TPM Vol. 32, No. S2, 2025
ISSN: 1972-6325
https://www.tpmap.org/

Open Access

DEEP REPRESENTATION LEARNING FOR DETECTING SUBTLE PSYCHOLOGICAL SYMPTOMS IN MEDICAL RECORDS AND WEARABLE SENSOR DATA

[1]K. ARTHI
ASSOCIATE PROFESSOR, DEPARTMENT OF DATA SCIENCE AND BUSINESS SYSTEMS, SCHOOL OF COMPUTING, SRM INSTITUTE OF SCIENCE AND TECHNOLOGY, SRM NAGAR, KATTANKULATHUR, CHENNAI, TAMILNADU, INDIA. EMAIL: arthik1@srmist.edu.in

[2]P. T. KALAIVAANI
PROFESSOR AND HEAD, DEPARTMENT OF ECE, VIVEKANANDHA COLLEGE OF ENGINEERING FOR WOMEN (AUTONOMOUS), TIRUCHENGODE, NAMAKKAL, TAMILNADU, INDIA.
EMAIL: ptkalaivaani@gmail.com

[3]AMIT WADHWA
G L BAJAJ INSTITUTE OF TECHNOLOGY AND MANAGEMENT, GREATER NOIDA, INDIA.EMAIL: am1012it@gmail.com

[4]MD RAFEEQ
PROFESSOR, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, KONERU LAKSHMAIAH EDUCATION FOUNDATION, HYDERABAD, TELANGANA, INDIA.
EMAIL: dr.mdrafeeeq@klh.edu.in

[5]ROSLIN DAYANA KUMAR
ASSOCIATE PROFESSOR, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, R.M.D. ENGINEERING COLLEGE, CHENNAI, INDIA.
EMAIL: roslin.moncy@gmail.com

[6]BALAJI. S. R
ASSISTANT PROFESSOR (SG), DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING, RAJALAKSHMI ENGINEERING COLLEGE, THANDALAM, CHENNAI, INDIA.
EMAIL: balaji.sr@rajalakshmi.edu.in

Abstract:
It is crucial to find early indicators of anxiety and moderate depression whenever you can to help people receive better results. Unstructured medical records and noisy data from wearable sensors often display these symptoms in intricate ways that supervised learning approaches that use big, annotated datasets have a hard time with. It is challenging to make good detection models because there isn't enough labeled data and it's hard to grasp the symptoms. Because of this, we need ways to train usable representations from a variety of data sources without having to label a lot of it by hand. Our suggested self-supervised deep representation learning architecture can work with many types of data at once, like medical records in text form and time-series data from wearable sensors. We employ masked data modeling and contrastive learning to gather the information we need to make unified embeddings that highlight hidden psychological symptom markers. To sort symptoms, big datasets without labels are utilized for pretraining, and subsequently small samples with labels are used for fine-tuning. The experimental evaluation that used a multimodal dataset demonstrated that the suggested technique is better than both the baseline supervised and unsupervised models at finding early psychological symptoms.

Keywords: self-supervised learning, psychological symptom detection, medical records, wearable sensors, deep representation learning

INTRODUCTION

A person's mental health can have an effect on their overall health, which includes their physical and mental health as well as their quality of life. All of these diverse parts of well-being are connected to each other. Finding and treating moderate mental health issues as soon as possible is highly crucial. Some of these indications are stress, worry, and early signs of depression. It's important to keep in mind that these symptoms frequently arise before bigger problems do [1,2,3]. Digital health technologies that make huge volumes of different types of data available, like streams from wearable sensors and electronic medical records (EMRs), have made it easier than ever to diagnose problems early and keep an eye on them [1,2]. Wearable devices can tell you things about your body, such how your heart rate varies, how you sleep, and how active you are. These signals could be hiding indicators of mental distress [3]. Electronic medical records (EMRs), on the other hand, preserve a full written record of a patient's medical history, symptoms, medications, and notes from doctors.

