Journal of Computer and Communications, 2025, 13(3), 116-137
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

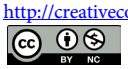
A Dual-Channel Prediction-Interpretation Framework with Pre-Trained Language Models and SHAP Explainability

Hui Nie, Xiaoyan Wu

School of Information Management, Sun Yat-sen University, Guangzhou, China
Email: issnh@mail.sysu.edu.cn

How to cite this paper: Nie, H. and Wu, X.Y. (2025) A Dual-Channel Prediction-Interpretation Framework with Pre-Trained Language Models and SHAP Explainability. *Journal of Computer and Communications*, **13**, 116-137.
https://doi.org/10.4236/jcc.2025.133009

Received: February 14, 2025
Accepted: March 23, 2025
Published: March 26, 2025



Abstract

This study addresses the challenges of data noise and model interpretability in depression diagnosis by proposing an intelligent diagnostic framework based on real-world medical scenarios. Utilizing a labeled dataset of 11,188 Chinese online consultation records, we developed a dual-channel architecture integrating BERT/RoBERTa pre-trained models with the SHAP interpretability framework for depression severity classification. Experimental results demonstrated that the BERT model achieved 92% overall accuracy, with 93% accuracy specifically for severe depression detection. SHAP analysis revealed the model's focus on clinically relevant features like suicidal tendencies and low mood, showing significant alignment with DSM-5 diagnostic criteria. The study confirms pre-trained models' capability in extracting pathological semantics from medical texts, while the "prediction-interpretation" framework provides a technical prototype for overcoming clinical application barriers of black-box models.

Keywords

Depression Prediction, Explainable Machine Learning, BERT Model, Patient Narrative, SHAP Analysis

1. Introduction

Depression, a pressing global public health issue, presents a significant challenge to human society. The World Health Organization (WHO) reports that the number of individuals affected by depression worldwide has surged to 280 million in 2022. The COVID-19 pandemic has further exacerbated this crisis, leading to over 53 million new cases and more than 700,000 related suicide incidents annually [1]

[2]. Despite the widespread recognition of its dangers, real-world diagnosis and treatment continue to face systemic challenges. These challenges include patient stigma, uneven distribution of medical resources, and a shortage of professionals [3]. In this context, the emergence of AI-based automated depression detection technology has become a crucial research direction to overcome the barriers in diagnosis and treatment.

In recent years, text analysis of social media has provided new insights for predicting depression. Multiple studies have identified high-risk populations by mining user-generated content (e.g., Reddit posts), achieving model accuracies comparable to clinical screening levels [4]-[6]. The CLEF eRisk competition has even attempted to automate the scoring of the Beck Depression Inventory (BDI) directly from social media [7]. However, excessive unrelated content in social media data limits the clinical applicability of model prediction accuracy [7]-[10]. Meanwhile, most research has focused on feature engineering and algorithm optimization of predictive models, lacking in-depth exploration of model decision interpretability [11]. To address this research gap, our study poses two important questions: 1) How can we leverage data from real medical scenarios to enhance the accuracy of depression severity predictions? 2) How can we elucidate the decision-making rationale of deep learning models to enhance their clinical credibility?

To address these questions, this study focuses on three key areas. First, we constructed a labeled dataset for depression severity from Chinese online consultations, featuring text directly reflecting real-world medical scenarios, thereby providing high signal-to-noise inputs for model training. Methodologically, we developed a dual-channel architecture for "prediction-semantics-level explanation" by integrating pre-trained language models (BERT [12], RoBERTa [13]) with the SHAP (SHapley Additive exPlanations), an interpretability framework rooted in game theory designed to elucidate the prediction outcomes of machine learning models [14], enabling simultaneous depression severity classification and key feature extraction. Finally, we systematically decoded the pathogenesis of depression in Chinese medical texts, establishing semantic features mappable to clinical diagnostic standards, aiming to provide interpretable decision support for intelligent assisted diagnosis.

The core value of this research lies in theoretically validating the pathological semantic encoding capability of deep learning models for Chinese medical texts, offering novel insights for computational linguistic analysis of mental disorders. Crucially, the "algorithmic decision-clinical experience" corroboration mechanism enabled by the interpretability framework provides a technical prototype to overcome the clinical application barriers of "black box models." Amid the rapidly growing demand for mental health services, this approach—combining predictive accuracy with interpretability—has the potential to redefine the collaboration between artificial intelligence and clinical medicine, paving the way for more effective and efficient mental health care.

2. Literature Review

2.1. Advances in Structured Data-Based Depression Prediction

Machine learning technologies have demonstrated significant potential in mental disorder prediction, with core breakthroughs manifested in two aspects: predictive performance enhancement and model interpretability improvement.

Early studies primarily relied on structured clinical data or questionnaire information to construct predictive models. Hueniken *et al.* [15] achieved dual breakthroughs in anxiety and depression identification (accuracy 85% - 88%) by integrating demographic features and mental health scale data using random forest and gradient boosting algorithms. Similarly, Nemesure *et al.* [16] developed a multimodal prediction framework based on college students' mental health records, achieving a 17% performance improvement over traditional statistical methods. Significant progress has also been made in special population studies: Nguyen *et al.* [17] developed a deep neural network model for elderly populations that achieved 89.9% accuracy in depression risk prediction, while Nordin's team [18] designed an ensemble learning model that successfully identified 86% of suicide-prone individuals.

Interpretability integration represents another key research trend. Amit *et al.* [19] identified gestational complications and sleep disorders as core predictors of postpartum depression through SHAP value analysis. Similarly, Nguyen *et al.* [17] demonstrated the critical impact of social isolation on geriatric depression using the LIME framework. These findings not only validate the clinical applicability of machine learning but also uncover potential disease mechanisms. However, limitations remain: 1) Heavy reliance on structured data (e.g., EHRs [19], questionnaires [16]) limits dynamic associations with patients' emotional states and life contexts; 2) Feature engineering depends extensively on prior medical knowledge, constraining models' ability to autonomously discover latent risk factors.

2.2. Deep Learning Exploration with Textual Data

As social media emerges as a critical platform for mental health information, researchers have increasingly explored depression detection through user-generated content. Current research can be categorized into two main approaches:

- Binary Risk Prediction

Eichstaedt *et al.* [4] pioneered the use of Facebook posts for depression prediction, achieving an AUC of 0.69, which highlights the predictive potential of linguistic features. Subsequent studies have focused on improving performance through algorithmic advancements. For example, Yates *et al.* [20] and Shrestha *et al.* [5] leveraged advanced deep learning frameworks, such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), on platforms like Reddit and various online forums to predict users' depressive tendencies. Shrestha *et al.* [5] enhanced the F1-score of RNN to 0.64 by integrating linguistic features with social network metrics. Additionally, Tadesse *et al.* [6] utilized Reddit user content for suicide risk prediction, further demonstrating the effectiveness of deep

learning models in such tasks. These findings underscore the strengths of deep learning in text representation. However, the clinical application of these models remains limited by inherent data challenges, including the presence of significant disease-irrelevant noise in social media texts and the compromised authenticity of data due to user anonymity.

- Multi-class Severity Assessment

The CLEF eRisk competition [7] first attempted BDI-based four-class prediction using Reddit posts, yet the best model achieved only 45% average accuracy. Even with BERT, Bucurer *et al.* [21] reported suboptimal classification performance. In Chinese contexts, Yang *et al.* [8] developed a Weibo-based severe depression identification model with a 62.2% F1-score, though misclassification rates for mild/moderate cases exceeded 35%. This performance gap likely arises from weak correlations between social media content and disease markers, limiting models' ability to capture progressive severity features.

2.3. Critical Analysis

Current depression prediction research exhibits a dichotomy: models based on structured clinical data benefit from interpretability but are constrained by limited information dimensions [15]-[19], while social media-based deep learning models expand data sources but grapple with semantic noise and interpretability challenges [4]-[10] [20] [21]. This landscape highlights two critical issues:

- Data Modality Disparity.

