keep the model lightweight, enabling faster global model updates and more practical communication, especially when these updates are transferred via blockchain smart contracts. Minimising the complexity and size of the shared model helps reduce both communication overhead and blockchain smart contract fees, optimising the scalability and efficiency of the federated learning system.

Importantly, both training and result exchanges are performed on encrypted data, and all results transmitted between platforms remain encrypted throughout the process. This privacy-preserving approach ensures that raw user data is never exposed, even during model aggregation and synchronisation. During federated learning, only the encrypted shared block is exchanged and aggregated between platforms, enabling the construction of an enriched global model without ever compromising data privacy. The global model is then fine-tuned locally on each platform and connected to its respective customised input and output layers, optimising the prediction of suicidal and self-harm behaviours according to the specificities of each social network. This approach provides tailored personalisation, effective knowledge sharing, strict

Journal of Information Systems Engineering and Management
2025, 10(57s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

preservation of sensitive data privacy, and efficient, secure communication across all participating platforms.

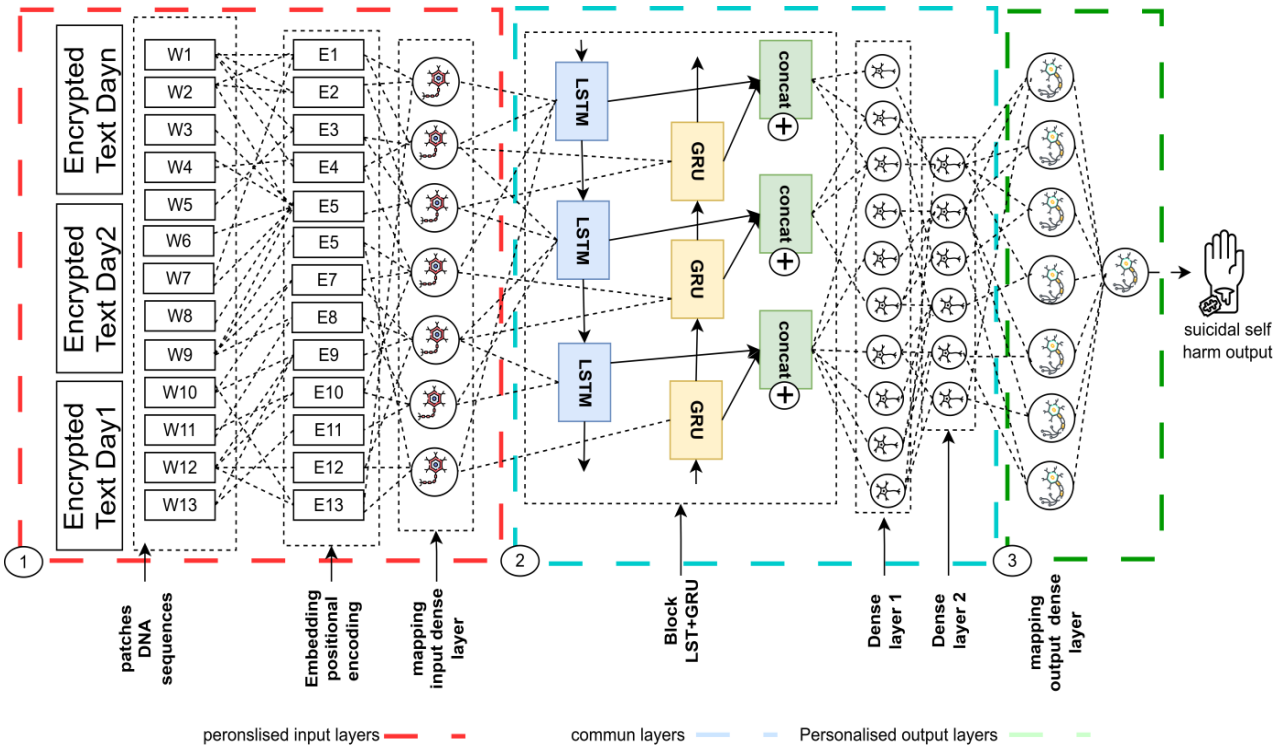


Fig. 2. personalised federated learning for suicidal self-harm detection. 1) personalized input layer to map input data to common block. 2) bidirectional LSTM+GRU to extract the features and relationships that exist between the user posts from left to right .3) Personalised output layers to predict the suicidal situation with an encrypted result for each platform.