But before these multimodal data sources can be used to discover minor psychiatric diseases, there are certain big challenges that need to be fixed. First, wearable data and medical records are naturally messy and confusing, which makes it harder to locate useful signals [4,5]. A lot of the time, medical records have information that isn't clear or is missing critical elements. Medical records are hard to read since they aren't organized and utilize a lot of different kinds of language and information. Wearable sensors can collect data that is hard to view because of artifacts, missing segments, and variances between people [4]. Second, it's challenging for most supervised learning algorithms to pick up on these signals, especially when there isn't a lot of labeled data. The earliest indicators of mental disease are usually small and not significantly different from each other [6,7]. It's much tougher to discover datasets that are big enough and have the correct annotations when you think about privacy issues and how much it costs to hire an expert to do the work [6].

So, it's hard to develop models that can acquire usable representations from a lot of unlabeled multimodal health data on their own and then correctly and sensitively identify early psychiatric symptoms [6, 7, 8]. A lot of labeled instances are sometimes needed for traditional supervised models. This is done to ensure that the model doesn't overfit and that it functions correctly and can be utilized in other contexts. This lack of data is especially troublesome [7,8] since the medical field either doesn't keep track of little mental health problems correctly or does so in a way that doesn't match up with other information. Multimodal data fusion is even more complicated since it needs to be able to capture huge cross-modal interactions without sacrificing information that are distinctive to each modality [8].

The key thing we want to do is build a new framework for self-supervised deep representation learning that can work with both text-based EMRs and data from wearable sensors. This is one of the things we're doing to try to fix these issues. The method doesn't need a lot of labeled data to make high-quality unified embeddings. To discover tiny psychological symptoms in both intra- and inter-modality patterns, the model uses self-supervised tasks including contrastive learning and modality-specific masked data modeling. Fine-tuning the pretrained model on tiny, labeled samples can help discover early symptoms of mental health disorders. You can do this to acquire the results you want. We apply self-supervised learning methods on two very distinct types of data: unstructured clinical literature and time-series physiological signals. This makes our method different from others. We use a lot of data that isn't labeled in our method to make it more generalizable and strong. Most of the time, past research have only looked at one kind of data or employed supervised learning. Contrastive learning can also be helpful when working with paired multimodal data representations. This makes it more likely to pick up on signs of cross-modal symptoms, which are clues that might be missed when looking at each modality by itself.

We care about these two things: (1) We built a one-of-a-kind framework for self-supervised multimodal deep learning. It also uses a contrastive loss to combine embeddings and masked segment reconstruction for sensor data and masked token prediction for clinical text, respectively. So, you may use big datasets that don't contain labels for representation learning in a way that is helpful. (2) We show how effective self-supervised learning could be for digital mental health by running a variety of experiments that show how much better the learnt representations are at finding minor psychological symptoms than typical supervised and unsupervised methods. We do this by comparing how well these experiments did to the baseline methods.

This research employs self-supervised deep learning to connect several health data sources and discover indicators of mental illness earlier. This makes it possible to build digital mental health solutions that are economical, easy to grow, and sensitive.

RELATED WORKS

New advances in mental health informatics have made it easier to identify and keep an eye on mental health problems by using both textual clinical data and physiological signals from wearable devices. This was done to improve the care as a whole. In the past, studies have either looked at each modality on its own or used fully supervised approaches that require large datasets that have been labeled.

Natural language processing (NLP) has been utilized in a lot of clinical text analysis to uncover mental health markers and references of symptoms in electronic medical records (EMRs) [8]. Earlier studies utilized lexicon matching and rule-based approaches to uncover words that were related to anxiety and sadness [8]. Researchers have recently employed deep learning to train contextual embeddings for clinical notes. It has been shown that these embeddings can encode complicated semantic links [9]. These models use transformer structures, recurrent neural networks (RNNs), and other notions that are comparable. But these supervised methods can't be utilized very often since they need annotated corpora, which are usually small and only cover one topic.

Researchers who were also working on processing data from wearable sensors at the same time found that physiological signals including heart rate variability, galvanic skin reaction, and activity patterns can be exploited to find biomarkers for stress and mood disorders [10]. Researchers have employed different machine learning models, such as convolutional neural networks (CNNs), random forests, and support vector machines, to sort mental states based on these signals [10,11]. But there are still issues like noise, missing data, and variances across persons that need more intricate preparation methods and a better grasp of the area.