Semantic gaps between structured clinical data and patients' authentic language expressions hinder a comprehensive understanding of disease mechanisms. For example, questionnaire-based "insomnia frequency" metrics cannot fully capture patients' descriptions of "sleepless nights with palpitations" [16].

- Interpretability Limitations.

Existing methods (e.g., SHAP) perform effectively with structured data [19] but often produce clinically incoherent explanations for unstructured text [11]. Additionally, the alignment between pre-trained language models' hierarchical attention mechanisms and medical knowledge systems remains unclear [5] [6] [8] [20] [21]. This study aims to address these limitations by: 1) Leveraging authentic patient narratives as data sources, achieving symptom-semantic alignment through professional annotation; and 2) Developing an interpretability framework tailored for Chinese medical texts to uncover deep learning models' encoding patterns of depressive pathology. By implementing a tripartite "data-algorithm-interpretation" research paradigm, we aim to provide novel theoretical frameworks and practical tools for intelligent mental health assessment."

3. Methodology

This study develops a depression pre-diagnosis model based on real-world online consultation records. The technical framework comprises two core modules: multi-stage data modeling and explainability analysis. To effectively extract deep seman-

tic features from consultation texts, two pre-trained language models, BERT [12] and RoBERTa (Robustly Optimized BERT approach) [13], were utilized to generate embedding vectors. Furthermore, the explainable machine learning method SHAP [14] was integrated to analyze the relationship between patient conditions and consultation texts, aiming to identify key features indicative of disease severity. The overall research process is illustrated in **Figure 1**.

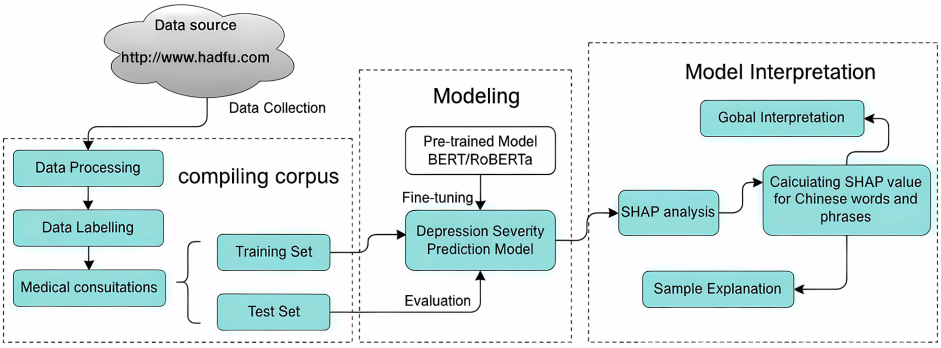


Figure 1. Research flow diagram.

Using this annotated corpus, we developed a depression pre-diagnosis model through fine-tuning strategies applied to the pre-trained models BERT and RoBERTa. The performance of the model was evaluated on a test set to identify the optimal parameter configuration. Additionally, SHAP analysis was employed to interpret the best-performing model, allowing us to explore common symptoms across different severity levels and gain insights into the model's decision-making process from both global and individual perspectives.

3.1. Corpus Construction

A corpus was constructed through a three-stage process: data collection, preprocessing, and annotation. Samples were extracted from detailed consultation records, focusing on the "patient narrative" section. The corpus was annotated by integrating automated recognition techniques with manual review, assigning each sample a class label corresponding to condition severity.

3.1.1. Data Source and Data

The well-known Chinese health and medical platform *Hao-Da-Fu Online* (https://www.haodf.com/) was utilized as the primary data source in the study. We initiated the data collection process in January 2023, amassing a dataset comprising approximately 300,000 user consultation records, with the temporal scope covering the period from 2018 to 2022. After data cleaning and deduplication, 71,654 depression-related records were obtained. Each record includes details such as disease name, patient gender, age, and consultation text, which comprises the patient narrative. The patient narrative encompasses symptoms, emotions, behaviors, treatment history, causes, and requests, forming the core corpus of this study. **Figure 2** illustrates the data collection interface.

Figure 2. Interface of a consultation record on *Hao-Da-Fu Online*.

3.1.2. Data Preprocessing and Annotation

For each patient narrative, preprocessing involved Chinese word segmentation using Python's Jieba, with numbers, punctuation, and non-semantic characters removed to reduce noise. Duplicates were filtered using regular expressions, and texts with insufficient or excessive information were excluded. Text length was standardized to 20 - 500 characters to ensure a high-quality input space.

Annotation followed DSM-5 criteria[1] with a multi-stage validation system: 1) Automated Screening: A rule-based engine classified cases using a keyword list of DSM-5 core symptoms (e.g., depressed mood, lack of interest) and additional symptoms (e.g., sleep disturbance, suicidal ideation). For example, reports explicitly stating "diagnosed with severe depression" or containing keywords like "suicidal" were tagged as severe depression. 2) Review: Three research assistants cross-validated the results, ensuring narratives met DSM-5 criteria, including at least two core symptoms, five total symptoms, and a minimum two-week duration. 3) Quality Control: Unclear cases were excluded, leveraging the ample volume of online consultation samples.

This process yielded 8,391 annotated cases: 3090 severe, 3,016 moderate, and 2285 mild depression. To complete the dataset, 2797 random non-mental health consultations were added as negative controls. The final dataset (11,188 records) includes two fields: patient narrative and depression severity, the latter classified as [0, 1, 2, 3] for non-depressed, mild, moderate, and severe depression, respectively.

3.2. Modeling

This study employs pre-trained language models, BERT and RoBERTa, to extract semantic features from patient narratives and fine-tunes these models to develop

[1]DSM-5 Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Chinese version https://www.mhealthu.com/index.php/list_liangbiao/120/267?page=3

a depression pre-diagnosis model. Additionally, SHAP is utilized to identify key narrative elements indicative of depression severity.

3.2.1. BERT and RoBERTa Models

BERT (Bidirectional Encoder Representations from Transformers) is a groundbreaking model in the field of natural language processing. It utilizes a bidirectional Transformer architecture, which captures contextual information from both the left and right sides of the text. This dual-direction understanding enhances BERT's ability to grasp word meanings and syntactic relationships, thereby improving its semantic comprehension accuracy. The BERT framework consists of two essential stages: Pre-Training and Fine-Tuning. In the pre-training phase, BERT acquires general language representations from extensive datasets. These pre-trained representations can then be fine-tuned with smaller, domain-specific corpora tailored for particular tasks. Typically, only the parameters of the final few layers require adjustment, resulting in a significant reduction in the computational cost associated with developing deep learning models.

This study utilizes BERT to generate semantic vectors for "patient narrative" texts. The semantic vectors derived from BERT more effectively represent content compared to text models based on word frequency. Specifically, we employ the Chinese BERT pre-trained model developed by Google and fine-tune it using the corpus created for this research [22]. In text classification tasks, BERT adds a special token, [CLS], at the beginning of the input text. Through iterative computations, the output at the [CLS] position, denoted as h, encapsulates a vectorized representation of the semantic content of the input. This semantic vector h is then linked to the *softmax* function to compute the probability $P(c|h)$ of the input text belonging to category c (see Equation (1)). W represents the parameter matrix of the classification model. During the fine-tuning process, the model is trained for downstream tasks by adjusting all parameters of BERT and W to maximize the logarithmic probability of the correct label.

$$P(c|h) = softmax(W^* h) \tag{1}$$

RoBERTa (A Robustly Optimized BERT Pretraining Approach) enhances and refines the training strategies and structural components of BERT. Unlike BERT, RoBERTa is trained in a more diverse linguistic context, which allows it to capture richer language patterns. By increasing the number of training iterations, RoBERTa achieves a more thorough data fitting, leading to improved learning outcomes. One of its most significant features is the implementation of a dynamic masking strategy, where masked positions are randomly chosen during pre-training. This method enables the model to better understand contextual information across various masking scenarios, thereby improving its adaptability to different linguistic contexts. In theory, RoBERTa is better suited for tasks that demand a deeper understanding of textual content.

