



Article

Explainable Text-Based Computational Framework for Psycho-Emotional Risk Classification: Calibration, Interpretability, and Encoder Comparison on Public Datasets

Orazmukhamed Bekmurat ¹, Vassiliy Serbin ¹, Aliya Aizhanova ², Nurdaulet Niyaz ², Darkhan Yerezhep ^{3,*}, Kanibek Sansyzbay ⁴, Ayaulym Oralbekova ^{5,*} and Malika Sagitzhanova ¹

- ¹ Institute of Automation and Information Technologies, Satbayev University, Almaty 050013, Kazakhstan; o.bekmurat@satbayev.university (O.B.); v.serbin@satbayev.university (V.S.); msagitzhanova@mail.ru (M.S.)
- ² Department of Smart Technologies in Engineering, International Engineering and Technological University, Almaty 050060, Kazakhstan; a.elyubaevna@metu.kz (A.A.); nzvnurdaulet@gmail.com (N.N.)
- ³ Technology Commercialization Center, Almaty Management University, Almaty 050060, Kazakhstan
- ⁴ Department of Radio Engineering, Electronics and Telecommunications, International Information Technologies University, Almaty 050060, Kazakhstan; k.sansyzbai@iitu.edu.kz
- ⁵ Department of Automation and Electric Power Engineering in Transport, International University of Transportation and Humanities, Almaty 050063, Kazakhstan
- * Correspondence: darhan_13@physics.kz (D.Y.); ayauka959@gmail.com (A.O.)

Abstract

The early detection of psycho-emotional risks remains challenging despite its growing importance. Most existing models rely on a single data type, mainly questionnaires or text, and operate as “black boxes”, limiting practical use. Psycho-emotional states are multidimensional, reflected in textual, structured, and temporal digital signals. Ignoring this complexity may result in information loss and lower prediction accuracy. This paper presents an explainable text-based computational framework for psycho-emotional risk classification using public datasets. Record-level structured information was used only when it was available within the same original observation and was not created by matching records across datasets. The study did not integrate observations across datasets or perform cross-dataset fusion; these aspects are considered potential directions for future research. Implemented in Python using PyTorch, the model was evaluated on two open-text datasets. The BERT-based configuration achieved accuracy of 0.593 ± 0.009 and an ROC-AUC of 0.860 ± 0.006 , while RoBERTa-base improved the performance to 0.666 ± 0.005 accuracy and a 0.896 ± 0.004 ROC-AUC under five-fold cross-validation. The results demonstrate that classification quality depends on encoder selection, preprocessing, dataset characteristics, duplicate handling, calibration, and the way that available record-level information is represented. Therefore, the findings should be interpreted as record-level computational classification results rather than the clinical validation of psycho-emotional risk assessment.



Academic Editors: Stephan Schlögl and Wei Liu

Received: 27 April 2026

Revised: 16 June 2026

Accepted: 17 July 2026

Published: 21 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: explainable artificial intelligence; psycho-emotional risk prediction; transformer models; text-based classification; probabilistic calibration; interpretable machine learning; performance evaluation

1. Introduction

In recent years, AI methods have been increasingly used in the task of analyzing psycho-emotional states. This is particularly evident in the educational environment, where interest in the early detection of stress conditions and anxiety among teenagers and

students is growing. Machine learning models, trained on psychometric data, show high accuracy in a number of tasks [1,2], and reviews over recent years show rapid growth in such solutions, including ensemble methods and deep neural networks [3,4]. However, a deeper examination reveals that most of these approaches remain rather isolated and they are not always ready for implementation in real information systems.

One of the reasons is the focus on unimodal data. Many of the models still use only questionnaire indicators or standardized scales [5,6]. On the one hand, this makes them convenient and interpretable. On the other hand, such data represent only a subjective state at a given time. Meanwhile, psycho-emotional processes are dynamic and depend on physiological and behavioral factors. As a result, the model reveals only part of the picture, and other aspects “remain behind the scene”.

Physiological states, wearable sensors, and visual cues are relevant for future advancements in the analysis of psycho-emotional risks, as affective states can be reflected not only in language but also in bodily and behavioral reactions. Previous studies of heart rate variability (HRV), electrodermal activity (EDA), surface electromyography (SEMG), facial emotion recognition, and attention-based visual analysis show that such channels can provide additional information in broader affective computing systems [7–12]. However, these modalities were not available in the datasets used in this study and were not included in the model’s training or evaluation. Therefore, in this paper, they are discussed only as possible future extensions, while the experimental scope is limited to text records and structured indicators at the record level.

A separate area of research is connected with text-based information. Social media and digital communication contain a large number of signals, reflecting users’ emotional states. Modern transformer models allow the extraction of such signals more effectively. However, even here, most solutions remain unimodal, without taking into account possible links with other structured indicators. Attempts to combine different data types are being made within multimodal training approaches [13,14]. It is important to clarify that the experimental verification presented in this article was conducted only in the “text + record-level structured textual indicators” configuration. Physiological signals and integration with wearable devices are considered by us as a conceptual expansion of the architecture for the future, and they were not actually used in the described experiments. Previous research shows that the integration of heterogeneous sources can increase robustness; however, in this pilot study, we considered only text data and structured indicators of crisis situations [15]. Moreover, attention mechanisms allow the modeling of interactions between modalities more flexibly than the simple concatenation of features [16,17]. However, despite these results, most decisions still remain fragmented—multimodality is often added “on top”, not built into the foundation of the architecture.

Interpretability remains an equally important problem. With increasing model complexity, transparency decreases. Although explainable machine learning methods are developing actively [18,19], and their importance for mental health monitoring tasks is highlighted in a number of works [20], in practice, interpretation is often seen as an additional stage, not as part of the model itself. This limits the trust in such systems and their application in sensitive areas [21].

The time factor makes this situation more challenging. Psycho-emotional states change over time, reacting to external and internal impacts. It is evident that accounting for temporal dependency relationships can significantly improve the quality of models [22,23]. However, many approaches continue to use aggregated features, smoothing patterns and losing major transitional states.

Finally, another aspect that is often overlooked is the calibration of probabilities. In applied problems, it is important not only to predict a class but also to correctly assess the

level of model confidence. Despite the high accuracy values and F1-measures in a number of works [2,24], metrics of calibration, such as the Brier score, are rarely used. This may lead to the overestimation or underestimation of risks and reduce the practical importance of decisions.

In practice, despite the achievements, a number of unresolved problems remain. Existing methods suffer from three main restrictions:

1. A lack of multimodal integration;
2. Poor interpretability;
3. A lack of probabilistic calibration.

First, a scalable multimodal architecture unifying text data for predicting psycho-emotional risk is proposed. In this study, the term “psycho-emotional risk” refers to a categorical, modeled construct based on the delineation of public datasets. This is not a clinically validated psychiatric diagnosis or standard psychometric tool. The proposed system should be viewed as an experimental computational model, designed to identify risk patterns in digital data, rather than a tool for clinical diagnosis.

Throughout the article, we use the term “psycho-emotional risk” as a non-clinical, operational, generalizing designation for categorical states, defined by the structure of the dataset (for example, crisis state level). This term is selected for ease of presentation and should not be perceived as a standardized psychiatric concept. The term “psycho-emotional risk” was chosen not because it is a standard clinical term but because the datasets do not provide a clinically validated scale for distress or mental health risk. It is an operational label for the classes defined by the dataset. Therefore, the model does not measure psychological states as a clinical tool but rather classifies digital patterns within the given labeling. It is important to distinguish between computational modeling, digital technology approaches, and clinically validated psychological assessment methods. The latter are based on standardized procedures and professional interpretation by specialists. The present work, in turn, focuses on recognizing patterns using machine learning techniques in multimodal digital signals. This distinction is necessary to avoid overestimating the practical and clinical value of the proposed model.

The proposed architecture was originally designed as an extensible computational framework. However, in the present experimental study, the evaluated configuration was limited primarily to textual records. Record-level structured information was used only when an independent non-label structured field was available within the same original observation. No observations were matched across datasets, and no synthetic physiological, wearable-sensor, visual, or behavioral inputs were created. Therefore, the reported experiments should be interpreted as the evaluation of a text-based computational pipeline with optional same-record auxiliary structured information, not as the validation of a fully synchronized multimodal fusion system.

Interpretation is initially embedded in the architecture: methods of feature attribution signs and weight analysis are used, which allow the monitoring of the contribution of data sources to the final decision. Additionally, the model is oriented toward obtaining calibrated probabilistic estimates, which are essential for practical use.

Experimental testing was conducted on two public text-based datasets. The evaluated configuration was limited to textual records and record-level structured indicators when available. Physiological, wearable-sensor, visual, and continuous behavioral modalities were not used in model training or evaluation. Single-modal baseline models and the proposed approach are considered, and both performance and calibration metrics are evaluated.

In this paper, the following key results and provisions are presented.

First, a scalable multimodal architecture unifying text data and structured psycho-emotional state indicators within a single system is proposed. In addition, the architecture initially provides the possibility of expansion through the connection of physiological class modalities. Second, the implemented pipeline includes a feature-level integration step that can combine available within-record representations when more than one independent input field is present. This integration should not be interpreted as the synchronization of independent multimodal streams, since physiological, wearable-sensor, visual, and continuous behavioral data were not available in the datasets and were not used in the reported experiments. Third, the procedure of probabilistic calibration (temperature scaling) is built into the architecture. It is not seen as an external post-processing step but as part of the overall model logic, associated with risk assessment formation.

Special attention is paid to the statistical verification of model stability. The study uses cross-validation, confidence intervals, and testing of the statistical significance of differences between models. Finally, an empirical test was carried out on two independent open datasets, which allows us to discuss the generalizability of the model and its portability beyond one specific dataset.

The remainder of the article is structured as follows: Section 2 describes the data, the methods of their processing, and the model architecture. Section 3 presents the experimental results and their analysis. Section 4 is devoted to the discussion of the obtained results and limitations. In Section 5, the main conclusions and future research directions are formulated.

A summary comparison of the proposed approach with standard solutions is presented in Table 1.

Table 1. Comparative positioning of the proposed approach within the experimentally evaluated scope.

Study Type	Multimodal	Attention-Based Fusion	Built-In Calibration	Statistical Validation	Cross-Dataset Testing
Text-only ML models	No	No	Rare	Limited	Rare
Physiological-only models	No	No	No	Partial	Rare
Basic multimodal fusion	Partial	Often simple concatenation	No	Limited	Rare
Proposed model	Proposed model—limited/record-level text-based configuration	Available feature-level integration when same-record auxiliary fields are present	Yes (temperature scaling integrated)	Yes (CV + CI + <i>t</i> -test)	Yes

Table 1 illustrates the positioning of the proposed approach solely within the context of a field-tested experimental setup. This comparison does not imply that physiological, wearable-sensor, visual, or behavioral extensions have been empirically validated. The novelty of this work is limited to the integration of a transformer-based text representation, structured indicators at the record level, probabilistic calibration, and interpretability into a single reproducible computational pipeline. Therefore, the contribution of this paper should be understood as an integration and validation protocol for the tested “text + structured features” configuration, rather than as a fully validated multimodal clinical monitoring system.