3.4. SECURE FEDERATED LEARNING AND BLOCKCHAIN

After each of the five platforms has locally trained its model on its individual, augmented and encrypted dataset, only the weights of the shared layers block (the bidirectional LSTM-GRU and dense layers) are transmitted to the blockchain network. Here, the blockchain serves as a secure and transparent environment for aggregating the updates from all platforms. A smart contract is deployed on the blockchain to collect these weights and perform an averaging operation, also known as Federated Averaging. This process ensures that no raw data or sensitive information is ever exposed or shared between platforms only shred layers parameters are exchanged and aggregated. The resulting global model weights, representing the averaged knowledge from all local models, are then redistributed to each participating platform, still in encrypted form.

The use of blockchain in this phase provides an additional layer of trust, transparency, and tamper-resistance, ensuring that the federated averaging process is auditable and that the integrity of model updates is maintained. Once each platform receives the updated global model, it can proceed with the local fine-tuning and reconnection to its

Journal of Information Systems Engineering and Management
2025, 10(57s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

customised input and output layers for optimal performance in suicidal and self-harm risk detection. This secure and efficient federated learning process maximises collaborative intelligence across platforms, while strictly upholding data privacy and communication integrity.

3.5. SECURE MULTIPARTY COMPUTATION PROTOCOL

Only the user can use their private key to decrypt the encrypted results that are returned to them. All transactions, including the transmission of encrypted results, are carried out via a secure multiparty computation (SMC) protocol in order to further improve security and thwart any unwanted access. Because the calculations and exchanges are divided among several parties, no one entity can access the raw data or decrypted results, guaranteeing that sensitive information stays private throughout the entire process. As a result, user information security and privacy are rigorously maintained throughout.

4. EXPERIMENTATION AND COMPARATIVE STUDY

We provide a thorough analysis of our suggested solution in this section by methodically altering a number of crucial configurations and hyperparameters. In particular, we test various epoch counts, dropout rates, batch sizes, learning rates, and optimizers to see how they affect the overall performance of the model. Additionally, we examine several homomorphic encryption algorithms, such as BGV, CKKS, and FHEE, and assess how each impacts security and efficiency in order to better explore the privacy-preserving component of our framework.

Many well-known performance metrics, including F-measure, success rate, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Relative Error Estimate (REE), and execution time, are used in our experimental validation. This thorough approach enables us to evaluate our system's computational efficiency, privacy guarantees, predictive efficacy, and robustness in a variety of configurations. The analysis and results that follow shed light on the best configurations and compromises for implementing our privacy-preserving federated learning solution in practical situations. We evaluated our solution using the Microsoft SEAL library, using fully homomorphic encryption (FHE) schemes, namely CKKS and BGV (which is similar to BFV), instead of differential privacy. Table 1 and the following figures provide a summary of the results for MRE, RMSE, CEE, accuracy, encryption time, and encrypted data size.

- CKKS is a FHE scheme designed to handle real numbers. To evaluate these performances (see Table 1), we varied several key parameters: PolyModulusDegree (size of the polynomial modulus) with values 4096, 8192, and 16384, CoeffBitWidth (number of bits for the coefficients) with options 20, 30, 40, and 50, Scale (scale factor

Journal of Information Systems Engineering and Management
2025, 10(57s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

to control precision) with values 220, 230, and 240, and patch size with dimensions 3, 5, and 7 [2].

- BGV (Brakerski, Gentry, and Vaikuntanathan) A FHE scheme relies on the Learning With Errors (LWE) problem and its ring variant (Ring-LWE), thus offering robust security. In the context of our experimentation (see Table 1), we varied configuration parameters, including the security level (SL), the polynomial degree (PD), the modulus q, and the multiplication depth L [1].