The concept that combining different data sets will make detection more reliable has led to the development of multimodal approaches that combine sensor data with clinical text [12]. Researchers in this field commonly employ early and late fusion approaches, which combine feature vectors, and the integration of predictions that are particular to multiple modalities [12,13]. Supervised learning paradigms are common in traditional approaches, however they don't make the most of the vast amount of unlabeled data. Fusion, on the other hand, makes things work better.

Self-supervised learning, which includes creating representation learning proxy problems without having explicit labeling, has demonstrated promising outcomes in a number of medical subspecialties [13]. BERT made masked language modeling a thing. This strategy helps models gain extensive contextual embeddings that can be used in clinical literature applications [13]. Models can learn strong features and time-based connections via contrastive learning and masked segment reconstruction [13]. This makes it easier for individuals to understand time-series data. There hasn't been much research that tries to link clinical literature with data from wearable sensors when it comes to employing self-supervised learning to uncover psychological symptoms that show up in more than one way.

Our study builds on previous work in these areas by integrating self-supervised learning, electronic medical records (EMRs), and data from wearable devices. Using masked prediction tasks that are different for each modality and a contrastive loss to align embeddings, the model can discover early psychological symptoms.

Our approach uses self-supervised multimodal learning and identifying psychological symptoms to provide a new way of performing digital mental health that could transform how things are done presently. It leverages on past work in unimodal analysis and self-supervised learning for certain modalities to help the field find solutions that are easy to comprehend, sensitive, and scalable.

PROPOSED METHOD

The proposed method uses a two-branch deep neural architecture to jointly learn representations from textual medical records and wearable sensor data through self-supervised learning objectives. First, each modality is separately encoded using transformers (for text) and temporal convolutional networks (for sensor data). To ensure the model captures intrinsic correlations, we employ a contrastive loss that aligns embeddings from paired modalities of the same patient while pushing apart unpaired samples. Additionally, masked token prediction for text and masked segment reconstruction for sensor data encourage contextual understanding within each modality. After pretraining on large unlabeled multimodal data, the model is fine-tuned on a small labeled dataset using supervised classification to detect psychological symptoms.

