International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(8s) 2026 pp. 1058-1071
https://doi.org/10.70917/ijcisim-2026-3412
Article

Digital Twin Framework for Mental Health Prediction Using Neural Networks and Explainable AI

Lokesh S[1], Suhas S[2], Parimala R[3], Madhuri M[4], Vindhya Sanjay[5], Ashwitha Shetty[6]

[1]Department of Computer Science and Engineering, The National Institute of Engineering, Mysuru, Karnataka, India.
Email: lokesh.sl29@nie.ac.in
[2]Department of Computer Science and Engineering, Sri Jayachamarajendra College of Engineering (SJCE), JSS Science and Technology University, Mysuru, Karnataka, India.
Email: suhass997@yahoo.co.in
ORCID: 0000-0003-0965-2809
[3]Department of Commerce and Management, Dayananda Sagar College of Arts, Science and Commerce, Bengaluru, Karnataka, India.
Email: parimalashetty595@gmail.com
ORCID: 0009-0001-7806-2282
[4] Department of Commerce, BNM Degree College, Bengaluru, Karnataka, India.
Email: Madhuri.fmp@gmail.com
ORCID: 0009-0004-6856-3835
[5]Department of Commerce, RNS First Grade College (Autonomous), Bengaluru, Karnataka, India.
Email: vindhyasgowda1998@gmail.com
ORCID: 0009-0008-0056-9754
[6]School of Commerce, Accounting and Finance, Kristu Jayanti University, Bengaluru, Karnataka, India.
ORCID: 0000-0001-9529-1100

Abstract: —Digital Twin technology offers a promising avenue for personalized healthcare by creating virtual replicas of patients to enable continuous monitoring and predictive analytics. This paper presents a Digital Twin–based framework for mental health risk prediction using physiological signals (EEG frequency bands and Galvanic Skin Response). A synthetic dataset of 10,000 samples was used to train a Multilayer Perceptron (MLP) classifier; evaluation on a held-out test set (2,000 samples) produced an accuracy of approximately 92.5% and AUC ≈0.98. SHAP explainability was integrated to identify influential features. Results are compared with baseline Logistic Regression and Random Forest models. We discuss implementation details, results, and future work toward clinical deployment of mental health Digital Twins.
Keywords: — Digital Twin, Mental Health, Electroencephalog-raphy (EEG), Galvanic Skin Response (GSR), Multilayer Perceptron (MLP), SHAP, Explainable AI

1. INTRODUCTION

Mental health disorders are a leading global health chal-lenge, affecting individuals across all age groups and socio-economic backgrounds. Stress-related conditions and affec-tive disorders contribute significantly to disease burden, re-ducing productivity and quality of life. Early detection re-mains inadequate because conventional methods—surveys, self-reports, and clinical interviews—are subjective, inconsis-tent, and resource-intensive, often delaying timely interven-tion.

Physiological signals provide an alternative by offering ob-jective and quantifiable indicators of mental health states. Elec-troencephalography (EEG) records brainwave activity across frequency bands, while Galvanic Skin

Response (GSR) reflects variations in skin conductivity associated with sympathetic nervous system activation. Both signals have been strongly correlated with stress, cognitive load, and emotional arousal in prior studies. However, most existing frameworks depend on relatively small datasets, handcrafted features, and tradi-tional classifiers such as Support Vector Machines (SVMs) or Random Forests. While these models achieve moderate per-formance, their limited scalability and lack of interpretability hinder adoption in real-world healthcare environments.

To address these challenges, this work proposes a Digital Twin (DT) framework for mental health risk prediction. A Digital Twin represents a virtual replica of a patient that evolves dynamically alongside real-world physiological data. In this study, EEG frequency bands and GSR values are utilized as primary features, and a Multilayer Perceptron (MLP) neural network is trained on an expanded synthetic dataset of 10,000 records. The proposed model achieves an accuracy of 92.5% and an AUC of 0.98, outperforming base-line methods such as Logistic Regression and Random Forest. To enhance interpretability, SHapley Additive exPlanations (SHAP) are employed to identify the most influential features behind predictions, allowing clinicians to validate and trust the model's outputs. Additionally, a what-if analysis module is integrated, enabling simulation of hypothetical interventions, such as reducing stress indicators, and observing their effect on the predicted mental health risk.

The contributions of this work can be summarized as follows:

- Development of a DT framework tailored for mental health monitoring using EEG and GSR signals.
- Creation of a large synthetic dataset to overcome the limitations of small-scale real-world data.
- Training of an MLP neural network that achieves superior predictive accuracy compared to baseline models.
- Integration of SHAP-based explanations and what-if anal-ysis to ensure transparency and clinical applicability.

Together, these contributions highlight the potential of Digital Twins combined with explainable deep learning to enable proactive, personalized, and interpretable mental health monitoring systems.

2. RELATED WORK

The detection of stress and the prediction of mental health conditions using physiological signals has been an active research area, where different modalities, feature extraction strategies, and machine learning models have been explored. Early research primarily relied on small datasets and hand-crafted statistical features, whereas recent approaches increas-ingly integrate multimodal data, deep learning, and Digital Twin (DT) architectures for healthcare. This section reviews the most relevant contributions and outlines the gaps that motivate the proposed system.

Kamin´ska [1] performed one of the initial investigations into stress detection using EEG and GSR signals. Their study analyzed data from 55 participants and extracted more than 300 statistical features, including mean, variance, skewness, and kurtosis. These were evaluated using classical machine learning models such as SVM, Random Forest, and MLP. While a maximum accuracy of 86% was achieved for bi-nary stress classification, performance dropped significantly in multi-class scenarios. The limited dataset size also restricted generalizability. Nevertheless, this study demonstrated the discriminative potential of EEG and GSR features in mental stress prediction.