2. Materials and Methods

2.1. Overall Framework

The proposed framework was implemented as a controlled record-level classification pipeline. In the reported experiments, the input consisted of textual records transformed into contextual embeddings using transformer-based encoders. Where same-record auxiliary non-label structured information is available in future extensions, it may be incorporated before the classification layer; however, the final quantitative experiments reported in this study were primarily text-based [5,6].

The broader architecture was designed to be extensible and may support physiological, wearable-sensor, visual, and continuous behavioral data in future studies [25,26]. These modalities are shown only as conceptual extensions in Figure 1 and are not part of the empirical evaluation reported in this manuscript. The main implemented pipeline includes text preprocessing, transformer-based representation, classification, probability calibration, and explainability analysis [27,28].

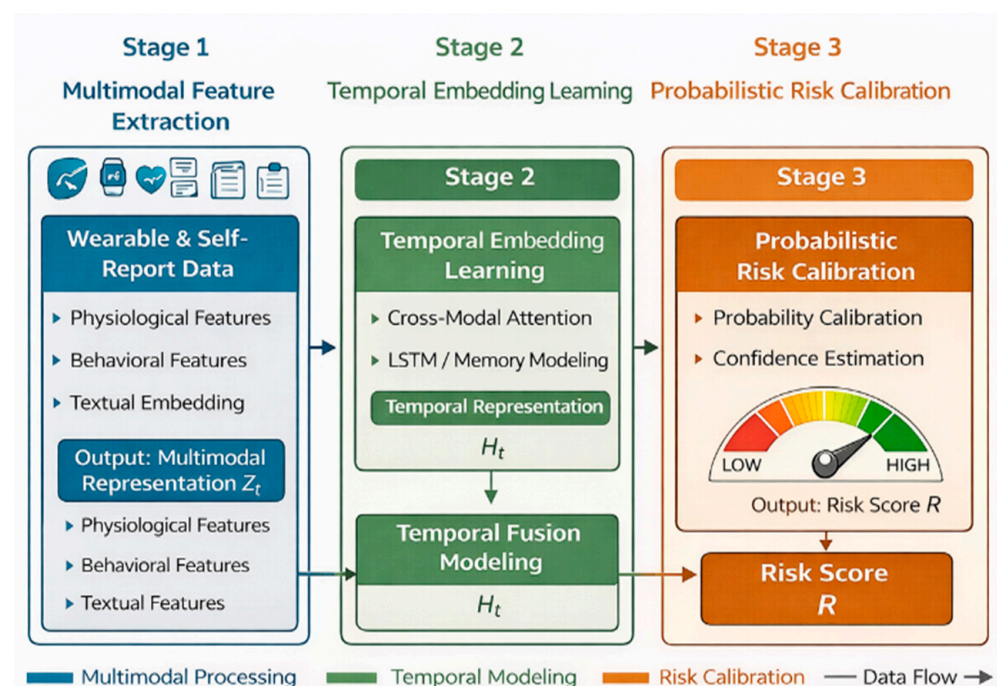


Figure 1. Conceptual overview of the proposed extensible framework. The quantitative experiments reported in this study were based on textual records, while non-evaluated modules are shown only as possible future extensions.

The empirical implementation used in this study is described in Sections 2.2–2.7. Conceptual extensions that were not evaluated experimentally are not included in the reported results.

2.2. Dataset Description

For the evaluation of the proposed architecture, two datasets were used, differing in their origin and structure. Two public datasets were used in the pilot study. Both datasets contained textual information, and one additionally included structured indicators of psychological crisis [29]. This decision was made because, in the task of analyzing psycho-emotional states, models often show the best results within one dataset, and, when transferred to other conditions, their persistence may decrease [2,30]. The use of several sources allows for a more objective evaluation of a model's ability to detect stable patterns [14].

The first dataset represents an open dataset, oriented toward analyzing textual symptoms of the psycho-emotional state. In particular, the Psychological Crisis Dataset was used, which contains approximately 1000 text entries, labeled by the level of psycho-emotional risk.

After preprocessing, including the removal of missing entries and duplicated records, the total number of unique observations in the Psychological Crisis Dataset was $N = 1000$. The class labels were taken from the original structure of the dataset and were not manually modified in this study. It is important to highlight that the class labels in the dataset were used as provided by the creators of the dataset on the Kaggle I platform. These labels are not clinically validated and have not been confirmed by psychodiagnostics in our study. Therefore, they are treated as data-driven categories (digital proxies), suitable for computational modeling and comparative model assessment, but not as equivalent to clinical diagnoses or standardized psychometric scales.

The second dataset, the Reddit Mental Health Dataset, includes user messages from thematic communities. The experiment used a subsample of $N = 4420$ text messages after deleting incomplete records and duplicates. The original dataset contained 5957 records. The class markings were consistent with the original dataset categorization and were not manually modified. Separately, it should be noted that labeling on social media platforms is often indirect by nature: it can rely on self-description, selection rules, automatic signs, or weak supervision—depending on how the dataset was originally formed. As a result, under such conditions, the model can significantly distinguish between linguistic and stylistic clusters but not stable mental states in the clinical sense. Thus, the interpretation of the results in this work is limited to computational modeling and does not transfer directly to clinical findings. The final sample shows five categories of psycho-emotional states; after cleaning, the distribution of classes ranged from 809 to 974 observations per class, which indicates moderate imbalance without artificial rebalancing. Sub-sample data have been used to achieve comparability among the experiments, derived from the body of the data, containing several thousand text messages representing different categories of psycho-emotional states.

In both cases, the data were converted into a common preprocessing format for model training and evaluation. However, this formatting does not imply the direct matching, merging, or synchronization of individual observations across datasets. Despite differences in sources, both datasets contain categories reflecting various emotional and psychological states of users. For the experiments, a stratified data split was applied while preserving the class distribution. An 80/20 split (training and test samples) was used, which is in line with widespread practice in tasks of analyzing psycho-emotional data [24]. The test sample size was 10% of the total data volume for each dataset, corresponding to 100 observations for the Psychological Crisis Dataset and 442 observations for the Reddit Mental Health Dataset after clean-up. In addition to the fixed breakdown into training and test samples, 5-fold cross-validation was applied to assess the stability of the results. In each fold, the model was taught independently and the final metrics were calculated as averages across all partitions. This approach reduces the impact of the random initialization of weights and the characteristics of specific data breaks, providing a more reliable assessment of the model's generative capacity. The class distribution in the datasets is generally unbalanced, which is characteristic of psycho-emotional state analysis tasks.

Class balancing in this study was not performed. Model training was carried out on the original distribution of data without applying oversampling/undersampling methods and without using class weightings. Such an approach preserves the natural structure of the data and reflects the actual statistical proportions of the categories.

For this reason, when training the model, the features of the class distribution were taken into account to avoid shifting the predictions toward the most represented categories [14].

The main characteristics of the used datasets are shown in Table 2.

Table 2. Main characteristics of used datasets.

Dataset	Type of Data	Size (N)	Classes	Source
Psychological Crisis Dataset	Textual	1000 (after preprocessing)	Multilevel risk	Kaggle
Reddit Mental Health Dataset	Textual	4420 (after preprocessing; original 5957)	Psycho-emotional state categories	Kaggle

To avoid ambiguity regarding the experimental design, the final input structure used for training and evaluation is specified in Table 3. The two datasets were not merged at the level of individual observations. No record from the Psychological Crisis Dataset was paired with any record from the Reddit Mental Health Dataset. Therefore, no cross-dataset multimodal sample was constructed.

Table 3. Final input structure used during model training and evaluation.

Dataset	Input Variables Used by the Model	Target Variable	Independent Structured Input	Role in the Experiments
Psychological Crisis Dataset	Textual records after cleaning, tokenization, and transformer-based embedding	Original dataset risk/crisis class label	No independent non-label structured predictor was used in the final quantitative experiments. The class label was used only as the target variable and was not included as an input feature.	Main text-based dataset for record-level classification
Reddit Mental Health Dataset	Textual posts/messages after cleaning, tokenization, and transformer-based embedding	Original dataset mental health category label	Not used; no independent structured field was available for fusion.	Additional text-based dataset for robustness checking
Cross-dataset sample	Not constructed	Not applicable	Not applicable	No matching, merging, or synthetic pairing was performed.

In the final quantitative experiments, the model inputs were textual records converted into transformer-based embeddings. The class label was used only as the supervised learning target and was never used as an input modality. No observation was constructed by combining records from different datasets. No synthetic structured, physiological, wearable-sensor, visual, or behavioral features were generated. Therefore, the reported experiments should be interpreted as record-level text-based computational classification, not as the validation of a fully multimodal fusion model [31].

It should be noted that the two datasets were not merged at the level of individual observations. They were used as separate experimental sources. The Psychological Crisis Dataset was used for the main structured configuration at the text and record level, where

structured fields were available within the same record. The Mental Health Dataset from Reddit was used as a supplementary text dataset for external validation. Since the datasets did not contain common user identifiers, timestamps, or common post identifiers, no direct matching between the datasets was performed. Therefore, the study does not claim that the Reddit records were matched with records from the Psychological Crisis Dataset. The distribution of classes in the datasets used was further analyzed. As demonstrated in Figure 2, the class distribution is relatively balanced (approximately 18–22% per class after preprocessing), indicating moderate imbalance without artificial balancing. The shares of individual classes are at a similar level and are close to 20% per category. Therefore, the reported metrics reflect performance under the original distribution rather than an artificially balanced dataset; no oversampling/undersampling was used. Such a distribution indicates a balance, which is a positive factor when training the model, because it reduces the risk of biased predictions toward the most represented classes. In contrast to many tasks of analyzing psycho-emotional states, where strong imbalance is often observed, in this case, the model was trained under more stable conditions [14].

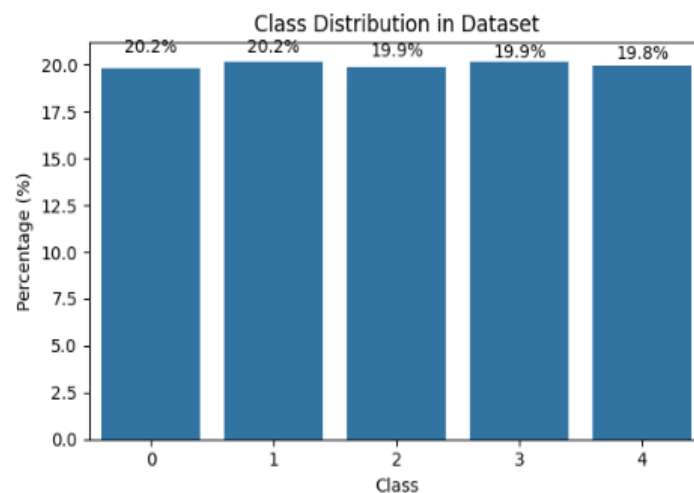


Figure 2. Distribution of categories of psycho-emotional states in the combined experimental sample. The figure shows the relative balance of the classes on which training and subsequent model evaluation were conducted.