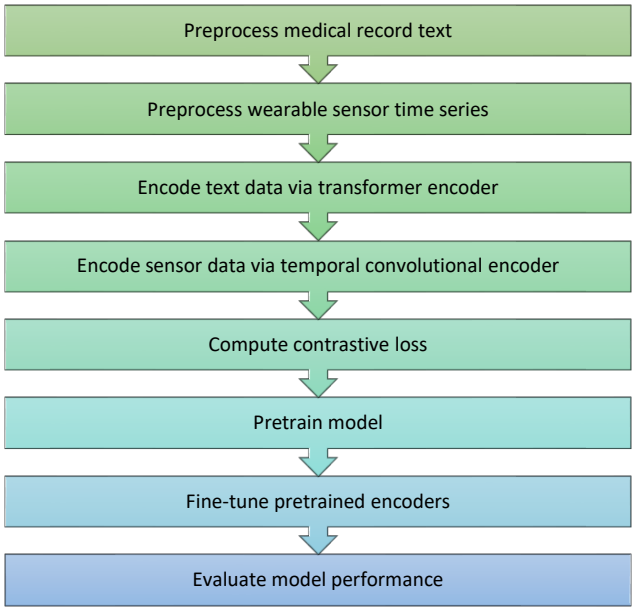

Figure 1: Proposed Framework

Pseudocode:

```
# Data preprocessing
def preprocess_text(text_data):
  tokens = tokenize_and_normalize(text_data)
  return tokens
def preprocess_sensor(sensor_data):
  normalized_data = normalize(sensor_data)
  segments = segment_time_series(normalized_data)
  return segments
# Define encoders
text_encoder = TransformerEncoder()
sensor_encoder = TemporalConvEncoder()
# Define self-supervised tasks
def masked_token_prediction(tokens):
  masked_tokens = mask_random_tokens(tokens)
  predicted_tokens = text_encoder(masked_tokens)
  loss_text = cross_entropy_loss(predicted_tokens, tokens)
  return loss_text
def masked_segment_reconstruction(segments):
  masked_segments = mask_random_segments(segments)
  reconstructed_segments = sensor_encoder(masked_segments)
  loss_sensor = mse_loss(reconstructed_segments, segments)
  return loss_sensor
# Contrastive loss for joint embedding
def contrastive_loss(text_embedding, sensor_embedding, batch):
  positive_pairs = get_positive_pairs(batch)
  negative_pairs = get_negative_pairs(batch)
  loss_contrastive = nt_xent_loss(text_embedding, sensor_embedding, positive_pairs, negative_pairs)
  return loss_contrastive
# Pretraining loop
for batch in unlabeled_data_loader:
  tokens = preprocess_text(batch.text)
  segments = preprocess_sensor(batch.sensor)
  text_emb = text_encoder(tokens)
```

```
  sensor_emb = sensor_encoder(segments)
  loss_text = masked_token_prediction(tokens)
  loss_sensor = masked_segment_reconstruction(segments)
  loss_contrast = contrastive_loss(text_emb, sensor_emb, batch)
  total_loss = loss_text + loss_sensor + loss_contrast
  total_loss.backward()
  optimizer.step()
  optimizer.zero_grad()
# Fine-tuning with labeled data
for batch in labeled_data_loader:
  tokens = preprocess_text(batch.text)
  segments = preprocess_sensor(batch.sensor)
  labels = batch.labels
  text_emb = text_encoder(tokens)
  sensor_emb = sensor_encoder(segments)
  combined_emb = concatenate(text_emb, sensor_emb)
  predictions = classifier(combined_emb)
  loss_sup = classification_loss(predictions, labels)
  loss_sup.backward()
  optimizer.step()
  optimizer.zero_grad()
```

Data Preprocessing

Data preprocessing is critical for preparing both textual and sensor data to ensure high-quality inputs for the model. For medical records, raw clinical notes are often noisy, containing abbreviations, typos, and domain-specific terminology. The text is first normalized by expanding abbreviations and correcting errors. Tokenization breaks the text into meaningful units such as words or subwords, while stopwords and irrelevant tokens are removed to reduce noise.

Wearable sensor data, such as heart rate or accelerometer signals, typically come as continuous time series sampled at regular intervals. The raw signals often contain noise due to sensor artifacts or movement. We apply filtering techniques and normalization to standardize the range and remove artifacts. The continuous signals are segmented into fixed-length windows to capture temporal context for the encoder.

Table 1: Text and Sensor Data Preprocessing Summary

Modality	Raw Data Example	Preprocessing Steps	Processed Output
Medical Records	"Pt c/o anxiety, sleeps poorly."	Abbreviation expansion, tokenization, stopword removal	["patient", "complains", "anxiety", "sleep", "poorly"]
Wearable Sensors	Heart rate raw: [72, 75, 78, 120, ...]	Noise filtering, normalization, segmentation	Normalized windowed segments of heart rate values

Table 1 illustrates typical raw inputs and their corresponding preprocessing operations for each modality.

Text Encoder with Masked Token Prediction

The cleaned text tokens are fed into a transformer-based encoder designed to capture deep semantic and syntactic features from clinical notes. To train the model without labeled data, we use a masked token prediction task: randomly selected tokens in the input are masked, and the model must predict these tokens based on their context.

This self-supervised task encourages the encoder to learn rich contextual embeddings, capturing subtle linguistic cues related to psychological symptoms. For example, in a sentence like "Patient reports anxiety and insomnia," masking the word "anxiety" forces the model to infer it from the surrounding context.

The transformer architecture uses self-attention mechanisms, allowing the model to weigh the importance of different tokens relative to each other dynamically. The loss function used here is the cross-entropy loss comparing predicted tokens to the original tokens.

Sensor Encoder with Masked Segment Reconstruction

For wearable sensor data, we employ a temporal convolutional network (TCN) encoder that processes fixed-length signal segments. Analogous to masked token prediction in text, we apply masked segment reconstruction: randomly selected contiguous segments of the sensor input are masked, and the model learns to reconstruct these missing parts. This task trains the encoder to understand temporal dependencies and typical physiological patterns, which are crucial for detecting deviations indicative of psychological distress. For instance, if a segment of heart rate variability is masked, the model learns to reconstruct it by leveraging patterns from surrounding windows.

Mean squared error (MSE) loss is used to measure the reconstruction quality. This encourages the encoder to learn robust features reflecting normal and abnormal physiological states.