In light of the characteristics of BERT and RoBERTa, along with the specific features of the corpus and the requirements of the task in this study, both pre-

trained models were chosen to develop the depression pre-diagnosis model. It is anticipated that RoBERTa may exhibit superior performance. The BERT model utilized in this research is "bert-base-chinese," a pre-trained model developed by Google based on the Chinese Wikipedia corpus. For the RoBERTa component, the "RoBERTa-base" pre-trained model from the Hugging Face Transformers library was employed.

3.2.2. Interpretable Model SHAP

The SHAP (SHapley Additive exPlanations) framework, rooted in cooperative game theory [23], leverages the mathematically rigorous Shapley value to interpret machine learning predictions. In coalition game theory, participants collaborate to achieve collective goals, and the subsequent reward distribution ("payout") is determined by their individual contributions to the outcome. Analogously, when interpreting predictive models through SHAP, features are conceptualized as cooperative players, with their collective influence generating the final prediction for a target instance $\boldsymbol{s}$. Formally, the Shapley value quantifies each feature's additive contribution by evaluating its average marginal effect across all possible feature combinations (coalitions). For a given instance $\boldsymbol{s}$, this involves calculating the deviation between the model's prediction for $\boldsymbol{s}$ and the baseline prediction (typically the dataset average). This deviation is then apportioned to individual features proportionally to their contribution across all possible feature subsets.

Shapley demonstrated that the Shapley value is the only attribution method that satisfies the properties of efficiency, symmetry, dummy feature compliance, and additivity [23]. The Shapley value of a feature is defined as its average marginal contribution across all possible coalitions (feature subsets). In practical applications, the Shapley value is typically computed using approximation methods, as exact calculations are computationally intensive. SHAP is an optimized algorithm designed to estimate Shapley values efficiently. Let ψ_j denote the Shapley value of feature j. Its computation is detailed in Equation (2) [23] [24]. Here, $\{x_1, \ldots, x_p\}$ represents the set of all input features, p is the number of input features, and $\{x_1, \ldots, x_p\}/\{x_j\}$ is the set of all possible input features excluding $\{x_j\}$. $fx(S)$ denotes the prediction value associated with the feature subset S.

$$\psi_j = \sum_{S \subseteq \{x_1, \cdots, x_p / x_j\}} \frac{|S|!\left(p - |S| - 1\right)!}{p!} \left(f_x\left(S \cup \{x_j\}\right) - f_x\left(S\right)\right) \tag{2}$$

SHAP inherits three foundational properties from Shapley attribution theory: local accuracy, missingness, and consistency. Local accuracy ensures that SHAP can precisely capture the discrepancy between the expected output and the actual prediction for a given instance, showcasing the effectiveness of Shapley values in attribution. Missingness guarantees that features not present in an instance are assigned a Shapley value of zero, explicitly indicating their lack of contribution to the prediction. Consistency ensures that if a feature's marginal contribution increases or remains stable due to changes in the model, its Shapley value does not decrease, reflecting the additive nature of Shapley values.

The primary advantage of SHAP lies in its model-agnostic nature [24] [25], enabling it to seamlessly integrate with various machine learning methods, including deep learning architectures. This adaptability makes SHAP a powerful tool for interpreting complex models. In this study, SHAP analysis is employed to interpret a deep learning-based depression pre-diagnosis model. The model takes patient narratives as input and predicts the severity of depression as output. By applying SHAP, we quantify the Shapley values of text-level features (characters, words, and sentences) derived from patient narratives. This analysis reveals critical textual indicators of depression severity, providing insights into the diagnostic reasoning of the pre-diagnosis model and enhancing its interpretability for clinical applications.

3.3. SHAP Value Calculation for Chinese Words and Sentences

Within the SHAP interpretability framework, feature influence is quantified by SHAP values. The Chinese pre-trained model BERT encodes individual characters into vector representations, and the SHAP model outputs SHAP values for each character. However, individual Chinese characters often lack semantic interpretability. To achieve linguistically coherent word- and sentence-level interpretations, we leverage the additive property of SHAP values, aggregating character-level SHAP values into meaningful word and sentence units. Specifically, the contribution of a word or sentence is measured as the sum of its constituent characters' SHAP values. For a given statement s in the patient narrative, its SHAP value is calculated as shown in Equation (3), where $token_i$ represents a single character.

$$\text{SHAP\_value}(s) = \sum_{i=1}^{n} \text{SHAP\_value}(token_i) \tag{3}$$

After obtaining word- and sentence-level SHAP values, we analyze the overall feature importance of the prediction model, enabling a comparative study of patient conditions across depression severity levels. Finally, we select representative cases to explore the reasoning behind the deep learning model's "black-box" disease interpretation, providing insights into its decision-making process.

4. Experiments and Results

4.1. Experimental Design

Following the research workflow (see **Figure 1**), we conducted two main experiments. Experiment 1 focused on developing a depression pre-diagnosis model to evaluate its effectiveness based on patient narratives. In this phase, patient narratives were transformed into semantic vectors using the pre-trained BERT and RoBERTa models, which were then fine-tuned to create a multi-classification model for predicting depression severity. Experiment 2 employed SHAP to analyze the optimal prediction model, uncovering key narrative elements that reflect patient conditions by investigating the model's decision-making process. Implementation was carried out using Python 3.6, with the deep learning model built on the Scikit-learn and PyTorch frameworks. The Deep SHAP algorithm (integrated within the

SHAP package) was used to calculate SHAP values.

4.2. Parameter Settings and Evaluation Metrics

The experimental corpus was split into training and testing sets at an 8:2 ratio. The model was fine-tuned on the training set, and its performance was evaluated on the testing set to identify the optimal parameter configuration. Key hyperparameters are detailed in Table 1. Among these, MAX_LEN—the maximum length of patient narratives—is a critical parameter. While BERT's default MAX_LEN is 512, this often demands significant computational resources, resulting in inefficiency. Conversely, setting MAX_LEN too low may degrade the quality of BERT's semantic feature extraction. To address this, MAX_LEN was determined based on the length distribution of patient narratives in the corpus. Four quartiles of the default maximum length (50%, 75%, 80%, and 85%) were tested, corresponding to MAX_LEN values of [128, 200, 256, 325].

Table 1. Key parameters of the model.

Parameters	Value assignment	Interpretation	Function
TRAIN_BATCH_SIZE	4, 8, 16, 32	Batch size during training iterations	Hyperparameter, affecting the training speed of the model, related to hardware resources
VALID_BATCH_SIZE	4, 8, 16, 32	Batch size during validation iterations	Hyperparameter, related to the computational efficiency of evaluation results and validation process
EPOCHS	1, 2, 3, 4, 5	Training epochs	Hyperparameter, appropriate number of EPOCHS can prevent overfitting or underfitting of the model
LEARNING_RATE	1e−5, 2e−5	Learning rate	Hyperparameter, determines the extent of each parameter update, selected through experimentation
MAX_LEN	128, 200, 256, 325	Maximum character length of input text	Model parameters, affecting model performance

For evaluation, the pre-diagnosis model was treated as a multi-classification task. Performance was assessed using metrics such as accuracy, precision, recall, and the macro-average F1 score. A confusion matrix was generated to visualize the predictive performance of the multi-classifier. Additionally, SHAP analysis was applied to interpret the model, with case studies used to illustrate the reasoning behind its predictions.

4.3. Experimental Results and Analysis

4.3.1. Depression Pre-Diagnosis Model Based on Patient Narratives

Guided by the parameter ranges outlined in Table 1, this study explored various parameter combinations to identify the optimal hyperparameter set. Detailed experimental results are presented in Table 2. Overall, the BERT model consistently outperformed the RoBERTa model. The length of input text, controlled by MAX_LEN, significantly impacted model performance. Longer input texts generally yielded better results, with the BERT model achieving its highest performance at MAX_LEN = 325, where the F1 score reached 0.92.