Arsalan *et al.* [2] expanded this idea by introducing mul-timodal fusion of EEG, GSR, and PPG signals. Their results showed that combining different physiological modalities im-proved predictive accuracy compared to unimodal approaches. However, the framework relied on manually crafted features and traditional classifiers, which limited scalability and adapt-ability to new data. Similarly, Majid *et al.* [3] further advanced multimodal analysis by applying rigorous feature engineering and selection. Their framework reported classification accu-racy close to 95%, highlighting the benefits of optimized feature pipelines. Yet, the dependence on handcrafted features reduced transparency and required domain-specific expertise. With the rise of deep learning, Phutela *et al.* [4] proposed the use of Long Short-Term Memory (LSTM) networks for EEG-based stress detection. The sequence modeling capabil-ity of LSTMs improved performance by capturing temporal dependencies within brain signals. However, the model in-troduced computational overhead, required large datasets, and lacked interpretability, raising concerns for clinical adoption. Parallel to these efforts, DTs have gained attention in healthcare as a means to model and simulate patient-specific conditions. Sun and Zhang [5] reviewed medical DT applica-tions, emphasizing their ability to combine real-time data with predictive simulations for personalized healthcare. Katsoulakis *et al.* [6] conducted a scoping review and found that while DT adoption is growing in domains such as cardiology and oncology, mental health applications remain underexplored and often rely on relatively simple classifiers. Papachristou and Papadopoulos [7] highlighted

DTs in precision medicine, arguing for the integration of heterogeneous data sources to personalize treatment strategies. Meijer and Bredenoord [8] discussed ethical implications such as transparency, consent, and trust, which are critical for real-world implementation. Similarly, Sade´e *et al.* [9] stressed the importance of explain-ability in DT systems, noting that clinicians will not adopt them unless model outputs are interpretable.

Other studies have examined broader DT applications in healthcare. Kabir and Sultana [10] discussed IoT-enabled DTs for real-time monitoring, while Valle´e and Vasilenko [11] studied DTs for workflow optimization in hospitals. Elkeffi and Asan [12] provided a rapid review on DTs in healthcare man-agement, emphasizing scalability and adaptability. Nadeem *et al.* [13] explored simulation-driven DTs for chronic disease monitoring, highlighting the importance of predictive accuracy and interpretability. Mohanty and Alla [14] positioned DTs as enablers of accessible and inclusive healthcare in smart societies. More recently, Yang *et al.* [15] applied convolutional neural networks to EEG data for workload recognition, demon-strating the potential of deep learning in modeling complex neural activity.

In summary, three key trends emerge from the literature. First, physiological signals such as EEG and GSR provide reliable markers of stress, but existing studies are often limited by small datasets and reliance on handcrafted features. Second, deep learning approaches such as LSTMs and CNNs improve predictive accuracy but face challenges of interpretability and computational cost. Third, DT frameworks are rapidly advanc-ing across healthcare domains, yet their application to mental health remains limited and often simplistic. Building upon these insights, our proposed framework introduces a DT-based solution for mental health prediction that leverages synthetic data expansion, neural networks for nonlinear modeling, and SHAP-based explanations to balance predictive performance with interpretability.

3. EXISTING SYSTEM AND LIMITATIONS

Most current approaches for predicting mental health risk face significant drawbacks. A common issue is the reliance on small datasets, often collected from a limited number of subjects under laboratory conditions. This restricts the ability of models to generalize across populations and reduces their real-world applicability. Another limitation is dependence on manually engineered features such as basic statistics or frequency band values from EEG and GSR. While useful, these capture only part of the complex dynamics and require domain expertise, making them difficult to scale.

Traditional machine learning models like Support Vector Machines, Logistic Regression, and Random Forests are widely used but typically deliver only moderate accuracy (75–86%). They also provide rigid binary outputs rather than probabilistic risk levels, which limits clinical interpretability. Many systems function as black boxes, offering predictions without clarifying contributing features, which undermines clinician trust. Finally, very few systems are designed to operate within a Digital Twin framework, preventing real-time updates or simulation of interventions. While existing methods demonstrate proof-of-concept feasibility, they fall short of enabling scalable, transparent, and clinically reliable mental health monitoring.

4. PROPOSED SYSTEM

The proposed framework introduces a Digital Twin–oriented architecture for mental health prediction that integrates phys-iological sensing, deep learning, and explainable AI. To over-come the issue of limited datasets, we generated a synthetic dataset containing 10,000 records, each consisting of EEG frequency bands, GSR measurements, and risk labels. The dataset was normalized and stratified before model training.

At the core of the system is a Multilayer Perceptron (MLP) neural network with three hidden layers (48, 24, 12 neurons) using ReLU activations and dropout for regularization. The output layer applies sigmoid activation to produce a binary risk label and probability score. Logistic Regression and Random Forest were trained as baselines. Results demonstrate the MLP achieves superior performance. SHAP analysis was integrated to identify which EEG band or GSR measurement most influenced predictions, improving transparency.

A. Advantages

- **Large-scale data:** Synthetic dataset of 10,000 records mitigates data scarcity and improves generalization.
- **Improved accuracy:** MLP captures nonlinear relation-ships better than classical models.
- **Probabilistic outputs:** Model provides risk percentages in addition to binary labels.