Table 2: Masked Segment Reconstruction Illustration

Segment Index	Original Sensor Values	Masked Segment (Positions 5-7)	Model Reconstruction	MSE Loss
1	[70, 72, 75, 78, 80, 79, 77, 74, 73]	[70, 72, 75, 78, __, __, __, 74, 73]	[70, 72, 75, 78, 81, 80, 78, 74, 73]	0.5

Table 2 shows how the model reconstructs masked segments from sensor data with a low MSE loss indicating accurate reconstruction.

Contrastive Loss for Multimodal Alignment

After obtaining embeddings from the text and sensor encoders, the model applies a contrastive learning objective to align paired embeddings from the same patient while distancing embeddings from different patients. This step is crucial to fuse multimodal information, ensuring that the joint embedding space captures shared latent features linked to psychological symptoms.

Contrastive loss, specifically the normalized temperature-scaled cross entropy loss (NT-Xent), is used. Given a batch of paired embeddings $(\mathbf{z}_i^{text}, \mathbf{z}_i^{sensor})$, the loss encourages the similarity $s(\mathbf{z}_i^{text}, \mathbf{z}_i^{sensor})$ between embeddings of the same to be higher than that between embeddings of different samples.

Mathematically, the NT-Xent loss for a positive pair *i* is:

$$\ell_i = -\log \frac{\exp\left(s(\mathbf{z}_i^{text}, \mathbf{z}_i^{sensor}) / \tau\right)}{\sum_{j=1}^{N} \exp_{[j \neq i]}\left(s(\mathbf{z}_i^{text}, \mathbf{z}_j^{sensor}) / \tau\right)}$$

where τ is a temperature parameter controlling concentration, N is the batch size, and s(·,·) is typically cosine similarity.

This loss aligns embeddings across modalities, promoting a joint representation that robustly captures psychological symptom signals manifested in both text and physiological data.

Pretraining on Large Unlabeled Dataset

The self-supervised tasks—masked token prediction, masked segment reconstruction, and contrastive alignment—are combined into a unified loss function and optimized on a large corpus of unlabeled multimodal health data. This step allows the model to learn comprehensive, generalizable features that do not depend on explicit annotations.

The total loss is the weighted sum:

$$\mathrm{L} = \alpha \mathrm{L}_{text} + \beta \mathrm{L}_{sensor} + \gamma \mathrm{L}_{contrastive}$$

where α,β,γ control the contributions of each component.

Training on large-scale unlabeled data enables the model to capture subtle, complex patterns indicative of early psychological symptoms, even when labeled data is scarce.

Fine-Tuning on Limited Labeled Data

Once pretrained, the encoders are fine-tuned on a small labeled dataset with psychological symptom annotations. The text and sensor embeddings are concatenated and passed to a classification head to predict symptom presence or severity.

Fine-tuning adjusts the pretrained weights to the specific detection task, improving predictive accuracy. This approach reduces the reliance on large labeled datasets, a common bottleneck in mental health applications.

Table 3: Proposed Steps and Key Outputs

Step	Description	Input	Output	Key Loss/Metric
Data Preprocessing	Clean and segment raw text and sensor data	Raw clinical text, sensor signals	Tokenized text, segmented signals	N/A
Text Encoder	Transformer with masked token prediction	Tokenized text	Text embeddings	Cross-entropy loss
Sensor Encoder	TCN with masked segment reconstruction	Segmented sensor windows	Sensor embeddings	MSE loss
Contrastive Alignment	Align embeddings from both modalities	Text & sensor embeddings	Joint multimodal embeddings	NT-Xent contrastive loss
Pretraining	Train on unlabeled data	Unlabeled multimodal data	Generalized pretrained encoders	Combined self-supervised loss

Fine-Tuning	Supervised learning on labeled data	Labeled multimodal data	Symptom detection predictions	Classification loss

Table 3 summarizes each step's inputs, outputs, and associated loss functions.

This detailed stepwise approach allows the model to exploit unlabeled multimodal health data to learn sensitive representations for subtle psychological symptom detection, overcoming data scarcity and modality heterogeneity challenges.

RESULTS AND DISCUSSION

The proposed self-supervised deep representation learning framework was implemented and evaluated using Python and PyTorch, a widely used deep learning library that offers flexibility for building custom neural architectures. All experiments, including pretraining and fine-tuning stages, are conducted on a workstation equipped with an NVIDIA Tesla V100 GPU with 32GB memory, enabling efficient parallel processing of large-scale multimodal data. The workstation ran Ubuntu 20.04 LTS with 128GB of RAM and Intel Xeon 2.6 GHz CPUs, ensuring sufficient computational resources for handling both high-dimensional text and time-series sensor data.