Table .1. Results of LifeGuardAI with variation of parameters and FHE in term of (CEE, MRE, RMSE, Encryption time

Models	CEE		MRE		RMSE		Encryption time (s)	FHE
	Train	Test	Train	Test	Train	Test	Train+test	
Platform 1	0.095	0.11	0.126	0.05	0.12	0.09	421.83	**CKKS**
Platform 2	0.095	0.34	0.11	0.06	0.11	0.16	68.163	**CKKS**
Platform 3	0.13	0.09	0.10	0.06	0.12	0.15	70.196	**FHEE**
Platform 4	0.11	0.09	0.10	0.07	0.06	0.17	87.806	**FHEE**
Platform 5	0.106	0.11	0.12	0.09	0.054	0.16	92.324	**BGV**
Global model	/	0.07	/	0.07	/	0.09	/	**CKKS**

The results summarized in Table 1 provide a comprehensive evaluation of the proposed federated learning framework across multiple social media platforms using various fully homomorphic encryption (FHE) schemes (CKKS, BGV, and FHEE). The performance metrics considered include CEE, MRE, and RMSE for both training and testing phases, as well as encryption time, providing insight into both model accuracy and computational overhead. As illustrated in this table , across individual platforms, the CEE (Cross-Entropy Error) values during testing remain consistently low, ranging from 0.09 to 0.34. Notably, Platform 2 shows a relatively higher CEE during testing (0.34), suggesting either a higher variability in the test set or a need for further model adaptation on this platform. The MRE (Mean Relative Error) and RMSE (Root Mean Squared Error) also generally indicate good model generalization, with most values during testing below 0.16, except for Platform 2 (0.16 RMSE) and Platform 5 (0.16 RMSE), which may reflect differences in data distribution or sample size.

In terms of encryption time, Platform 1 records the highest value (421.83 seconds), likely due to the volume or complexity of the encrypted operations under the CKKS scheme. In contrast, Platforms 2, 3, 4, and 5 show much lower encryption times,

Journal of Information Systems Engineering and Management
2025, 10(57s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

ranging from 68 to 91 seconds, suggesting that the computational overhead of the FHE schemes remains manageable for most use cases, especially when processing is distributed.

As evidence of the advantages of cross-platform generalization and collaborative learning, the global model aggregated using the CKKS scheme achieves the lowest test set errors (CEE: 0.07, MRE: 0.07, RMSE: 0.09). This demonstrates how federated aggregation adds value by preserving privacy while identifying common patterns across platforms. All things considered, these findings demonstrate that the suggested privacy-preserving method can detect suicide risk with high accuracy and resilience while maintaining an acceptable encryption overhead for practical use. The disparities in performance between platforms underscore the significance of additional fine-tuning and adaptation for every social media scenario, in addition to the selection of the best FHE schemes based on platform-specific attributes.

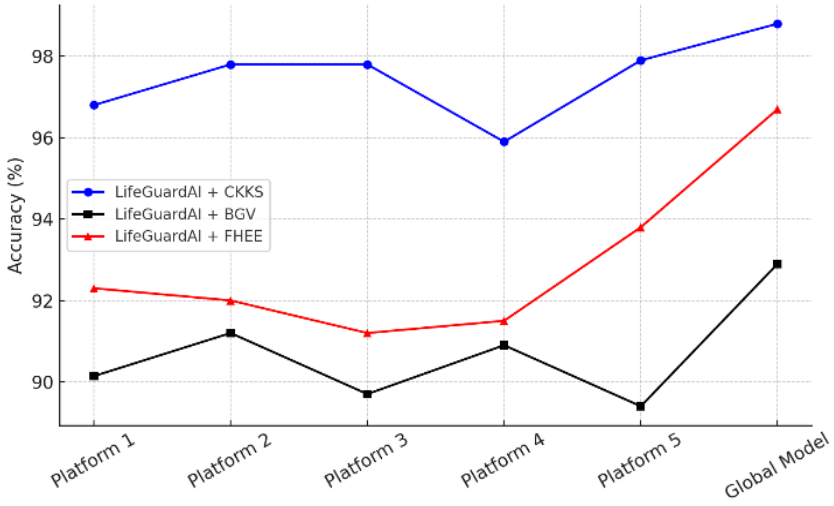


Fig .3. Accuracy of lifeGuardAi with variation of FHE for 5 platformes and the global model.