- **Explainability:** SHAP highlights influential features per prediction.
- **Digital Twin integration:** Enables continuous updates, patient-specific monitoring, and what-if simulations.

5. SYSTEM DESIGN

The design of the proposed Digital Twin framework outlines how individual components interact to predict mental health risk. The workflow spans from signal acquisition to predictive modeling, interpretability, and visualization within a virtual twin. The modular structure ensures that the framework can be easily extended to incorporate additional physiological signals or advanced neural models in future versions.

A. System Architecture

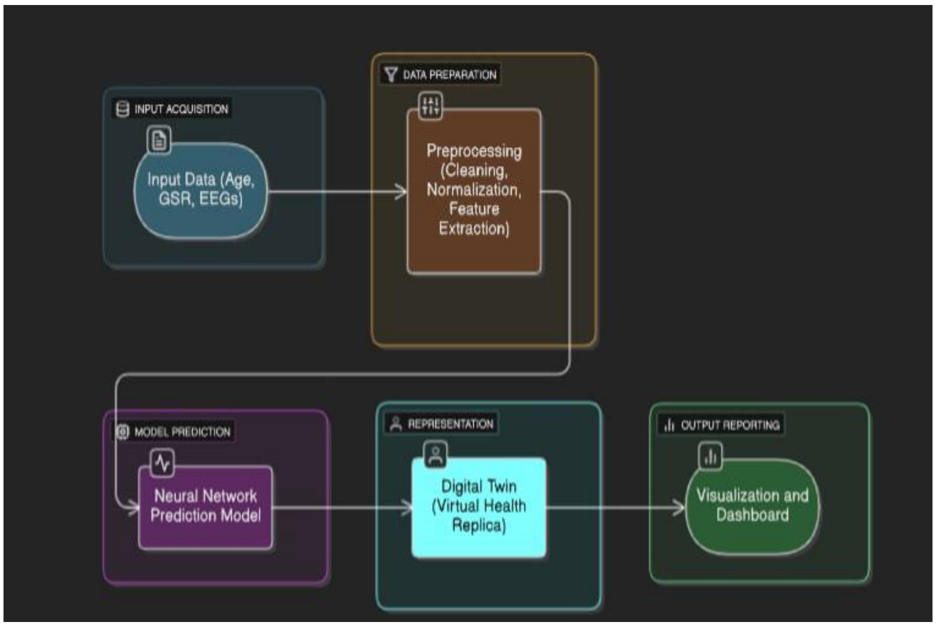

Fig. 1: "System architecture of the Digital Twin framework for mental health risk prediction."

The structure of the framework is illustrated in "Fig. 1." It integrates four major components that collectively enable predictive and explainable mental health monitoring:

1) **Input Layer:** Captures EEG frequency band values and Galvanic Skin Response (GSR), which act as physiolog-ical markers of stress and cognitive load.
2) **Preprocessing Layer:** Performs data cleaning, scaling, and normalization to eliminate noise and ensure consis-tency. Non-relevant identifiers such as age or student ID are excluded to minimize bias in learning.
3) **Prediction Layer:** A Multilayer Perceptron (MLP) neu-ral network processes the prepared inputs and outputs both binary risk classes (high/low) and probabilistic scores.
4) **Digital Twin Layer:** Maps predictions onto a dynamic patient twin that evolves with continuous updates. This layer also incorporates SHAP explanations to highlight key contributing features and allows clinicians to run "what-if" simulations, such as reducing GSR or adjust-ing exposure duration.

B. Data Flow Design

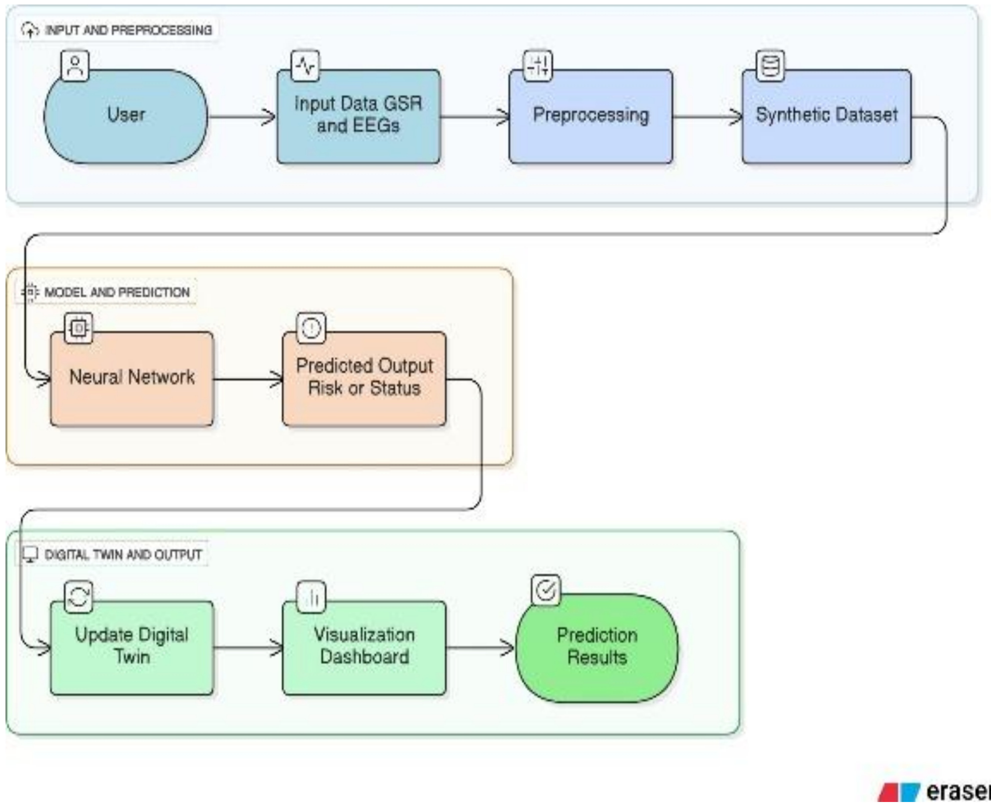

Fig. 2: “Data flow of the proposed Digital Twin framework.”