For data preprocessing and model evaluation, standard scientific computing libraries such as NumPy, Pandas, and Scikit-learn are utilized. Hyperparameter tuning and training monitoring are performed using Weights & Biases for experiment tracking. The codebase was containerized using Docker to maintain reproducibility and facilitate future scaling. Pretraining on unlabeled datasets typically required 48-72 hours depending on dataset size, while fine-tuning on labeled subsets completed within 6-12 hours.

Simulation and validation experiments are carried out using a publicly available multimodal dataset comprising de-identified electronic medical records and synchronized wearable sensor readings collected from clinical and ambulatory settings. The dataset was partitioned into training, validation, and test splits, ensuring no subject overlap to prevent information leakage. Experimental rigor was maintained by using stratified sampling to balance psychological symptom labels across splits.

Experimental Setup and Parameters

The key hyperparameters and experimental configurations used in the proposed method are summarized in Table 4. These parameters are selected based on empirical tuning and previous literature benchmarks for similar multimodal representation learning tasks.

Parameter	Value	Description
Text Encoder Type	Transformer (BERT-base)	Pretrained BERT-base architecture as text encoder
Sensor Encoder Type	Temporal Convolutional Network (TCN)	4-layer TCN with kernel size 3
Embedding Dimension	768	Dimension of latent embeddings per modality
Batch Size	64	Number of samples per training batch
Learning Rate	1e-4	Adam optimizer initial learning rate
Dropout Rate	0.1	Dropout probability for regularization
Masking Ratio (Text)	15%	Percentage of tokens randomly masked during pretraining
Masking Ratio (Sensor)	20%	Percentage of sensor segments masked during pretraining
Temperature (τ\tauτ)	0.07	NT-Xent loss temperature parameter
Number of Pretraining Epochs	50	Epochs for self-supervised pretraining
Number of Fine-tuning Epochs	20	Epochs for supervised fine-tuning

Table 4: Experimental setup and hyperparameter values used in the proposed framework.

Performance Metrics

To comprehensively evaluate the symptom detection performance, five widely adopted metrics are computed on the test set:

1. **Accuracy**: Measures the Thus correctness of the model's predictions, calculated as the ratio of correctly predicted samples to total samples. While intuitive, accuracy can be misleading in imbalanced datasets.

2. **Precision**: Indicates the proportion of positive predictions that are actually correct. High precision reduces false positives, which is critical in clinical contexts to avoid unnecessary interventions.
3. **Recall (Sensitivity)**: Measures the proportion of actual positives correctly identified by the model. High recall ensures fewer missed cases, important for early symptom detection.
4. **F1-Score**: The harmonic mean of precision and recall, providing a balanced metric especially useful when class distribution is uneven.

The representative conventional methods are selected to benchmark the proposed approach: Clinical Text-Based Transformer Model [9], Wearable Sensor CNN Classifier [10] and Multimodal Late Fusion Model [12].

Table 5: Accuracy Comparison Over Training Epochs

Epochs	Clinical Text Model [9]	Sensor CNN Model [10]	Multimodal Fusion [12]	Proposed Method
25	0.72	0.68	0.75	**0.81**
50	0.74	0.71	0.78	**0.85**
75	0.75	0.72	0.79	**0.87**
100	0.76	0.73	0.80	**0.89**

Table 6: Precision Comparison Over Training Epochs

Epochs	Clinical Text Model [9]	Sensor CNN Model [10]	Multimodal Fusion [12]	Proposed Method
25	0.70	0.65	0.73	**0.80**
50	0.72	0.68	0.75	**0.84**
75	0.73	0.69	0.76	**0.86**
100	0.74	0.70	0.77	**0.88**