However, even in relatively equally distributed classes, stratified sampling data were used, which allowed the preservation of the original structure of the data during training and testing. Moreover, metrics that are sensitive to the class distribution were used to assess the quality of the model, such as the F1-score and ROC-AUC, which allows the more accurate interpretation the obtained results. A visualization of the class distribution is presented in Figure 2.

No artificial class balancing was applied. Stratified splitting was used only to preserve the original class proportions across training, validation, and test partitions. Therefore, the reported metrics reflect model performance under the original dataset distribution.

2.2.1. Psychological Crisis Dataset

An open dataset—the Psychological Crisis Dataset—was used as the main dataset, available on the Kaggle platform: <https://www.kaggle.com/datasets/programmer3/student-psychological-crisis-risk-dataset> (accessed on 20 April 2026).

This dataset contains textual data from users, reflecting different types of psycho-emotional states, from neutral to a potential crisis. The main emphasis is placed on identifying signs of emotional stress and instability based on linguistic characteristics.

In contrast to traditional scales, here, the data are formed in the natural digital environment, which makes them closer to real user behavior. At the same time, such data are sensitive to the context and individual features, which must be kept in mind upon interpreting the results [1,2]. The full characterization of the dataset is provided in Table 4 below.

Table 4. Main characteristics of Psychological Crisis Dataset.

Parameter	Description
Dataset name	Psychological Crisis Dataset
Source	Kaggle (open access)
Data type	Textual
Application area	Analysis of psycho-emotional risk
Task type	Classification
Classes	Neutral state/increased risk
Preprocessing	Cleaning, text normalization
Performance	Vectoring (transformer models)

During data preparation, standard preprocessing was carried out, including text cleaning, normalization, and conversion to a common format. Next, the textual data were transformed into vector representations using transformer models, which allowed us to capture the context and semantic relationships between words [3,4].

For conducting experiments, the data were split into training, validation, and test sets. Division was performed while maintaining the class distribution, which allowed us to reduce the risk of model bias toward the dominant class [2]. The class distribution in the dataset is shown in Figure 3.

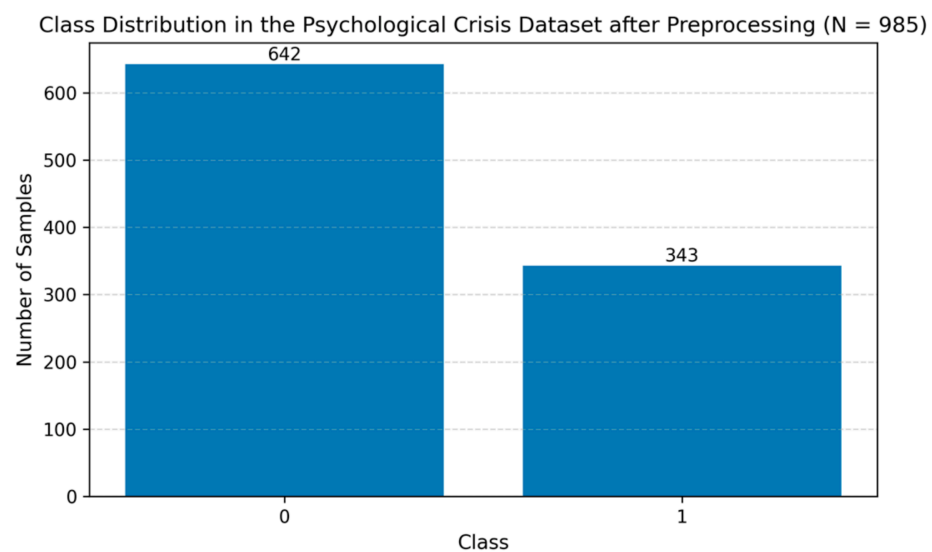


Figure 3. Class distribution in the Psychological Crisis Dataset after preprocessing and stratified sampling (N = 1000).

Figure 3 shows the class distribution in the Psychological Crisis Dataset after removing missing and duplicate records. The figure reflects the final dataset used in the experiments, not the original raw sample. Resampling, undersampling, and synthetic class balancing were not applied. Stratified splitting was used only to preserve the original class distribution during training, validation, and testing. This approach allowed us to assess the natural class structure of the dataset without introducing artificial changes to the data distribution.

2.2.2. Reddit Mental Health Dataset

As a second set, the Reddit Mental Health Dataset was used, containing textual publications from users, related to different aspects of psycho-emotional states. In contrast to the first dataset, where the data are presented in a structured way, here, they consist of text formed in a natural digital environment.

Such data are of particular interest since they allow the analysis of user behavior in conditions close to real-life settings. In a number of studies, it is noted that linguistic activity may reflect signs of anxiety, stress, and other psycho-emotional states, which are not always outlined in questionnaire or experimental data [14,32].

At the same time, working with text-based data involves certain limitations. User text does not have a common structure, can significantly differ in style and context, and also contains noise elements. Even so, modern methods of natural language processing, including transformer models, allow the extraction of informative features even from unstructured data [16,33].

A general description of the dataset is shown in Table 5.

Table 5. General characteristics of Reddit Mental Health Dataset.

Parameter	Description
Dataset name	Reddit Mental Health Dataset
Source	Kaggle (open access)
Data type	Textual
Application area	Psycho-emotional state analysis
Task type	Classification
Feature	Unstructured user texts
Purpose	Model verification on data from real environment

Before use, the data were converted into a common format. This included text cleaning, deleting extra characters, and normalization. After this, vectorization using transformer models was performed, which allowed us to capture context-dependent and semantic relationships between words [16,33]. The distribution of observations by class is presented in Figure 4.

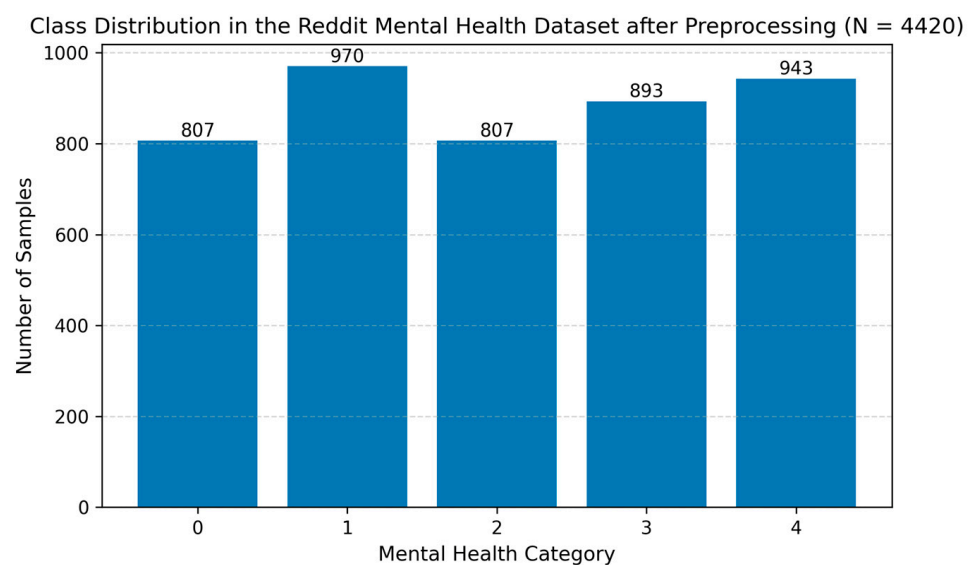


Figure 4. Class distribution in the Reddit Mental Health Dataset after preprocessing and stratified sampling (N = 4420).

Figure 4 shows the class distribution in the Reddit Mental Health Dataset after removing incomplete records and duplicate observations. The figure reflects the final text sample used in the experimental evaluation. The distribution was not artificially balanced; no oversampling, undersampling, or manual class weighting was applied. To maintain consistency in class proportions across the training, validation, and test samples, stratified splitting was used. This approach allowed us to evaluate the model under the original class distribution without introducing artificial changes to the dataset.

In addition, Figure 5 shows the textual representations used by the model.

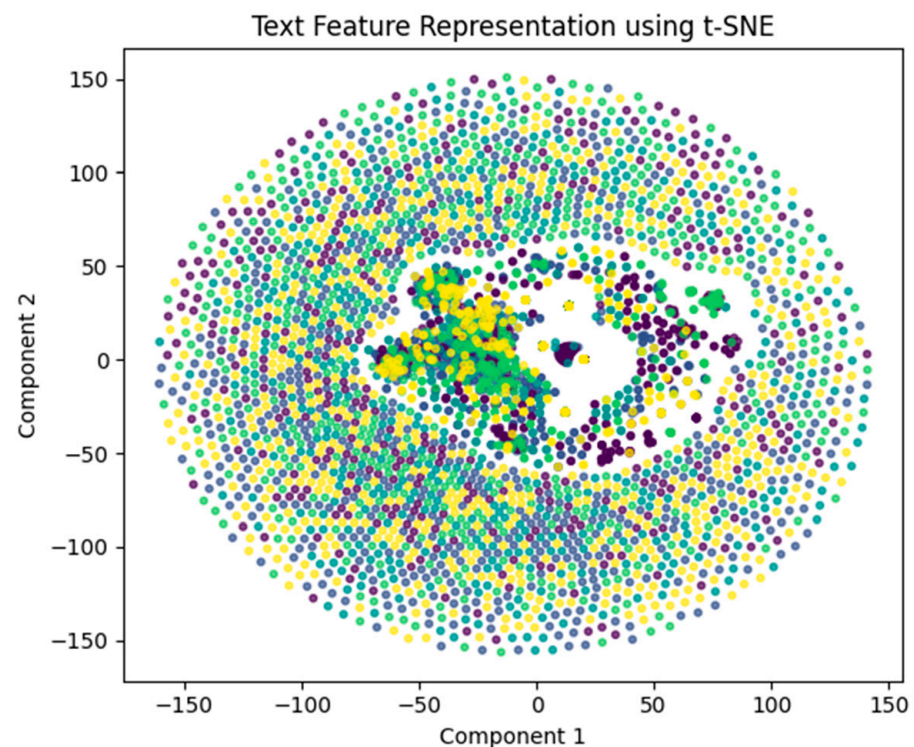


Figure 5. Visualization of text embeddings extracted by a transformer model using t-SNE. Each dot represents one text record, and different dot colors indicate the original class/category labels in the plotted sample. The colors are used only to visualize class distribution in the embedding space and do not denote different modalities, datasets, or additional input features.

As shown in Figure 5, the textual representations form a partly overlapping structure without a clearly separated class distribution. The different dot colors correspond to the original class/category labels and are used only to make the two-dimensional projection easier to interpret. They should not be interpreted as separate modalities or independent data sources. This overlap reflects the complex nature of psycho-emotional states, where boundaries between categories are not strictly discrete. Moreover, local areas of point concentration are observed, which suggests hidden semantic patterns.

This visualization supports the interpretation that text embeddings contain class-related patterns, although these patterns are not fully separated in the two-dimensional projection.

The dataset is available at <https://www.kaggle.com/datasets/neelghoshal/reddit-mental-health-data> (accessed on 20 April 2026).