Table 2. Model performances.

Model	EPOCH	MAX_LEN	Accuracy	Precision	Recall	F1-score
BERT	3	128	0.89	0.89	0.90	0.90
	2	256	0.91	0.91	0.91	0.91
	3	325	0.92	0.91	0.92	**0.92**
RoBERT	5	128	0.76	0.76	0.75	0.75
	2	256	0.80	0.81	0.78	0.79
	3	325	0.90	0.89	0.90	0.89

Note: LEARNING_RATE = 1e-05, TRAIN_BATCH_SIZE = 8, VALID_BATCH_SIZE = 4.

The performance of the optimal BERT model across different depression severity categories is depicted in **Figure 3**. Analysis of the confusion matrix yielded the following key observations: 1) Strong Overall Predictive Performance: The model achieved an accuracy exceeding 0.85 for each category, highlighting its robustness. 2) High Accuracy for Non-Depression Cases (label = none): The prediction accuracy for non-depression cases reached 0.98, suggesting that the content of narrative from non-depressed patients markedly differs from that of depressed patients. 3) Reliable Diagnosis of Severe Depression (label = severe): The accuracy for severe depression cases was 0.93, with only 3% misclassified as moderate or mild. 4) Moderate Depression Diagnosis (label = moderate): The accuracy for moderate depression cases was 0.86, with 7% misclassified as mild and 6% as severe. 5) Mild Depression Diagnosis (label = mild): The accuracy for mild depression cases was 0.90, with 5% misclassified as moderate.

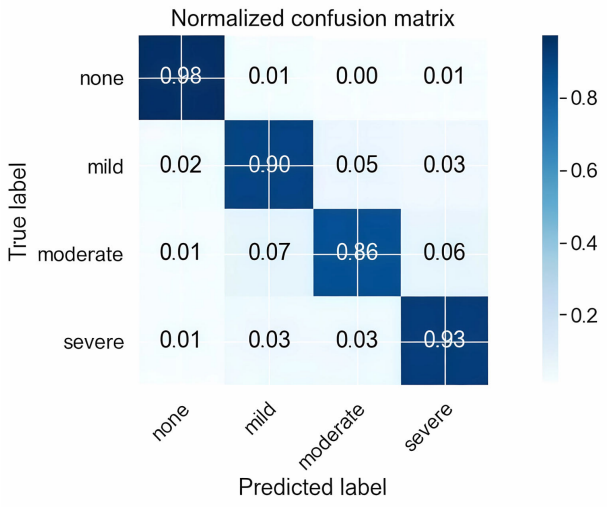


Figure 3. Confusion matrix of the depression severity prediction model (Running Screenshot).

In summary, leveraging patient narratives, the BERT model achieved exceptional accuracy, exceeding 90%, in predicting depression risk and severity. While minor discrepancies between mild and moderate cases persist, this reflects the in-

herent ambiguity in clinical diagnosis, where intermediate conditions often exhibit overlapping symptoms.

4.3.2. Feature Importance Analysis Based on SHAP Values

To gain a global perspective on feature importance, we applied SHAP analysis to the optimal model, extracting keywords or phrases from patient narratives that significantly influence prediction outcomes (*i.e.*, depression severity). This analysis aimed to identify typical symptoms associated with varying levels of depression.

We began by identifying words with SHAP values above zero from the reports of patients across all four depression categories, sorted by descending frequency (see Figure 4). Positive SHAP values indicate that these words support the prediction of severe depression, and distinct patterns emerged when comparing the narratives of depressed and non-depressed patients. High-frequency words in non-depressed patients' narratives lacked depressive markers, highlighting a linguistic divergence from depression cases. In contrast, depression cases were characterized by narratives consistently featuring elements directly associated with depressive disorders, e.g., "抑郁" (depression), "抑郁症" (depressive disorder), "自己" (self), "焦虑" (anxiety), and "情绪" (mood), reflecting common psychological and emotional states of depression.

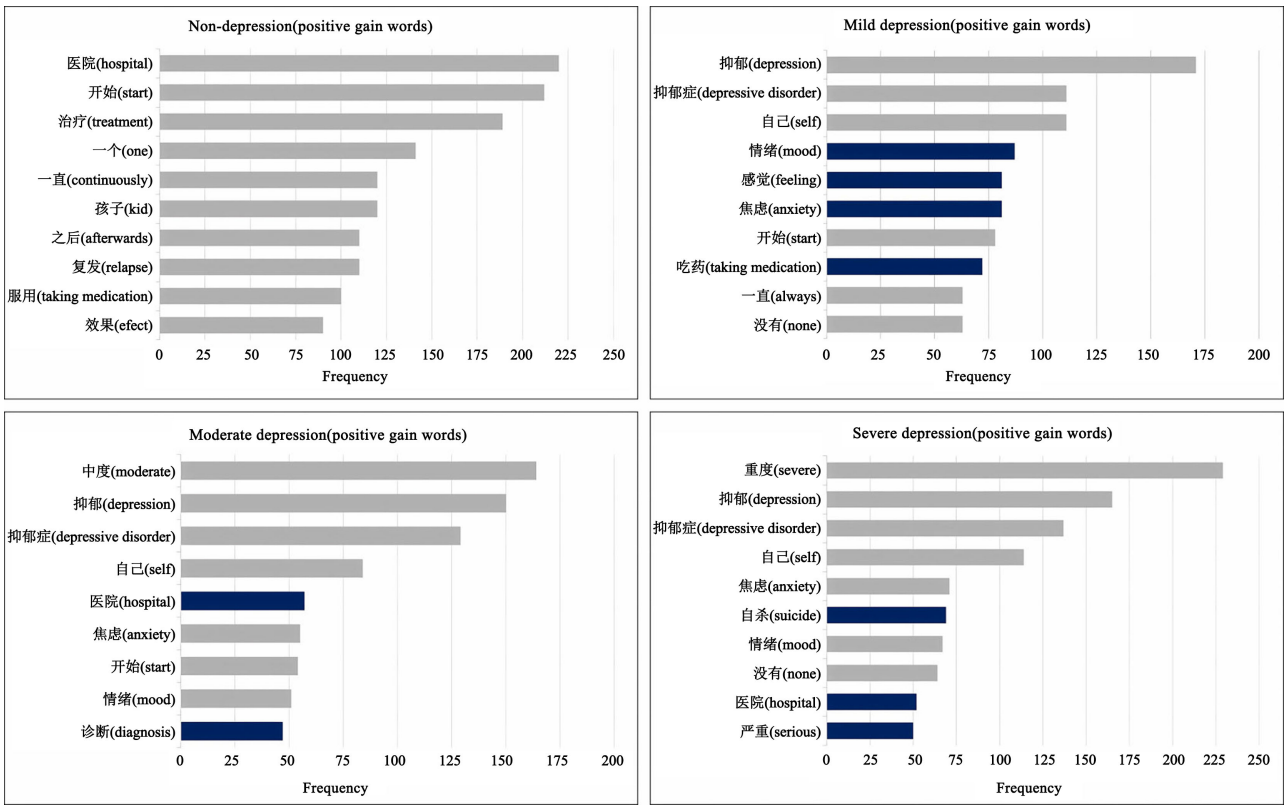


Figure 4. Words Positively Correlated with Depression Severity (Based on Frequency Ranking, Top 10).

Further analysis revealed nuanced variations in language use across different

severity levels. Patients with mild depression frequently employed words such as "情绪" (mood), "感觉" (feeling), "焦虑" (anxiety), and "吃药" (taking medication), reflecting a focus on subjective experiences and self-directed actions. In moderate to severe depression, terms like "诊断" (diagnosis) and "医院" (hospital) were prominent, indicating active medical engagement and heightened levels of distress. For severe depression, the recurrent use of words such as "自杀" (suicide) and "严重" (serious) highlighted the critical, often life-threatening nature of their condition. These findings illustrate systematic language variations across depression severity levels, offering valuable insights into patients' psychological states and help-seeking behaviors.