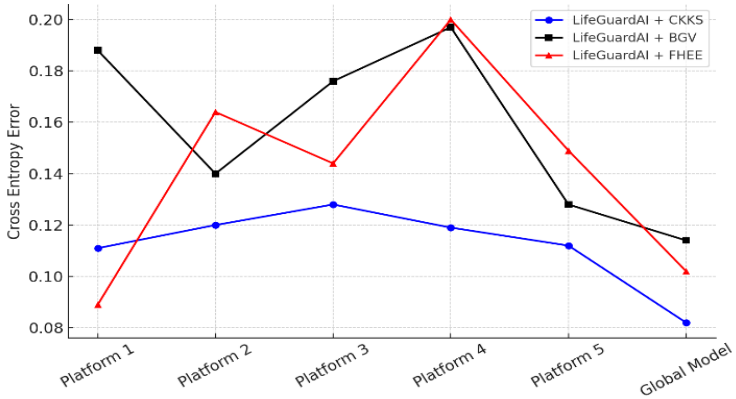


Fig. 4. Cross Entropy Error of lifeGuardAI with variation of FHE algorithms and 5 platformes and global model.

Journal of Information Systems Engineering and Management
2025, 10(57s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The results in figures 3 and 4 show that LifeGuardAI combined with the CKKS scheme consistently achieve the highest accuracy across all platforms and the global model, outperforming BGV and FHEE. Conversely, CKKS also delivers the lowest cross-entropy error, indicating more reliable and stable predictions. Overall, these findings highlight CKKS as the most effective homomorphic encryption scheme for preserving both accuracy and prediction quality in privacy-preserving federated learning scenarios.

To evaluate the effectiveness of our proposed approach, we conducted a comprehensive comparative study with several state-of-the-art models reported in recent literature. The results, as summarized in the table below, reveal significant advancements in suicide and self-harm risk detection performance over time. Earlier machine learning models, such as Random Forest [13] and Decision Tree [12], achieved moderate F1 scores of 53.6% and 90.8% respectively. Traditional deep learning models, including C-LSTM and Contextual CNN, showed noticeable improvements, with F1 scores reaching up to 72.9%. Transformer-based models and advanced architectures like DualContextBert and STATENet [13] pushed the boundaries further, achieving F1 scores of 76.7% and 79.9%, respectively, and accuracies exceeding 80%. Among classical approaches, AdaBoost and Logistic Regression reported impressive F1 scores of 95.9% and 95.5% [12]. More recent advances leveraging hybrid and deep architectures (such as BiLSTM, CNN-LSTM, and Pretrained CNNs) have delivered consistently high performance, with models like C-BiLSTM [7], LSTM [3], and CNN-LSTM [10] reporting F1 scores and accuracies above 90%. Notably, transformer-based and large pre-trained models have set new benchmarks, as illustrated by the Transformer [5] with an accuracy of 97.4% and GPT [7] achieving up to 97.69%. The highest performances were observed with deep CNN variants such as VGG16 2D [6], which attained an F1 score of 98.51% and accuracy of 98.73%, and ResNet 50 2D, which similarly reported outstanding results.

This progression highlights the impact of advanced architectures, data augmentation techniques, and pre-training strategies in boosting model performance for this challenging task. The comparison also underscores the effectiveness of combining sequential models like LSTM and GRU with convolutional or transformer-based architectures. Overall, the study demonstrates that modern deep learning and hybrid models significantly outperform traditional approaches, offering higher accuracy, F1 scores, and recall rates in suicide and self-harm detection on social media. These findings validate our decision to adopt a hybrid federated learning framework leveraging LSTM-GRU blocks, which strikes a balance between model performance, scalability, and privacy.