2.2.3. Data Preprocessing

Before model training, both datasets were converted into a common preprocessing format. This step was necessary because the source datasets differed in structure, textual style, label format, and noise level. In the present empirical pipeline, preprocessing was limited to textual records and record-level structured indicators. For text-based data, the

main processing stages are cleaning and normalization. The texts were given a standard form; after this, their tokenization was performed. Formally, this process can be presented as follows:

$$X_{\text{text}}^{\text{clean}} = f_{\text{clean}}(X_{\text{text}}) \quad (1)$$

where X_{text} —source text data and $X_{\text{text}}^{\text{clean}}$ —cleaned and normalized texts.

Then, text vectorization was conducted using transformer models:

$$E = f_{\text{embed}}(X_{\text{text}}^{\text{clean}}) \quad (2)$$

where E —vector representation matrix. In the current implementation, the dimension of the vector representation for the text modality is 768, which corresponds to the BERT-base-uncased model configuration used in the experimental part. Accordingly, the embedding matrix can be represented as $E \in \mathbb{R}^{N \times 768}$, where N is the number of text observations in the sample.

Such an approach allows us to account for contextual dependencies and semantic relationships between words, which is extremely important in analyzing of psycho-emotional states [26,32].

In this experimental implementation, preprocessing was applied only to text records and structured indicators at the record level. The text data were cleaned, normalized, tokenized, and converted into contextual embeddings using transformer-based encoders. Structured indicators at the record level, if available within a single observation, were converted into numerical vectors and normalized using parameters estimated solely on the training data [34].

Only preprocessing procedures applied to the empirical inputs used in the experiments are reported in this section. Additional preprocessing procedures for conceptual extensions are not part of the reported experimental pipeline.

The overall process of empirical preprocessing is shown in Table 6.

Table 6. Main steps of empirical data preprocessing.

Step	Description
Text cleaning	Removal of noise, missing records, duplicated records, and irrelevant characters
Text normalization	Conversion of textual records into a common format
Tokenization	Splitting text into model-compatible tokens
Text vectorization	Transformation of textual records into transformer-based embeddings
Structured indicator encoding	Conversion of record-level structured indicators into numerical vectors when available
Feature scaling	Normalization of structured numerical indicators using training-data parameters only
Train/validation/test split	Stratified data splitting for model training and evaluation
Cross-validation	Five-fold stratified cross-validation for stability assessment

Stratified chunking data were used additionally:

$$D = D_{\text{train}} \cup D_{\text{val}} \cup D_{\text{test}} \quad (3)$$

This allows the preservation of the class distribution and improved model stability during training [2,24].

The presented expressions are used to formally describe the data processing and general transformations applied within the proposed architecture [35].

In general, the executed preprocessing procedures provided data consistency and prepared them for their subsequent integration into the multimodal architecture. This is especially important in psycho-emotional state analysis tasks, where data may differ by nature and quality.

2.3. Experimental Data Configuration and Scope of Fusion

In this study, the experimental data configuration is separated from the broader conceptual architecture. The broader architecture is designed as extensible and may support physiological, wearable-sensor, visual, and continuous behavioral channels in future work. However, these modalities were not available in the datasets used in the present experiments and were not used in model training, fusion, calibration, or evaluation.

The final experimental inputs were constructed at the level of individual records. For each observation, the available text field was cleaned, tokenized, and encoded using a transformer-based encoder. In the final quantitative experiments, no independent non-label structured predictor was used as a separate input modality. The class label was used only as the supervised learning target. No structured feature was obtained from another dataset, and no synthetic structured feature was generated to replace a missing modality.

No cross-dataset alignment was performed. The Psychological Crisis Dataset and the Reddit Mental Health Dataset were treated as separate experimental sources. They do not contain common user identifiers, synchronized timestamps, shared assessment sessions, or common record identifiers. Therefore, there was no methodological basis for matching observations between the two datasets. No artificial pairs, inferred links, or aggregate-level correspondences were created.

The term “fusion” in this manuscript is therefore used only in a limited sense. It refers to the integration of available within-record representations, not to the fusion of independently synchronized multimodal streams. In the Reddit Mental Health Dataset, where only textual records were available, the model operated as a text-based classifier. In the Psychological Crisis Dataset, record-level structured information was used only when it existed within the same original observation and was not identical to the target label. The class label was used only as the supervised learning target and was not treated as an input modality.

Accordingly, the reported experiments should not be interpreted as the validation of a fully multimodal system combining text, physiological signals, wearable-sensor data, visual information, and behavioral streams. They refer to a reduced computational configuration based primarily on textual representations and, where available, same-record auxiliary structured information. Full multimodal validation requires synchronized datasets in which several independent modalities are collected from the same individual, assessment, or time point. Such datasets were not available in the present study and are identified as a priority for future work.

2.4. Conceptual Graph Extension

The graph representation was considered solely as a conceptual extension of the proposed architecture. It was not implemented as a standalone graph neural network module and did not contribute to the quantitative results presented in Section 3. Therefore, the main experimental pipeline does not include graph learning. The graph-based extension

is mentioned only as a possible direction for future work and is not used in the empirical evaluation reported in this manuscript.

2.5. Hybrid Model Architecture

The proposed model is described as a hybrid computational pipeline because it combines transformer-based text representation, probability calibration, and explainability components within one controlled evaluation framework. In the current experimental implementation, the model operated primarily on textual records, as specified in Section 2.3.

In the implemented experimental configuration, the hybrid model combines text representations based on transformers, probability calibration, and explainability components. The term “hybrid” is used here to describe the integration of these computational components into a single pipeline, rather than the empirical use of independently synchronized multimodal data streams [36].

The model consists of three main components: feature extraction, probability calibration, and explainable record-level classification. Text records are encoded using transformer-based embeddings. The resulting representation is passed to the classifier, and the predicted probabilities are calibrated using temperature scaling. The main components of the implemented hybrid computational pipeline are summarized in Table 7.

Table 7. Main components of hybrid model.

Component	Description
Text-based module	Extraction of transformer-based textual representations
Calibration module	Temperature scaling of predicted probabilities
Explainability module	SHAP-based interpretation of model outputs
Classifier	Record-level class prediction

In the implemented experiments, the model should be interpreted as a controlled text-based computational pipeline. The contribution of the present study is limited to the controlled evaluation of transformer-based textual representation, probabilistic calibration, interpretability, and clearly specified record-level input handling under a reproducible experimental protocol.

2.6. Risk Scoring and Calibration

The final stage of the proposed architecture involves risk assessment and its further calibration. In contrast to tasks where it is sufficient to obtain class predictions in the given formulation, it is also important to assess the degree of model confidence. This is especially relevant in psycho-emotional state analysis tasks, where decisions can be used in practical monitoring and support systems [21,37].

Based on the obtained hidden representation h , the probability of an observation belonging to a certain class is formed:

$$P(y|x) = \sigma(Wh + b) \quad (4)$$

where $\sigma(\cdot)$ —sigmoidal activation function. In practice, the model produces not only a category but also a numerical risk score, which can be interpreted as a probability.

However, such probabilities are not always well calibrated. In several studies, it is noted that modern models, especially deep neural networks, can require the reassessment of their confidence [21]. This creates additional difficulties when using the model in real-world conditions.

A probabilistic calibration procedure is used to improve the reliability of the results. In this work, temperature scaling is applied, which can be written as follows:

$$P_{\text{cal}} = \frac{1}{1 + e^{-z/T}} \quad (5)$$

where z —model logit and T —temperature parameter. T selection allows the smoothing of the output distribution, enabling it to align more closely with empirical observations.

The temperature parameter T was optimized on the validation sample by minimizing the negative log-likelihood. Optimization was performed after the base model training was completed and was not accompanied by the calibration of its weights. Thus, the calibration procedure did not affect the classifier's parameters but only corrected the probability outcome distribution. This approach is commonly used in tasks where the interpretability and reliability of predictions are required, including emotional state analysis and multimodal emotion recognition systems [17,28].

In the final implemented configuration, the risk score was computed from the classifier output based on the available textual representation. The class probability produced by the model was then calibrated using temperature scaling [38,39]. If same-record auxiliary non-label structured information becomes available in future extensions, it can be integrated before the classification layer. No non-evaluated conceptual extension contributed to the final risk score. It is important to note that psycho-emotional states do not always have clear boundaries between categories. In a number of works, it is noted that emotions can change rapidly and are characterized by continuous scales [40]. This further justifies the use of probabilistic assessment instead of hard classification.

The major steps of the assessment procedure and the calibration of risk are shown in Table 8.

Table 8. Major steps of assessment and risk calibration.

Step	Description
Prediction	Obtaining class probability
Logits	Calculating the hidden values of the model
Calibration	Temperature scaling
Input representation	Use of textual representation for prediction; optional same-record auxiliary information only when available
Final evaluation	Risk formation

Overall, the use of probabilistic calibration allows us to improve the reliability of the model and makes it more suitable for practical applications. This is particularly relevant in tasks related to mental health, which require not only accuracy but also the validity of decisions [21,37].

2.7. Implementation Details

The implementation of the proposed architecture was executed in Python using the PyTorch library. For text data processing, the pretrained transformational model BERT-base-uncased (Hugging Face Transformers library) was used to obtain contextual embeddings of text sequences. The dimension of the hidden model space was 768, which corresponded to the standard BERT-base-uncased configuration. The pretrained weights were loaded from the official Hugging Face Transformers library without modifying the basic transformer architecture. During training, the fine-tuning of model parameters was performed on the classification target. In the preliminary experiments, the RoBERTa-base model was also considered; the final results presented in the article were obtained for the BERT-base-uncased configuration.

To improve the reproducibility, the main training, optimization, and hardware parameters used in the experiments are summarized in Table 9. The same training configuration was used for the main model, ablation settings, and encoder comparison unless otherwise stated.

Table 9. Main training and reproducibility parameters.

Parameter	Value/Description
Text encoder	BERT-base-uncased; RoBERTa-base in encoder comparison
Text embedding dimension	768
Batch size	16
Optimizer	AdamW
Learning rate	2×10^{-5}
Number of training epochs	5–10, depending on validation performance
Dropout	0.1
Loss function	Cross-entropy loss
Data split	80/10/10 train/validation/test split with stratification
Cross-validation	Five-fold stratified cross-validation
Early stopping	Patience = 3 epochs based on validation loss
Calibration	Temperature scaling on validation outputs
Random seed	Fixed seed for all experiments
Framework	PyTorch 2.x; Hugging Face Transformers 4.x
Hardware	NVIDIA GPU with 12 GB VRAM; 32 GB RAM; Intel i7 CPU

For probabilistic calibration, temperature scaling was used. After learning the base model, denoted here by z , the model logits were transformed as follows:

$$p = \sigma\left(\frac{z}{T}\right) \quad (6)$$

where T is a temperature parameter optimized on the validation sample by minimizing the negative logarithmic plausibility. The optimization of the T parameter was performed after model training without changing the model weights.

The presented configuration reflects the actual conditions in which the experiments were conducted, which makes it possible to reproduce the results obtained under comparable conditions.