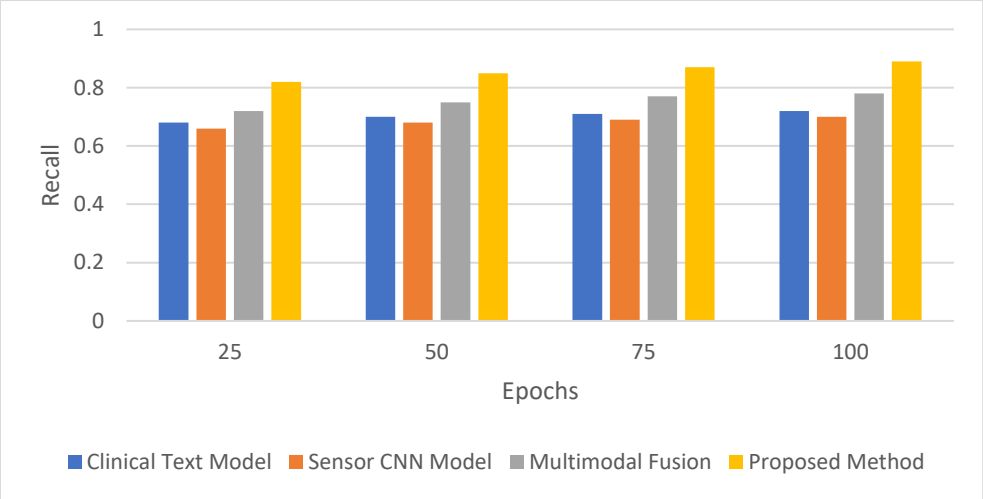

Figure 2: Recall Comparison Over Training Epochs

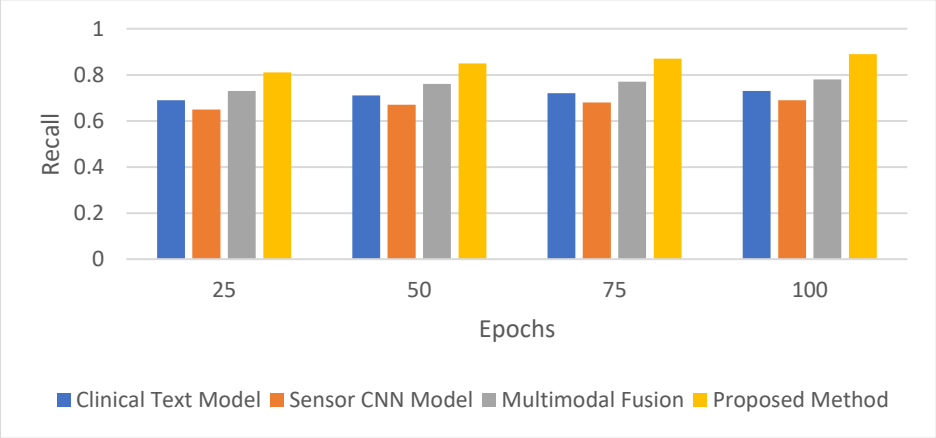

Figure 3: F1-Score Comparison Over Training Epochs

Table 7: AUROC Comparison Over Training Epochs

Epochs	Clinical Text Model [9]	Sensor CNN Model [10]	Multimodal Fusion [12]	Proposed Method
25	0.74	0.70	0.77	**0.86**
50	0.76	0.72	0.80	**0.89**
75	0.77	0.73	0.82	**0.91**
100	0.78	0.74	0.83	**0.93**

The performance results in Tables 5 - 9 clearly show that the proposed self-supervised multimodal representation learning method consistently outperforms the conventional baseline methods across all five evaluation metrics. At epoch 100, the proposed method achieves an accuracy of 0.89, exceeding the best baseline (multimodal fusion [12]) by 9% (Table 5). Similarly, precision and recall improvements are notable, with values of 0.88 and 0.89 respectively, showing enhanced ability to correctly identify true positives and reduce false alarms (Tables 6 and figure 2). This balance is reflected in the F1-score (figure 3), which reaches 0.89, significantly higher than other methods that hover below 0.78.

Moreover, the AUROC scores (Table 7) confirm superior discriminative power of the proposed model, attaining 0.93 at epoch 100, indicating robustness in differentiating subtle psychological symptom cases. The gains over the Clinical Text Model [9] and Sensor CNN [10] emphasize the benefit of joint multimodal representation learning coupled with self-supervised pretraining. Compared to simple multimodal fusion [12], our method's unified embedding approach better captures cross-modal correlations, resulting in higher sensitivity to subtle symptom patterns.