Journal of Information Systems Engineering and Management
2025, 10(57s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Table 2. A comparative study between our solution RoboMediTrust and different techniques existed in literature in term of accuracy, specificity and sensitivity

Approach	F1 Score	Accuracy	Rappel
Random forest	53.6%	54.8%	51.3%
C-LSTM [13]	58.8%	60.2%	59.7%
Contextual CNN [13]	72.9%	80.3%	58.7%
DualContextBert [13]	76.7%	82.3%	78.6%
STATENet [13]	79.9%	85.1%	81%
Logistic Regression [12]	95.5%	95.47%	95.5
Decision tree [12]	90.8%	90.78%	90.8%
AdaBoost [12]	95.9	95.2	95.9
Nive bayes [12]	81.7%	80.9%	78.6%
SVM [8]	88.96%	88.96%	80.47%
XGboosting [8]	89.99%	89.99%	83.62%
Stacking Classifier [8]	89.25%	89.25%	83.31%
C-BiLSTM [7]	/	94.5%	/
GPT[7]	/	97.69%	/
Pretrained CNN [4]	91.5%	87%	
CNN-LSTM [10]	92.26%	/	/
Transformer [5]	/	97.4%	/
LSTM [3]	94%	/	/
Chat GPT-3 [6]	79.74	81.73	82.88
BiLSTM-CNN [6]	69.19	69.83	69.90
AlexNet 2D [6]	94.38%	94.68%	94.96%
ResNet 50 2D [6]	95.27%	97.59%	98.45%
VGG 16 2D [6]	98.51%	98.73%	98.05%
BiGRU-CNN [9]	93.07%	93.07%	
Ours solution	98.9%	98.8%	97.5%

5. CONCLUSION

In conclusion, our work shows how to build a strong, privacy-protecting system for finding people who are at risk of suicide or self-harm on many social media sites, such as Facebook, Twitter/X, Reddit, and others. Our solution directly addresses some of the most important problems in this field: multimodal analysis (by combining text, images, and multilingual content), generalization (through federated learning and data augmentation across different datasets), collaborative intelligence (through personalized local models and a shared lightweight core), user history analysis (by

keeping track of patterns in user activity over time), and privacy protection (by making sure that all data and model updates are encrypted and processed locally). LifeGuardAI also handles the problem of users with multiple accounts in a unique way by combining signals from different platforms. This lets for a complete and unified evaluation of each person's mental health.

We can fully test our federated learning approach by splitting the dataset into five separate training subsets and keeping one test set. Combining federated learning with blockchain technology makes the model aggregation process more secure, open, and trustworthy. Blockchain-based smart contracts make sure that global model updates are correct and can be audited.

LifeGuardAI's main goal is not just to find people who are in danger, but also to stop them before they get hurt. When the system finds someone who is at risk, it automatically alerts their close contacts, encouraging them to get help and intervene quickly. This design finds a middle ground between personalization and collective intelligence, so that each platform can use shared knowledge without giving up privacy. So, federated learning, blockchain, and advanced deep learning together provide a solution for sensitive prediction tasks on social media that is scalable, effective, and ethical.

Lastly, we suggest that social network providers add features that let users get status updates that show whether someone seems to be in a normal or depressed state and give them advice on things like seeing a mental health professional, learning more about depression and its treatments, and living a healthy lifestyle. Also, looking at social media data over time, especially from sites like Twitter, can help us learn more about how online conversations affect health behaviors, which can help us come up with better ways to prevent and support people.

REFRENCES

[1] Brakerski, Z., Gentry, C., & Vaikuntanathan, V. (2014). (Leveled) fully homomorphic encryption without bootstrapping. *ACM Transactions on Computation Theory (TOCT)*, *6*(3), 1-36.