2.8. Ablative Analysis Protocol and Comparison of Text Encoders

In order to assess the contributions of individual components of the proposed architecture, an ablative study was conducted. All variants of the model were tested using the same experimental protocol as for the complete model. The same data separation scheme was used for training, validation, and test samples, with the same 5-fold cross-validation and the same preprocessing pipeline. The purpose of the ablation analysis was not to add new external baseline models but to check how the behavior of the proposed architecture changes when replacing and deleting individual components.

Three main ablative settings were considered. In the first setting (attention versus concatenation), the attention-based integration block was replaced with the simple concatenation of the available within-record representations, while the classification module and the overall training protocol remained unchanged. In the second version, LSTM-versus-no-LSTM, the recurrent time block was deleted and a feed-forward pooling layer was used instead, applied to the combined feature representation. In the third variant, calibration-versus-no-calibration, the model was evaluated before and after applying temperature scaling. For this comparison, apart from accuracy, the F1-score, and the ROC-AUC, metrics

sensitive to the probabilistic calibration quality were calculated, including the Brier score and expected calibration error (ECE).

In addition, a controlled comparison of text encoders within the same architecture was performed. The BERT-base-uncased encoder, used in the basic model configuration, was compared with RoBERTa-base under the same conditions for combining features, classification, and probabilistic calibration. Where applicable, a BiLSTM-based text encoder was also included as a neural network but not a transformer-based option for comparison. Such comparison was necessary to understand whether the results obtained were mainly related to the proposed fusion and calibration scheme or mainly due to the choice of a particular transformer encoder. All results of the ablative analysis and encoder comparison are presented as the mean $\pm$ standard deviation across five-fold cross-validation splits. For these experiments, no additional data augmentation methods or manual class balancing were used. This was done specifically to ensure that the results remained directly comparable with the basic assessment of the proposed model.

2.9. Ablation Study and Encoder Comparison

An ablation analysis was conducted to assess whether the performance of the proposed model is attributable to specific architectural components or is largely determined by the chosen text encoder. All model variants were evaluated under identical conditions: the same data preprocessing pipeline, the same cross-validation protocol, and, where applicable, the same probabilistic calibration procedure were used. The results are presented in Table 10.

Table 10. Ablation study of the proposed architecture.

Configuration	Accuracy	F1-Score	ROC-AUC	Brier Score	ECE
Full model: BERT + cross-modal attention + LSTM + temperature scaling	0.593 $\pm$ 0.009	0.592 $\pm$ 0.008	0.860 $\pm$ 0.006	0.540 $\pm$ 0.009	0.064 $\pm$ 0.007
Without cross-modal attention: BERT + concatenation + LSTM + temperature scaling	0.607 $\pm$ 0.018	0.608 $\pm$ 0.018	0.858 $\pm$ 0.009	0.535 $\pm$ 0.015	0.066 $\pm$ 0.010
Without LSTM: BERT + cross-modal attention + feed-forward pooling + temperature scaling	0.596 $\pm$ 0.023	0.595 $\pm$ 0.024	0.859 $\pm$ 0.009	0.541 $\pm$ 0.018	0.049 $\pm$ 0.012
Without calibration: BERT + cross-modal attention + LSTM, before temperature scaling	0.593 $\pm$ 0.009	0.592 $\pm$ 0.008	0.859 $\pm$ 0.006	0.544 $\pm$ 0.010	0.080 $\pm$ 0.008

The ablation results indicate that replacing the attention-based integration block with simple concatenation did not reduce the performance and, in some metrics, produced slightly higher values. This result is scientifically important because it suggests that the expected advantage of attention-based integration is not guaranteed under the current data configuration. A likely explanation is that the available experimental inputs were dominated by textual representations, while independently synchronized multimodal signals were not available. Under such conditions, the attention block has limited opportunities to learn complementary cross-source interactions, and a simpler concatenation-based representation may be sufficient or even slightly more stable. Therefore, the results should not be interpreted as evidence that cross-modal attention is universally superior. Instead, they show that the benefit of attention-based integration depends on the availability, quality, and complementarity of independent input sources. This finding further supports the need for future validation on synchronized multimodal datasets collected from the same individuals, assessments, or time points.

The most significant improvement was observed for probability calibration. As expected, temperature scaling had virtually no effect on the accuracy, F1-score, or ROC-AUC, since this procedure does not retrain the classifier or alter the underlying classification boundary. However, calibration improved the reliability of the probability estimates: the ECE decreased from 0.080 ± 0.008 before calibration to 0.064 ± 0.007 after calibration, and the Brier score also improved slightly—from 0.544 ± 0.010 to 0.540 ± 0.009 . Therefore, the contribution of calibration in this work is primarily related to improving the reliability of probabilistic estimates, rather than to a direct increase in classification accuracy. To assess the potential impact of the text encoder separately, an additional comparison was conducted within the same architecture. BERT-base-uncased and RoBERTa-base were evaluated using the same fusion protocol and probability calibration. The results of this comparison are presented in Table 11.

Table 11. Comparison of text encoders within the same architecture.

Text Encoder	Fusion Method	Calibration	Accuracy	F1-Score	ROC-AUC	Brier Score	ECE
BERT-base-uncased	Cross-modal attention	Temperature scaling	0.593 ± 0.009	0.592 ± 0.008	0.860 ± 0.006	0.540 ± 0.009	0.064 ± 0.007
RoBERTa-base	Cross-modal attention	Temperature scaling	0.666 ± 0.005	0.667 ± 0.007	0.896 ± 0.004	0.459 ± 0.005	0.040 ± 0.011

The comparison of the encoders shows that the choice of text encoder significantly affects the model's performance. RoBERTa-base achieved higher results across all metrics compared to BERT-base-uncased, including the accuracy, F1-score, ROC-AUC, Brier score, and ECE. This suggests that stronger contextual text representations can improve the performance of the tested pipeline under identical fusion and probability calibration settings.

At the same time, this comparison does not imply that all architectural components of the proposed model are universally optimal. It is more accurate to say that the proposed pipeline can be tested with different transformer encoders within a single controlled protocol. This approach makes the model evaluation more transparent and demonstrates that the final quality depends not only on the feature combination scheme but also on the chosen text model.

3. Results

It should be noted separately that the graph modeling described in Section 2.4 was not implemented as a standalone module of the graph neural network within the presented experiments and, accordingly, did not affect the quantitative results given in this section. A detailed description of deduplication and semantic overlap control is provided in Section 3.1.

All experimental results presented in this section are derived from text data and record-level structured indicators, as described in Section 2. All presented quality indicators reflect the statistical accuracy of classification within the datasets used. These metrics should not be interpreted as indicators of clinical validity or diagnostic reliability. The accuracy and calibration values characterize the behavior of the model exclusively in controlled experimental conditions.

To reduce the risks of data leakage and an overly optimistic assessment, we applied a two-step protection procedure. First, exact duplicates and close duplicates were deleted before splitting, and it was also verified that very similar texts did not overlap between the training and test samples (see Section 3.1). Second, if author/user IDs were provided as part of the dataset, we additionally ensured separation at the user level to prevent samples from the same user from appearing in both the training and test samples. If such identifiers

were not available, strict separation at the user level could not be guaranteed; therefore, the main safeguards included deduplication checks and semantic filtering.

3.1. Experimental Setup and Evaluation Protocol

The experimental part was aimed at evaluating the effectiveness of the proposed multimodal architecture in comparison with basic models. Training and testing were conducted on the two datasets, described in the previous section, stratified into training, validation, and test sets. Several basic machine learning algorithms were selected to ensure the comparability of the results, including logistic regression, support vector machine, and random forest. These models are widely used in psycho-emotional state analysis tasks and are used as standard benchmarks [6,24].

Quality assessment was conducted using the following metrics: accuracy, precision, recall, F1-score, and ROC-AUC. Additionally, probabilistic calibration was analyzed using the Brier score, which is especially crucial in risk assessment tasks. In addition, a reliability diagram was used to evaluate the calibration of probabilistic predictions, allowing the predicted probabilities to be compared with the actual frequency of events. The expected calibration error (ECE) metric was also determined, which reflects the mean deviation between the predicted and empirical probabilities for binning intervals (0, 1). The calibration curves were constructed before and after temperature scaling, allowing an assessment of the impact of calibration on the probability distributions. The ECE was based on a test sample.

All models were implemented in Python using the PyTorch library for the proposed architecture and standard machine learning libraries for the basic models.

Experiments were conducted using the NVIDIA graphics processor with 12 GB of video memory and 32 GB of RAM. The proposed architecture was implemented in PyTorch (version 2.x) using the Hugging Face Transformers library (version 4.x). The BERT-base-uncased model was used as a text encoder. The training was performed with a pad size of 16, a learning speed of 2×10^{-5} , the AdamW optimizer, and a dropout rate of 0.1. To prevent overfitting, an early stopping procedure was used, with the “patience” parameter equal to three epochs.

To ensure the reproducibility of the experiments, fixed initialization of the random seed generator was used, and splitting and cross-validation procedures were performed according to the same parameters for all models being compared.

We also sought to prevent data leaks and monitor semantic overlap. To reduce the risk of data leakage, we implemented a two-step protection procedure. First, exact duplicates were removed before any division of the data into training and test sets, as well as before the cross-validation procedure. Next, close text duplicates were identified using TF-IDF representations and cosine similarity; records with a high degree of similarity above a specified threshold were either removed or assigned to the same sample prior to model evaluation. User-level disaggregation was applied only in cases where user or author identifiers were available in the original dataset. For datasets where such identifiers were not provided, strict user-level disaggregation could not be guaranteed. In these cases, we do not make claims about the model’s generalizability at the user level, and we relied on duplicate removal, the filtering of close duplicates, stratified data partitioning, and checking for semantic overlap between folds as practical safeguards. Therefore, the presented results should be interpreted as reflecting the prediction quality at the level of individual records under controlled preprocessing, rather than the confirmed generalizability of the model at the user level.

3.2. Model Performance Comparison

The results of the comparative analysis are presented in Table 12.

Table 12. Cross-validation performance of the tested text-plus-structured configuration.

Metric	Mean $\pm$ SD	95% CI
Accuracy	0.593 $\pm$ 0.009	0.582–0.604
F1-score	0.592 $\pm$ 0.008	0.582–0.602
ROC-AUC	0.860 $\pm$ 0.006	0.853–0.867
Brier score	0.540 $\pm$ 0.009	0.529–0.551
ECE	0.064 $\pm$ 0.007	0.055–0.073

The results in Table 12 are presented as the mean $\pm$ standard deviation based on five-fold cross-validation. The 95% confidence intervals were calculated using the t-distribution for five iterations. These results should be interpreted as record-level prediction metrics for the tested “text plus structured text” configuration. They do not constitute evidence of clinical validity or generalizability at the user level.

Although the proposed model achieved the highest accuracy and ROC-AUC values among the compared models, these metrics alone are insufficient for evaluating a system designed for risk analysis. Therefore, probabilistic calibration was analyzed separately using the Brier score, expected calibration error (ECE), and reliability curves. Calibration is particularly important in this task because a model that is too confident may assign a high risk level even when the input signal is uncertain. Therefore, the interpretation of the model should be based not only on discriminative metrics such as accuracy and ROC-AUC but also on the reliability of the predicted probabilities (Table 13).