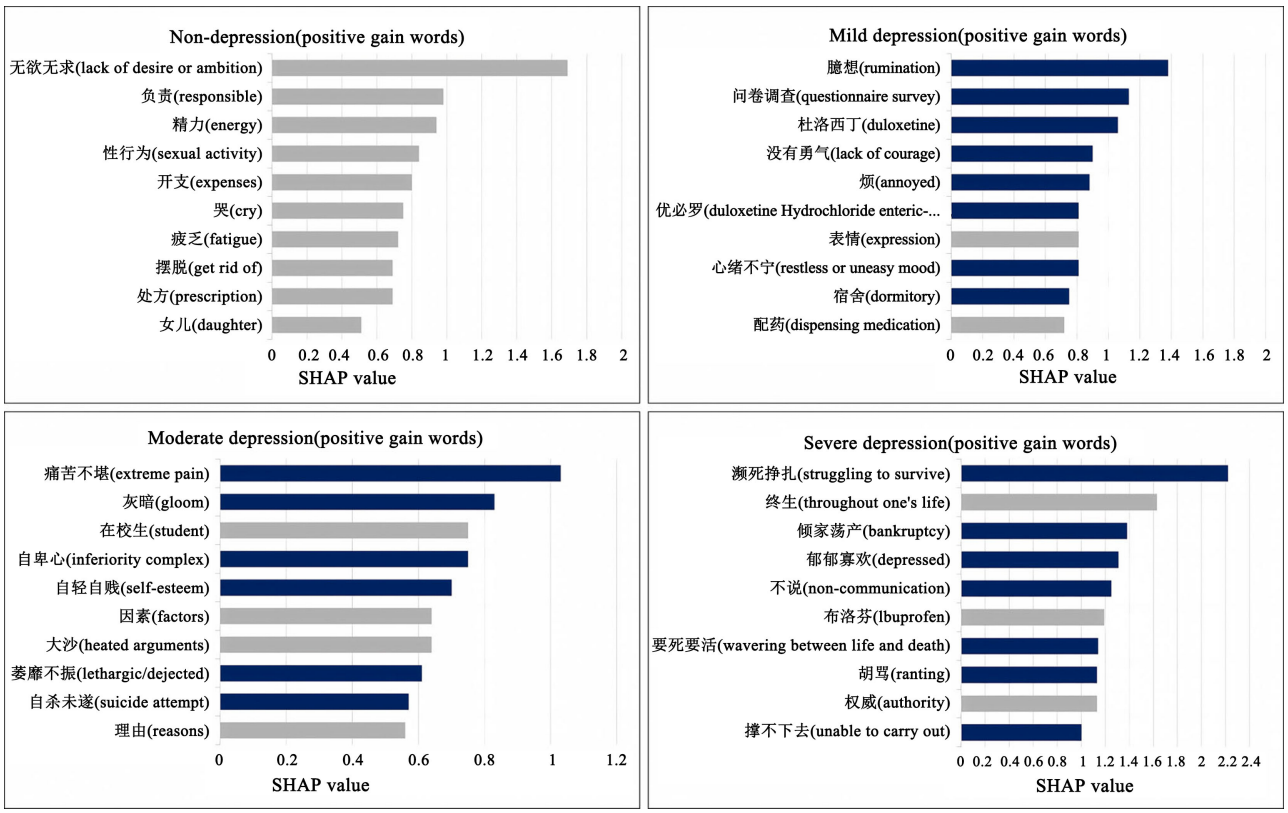


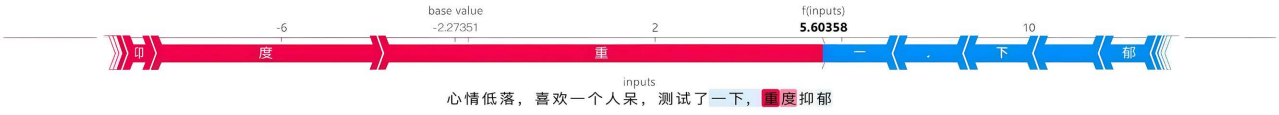
Figure 5. Words positively correlated with depression severity (Based on SHAP value Ranking, Top 10).

Furthermore, we ranked words/phrases by their SHAP values to identify features driving the model's diagnostic decisions (**Figure 5**). Unlike frequency-based rankings, SHAP-selected terms better captured clinically meaningful patterns. For mild depression, the model focused on personal emotions (e.g., 臆想 rumination, 没有勇气 lack of courage), self-care actions (e.g., 问卷调查自测 self-assessment, 杜洛西汀/优必罗 duloxetine/vortioxetine use), and identity cues (e.g., 宿舍 dormitory). Moderate depression narratives emphasized intensified negativity (e.g., 痛苦不堪 extreme pain, 昏暗 gloom, 自卑心 inferiority complex, 自轻自贱 self-esteem) and extreme acts (e.g., 自杀未遂 suicide attempt). Severe depression accounts featured terminal despair (e.g., 郁郁寡欢 depressed,

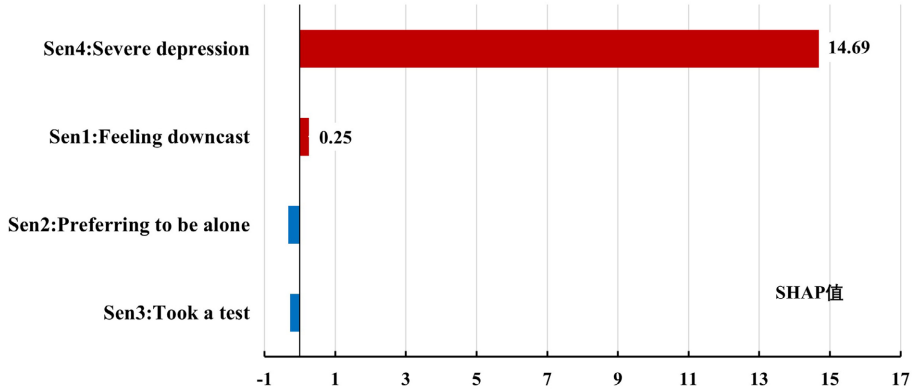
濒死挣扎 struggling to survive), behavioral dysregulation (e.g., 不说 non-communication, 胡骂 ranting), and potential triggers (e.g., 倾家荡产 bankruptcy). Non-depressed individuals' language showed minimal overlap with depressed groups. Many of these SHAP-interpreted latent features, extracted via deep learning, align with DSM-5 core symptoms.

4.3.3. Case-Specific Interpretation

This study applied SHAP interpretability modeling to measure how different text elements (characters, words, sentences) in patient narratives influence predictions. The visualization system (**Figure 6(a)**, **Figure 6(b)**) clarifies the model's reasoning. In **Figure 6(a)**'s force diagram, each Chinese character is shown as an arrow whose length reflects its impact strength (absolute SHAP value), and red/blue arrows denote positive/negative contributions (SHAP > 0/<0), with color intensity matching effect magnitude. For Case 1, the model output $f(x)$ = 5.603 (99.6% severe depression probability) highlights the characters "重" and "度" as key drivers, while others show weak impacts (faded colors).