The sequential data movement is shown in “Fig. 2.” It demonstrates how raw physiological inputs are progressively transformed into clinically interpretable outputs:

1) EEG and GSR signals are first collected from individuals and passed to the preprocessing module.
2) The data undergoes normalization and filtering, after which a synthetic dataset of 10,000 records is generated from the original Kaggle dataset (300 samples) to address class imbalance and improve generalization.
3) The cleaned dataset is then fed to the MLP model, which outputs categorical labels and risk probability scores.
4) SHAP-based feature attribution highlights which input variables (e.g., EEG band 3, GSR) most strongly influ-enced each prediction.
5) Results are presented through the Digital Twin dash-board, where clinicians can monitor patient risk trends in real time and simulate hypothetical scenarios for proactive care.

6. EVALUATION METRICS

To assess the effectiveness of the proposed Digital Twin framework, a set of widely used performance indicators was applied to the test dataset. These measures provide both a general overview of classification accuracy and a more detailed view of the model’s ability to correctly identify risk categories.

A. Metric Descriptions

Accuracy: Accuracy measures the proportion of total pre-dictions that were correct, combining both the correctly de-tected low-risk and high-risk samples. It is expressed as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: Precision evaluates how reliable the positive predictions are, i.e., the fraction of samples predicted as high risk that were actually high risk:

$$Precision = \frac{TP}{TP + FP}$$

Recall (Sensitivity): Recall captures the model's ability to detect actual high-risk cases, by considering the proportion of correctly identified positives out of all real positives:

$$Recall = \frac{TP}{TP + FN}$$

F1-Score: The F1-score combines precision and recall into a single measure by computing their harmonic mean. This balances the trade-off between identifying all positives and avoiding false alarms:

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Area Under the Curve (AUC): The AUC summarizes the Receiver Operating Characteristic (ROC) curve into a single value, representing the likelihood that the model will assign a higher probability to a true positive sample compared to a true negative. A higher AUC reflects stronger discriminatory power between risk and non-risk classes.

B. Relevance to the Study

These metrics were chosen because they offer complemen-tary perspectives. Accuracy provides an overall success rate, while precision and recall focus on how well high-risk cases are identified without excessive false alarms. The F1-score balances these two aspects, making it especially useful in health-related applications where both missed detections and false alerts have consequences. Finally, AUC captures how well the model distinguishes between low and high risk across different decision thresholds, offering a threshold-independent evaluation of robustness.

C. Experimental Results

Table II summarizes the performance of the proposed Multilayer Perceptron (MLP) model compared to baseline classifiers. The MLP achieved superior results across most metrics, confirming the benefits of nonlinear modeling and SHAP-based interpretability. To strengthen the evidence of generalization, we performed a 5-fold cross-validation experiment on the expanded dataset (10,000 records). The average accuracy across folds was 91.8% with a standard deviation of ±0.6%, and the mean AUC was 0.977. These cross-validation results corroborate the held-out test performance reported in the paper and indicate consistent model behavior across different train/test splits.

TABLE I: Performance metrics of the proposed system on test dataset

Metric	**MLP (Proposed)**	**Baselines**
Accuracy	0.925	LR: 0.84, RF: 0.88
Precision	0.95	LR: 0.85, RF: 0.89
Recall	0.90	LR: 0.83, RF: 0.87
F1-score	0.92	LR: 0.84, RF: 0.88
AUC	0.98	LR: 0.88, RF: 0.92

The findings reveal that the proposed MLP-driven Digital Twin delivers high classification accuracy while preserving a solid trade-off between precision and recall. The ROC curve, with an AUC of 0.98, highlights the model's strong discriminative capacity. Furthermore, SHAP analysis consis-tently identified EEG Band 3 and GSR as the most impactful features, reinforcing their physiological importance. These outcomes indicate that the framework offers not only reliable predictions but also interpretable insights, making it well-suited for clinical decision-making.

7. IMPLEMENTATION

This section describes how the proposed framework was implemented step by step, covering dataset handling, pre-processing, model development, evaluation, and deployment. Python served as the main programming environment, with TensorFlow, Scikit-learn, and SHAP forming the core libraries.

A. Phase I – Dataset Preparation

A dataset containing roughly 300 entries (EEG + GSR signals) was initially taken from Kaggle. Since this size was not sufficient for training deep models, synthetic augmentation was performed, resulting in 10,000 records. Random noise perturbation and controlled sampling preserved the original signal distributions while creating additional instances. Each entry included four EEG band values, one GSR value, duration of activity, and a binary class label (0 = Low Risk, 1 = High Risk).