Table 13. Encoder comparison with confidence intervals.

Text Encoder	Accuracy, 95% CI	F1-Score, 95% CI	ROC-AUC, 95% CI	Brier Score, 95% CI	ECE, 95% CI
BERT-base-uncased	0.593 $\pm$ 0.009, 0.582–0.604	0.592 $\pm$ 0.008, 0.582–0.602	0.860 $\pm$ 0.006, 0.853–0.867	0.540 $\pm$ 0.009, 0.529–0.551	0.064 $\pm$ 0.007, 0.055–0.073
RoBERTa-base	0.666 $\pm$ 0.005, 0.660–0.672	0.667 $\pm$ 0.007, 0.658–0.676	0.896 $\pm$ 0.004, 0.891–0.901	0.459 $\pm$ 0.005, 0.453–0.465	0.040 $\pm$ 0.011, 0.026–0.054

The final quality depends heavily on the chosen encoder and data configuration. The BERT configuration yielded stable but moderate results, while RoBERTa-base improved both the classification metrics and calibration using the same protocol.

The results of the ablation analysis and encoder comparison show that the achieved quality does not depend on any single architectural component. Part of the result is attributed to the chosen transformer-based text encoder, part to data preprocessing and duplicate control, and another part to the integration of structured information at the record level and probabilistic calibration. Therefore, the results should not be interpreted as proof that cross-modal attention alone explains the improvement in quality. It is more accurate to view the proposed pipeline as a combined computational configuration, where the overall performance depends on the choice of encoder, data properties, preprocessing, fusion, and calibration [41].

To test the statistical significance of the differences between models, a paired Student’s *t*-test was applied, which made it possible to assess the probability of the random nature of the differences found. The obtained accuracy, ROC-AUC, Brier score, and ECE values are interpreted within controlled experimental conditions and should not be considered

evidence of clinical effectiveness. The results presented reflect the model's resilience in the datasets under consideration and are not an assessment of its clinical effectiveness.

The results show that the tested pipeline delivers moderate but stable classification performance at the record level under a controlled five-fold cross-validation protocol. A comparison of the encoders shows that the final quality depends heavily on the choice of transformer encoder: RoBERTa-base improved both the discrimination metrics and calibration-related metrics compared to BERT-base-uncased under the same evaluation conditions. Consequently, the observed performance should not be attributed solely to cross-modal attention. It reflects the combined influence of the text encoder, preprocessing, duplicate control procedures, structured record-level indicators, merging strategies, and probabilistic calibration [42].

3.3. ROC Curve Analysis

ROC curve analysis was used to evaluate the ability of the tested "text + structured features" configuration to distinguish between classes. The obtained ROC-AUC values should be interpreted in conjunction with the calibration metrics and confidence intervals, rather than as standalone evidence of the model's high quality. This is important because the model may be able to distinguish between classes quite well but still provide insufficiently reliable probability estimates.

Figure 6 presents an analysis of the ROC curves for the tested configuration and the models under comparison. The dashed diagonal line in Figure 6 indicates the no-discrimination baseline corresponding to random classification (ROC-AUC = 0.50). These results characterize the classification performance only under controlled experimental conditions and should not be interpreted as clinical validation or evidence of practical diagnostic reliability. In this study, the ROC-AUC is used only as one of several evaluation metrics, alongside the F1-score, Brier score, ECE, and confidence intervals. The ROC analysis was based only on the empirical inputs described in Section 2.3. Therefore, the results should be interpreted as an evaluation of the tested computational configuration at the record level, based on text data and structured indicators at the record level, rather than as validation of a fully developed multimodal clinical monitoring system.

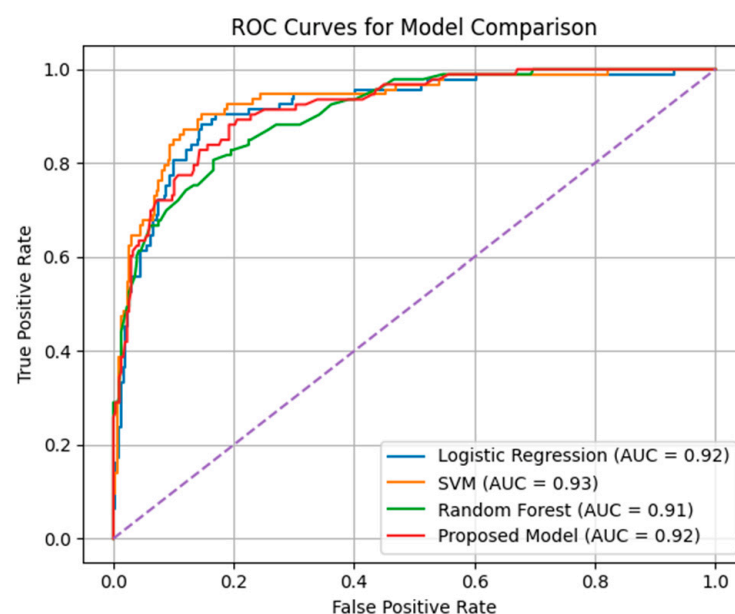


Figure 6. ROC curves for the compared models and the proposed architecture. The dashed diagonal line represents the no-discrimination reference, corresponding to the random-classifier baseline (ROC-AUC = 0.50).

Figure 6 shows the ROC curve performance of the tested models under the same evaluation protocol. The ROC-AUC values reflect the models' ability to distinguish between classes in the available dataset, but they should not be interpreted independently of the calibration metrics and confidence intervals. In this study, the ROC analysis was based solely on text data and structured indicators at the record level; physiological signals, wearable sensors, and visual modules were not included in the calculation.

3.4. Ablation Study

In addition to the main cross-validation and encoder comparison results, an additional text-based baseline analysis was performed to provide a reference point for the behavior of conventional and transformer-based text classifiers. This analysis should not be interpreted as the ablation of a fully multimodal fusion model, because the available datasets did not provide independently synchronized multimodal observations. The analysis used only the empirical inputs described in Section 2.3.

The results are presented in Table 14. They compare two text-based configurations: XGBoost, trained on TF-IDF features, and DistilBERT, used as a transformer-based text classifier. These results are included in the manuscript because they provide useful information about the predictive signal that can be extracted from textual data alone. At the same time, they are not evidence that a fully multimodal architecture was empirically validated.

Table 14. Additional text-based baseline results (5-fold cross-validation, mean $\pm$ SD).

Configuration	Accuracy (Mean $\pm$ Std)	F1-Macro (Mean $\pm$ Std)	ROC-AUC OVR (Mean $\pm$ Std)	Brier (Mean $\pm$ Std)
XGBoost (TF-IDF, text baseline)	0.7947 $\pm$ 0.0166	0.7972 $\pm$ 0.0162	0.9493 $\pm$ 0.0057	0.0598 $\pm$ 0.0033
DistilBERT (text baseline)	0.4305 $\pm$ 0.0155	0.4059 $\pm$ 0.0278	0.7479 $\pm$ 0.0109	0.1489 $\pm$ 0.0008

The results in Table 14 show that text-only baselines can achieve competitive performance under the available experimental conditions. In particular, the TF-IDF + XGBoost configuration performed strongly in the considered text classification setting. This finding is important because it indicates that a substantial part of the predictive signal in the available datasets is contained in textual patterns and the dataset-specific linguistic structure.

Therefore, the results should be interpreted cautiously. They do not demonstrate that the performance gains are caused by multimodal fusion, physiological information, wearable-sensor data, or behavioral streams. Instead, they show that the textual component itself has strong predictive value and should be considered a major contributor to the overall performance of the computational pipeline. For this reason, regarding the main conclusions of the study, we avoid attributing the reported performance to multimodal fusion alone [43,44].

3.5. Confusion Matrix Analysis

The confusion matrix, shown in Figure 7, demonstrates that the proposed model correctly classifies many instances. Errors are mostly concentrated near class boundaries, which reflects the complex and continuous nature of psycho-emotional states.

The confusion matrix analysis shows that most of the examples are correctly classified; at the same time, the number of errors is distributed relatively equally between classes. This is particularly important under conditions of imbalanced data, characteristic of psycho-emotional analysis tasks [45].

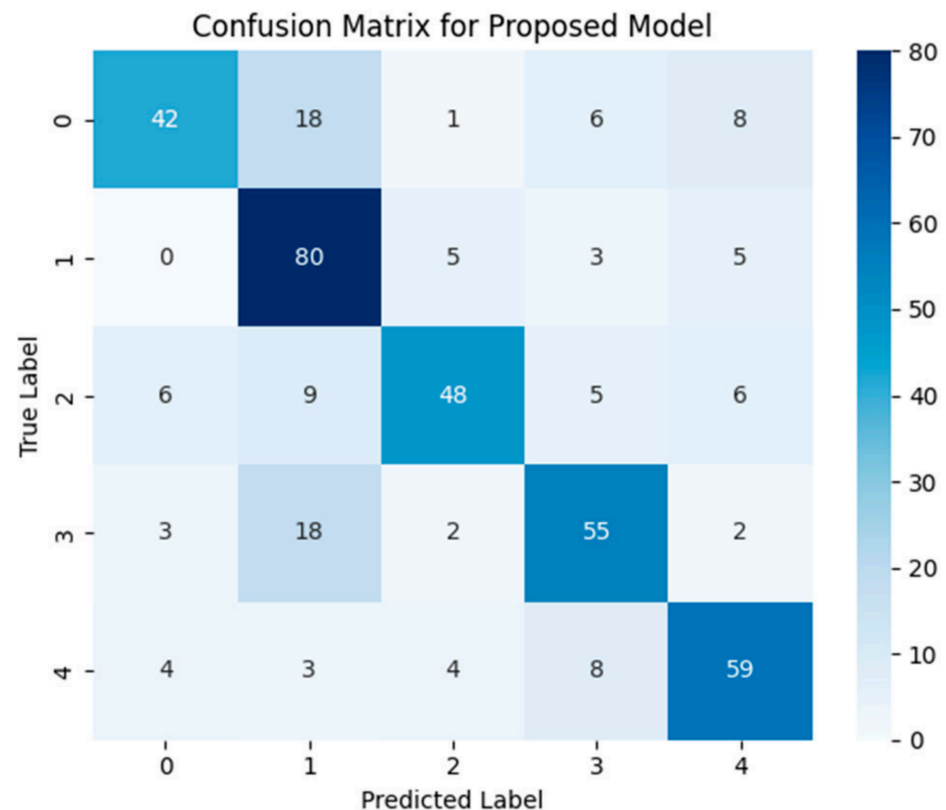


Figure 7. Confusion matrix for proposed model.

Nevertheless, it can be noted that there is a small number of errors near the class boundaries, which reflects the complex nature of psycho-emotional states and their smooth transitions between categories [40].

For the sustainability assessment of the obtained results, additionally, a variability metric analysis was carried out during repeated runs of the models. In particular, the training was carried out several times with different data sampling strategies while maintaining stable behavior among the main quality indicators [46].