[2] Sathishkumar, P., Pugalarasan, K., Ponnparamaguru, C., & Vasanthkumar, M. (2024, April). Improving healthcare data security using cheon-kim-kim-song (ckks) homomorphic encryption. In *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)* (Vol. 1, pp. 1-6). IEEE.

[3] Bhattacharya, D., & Shahina, A. (2021, November). Early detection of suicidal tendencies from text data using LSTM. In *2021 Innovations in Power and Advanced Computing Technologies (i-PACT)* (pp. 1-5). IEEE.

[4] Skaik, R., & Inkpen, D. (2021). Suicide ideation estimators within Canadian provinces using machine learning tools on social media text. *Journal of Advances in Information Technology Vol*, *12*(4).

[5] Amin, M. M., Cambria, E., & Schuller, B. W. (2023). Will affective computing emerge from foundation models and general artificial intelligence? A first evaluation of ChatGPT. *IEEE Intelligent Systems*, *38*(2), 15-23.

[6] Esmi, N., Shahbahrami, A., Gaydadjiev, G., & de Jonge, P. (2025). Suicide ideation detection based on documents dimensionality expansion. *Computers in biology and medicine*, *192*, 110266.

[7] Qorich, M., & El Ouazzani, R. (2024). Advanced deep learning and large language models for suicide ideation detection on social media. *Progress in Artificial Intelligence*, *13*(2), 135-147.

[8] Dewangan, D., Selot, S., & Panicker, S. 2024 Detecting Self-harm Content and Behavior in Tweets with SVM and Ensemble Classifiers: A Comparative Study. *International Journal of Computer Applications*, *975*, 8887.

[9] Anika, S., Dewanjee, S., & Muntaha, S. (2024). Analyzing Multiple Data Sources for Suicide Risk Detection: A Deep Learning Hybrid Approach. *International Journal of Advanced Computer Science and Applications*, *15*(2). https://doi.org/10.14569/ijacsa.2024.0150270

[10] Gupta, H., Gola, K. K., Kumar, S., Kumar, P., & Jee, N. (2023). *Suicide Ideation Detection: Harnessing Machine and Deep Learning for Early Risk Identification*. *6*, 1555–1562. https://doi.org/10.1109/ic3i59117.2023.10397942

[11] Aldhyani, T. H. H., Alsubari, S. N., Alshebami, A. S., Alkahtani, H., & Ahmed, Z. A. T. (2022). Detecting and Analyzing Suicidal Ideation on Social Media Using Deep Learning and Machine Learning Models. *International Journal of Environmental Research and Public Health*, *19*(19), 12635. https://doi.org/10.3390/ijerph191912635

[12] Rabani, S. T., Khan, Q. R., & Khanday, A. M. U. D. (2020). Detection of suicidal ideation on Twitter using machine learning & ensemble approaches. Baghdad science journal, 17(4), 1328-1328. DOI: 10.21123/bsj.2020.17.4.1328

[13] Sawhney, R., Joshi, H., Gandhi, S., & Shah, R. (2020, November). A time-aware transformer based model for suicide ideation detection on social media. In Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP) (pp. 7685-7697). DOI: 10.18653/v1/2020.emnlp-main.619

[14] Knipe, D., Padmanathan, P., Newton-Howes, G., Chan, L. F., & Kapur, N. (2022). Suicide and self-harm. *The Lancet*, *399*(10338), 1903-1916.

[15] McMahon, E. M., Cully, G., Corcoran, P., Arensman, E., & Griffin, E. (2024). Advancing early detection of suicide? A national study examining socio-demographic factors, antecedent stressors and long-term history of self-harm. *Journal of affective disorders*, *350*, 372-378.

[16] Martinez-Romo, J., Araujo, L., & Reneses, B. (2025). Guardian-bert: Early detection of self-injury and suicidal signs with language technologies in electronic health reports. *Computers in Biology and Medicine*, *186*, 109701.

[17] Chen, Q., & Hong, Y. (2024). Medblip: Bootstrapping language-image pre-training from 3d medical images and texts. In *Proceedings of the Asian conference on computer vision* (pp. 2404-2420).