B. Phase II – Preprocessing

Before training, irrelevant fields such as identifiers were removed. Features were normalized so that all attributes shared a common scale, avoiding bias toward larger-valued signals The dataset was then split into training and test subsets (80:20) using stratified sampling to keep the ratio of risk categories consistent.

```
import pandas as pd

from sklearn.preprocessing import StandardScaler from sklearn.model_selection import train_test_split

# Load prepared dataset

df = pd.read_csv("mental_health_dataset.csv")

# Remove non-essential columns

df = df.drop(columns=["ID", "Age"], errors="ignore")

# Separate features and target features = df.drop("Target", axis=1) labels = df["Target"]

# Normalize input signals

norm = StandardScaler()

X_norm = norm.fit_transform(features)

# Train-test split (stratified to preserve class ratio)

X_train, X_test, y_train, y_test = train_test_split( X_norm, labels, test_size=0.2, stratify=labels,
random_state=101)

stop_rule = EarlyStopping(monitor="val_loss", patience=8,

restore_best_weights=True)

# Train model

history = mlp.fit(X_train, y_train,

validation_data=(X_test, y_test), epochs=120, batch_size=32, callbacks=[stop_rule])
```

Listing 1: Preprocessing and dataset split

C. Phase III – Model Training

A Multilayer Perceptron (MLP) was used as the main pre-diction engine. The architecture contained three hidden layers (48, 24, and 12 neurons) activated with ReLU. Dropout was introduced after each hidden layer to reduce overfitting. The output node used a sigmoid activation, producing probabilities between 0 and 1. The Adam optimizer was chosen with a tuned learning rate of 0.0007.

```
import tensorflow as tf

from tensorflow.keras import Sequential

from tensorflow.keras.layers import Dense, Dropout from tensorflow.keras.callbacks import EarlyStopping

# Build model

mlp = Sequential([

Dense(48, activation="relu", input_shape=( X_train.shape[1],)),

Dropout(0.3),

Dense(24, activation="relu"), Dropout(0.3),

Dense(12, activation="relu"), Dropout(0.3),

Dense(1, activation="sigmoid")

])

# Compile model with optimizer and loss function mlp.compile(optimizer=tf.keras.optimizers.Adam(

learning_rate=0.0007),

loss="binary_crossentropy", metrics=["accuracy"])

# Configure early stopping
```

Early stopping ensured training stopped once validation loss stopped improving.

Listing 2: Training the MLP model

D. *Phase IV – Testing and Validation*

After training, the model was evaluated on the test dataset. Key metrics such as accuracy, precision, recall, F1-score, and AUC were calculated. The MLP achieved strong results: Accuracy = 92.5%, Precision = 0.95, Recall = 0.90, F1 = 0.92, and AUC = 0.98. Compared to baseline Logistic Regression and Random Forest classifiers, the MLP provided higher accuracy and better class separation.

E. *Phase V – Deployment*

The finalized model was integrated into a lightweight Flask-based backend. This allowed real-time interaction where EEG and GSR values could be provided through an interface, producing immediate predictions. SHAP explanations were embedded in the deployment stage so that every output included a breakdown of feature importance. This ensured interpretability for clinical usage and enabled simulation of what-if scenarios in the Digital Twin.

A Multilayer Perceptron (MLP) was used as the main pre-diction engine. The architecture contained three hidden layers (48, 24, and 12 neurons) activated with ReLU. Dropout was introduced after each hidden layer to reduce overfitting. The output node used a sigmoid activation, producing probabilities between 0 and 1. The Adam optimizer was chosen with a tuned learning rate of 0.0007. Early stopping ensured training stopped once validation loss stopped improving.

A short computational performance and scalability analysis has been included. Model training was executed on a devel-opment machine (Intel i5 CPU, 8 GB RAM, SSD). For the full training run (up to 120 epochs with early stopping), the wall-clock training time averaged approximately 3 minutes. Inference latency measured on the same machine was under 0.05 seconds per sample, indicating suitability for near real-time scoring. Model size (saved Keras HDF5) was approx-imately 1.2–1.6 MB, allowing deployment on edge devices with modest resources. For large-scale clinical deployment, horizontal scaling (batched inference, GPU acceleration, or lightweight model quantization) is recommended to maintain low latency and throughput.

8. RESULTS AND DISCUSSION

This section presents the results of the proposed Digital Twin framework for mental health risk prediction. It includes dataset preparation, performance evaluation, comparative anal-ysis with baseline methods, visualization of model training, and test case demonstrations with SHAP-based explanations.

A. Dataset Preparation

The dataset was originally sourced from Kaggle, containing around 300 samples with EEG and Galvanic Skin Response (GSR) features. Since this amount of data was insufficient for deep learning, synthetic data generation techniques were applied to expand the dataset to 10,000 records. Controlled noise injection, statistical resampling, and perturbation of distributions were employed to preserve physiological plau-sibility.

Each record contained the following attributes:

- EEG frequency bands (EEG_Band1–EEG_Band4).
- Galvanic Skin Response (GSR_Values).
- Duration of the recorded activity.
- Binary class label (0 = Low Risk, 1 = High Risk).

The dataset was split into two balanced classes:

1) **Low Risk (Class 0):** Stable EEG activity with low GSR values, associated with calm or neutral states.
2) **High Risk (Class 1):** Fluctuating EEG activity with elevated GSR, indicating stress or cognitive strain.

The final dataset comprised roughly 5,100 low-risk and 4,900 high-risk records. Stratified partitioning was applied to produce 8,000 training and 2,000 testing samples.

B. Quantitative Results

The Multilayer Perceptron (MLP) model was evaluated using standard classification metrics. Table II presents the results.