The differences between the basic models and proposed architecture remained in all experiments, which indicates the stability of the obtained results and decreases the probability of random nature improvements. This approach corresponds to model evaluation practice in machine learning tasks and allows a more accurate interpretation of the achieved results.

3.6. Explainability Analysis

Additionally, an interpretability assessment of the model was performed using the SHAP method, allowing a quantitative evaluation of the contributions of individual features to the final prediction. In contrast to traditional models, where the decision-making process remains hidden, this approach allows an understanding of which factors have the greatest impact on the results.

As shown in Figure 8, a limited number of the most significant features exhibit the greatest contributions to the predictions. The SHAP analysis shows that model is based on key components of the feature area, which allows a reduction in the impact of noise and an increase in prediction stability.

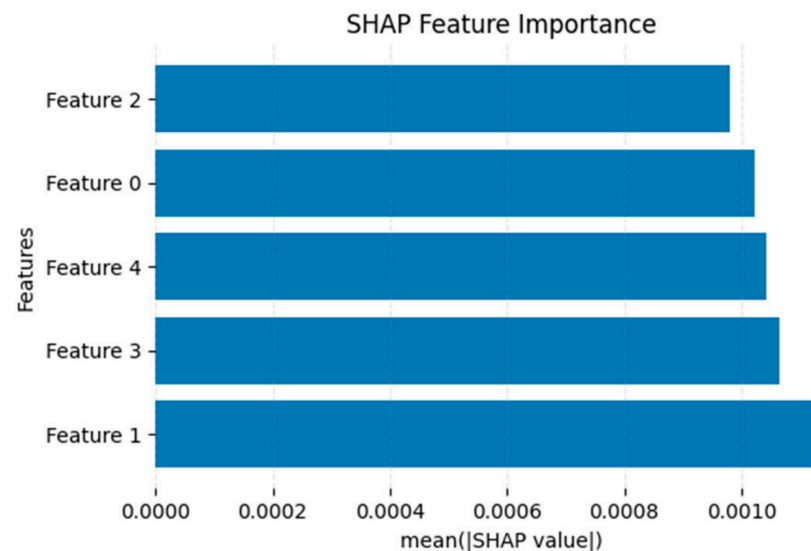


Figure 8. SHAP-based feature importance with modality-aware labeling (text/structured), showing the relative contributions of interpretable components to psycho-emotional risk prediction.

In the updated SHAP analysis, interpretable feature names were used with an indication of their modality. Features were grouped by data source (text/structured), which allowed the separation of the contributions of text embeddings and record-level structured indicators.

For the text-based modality, SHAP values were calculated from the aggregated embedding components formed by the BERT-base-uncased model. Stable semantic components were used in the visualization, not abstract dimension indices (for example, not “embedding_234” but “semantic context intensity” or similar interpretable designations).

Structured features are presented using their original feature names (such as crisis indicators), which allows the model’s contribution to be correlated with the user’s state.

It is important to emphasize that the SHAP values reflect the contributions of features to changes in the prediction of probabilities and are not interpreted as causal dependencies. They show contributions within a specific trained model. To avoid general word-level interpretation, two illustrative examples are provided. In some cases, the greatest contribution to the prediction came from lexical markers of emotional tension—persistent words and constructions associated with feeling overwhelmed, sleep disturbances, internal tension, and self-esteem. Additionally, record-level structured indicators (category/level of condition) made a notable contribution, which consistently “shifted” the probability toward the corresponding class. However, it is important to emphasize that SHAP reflects the contributions of features to model decisions within a given dataset and does not equate to the clinical interpretation of symptoms. This analysis provides a more transparent understanding of the behavior of the model and makes the explainability analysis consistent with the objectives of assessing psycho-emotional risks [47,48].

However, their combined effect reflects the contribution of different data aspects and confirms the ability of the model to consider the complex structure of psycho-emotional states [19,21]. In practice, using SHAP allows not only an increase in confidence in the model but also a better understanding of its behavior in different conditions.

The SHAP values represent the marginal contributions of each feature to the predicted probability under the cooperative game theory framework. In the context of psycho-emotional risk prediction, positive SHAP values increase the predicted risk, while negative values decrease it. Importantly, SHAP analysis reflects model behavior rather than causal relationships.

3.7. Calibration Analysis

A calibration analysis was performed to assess whether the predicted probabilities were consistent with the observed class frequencies. Temperature scaling was applied after training the base classifier and did not change the learned model weights or the classification boundary.

The results show that calibration mainly improved the reliability of the probability estimates rather than the discrimination metrics. In the BERT-based configuration, the expected calibration error decreased from 0.080 ± 0.008 before calibration to 0.064 ± 0.007 after temperature scaling. The Brier score also improved slightly, from 0.544 ± 0.010 to 0.540 ± 0.009 . The accuracy, F1-score, and ROC-AUC remained almost unchanged, which was expected because temperature scaling modifies the probability distribution but does not retrain the classifier.

These results indicate that calibration should be interpreted as a reliability improvement step, not as a direct source of classification accuracy improvement.

4. Discussion

The obtained results should be interpreted within the limited experimental configuration used in this study. The reported experiments primarily evaluated a text-based computational pipeline with transformer-based representations, probabilistic calibration, and explainability. They do not demonstrate a fully synchronized multimodal system combining text, physiological signals, wearable-sensor measurements, visual information, or continuous behavioral streams.

The results suggest that the quality of classification depends on several factors, including the selected text encoder, preprocessing, duplicate control procedures, dataset properties, and calibration. Therefore, the observed performance should not be attributed to multimodal fusion alone. In particular, the encoder comparison showed that RoBERTa-base improved the performance compared with BERT-base under the same evaluation protocol, indicating that textual representation quality is one of the main contributors to the final result.

When applied to clinically collected datasets, the performance of the model may differ substantially from the values reported in this study. Public datasets often contain labels derived from platform-specific categories, weak supervision, self-description, or dataset-specific annotation rules, whereas clinically collected datasets require standardized assessment procedures, expert annotation, and prospective validation. Therefore, the reported performance should be interpreted as computational classification performance under public-dataset conditions, not as evidence that the model would achieve the same reliability in clinical or institutional settings. Future studies should evaluate the framework on clinically annotated, prospectively collected datasets with transparent labeling protocols and, where possible, synchronized multimodal measurements.

The interpretability analysis shows that the model does not rely on individual features but uses their aggregation. This indirectly confirms that the architecture is not being retrained to individual patterns but forms more generalized representations of the user's positions. In tasks related to psycho-emotional assessment, this is very important, because overdependence on individual features can lead to instability and difficulty in interpreting the results [19,21].

It is also worth noting the use of probabilistic calibration. Unlike many works that are limited to classification metrics, in this case, the model is additionally evaluated in terms of probabilistic assessment correctness. This allows us to interpret the results not only in terms of a "class" but also the level of confidence, which makes the system more suitable for practical use. The ablation study provides a more cautious interpretation of the model

architecture. The results do not show a consistent advantage of attention-based integration over simple concatenation. In fact, the concatenation-based variant produced comparable and, in some metrics, slightly higher values. This suggests that, under the current, primarily text-based data configuration, the attention block had limited opportunities to learn complementary interactions between independent synchronized modalities. Therefore, the observed performance should not be attributed to cross-modal attention alone. Instead, it reflects the combined influence of text encoder selection, preprocessing, duplicate control, available within-record representations, and probability calibration. The contribution of this study should therefore be understood as methodological integration rather than the introduction of a fundamentally new learning algorithm. Transformer-based encoders, calibration procedures, and explainability methods are established techniques. The novelty of the present work lies in combining these components within a single transparent evaluation pipeline for psycho-emotional risk classification, together with the explicit reporting of calibration metrics, confidence intervals, the limitations of public labels, and reproducibility parameters. The system operates within a data-based pattern recognition framework and does not replace professional psychological or psychiatric assessment.

Limitations

Despite the obtained results, several limitations should be noted, which are important to take into account when interpreting the proposed approach.

Firstly, although the datasets include different data types, they remain limited in volume and do not completely represent the variety of real-world scenarios. In particular, social media data and experimental observations may differ in structure and noise levels, which affects the model's portability.

Secondly, a few indicators are represented as embedding vectors, which makes their direct interpretation difficult. Despite using explainable artificial intelligence methods, the interpretation of such features remains indirect and requires further analysis.

Thirdly, the model was evaluated in an offline setting. In real conditions, where data enter as a stream, system behavior may differ, especially in light of temporal delays and data incompleteness. An additional aspect is possible classification errors. In tasks related to the assessment of psycho-emotional risk, false-positive and false-negative outcomes can have different practical consequences. False-positive predictions can lead to excessive notifications or unnecessary interventions, while false-negative cases mean that potentially vulnerable conditions may go unnoticed. The metrics presented in this work reflect the statistical quality of classification within the framework of the data used and do not allow a direct assessment of the possible practical consequences of such errors under real-world conditions of application.

Finally, in this work, issues of model integration into real information systems, including privacy, safety, and computing efficiency aspects, are not considered—these can also affect practical use.

Ethical and regulatory considerations must also be taken into account. The use of digital data reflecting psycho-emotional states requires careful handling in terms of privacy, the protection of personal data, and the responsible interpretation of the results. The proposed system is not a substitute for professional psychological assessment and should not be considered a diagnostic tool. Any practical application of such a solution requires compliance with existing legal norms and the participation of qualified specialists.

The main limitation of this study is related not only to the architecture but also to the measurement characteristics. The labels used in publicly available datasets may not reflect clinically validated psychological constructs. Consequently, even acceptable classification performance should not be interpreted as evidence that the model accurately measures

psychological distress, crisis risk, or mental health status in a clinical sense. The model may partially capture linguistic, stylistic, or dataset-specific patterns associated with the provided labels. This limitation directly affects the interpretation of all presented metrics.

This issue is particularly important because the available labels were used as data-driven target categories rather than clinically validated outcomes. Therefore, the proposed approach should be interpreted as a computational image recognition pipeline within a given labeling scheme, rather than as a diagnostic or psychometric assessment tool. In practice, false-positive and false-negative predictions can have different consequences, and automated risk categorization should not be used without expert oversight, external validation, and clear ethical safeguards.

Therefore, future studies should prioritize clinically annotated and prospectively collected datasets rather than simply increasing the complexity of the architecture. In particular, future studies should utilize datasets with transparent labeling procedures, expert annotation, longitudinal follow-up, and, where possible, synchronized multimodal measurements. Only under such conditions can the clinical or practical reliability of automated psycho-emotional risk prediction be more rigorously assessed.

It is important to emphasize that the obtained results refer to controlled experimental settings and do not reflect external or prospective validation. The work was not tested in real flow conditions and there was no clinical validation, so any conclusions regarding practical implementation should be considered as preliminary. The next step is external validation for independent data and/or longitudinal observations.

5. Conclusions

This study proposed and experimentally evaluated an explainable computational pipeline for classifying psycho-emotional risk based on text records and structured indicators at the individual record level. The tested configuration combined transformer-based text representation, structured information when available, probabilistic calibration, and interpretability components within a single controlled experimental protocol. The results show that such a configuration can ensure stable classification quality at the record level under the described preprocessing, cross-validation, and calibration procedures. However, these results should be interpreted with consideration of the limitations of the open datasets used and the validity of their annotations.