(a)



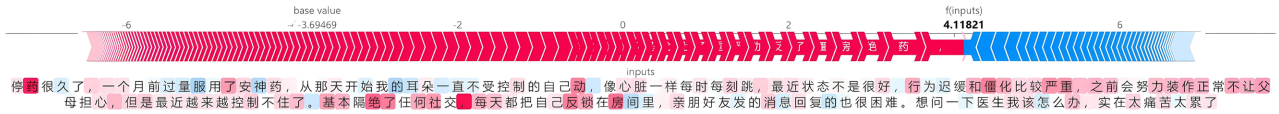
Chinese//English Sentences
Sen1:心情低落//Feeling downcast
Sen2:喜欢一个人呆//Preferring to be alone
Sen3:测试了一下//Took a test
Sen4:重度抑郁//Severe depression

(b)

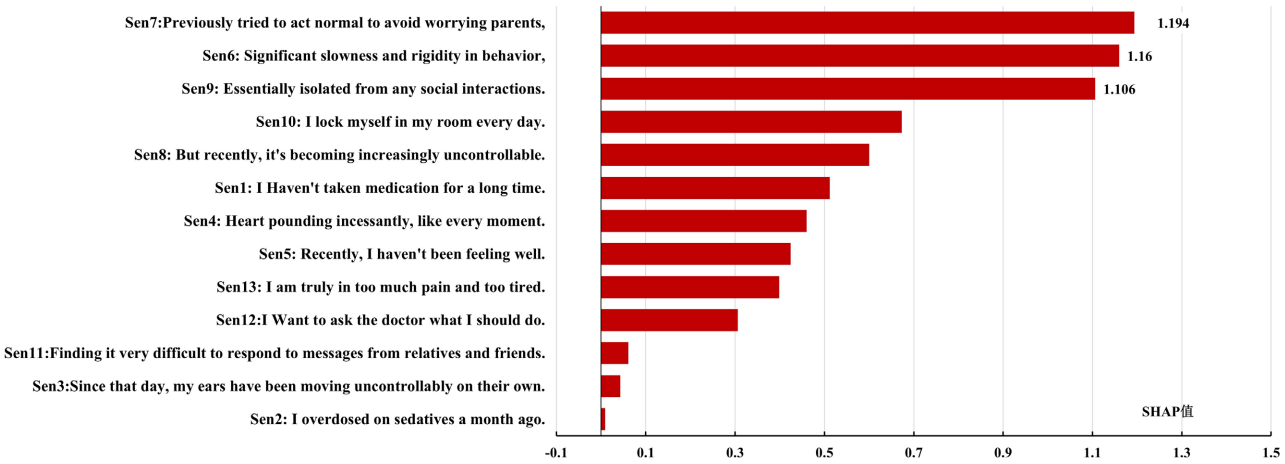
Figure 6. (a) Force Plot for Case 1 Derived from SHAP Text Analysis (Predicted Severe Depression Risk: 99.6%); (b) Sentence Importance Analysis Based on SHAP Values (Case 1, Severe Depression Risk Value 99.6%).

Whereas, BERT's character-level tokenization of Chinese text introduces semantic fragmentation in SHAP analyses, compromising attribution clarity. To address this, we employ the aggregation algorithm (Section 3.3.3) to quantify

and visualize SHAP values at the sentence level. As demonstrated in Figure 6(b), the diagnostic phrase "Sen4:重度抑郁//Severe depression" exerts the strongest predictive influence (SHAP 14.69), while the symptom descriptor "Sen1: 心情低落//Feeling downcast" contributes modestly yet remains clinically coherent (SHAP 0.25). Otherwise, the phrase "Sen3: 测试一下//Took a test" exhibits counter_predictive effects, contrasting sharply with "重度抑郁" (ΔSHAP 17.6). Clearly, the model prioritizes explicit pathological markers like "重度抑郁//Severe depression" while contextualizing auxiliary symptoms, achieving diagnostic conclusions that align precisely with clinical terminology in patient narratives.



(a)



Chinese//English Sentences
Sen1:停药很久了//I Haven't taken medication for a long time.
Sen2:一个月前过量服用了安神药//I overdosed on sedatives a month ago.
Sen3:从那天开始我的耳朵一直不受控制的自己动//Since that day, my ears have been moving uncontrollably on their own.
Sen4:像心脏一样每时每刻跳//Heart pounding incessantly, like every moment.
Sen5:最近状态不是很好//Recently, I haven't been feeling well.
Sen6:行为迟缓和僵化比较严重//Significant slowness and rigidity in behavior.
Sen7:之前会努力装作正常不让父母担心//Previously tried to act normal to avoid worrying parents.
Sen8:但是最近越来越控制不住了//But recently, it's becoming increasingly uncontrollable.
Sen9:基本隔绝了任何社交//Essentially isolated from any social interaction.
Sen10:每天都把自己反锁在房间力//I lock myself in my room every day.
Sen11:亲朋好友发的消息回复的也很困难//Finding it very difficult to respond to messages from relatives and friends
Sen12:想问一下医生我该怎么办//I want to ask the doctor what I should do.
Sen13:实在太痛苦太累了//I am truly in too much pain and too tired.

(b)

Figure 7. (a) Force Plot for Case 421 Derived from SHAP Text Analysis (Predicted Severe Depression Risk: 98.3%); (b) Sentence Importance Analysis Based on SHAP Values (Case 421, Severe Depression Risk: 98.3%).

Figure 7(a) and Figure 7(b) present BERT's interpretability analysis for Sample 421, where the model predicts severe depression with 98.3% confidence. In the SHAP force diagram (Figure 7(a)), positive predictors (red) significantly outnumber negative ones (blue), showing that clinically meaningful symptoms drive the model's decisions. The sentence-level analysis (Figure 7(b)) confirms this pattern—all text segments contribute positively, aligning strongly with diagnostic criteria for depression. Clinically, many of these text segments mirror DSM-5 standards directly, for instance.

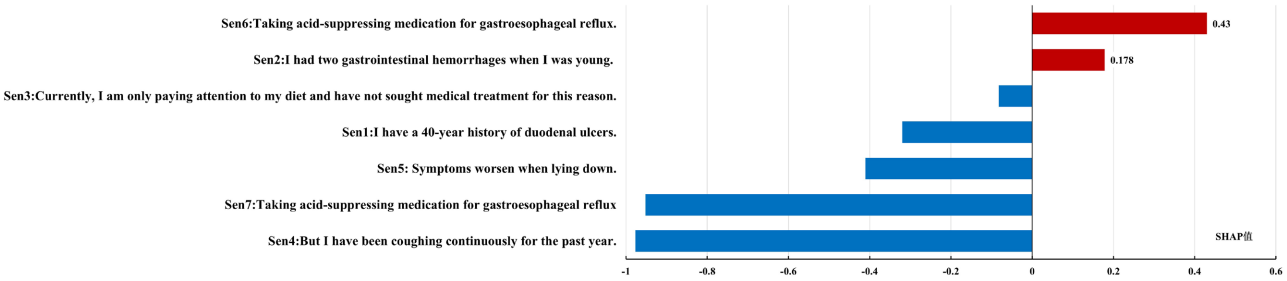
"Sen7: …tried to act normal." → Criterion B (social impairment)

"Sen6: …slowness and rigidity in behavior." → Criterion A5 (psychomotor retardation)

"Sen13: …too much pain and too tired." → Criterion A1 (persistent low mood)

Clearly, for this case, BERT detects depression through symptom semantics rather than explicit mentions. By analyzing implicit patterns (e.g., social withdrawal cues in "Sen9: 基本隔绝了任何社交//Essentially isolated from any social interaction ", the model establishes diagnostic alignment with DSM-5 without direct disease references. This symptom-driven reasoning—employing latent clinical markers over overt declarations only—demonstrates AI's capacity to augment mental health evaluations, particularly for cases with subtle or context-dependent symptom presentations.

(a)



Chinese//English Sentence
Sen1: 既往有 40 年十二指肠溃疡史//I have a 40-year history of duodenal ulcers.
Sen2: 年轻时先后消化道大出血两次//I had two gastrointestinal hemorrhages when I was young.
Sen3: 目前只是自己注意饮食，未因此做医疗//Currently, I am only paying attention to my diet and have not sought medical treatment for this reason.
Sen4: 但近一年来咳嗽不断//But I have been coughing continuously for the past year.
Sen5: 卧位加重//Symptoms worsen when lying down
Sen6: 咽喉痛//Sore throat
Sen7: 按食管反流服用抑酸药//Taking acid-suppressing medication for gastroesophageal reflux

(b)

Figure 8. (a) Force Plot for Case 180 Derived from SHAP Text Analysis (Predicted Severe Depression Risk: 0.27%); (b) Sentence Importance Analysis Based on SHAP value (Case 180, Severe Depression Risk 0.27%).