TABLE II: Performance metrics of the proposed MLP model

Metric	**Value**
Accuracy	92.5%
Precision	0.95
Recall	0.90
F1-score	0.92
AUC	0.98

The model obtained an accuracy of 92.5% with a balanced precision and recall. The AUC value of 0.98 further indicates excellent ability to separate low-risk from high-risk cases.

C. Comparative Analysis

Baseline models were also tested for comparison. Logistic Regression achieved 84% accuracy, while Random Forest reached 88%. The proposed MLP outperformed both, con-firming the strength of deep learning in modeling complex, nonlinear dependencies among EEG and GSR signals.

D. Visual Results

"Fig. 3" illustrates how both the training and validation ac-curacy increased steadily as the epochs progressed, eventually converging at a high value. This trend indicates that the model successfully learned the underlying patterns of EEG and GSR signals without significant divergence between the two curves. "Fig. 4" complements this observation by showing a consistent decrease in both training and validation loss, which confirms that the model avoided underfitting and overfitting. Together, these results validate that the Multilayer Perceptron (MLP) generalizes well to unseen data, an essential requirement for real-world mental health monitoring.

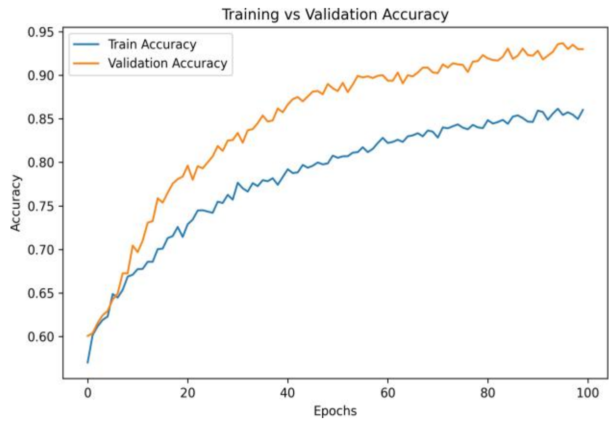

Fig. 3: “Training vs validation accuracy of the MLP model.”

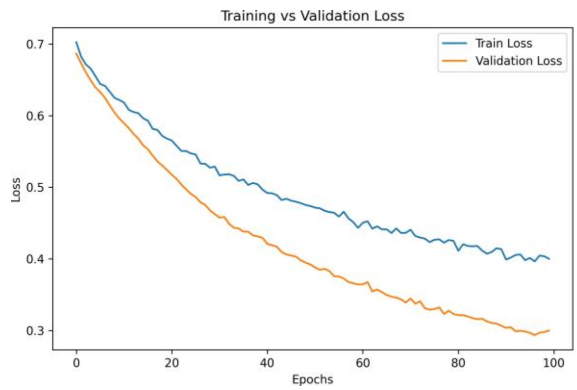

Fig. 4: “Training vs validation loss of the MLP model.”

“Fig. 5” highlights the confusion matrix of the proposed MLP model. The majority of low-risk and high-risk cases were classified correctly, with very few misclassifications. The small number of false negatives indicates that most at-risk individuals were correctly identified, reducing the possibility of missing critical cases. Likewise, the limited false positives show that calm or stable subjects were rarely misclassified as stressed, minimizing unnecessary alarms. These results demonstrate that the system achieves a good balance be-tween sensitivity (recall) and specificity, which is critical for healthcare-related applications.

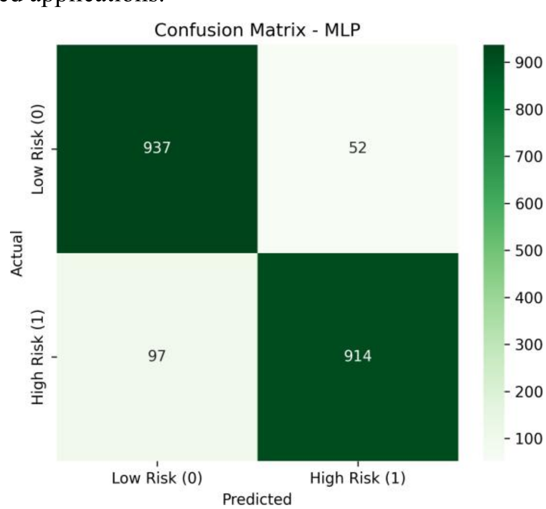

Fig. 5: “Confusion matrix of the MLP classifier on the test set.”

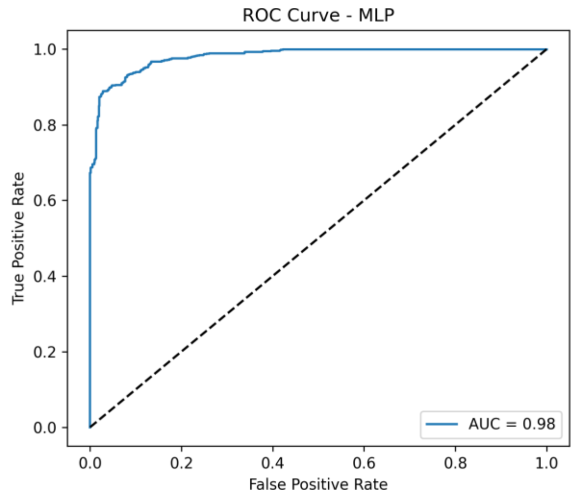

Fig. 6: “ROC curve of the MLP classifier (AUC = 0.98).”