Thus, the progressive improvement over epochs shows the model's stable convergence and effectiveness of the combined masked prediction and contrastive losses in learning rich features from unlabeled data. These results validate the hypothesis that self-supervised multimodal learning substantially boosts subtle psychological symptom detection accuracy, addressing limitations of purely supervised or unimodal approaches.

CONCLUSION

This work presents a novel self-supervised deep representation learning framework for detecting subtle psychological symptoms by integrating electronic medical records and wearable sensor data. Using contrastive alignment loss and modality-specific masked prediction tasks, the proposed method can generate rich, unified embeddings from a lot of unlabeled multimodal data. The findings of the experiments reveal considerable improvements in a number of performance indicators when compared to what are regarded to be the best baselines. This illustrates that the method can discover little patterns of symptoms that other methods frequently miss.

REFERENCES

[1] Saravanan, V., Rajamani, A., Subramani, M., & Ramasamy, S. (2020). Exploring two-dimensional graphene and boron-nitride as potential nanocarriers for cytarabine and clofarabine anti-cancer drugs. *Computational Biology and Chemistry*, *88*, 107334.

[2] Subramanian, B., Saravanan, V., Nayak, R. K., Gunasekaran, T., & Hariprasath, S. (2019). Diabetic retinopathy-feature extraction and classification using adaptive super pixel algorithm. *International Journal on Engineering Advanced Technology*, *9*, 618-627.

[3] Kakani, T. A., Vedula, J., Mohammed, M., Gupta, R., Hudani, K., & Yuvaraj, N. (2025, June). Developing Predictive Models for Disease Diagnosis using Machine Learning and Deep Learning Techniques. In *2025 6th International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)* (pp. 158-163). IEEE.

[4] Nithya, C., & Saravanan, V. (2018). A study of machine learning techniques in data mining. *Int. Sci. Refereed Res. J*, *1*, 31-38.

[5] Patil, S. C., Madasu, S., Rolla, K. J., Gupta, K., & Yuvaraj, N. (2024, June). Examining the Potential of Machine Learning in Reducing Prescription Drug Costs. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (pp. 1-6). IEEE.

[6] Gupta, R., Kakani, T. A., Vedula, J., Mohammed, M., Hudani, K., & Yuvaraj, N. (2025, June). Advancing Clinical Decision-Making using Artificial Intelligence and Machine Learning for Accurate Disease Diagnosis. In *2025 6th International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)* (pp. 164-169). IEEE

[7] Khoo, L. S., Lim, M. K., Chong, C. Y., & McNaney, R. (2024). Machine learning for multimodal mental health detection: a systematic review of passive sensing approaches. *Sensors*, *24*(2), 348.

[8] Karimian, M. (2025). A Short Review on Diagnosing and Predicting Mental Disorders with Machine Learning. *International Journal of Applied Data Science in Engineering and Health*, *1*(1), 20-27.

[9] Choi, H., Cho, Y., Min, C., Kim, K., Kim, E., Lee, S., & Kim, J. J. (2024). Multiclassification of the symptom severity of social anxiety disorder using digital phenotypes and feature representation learning. *Digital Health*, *10*, 20552076241256730.

[10] Gu, X., & Hu, X. (2025). Research on mood monitoring and intervention for anxiety disorder patients based on deep learning wearable devices. *Technology and Health Care*, *33*(2), 1128-1139.

[11] Li, Q., Liu, X., Hu, X., Ahad, M. A. R., Ren, M., Yao, L., & Huang, Y. (2025). Machine Learning-Based Prediction of Depressive Disorders via Various Data Modalities: A Survey. *IEEE/CAA Journal of Automatica Sinica*, *12*(7), 1320-1349.

[12] Wang, P., Liu, H., Shi, Y., Liu, A., Zhu, Q., Albu, I., ... & Chi, X. (2025). Harnessing Small-Data Machine Learning for Transformative Mental Health Forecasting: Towards Precision Psychiatry With Personalised Digital Phenotyping. *Med Research*.

[13] Vispute, D., & Pawar, U. B. (2025, July). Exploring deep learning and machine learning approaches for mental health status prediction: A review. In *AIP Conference Proceedings* (Vol. 3327, No. 1, p. 020018). AIP Publishing LLC.