The model's interpretability analysis for Sample 180 (**Figure 8(a)** and **Figure 8(b)**) clarifies its reasoning for ruling out depression. With an extraordinarily low predicted risk (0.27%, 99.73% confidence), the case demonstrates how textual semantics drive non-depression classifications. Across both character and sentence levels (**Figure 8(a)** and **Figure 8(b)**), dominant blue regions indicate widespread negative feature impacts, for instance, descriptions of non-psychiatric conditions like "Sen4: 但近一年来咳嗽不断//But I have been coughing continuously for the past year." and "Sen7: 按食管反流服用抑酸药//Taking acid-suppressing medication for gastroesophageal reflux" dilute depression likelihood. Limited positive contributions (red features) include "Sen6: 咽喉痛//Sore throat" (SHAP = 0.43) and "Sen2: 年轻时先后消化道大出血两次//I had two gastrointestinal hemorrhages when I was young." (SHAP = 0.178), weakly reflecting potential somatic depression correlates or stress-related physiological responses. Crucially, the complete absence of core diagnostic markers (depression, low mood, anxiety; frequency = 0) and the overall semantic focus on non-affective health issues form the decisive framework for the model's negative prediction. This dual analytical approach—sensitivity to subclinical indicators alongside strict semantic alignment with diagnostic standards—validates the model's precision in distinguishing non-depressive cases through integrated symptom prioritization and terminology-based exclusion criteria.

5. Discussion

5.1. Pre-Trained Language Models for Automatic Depression Diagnosis

This study demonstrates the viability of BERT-based deep learning for automated depression diagnosis using clinical narratives from online healthcare platforms. Our fine-tuned multi-class model achieved 92% overall accuracy in severity assessment, with class-specific performance reaching 98% (non-depression), 90% (mild), 86% (moderate), and 93% (severe). These results significantly outperform conventional ML approaches and establish the clinical validity of pre-trained language models in psychiatric evaluation.

A key innovation of this study is the use of medical-contextualized data—patient narratives—instead of social media content for analysis. While BERT achieved only a 62.2% F1 score on non-clinical texts (e.g., Yang *et al.*'s social media analysis [7]), our medical-context model reached a 92% F1 score, highlighting the diagnostic superiority of patient narratives in medical settings. Although data limitations prevented direct access to patients' demographic and socioeconomic characteristics—critical for depression diagnosis—our model's 92% accuracy suggests it captures these features implicitly through contextual analysis, encoded as latent semantic vectors. In our analysis, terms like "student," "dormitory," and "bankruptcy" carried significant SHAP values. Thus, even without explicitly incorporating demographic or socioeconomic data, the BERT model can extract and leverage such information from patient narratives through contextual analysis,

enabling high predictive precision.

Architectural comparisons revealed unexpected performance about models: while RoBERTa theoretically outperforms BERT in general contexts, the Chinese BERT variant demonstrated superior effectiveness compared to the non-specialized RoBERTa implementation. Achieving optimal model performance necessitated balancing computational constraints with clinical requirements—setting the input length threshold at 325 characters (85th percentile of text length distribution) maintained 92% diagnostic accuracy while optimizing processing efficiency. However, this configuration inherently truncates 15% of patient narratives exceeding this limit, potentially omitting clinically significant information contained in extended descriptions. This fundamental conflict between algorithmic efficiency and comprehensive clinical data integration represents a critical implementation barrier for AI-driven mental health assessment systems in practical healthcare settings.

5.2. Insights from SHAP Interpretable Machine Learning Models

This study advances the interpretability of deep learning models for Chinese text analysis through innovative SHAP-based methods. While SHAP has become a standard tool for model interpretation, its application in understanding text-prediction relationships remains limited. We developed a specialized SHAP framework tailored for Chinese corpora, enabling word- and sentence-level analysis of feature contributions. When applied to depression diagnosis, this approach reveals BERT's sensitivity to key linguistic patterns: descriptions of mood states, behavioral changes, and treatment experiences consistently exhibit high SHAP values, strongly correlating with depression risk.

Moreover, the analysis uncovers clinically meaningful patterns across depression severity levels. As severity increases, expressions of negative affect and extreme behaviors gain prominence in SHAP rankings. The model also identifies demographic markers (e.g., "宿舍//dormitory," "学生//students") and socioeconomic factors (e.g., "倾家荡产//bankruptcy") as predictive features, suggesting underlying associations between social context and mental health. Notably, the SHAP-derived features align closely with DSM-5 diagnostic criteria, with phrases such as "濒死挣扎//struggling to survive" and "自杀未遂//suicide attempts" directly mapping to core depression symptoms.

These findings demonstrate BERT's capability to capture clinically relevant semantic patterns while validating SHAP's utility for explaining model decisions in mental health applications. By bridging the gap between machine learning outputs and clinical diagnostic standards, this work provides robust evidence supporting the reliability and interpretability of AI-driven diagnostic tools.

5.3. Limitations and Future Directions

This study advances the interpretability of text classification models using SHAP methods, yet several challenges warrant further exploration. A key limitation

arises from BERT's constrained text length capacity, set at 325 characters due to computational resource constraints. Given that depression-related medical records often exceed this threshold, approximately 15% of the text was truncated, potentially compromising prediction accuracy. This highlights the need for more effective semantic modeling of long texts in medical applications.

Furthermore, the proposed Chinese word- and sentence-level SHAP value computation method, although experimentally validated, is constrained by basic tokenization techniques that may fail to capture domain-specific medical terminology. This limitation can significantly affect SHAP analysis and overall model interpretation, as missing key medical terms may reduce accuracy and reliability in clinical settings. Exploring alternative tokenization strategies, such as subword tokenization or rule-based methods, could greatly improve the identification of relevant medical terms and represent a future optimization direction for this study.

Additionally, the use of general-purpose models (BERT and RoBERTa) for analyzing online patient-generated content raises questions about the potential advantages of medical domain-specific large language models in processing disease-related texts. These concerns emphasize the importance of tailoring models for specific medical contexts to enhance analysis accuracy.

Overall, these limitations point to critical avenues for future research, including corpus optimization, the development of medical-contextualized models, and enhanced semantic processing capabilities for long and complex texts. Addressing these challenges is essential for improving model applicability and predictive accuracy in clinical settings.

6. Conclusions

This study establishes an integrated "data-prediction-interpretation" framework to advance intelligent depression diagnosis using Chinese medical texts. The key findings are as follows:

- The deep learning model trained on real-world clinical data demonstrated superior predictive performance. Leveraging high-quality patient narratives from online clinical consultation platforms, the BERT model achieved 92% overall accuracy in a four-class classification task, significantly outperforming social media-based approaches. This performance validates the semantic richness and clinical relevance of medical context data. Notably, the model achieved 93% accuracy in identifying severe depression cases, with misclassifications primarily occurring in the mild-to-moderate transition zone, reflecting the inherent complexity of clinical diagnosis.
- The study's interpretability framework successfully decoded the model's pathological semantic encoding mechanisms. Using an improved SHAP value computation method, it achieved the first quantitative analysis of word- and sentence-level feature contributions in Chinese medical texts. The results revealed a direct mapping between model decisions and DSM-5 diagnostic cri-

teria. For instance, phrase like "濒死挣扎" (struggling to survive) in severe depression cases exhibited significantly higher mean SHAP values (2.22) compared to terms such as "自杀未遂" (attempted suicide) in moderate cases (mean SHAP: 0.57). This alignment between algorithmic decisions and clinical standards addresses the "black box" challenge in mental health applications.