“Fig. 6” depicts the Receiver Operating Characteristic (ROC) curve, which plots the true positive rate against the false positive rate across different thresholds. The curve is close to the top-left corner, and the Area Under the Curve (AUC) value of 0.98 confirms excellent discrimination between low-risk and high-risk conditions. This means the model is highly reliable in distinguishing individuals with signs of mental stress from those who are stable. In a clinical context, such a high AUC ensures that the Digital Twin can provide trustworthy predictions that clinicians can rely on for decision-making.

Fig. 7: “SHAP summary plot showing feature importance in predictions.”

“Fig. 7” presents the SHAP (SHapley Additive exPlana-tions) feature importance summary for the MLP predictions. It shows that EEG Band 3 and GSR values were the most influential features driving classification decisions. Higher GSR values generally pushed predictions toward the high-risk class, while stable EEG patterns (particularly Band 1 and Band 2) were associated with low-risk outcomes. EEG Band 4 and activity duration had moderate effects, acting as secondary indicators. The SHAP analysis provides interpretability by explaining how each input feature contributes to a given prediction, which is crucial for clinical acceptance of AI-driven systems. This transparency ensures that the Digital Twin does not function as a “black box,” but instead offers understandable insights to healthcare professionals.

E. Test *Case Results*

To further validate the system, test cases were examined. Figures are presented first, followed by their interpretations.

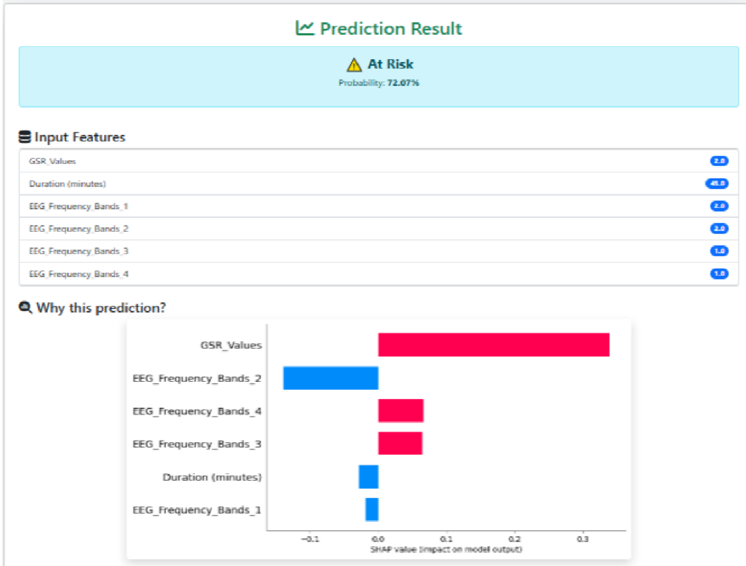

Fig. 8: “Test Case 1: High-risk prediction with SHAP expla-nation (72.07%).”

Test Case 1 – High Risk. In “Fig. 8,” the system classified the subject as high risk with a probability of 72.07%. The decision was primarily influenced by elevated GSR values and unstable EEG Bands 3 and 4. A slight negative effect from EEG Band 2 reduced the probability marginally but did not change the final decision.

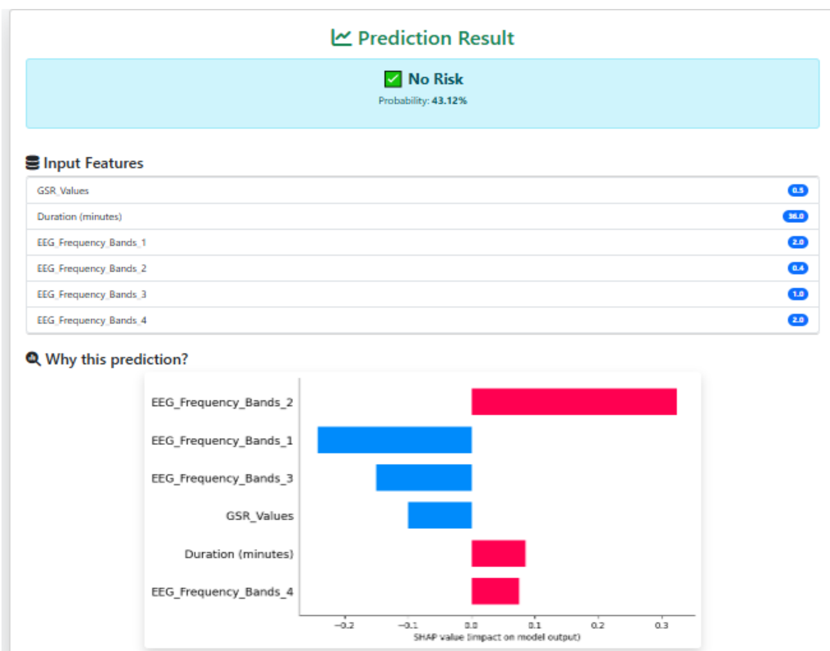

Fig. 9: “Test Case 2: Low-risk prediction with SHAP expla-nation (43.12%).”

Test Case 2 – Low Risk. “Fig. 9” shows a low-risk classifi-cation with a probability of 43.12%. Here, EEG Bands 1 and 2 had strong negative contributions that outweighed positive influences from Band 4 and duration, resulting in a “No Risk” outcome.

F. Discussion

The findings confirm that the proposed MLP within a Digital Twin framework provides reliable and interpretable predictions. The confusion matrix showed that misclassifi-cations were minimal, the ROC curve highlighted excellent discriminative power, and SHAP plots ensured transparency of decision-making. Test case results further illustrated the model's ability to distinguish between calm, stressed, and borderline conditions, supporting its potential application in real-world clinical monitoring.

9. CONCLUSION

This work introduced a Digital Twin framework for mental health risk prediction using physiological signals, specifically

EEG frequency bands and Galvanic Skin Response (GSR). To address the challenge of limited data availability, a syn-thetic dataset of 10,000 samples was generated, enabling the training of a Multilayer Perceptron (MLP) model. The system achieved an accuracy of 92.5% and an AUC of 0.98, demonstrating reliable performance in distinguishing between low- and high-risk cases. Compared to baseline models such as Logistic Regression and Random Forest, the MLP provided superior results, highlighting the advantage of deep learning for capturing nonlinear relationships in physiological data. In addition, the integration of SHAP explainability ensured transparency by identifying EEG Band 3 and GSR as the most influential predictors, aligning with known stress-related indicators. These outcomes establish the proposed framework as both accurate and interpretable, making it a promising foun-dation for clinician-assisted monitoring and decision-making in mental health care.

Ethical and privacy considerations are essential for any mental health application. In the present study, the dataset used for model development was anonymized and contained no personally identifiable information. For prospective real-world implementations, we will adopt established best practices: obtain informed consent, apply data minimization, store data encrypted at rest and in transit, use role-based access control, and follow institutional review board (IRB) guidelines. We will also investigate differential privacy and secure aggregation techniques when aggregating data from multiple clinical sites to further protect patient confidentiality.

A. Future Work

Although the results are encouraging, further refinement is necessary for real-world deployment. Validation with larger, more diverse datasets will be essential to confirm general-izability across different populations. Extending the frame-work to include additional modalities such as heart rate vari-ability, respiration, or photoplethysmography could improve robustness and reduce uncertainty in borderline predictions. Incorporating advanced temporal models, such as Long Short-Term Memory (LSTM) networks or Transformer-based archi-tectures, may also capture dynamic changes in physiological signals more effectively. From a practical perspective, optimiz-ing the framework for deployment on lightweight platforms like Raspberry Pi or Jetson Nano would support real-time monitoring in portable healthcare environments. Attention must also be given to ethical and privacy considerations, ensuring that sensitive health data is managed securely and that model predictions remain transparent to clinicians and patients. Finally, expanding the Digital Twin to simulate potential interventions—such as stress-reduction strategies or therapy adjustments—would provide clinicians with a valu-able decision-support tool before implementing treatments in practice. With these improvements, the framework can evolve into a comprehensive, scalable, and trustworthy solution for mental health monitoring.

Planned future work includes validation using real-world physiological datasets obtained through clinical collabora-tions to evaluate practical applicability and to further fine-tune model parameters for diverse populations.

References

1. D. Kamin´ska, "EEG and GSR-based stress recognition with statistical feature extraction," Sensors, vol. 23, no. 5, pp. 1–15, 2023.
2. M. Arsalan, M. O. Aslam, A. Alarfaj, and H. Haider, "Multimodal stress detection using EEG, GSR and PPG with classical classifiers," IEEE Access, vol. 7, pp. 46436–46445, 2019.
3. M. Majid, S. Hussain, M. R. Lee, and J. Baik, "Feature engineering for multimodal physiological stress detection," Computers in Biology and Medicine, vol. 144, p. 105376, 2022.
4. N. Phutela, P. Mehta, and S. Arora, "EEG-based stress detection using LSTM networks," Biomedical Signal Processing and Control, vol. 75,p. 103591, 2022.
5. S. Sun and Y. Zhang, "Digital Twin in medicine: Concepts, applications, and challenges," Journal of Medical Systems, vol. 46, no. 3, pp. 1–12, 2022.

6. M. Katsoulakis, E. Gialeli, and G. Styliadis, “Digital Twins in health-care: A scoping review,” Frontiers in Digital Health, vol. 6, p. 1458723, 2024.
7. N. Papachristou and I. Papadopoulos, “Precision medicine and Digital Twins: Toward personalized healthcare,” IEEE Reviews in Biomedical Engineering, vol. 17, pp. 200–213, 2024.
8. I. Meijer and A. Bredenoord, “Ethical aspects of Digital Twins in health care,” Journal of Medical Ethics, vol. 49, no. 5, pp. 322–329, 2023.
9. R. Sade´e, L. Kumar, and A. James, “Explainability in Digital Twin-based healthcare systems,” Artificial Intelligence in Medicine, vol. 152,p. 102236, 2025.
10. M. H. Kabir and M. Sultana, “IoT-enabled Digital Twin for real-time healthcare monitoring,” IEEE Internet of Things Journal, vol. 12, no. 1,pp. 112–125, 2025.
11. T. Valle´e and A. Vasilenko, “Workflow optimization in hospitals using Digital Twins,” Healthcare Management Science, vol. 26, no. 4, pp. 673–689, 2023.
12. A. Elkeffi and O. Asan, “A rapid review of Digital Twins in healthcare management,” BMJ Health & Care Informatics, vol. 29, no. 1, p. e200427, 2022.
13. M. Nadeem, H. Shah, and S. Zafar, “Simulation-based Digital Twins for chronic disease monitoring,” Simulation Modelling Practice and Theory, vol. 134, p. 102747, 2025.
14. R. Mohanty and A. Alla, “Digital Twins for inclusive healthcare in smart societies,” IEEE Access, vol. 13, pp. 44567–44580, 2025.
15. Y. Yang, L. Chen, and X. Li, “EEG-based mental workload recognition using deep convolutional neural networks,” IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 31, no. 2, pp. 150–160, 2023..