- The study uncovered a novel collaborative paradigm between intelligent diagnostic systems and clinical practice. Beyond explicit diagnostic statements, the model demonstrated the ability to capture implicit symptom features through semantic associations. In cases lacking diagnostic terminology, it is classified based on descriptions like "social isolation" and "psychomotor retardation," showcasing a deep semantic understanding that transcends traditional keyword matching. This capability establishes a novel collaborative paradigm between AI systems and clinical practice, providing clinicians with a multidimensional decision-making framework.

In summary, the proposed interpretable diagnostic framework provides empirical evidence and technical prototypes for building trustworthy intelligent mental health assessment systems. Future research will focus on developing dynamic encoding algorithms for long texts, optimizing feature extraction using medical ontologies, and exploring multimodal data integration pathways.

Acknowledgements

This research was supported by the 2022 Guangzhou (China) Social Science Foundation Project (Grant No. 10000-42220402). We thank Mr. Long Zhaohui for his assistance and all contributors for their support.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Institute of Health Metrics and Evaluation (2023) Global Health Data Exchange (GHDx). https://vizhub.healthdata.org/gbd-results/

[2] COVID-19 Mental Disorders Collaborators (2021) Global Prevalence and Burden of Depressive and Anxiety Disorders in 204 Countries and Territories in 2020 Due to the COVID-19 Pandemic. *The Lancet*, **398**, 1700-1712.

[3] Byeon, H. (2023) Advances in Machine Learning and Explainable Artificial Intelligence for Depression Prediction. *International Journal of Advanced Computer Science and Applications*, **14**, 520-526. https://doi.org/10.14569/ijacsa.2023.0140656

[4] Eichstaedt, J.C., Smith, R.J., Merchant, R.M., Ungar, L.H., Crutchley, P., Preoțiuc-Pietro, D., *et al.* (2018) Facebook Language Predicts Depression in Medical Records. *Proceedings of the National Academy of Sciences*, **115**, 11203-11208. https://doi.org/10.1073/pnas.1802331115

[5] Shrestha, A., Serra, E. and Spezzano, F. (2020) Multi-Modal Social and Psycho-Linguistic Embedding via Recurrent Neural Networks to Identify Depressed Users in Online Forums. *Network Modeling Analysis in Health Informatics and Bioinformatics*, **9**, Article No. 22. https://doi.org/10.1007/s13721-020-0226-0

[6] Tadesse, M.M., Lin, H., Xu, B. and Yang, L. (2019) Detection of Suicide Ideation in Social Media Forums Using Deep Learning. *Algorithms*, **13**, Article 7. https://doi.org/10.3390/a13010007

[7] Parapar, J., Martín-Rodilla, P., Losada, D.E. and Crestani, F. (2021) Overview of eRisk 2021: Early Risk Prediction on the Internet. In: Candan, K.S., Ionescu, B., Goeuriot, L., *et al.*, Eds., *Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF* 2021. *Lecture Notes in Computer Science*, Springer, 324-344. https://doi.org/10.1007/978-3-030-85251-1_22

[8] Yang, T.T., Li, F., Ji, D.H., Liang, X.H., Xie, T., Tian, S.W., *et al.* (2021) Fine-Grained Depression Analysis Based on Chinese Micro-Blog Reviews. *Information Processing & Management*, **58**, Article 102681. https://doi.org/10.1016/j.ipm.2021.102681

[9] Burdisso, S.G., Errecalde, M. and Montes-y-Gómez, M. (2021) Using Text Classification to Estimate the Depression Level of Reddit Users. *Journal of Computer Science and Technology*, **21**, e1. https://doi.org/10.24215/16666038.21.e1

[10] Abed-Esfahani, P., Howard, D., Maslej, M., *et al.* (2019) Transfer Learning for Depression: Early Detection and Severity Prediction from Social Media Postings. CLEF (Working Notes), 1-6. https://ceur-ws.org/Vol-2380/paper_102.pdf

[11] Mi, J., Li, A. and Zhou, L. (2020) Review Study of Interpretation Methods for Future Interpretable Machine Learning. *IEEE Access*, **8**, 191969-191985. https://doi.org/10.1109/access.2020.3032756

[12] Devlin, J., Chang, M.W., Lee, K., *et al.* (2019) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the* 2019 *Conference of the North American Chapter of the Association for Computational Lin-Guistics*: *Human Language Technologies*, Minnesota, 2-7 June 2019, 4171-4186.

[13] Liu, Y., Ott, M., Goyal, N., *et al.* (2019) RoBERTa: A Robustly Optimized BERT Pretraining Approach.

[14] Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., *et al.* (2020) From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature Machine Intelligence*, **2**, 56-67. https://doi.org/10.1038/s42256-019-0138-9

[15] Hueniken, K., Somé, N.H., Abdelhack, M., Taylor, G., Elton Marshall, T., Wickens, C.M., *et al.* (2021) Machine Learning-Based Predictive Modeling of Anxiety and Depressive Symptoms during 8 Months of the COVID-19 Global Pandemic: Repeated Cross-Sectional Survey Study. *JMIR Mental Health*, **8**, e32876. https://doi.org/10.2196/32876

[16] Nemesure, M.D., Heinz, M.V., Huang, R. and Jacobson, N.C. (2021) Predictive Modeling of Depression and Anxiety Using Electronic Health Records and a Novel Machine Learning Approach with Artificial Intelligence. *Scientific Reports*, **11**, Article No. 1980. https://doi.org/10.1038/s41598-021-81368-4

[17] Nguyen, H.V. and Byeon, H. (2022) Explainable Deep-Learning-Based Depression Modeling of Elderly Community after COVID-19 Pandemic. *Mathematics*, **10**, Article 4408. https://doi.org/10.3390/math10234408

[18] Nordin, N., Zainol, Z., Mohd Noor, M.H. and Chan, L.F. (2023) An Explainable Predictive Model for Suicide Attempt Risk Using an Ensemble Learning and Shapley Additive Explanations (SHAP) Approach. *Asian Journal of Psychiatry*, **79**, Article 103316. https://doi.org/10.1016/j.ajp.2022.103316

[19] Do, H.P., Baker, P.R.A., Van Vo, T., Murray, A., Murray, L., Valdebenito, S., *et al.* (2021) Intergenerational Effects of Violence on Women's Perinatal Wellbeing and Infant Health Outcomes: Evidence from a Birth Cohort Study in Central Vietnam. *BMC Pregnancy and Childbirth*, **21**, Article No. 648.

https://doi.org/10.1186/s12884-021-04097-6

[20] Yates, A., Cohan, A. and Goharian, N. (2017) Depression and Self-Harm Risk Assessment in Online Forums. In: Martha, P. and Rebecca, H. and Sebastian, R. Eds., *Proceedings of the* 2017 *Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2968-2978. https://doi.org/10.18653/v1/d17-1322

[21] Bucur, A., Cosma, A. and Dinu, P.L. (2021) Early Risk Detection of Pathological Gambling, Self-Harm and Depression Using BERT. http://dx.doi.org/10.13140/RG.2.2.25060.50567

[22] Sun, C., Qiu, X.P., Xu, Y. and Huang, X.J. (2019) How to Fine-Tune BERT for Text Classification? In: *Lecture Notes in Computer Science*, Springer International Publishing, 194-206. https://doi.org/10.1007/978-3-030-32381-3_16

[23] Shapley, L.S. (1953) 17. A Value for N-Person Games. In: Kuhn, H.W. and Tucker, A.W., Eds., *Contributions to the Theory of Games* (*AM*-28), *Volume II*, Princeton University Press, 307-318. https://doi.org/10.1515/9781400881970-018

[24] Adadi, A. and Berrada, M. (2018) Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, **6**, 52138-52160. https://doi.org/10.1109/access.2018.2870052

[25] Azpiazu, C., Bosch, J., Bortolotti, L., Medrzycki, P., Teper, D., Molowny-Horas, R., *et al.* (2021) Toxicity of the Insecticide Sulfoxaflor Alone and in Combination with the Fungicide Fluxapyroxad in Three Bee Species. *Scientific Reports*, **11**, Article No. 6821. https://doi.org/10.1038/s41598-021-86036-1