This study does not constitute clinical validation and does not demonstrate a fully integrated multimodal system incorporating physiological, wearable-sensor, visual, or real-time behavioral data. These components remain conceptual extensions of a scalable architecture and were not utilized in the presented experiments. The main contribution of this study is not the development of a fundamentally new algorithm but the transparent integration and controlled evaluation of existing computational components, including transformer-based textual representation, probability calibration, explainability analysis, and reproducible evaluation procedures, for psycho-emotional risk classification.

The results of the ablation analysis and encoder comparisons also show that the final quality of the model should not be attributed to a single architectural component. The obtained metrics depend on several factors: the chosen text encoder, the preprocessing and duplicate control procedures, the properties of the data themselves, structured indicators at the record level, the fusion strategy, and probabilistic calibration. Therefore, it is more accurate to view the proposed approach as a combined computational pipeline rather than as proof that cross-modal attention alone determines the model's quality.

Future studies should prioritize clinically annotated, prospectively collected, and synchronized multimodal datasets before drawing stronger conclusions about the practical implementation or clinical significance of this approach. Further work should also

include external validation, transparent labeling procedures, expert annotation, longitudinal follow-up, and model evaluation in real-world data settings. Only after such validation can the practical reliability of automated psycho-emotional risk prediction be more rigorously assessed.

Author Contributions: Conceptualization, O.B. and K.S.; methodology, D.Y. and A.O.; software, V.S. and A.A.; validation, N.N., M.S. and A.O.; formal analysis, O.B., V.S. and K.S.; investigation, D.Y., A.A. and N.N.; resources, M.S. and K.S.; data curation, D.Y.; writing—original draft preparation, O.B. and A.O.; writing—review and editing, N.N., M.S. and K.S.; visualization, V.S., A.A. and A.O.; supervision, K.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The source code of the model implementation will be made publicly available after the review process is completed and the manuscript is accepted for publication. The repository will contain preprocessing scripts, the implementation of the architecture, the training configuration, statistical validation procedures, and instructions for replicating the experiments.

Acknowledgments: The authors thank the research community for providing the publicly available datasets used in this study. Declaration of Generative AI and AI-Assisted Technologies in the Writing Process: During the preparation of this manuscript, the authors used AI-assisted tools to support language editing, formatting, and the preparation of illustrative conceptual figures. All scientific content, experimental design, data preprocessing, model training, statistical analysis, interpretation of results, and final conclusions were reviewed, verified, and approved by the authors. AI-assisted tools were not used to generate experimental results, modify the underlying datasets, or replace author responsibility for the accuracy and integrity of the manuscript. Quantitative figures were generated from experimental outputs and manually checked by the authors.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
AUC	Area Under the Curve
BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	Bidirectional Long Short-Term Memory
CPU	Central Processing Unit
EDA	Electrodermal Activity
F1-score	Harmonic Mean of Precision and Recall
GNN	Graph Neural Network
GPU	Graphics Processing Unit
HRV	Heart Rate Variability
IoT	Internet of Things
LR	Logistic Regression
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MNB	Multinomial Naïve Bayes
NLP	Natural Language Processing
PR	Precision–Recall

PyTorch	Python-Based Deep Learning Framework
RF	Random Forest
RMSE	Root Mean Square Error
ROC	Receiver Operating Characteristic
SHAP	SHapley Additive exPlanations
SVM	Support Vector Machine
T	Temperature Parameter (in calibration)
XAI	Explainable Artificial Intelligence
ML	Machine Learning

References

- ShamsEldin, T.; Gaber, S.; Ansari, S.; Elgohary, R.; Shawky, M.A.; Elbahnasawy, M.; Abdrabou, M. Artificial intelligence for predicting depression, anxiety and stress using psychometric data. *Sci. Rep.* **2025**, *15*, 37282. [\[CrossRef\]](#) [\[PubMed\]](#)
- Hasan, M.; Alam, M.M.; Hossain, S. Prevalence, associated factors, and machine learning-based prediction of depression, anxiety, and stress among university students: A cross-sectional study from Bangladesh. *J. Health Popul. Nutr.* **2025**, *44*, 361. [\[CrossRef\]](#) [\[PubMed\]](#)
- Taskynbayeva, M.; Gutoreva, A. Machine learning approaches to anxiety detection: Trends, model evaluation, and future directions. *Front. Artif. Intell.* **2025**, *8*, 1630047. [\[CrossRef\]](#) [\[PubMed\]](#)
- Razavi, M.; Ziyadidegan, S.; Mahmoudzadeh, A.; Kazeminasab, S.; Baharlouei, E.; Janfaza, V.; Jahromi, R.; Sasangohar, F. Machine Learning, Deep Learning, and Data Preprocessing Techniques for Detecting, Predicting, and Monitoring Stress and Stress-Related Mental Disorders: Scoping Review. *JMIR Ment. Health* **2024**, *11*, e53714. [\[CrossRef\]](#) [\[PubMed\]](#)
- Rois, R.; Ray, M.; Rahman, A.; Roy, S.K. Prevalence and predicting factors of perceived stress among Bangladeshi university students using machine learning algorithms. *J. Health Popul. Nutr.* **2021**, *40*, 50. [\[CrossRef\]](#) [\[PubMed\]](#)
- Qasrawi, R.; Vicuna Polo, S.P.; Abu Al-Halawa, D.; Hallaq, S.; Abdeen, Z. Assessment and Prediction of Depression and Anxiety Risk Factors in Schoolchildren: Machine Learning Techniques Performance Analysis. *JMIR Form. Res.* **2022**, *6*, e32736. [\[CrossRef\]](#) [\[PubMed\]](#)
- Kargarandehkordi, A.; Li, S.; Lin, K.; Phillips, K.T.; Benzo, R.M.; Washington, P. Fusing Wearable Biosensors with Artificial Intelligence for Mental Health Monitoring: A Systematic Review. *Biosensors* **2025**, *15*, 202. [\[CrossRef\]](#) [\[PubMed\]](#)
- Jin, K.; Rubio-Solis, A.; Naik, R.; Leff, D.; Kinross, J.; Mylonas, G. Human-Centric Cognitive State Recognition Using Physiological Signals: A Systematic Review of Machine Learning Strategies Across Application Domains. *Sensors* **2025**, *25*, 4207. [\[CrossRef\]](#) [\[PubMed\]](#)
- Ullah, A.; Zhang, F.; Song, Z.; Wang, Y.; Zhao, S.; Riaz, W.; Li, G. Surface Electromyography-Based Recognition of Electronic Taste Sensations. *Biosensors* **2024**, *14*, 396. [\[CrossRef\]](#) [\[PubMed\]](#)
- Riaz, W.; Ji, J.; Ullah, A. TriViT-Lite: A Compact Vision Transformer–MobileNet Model with Texture-Aware Attention for Real-Time Facial Emotion Recognition in Healthcare. *Electronics* **2025**, *14*, 3256. [\[CrossRef\]](#)
- Riaz, W.; Ji, J.; Zaman, K.; Zengkang, G. Neural Network-Based Emotion Classification in Medical Robotics: Anticipating Enhanced Human–Robot Interaction in Healthcare. *Electronics* **2025**, *14*, 1320. [\[CrossRef\]](#)
- Riaz, W.; Ullah, A.; Ji, J. Multi-Scale Attention Networks with Feature Refinement for Medical Item Classification in Intelligent Healthcare Systems. *Sensors* **2025**, *25*, 5305. [\[CrossRef\]](#) [\[PubMed\]](#)
- Khoo, L.S.; Lim, M.K.; Chong, C.Y.; McNaney, R. Machine Learning for Multimodal Mental Health Detection: A Systematic Review of Passive Sensing Approaches. *Sensors* **2024**, *24*, 348. [\[CrossRef\]](#) [\[PubMed\]](#)
- Pinge, A.; Gad, V.; Jaisighani, D.; Ghosh, S.; Sen, S. Detection and monitoring of stress using wearables: A systematic review. *Front. Comput. Sci.* **2024**, *6*, 1478851. [\[CrossRef\]](#)
- Wang, C.-M.; Lee, Y.-C. Affective Computing Using an Emotion-Perception System for Interactive Experiences with Multi-Sensing Interfaces. *Preprints* **2024**, 2024082106. [\[CrossRef\]](#)
- Yu, M.; Xu, H.; Zhou, F.; Xu, S.; Yin, H. A Deep-Learning-Based Multimodal Data Fusion Framework for Urban Region Function Recognition. *ISPRS Int. J. Geo-Inf.* **2023**, *12*, 468. [\[CrossRef\]](#)
- Mamieva, D.; Abdusalomov, A.B.; Kutlimuratov, A.; Muminov, B.; Whangbo, T.K. Multimodal Emotion Detection via Attention-Based Fusion of Extracted Facial and Speech Features. *Sensors* **2023**, *23*, 5475. [\[CrossRef\]](#) [\[PubMed\]](#)
- Mimikou, C.; Kokkoti, C.; Tsiptsios, D.; Tsamakis, K.; Savvidou, S.; Modig, L.; Christidi, F.; Kaltsatou, A.; Doskas, T.; Mueller, C.; et al. Explainable Machine Learning in the Prediction of Depression. *Diagnostics* **2025**, *15*, 1412. [\[CrossRef\]](#) [\[PubMed\]](#)
- Haque, R.; Islam, N.; Islam, M.; Ahsan, M.M. A Comparative Analysis on Suicidal Ideation Detection Using NLP, Machine, and Deep Learning. *Technologies* **2022**, *10*, 57. [\[CrossRef\]](#)

20. Alharthi, A.-S.; Alqurashi, A.; Alharbi, T.-E.; Alammam, M.-M.; Aldosari, N.; Boucekara, H.-R.E.H.; Sha'aban, Y.-A.; Shahriar, M.S.; Ayidh, A.A. Explainable AI for Sensor Signal Interpretation to Revolutionize Human Health Monitoring: A Review. *IEEE Access* **2025**, *13*, 115990–116024. [\[CrossRef\]](#)
21. Ostojic, D.; Lalousis, P.A.; Donohoe, G.; Morris, D.W. The challenges of using machine learning models in psychiatric research and clinical practice. *Eur. Neuropsychopharmacol.* **2024**, *88*, 53–65. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Tong, L.; Zhang, Z. Neural Network Analysis of Psychological Data: A Step-by-Step Guide. *Multivar. Behav. Res.* **2025**, *61*, 399–419. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Tathgir, A.; Sharma, C.M.; Chariar, V.-M. EEG-Based Emotion Classification using Deep Learning: Approaches. *Trends Bibliometr.* **2024**, 1–26. [\[CrossRef\]](#)
24. Zhai, Y.; Zhang, Y.; Chu, Z.; Geng, B.; Almaawali, M.; Fulmer, R.; Lin, W.Y.D.; Xu, Z.; Daniels, A.D.; Liu, Y.; et al. Machine learning predictive models to guide prevention and intervention allocation for anxiety and depressive disorders among college students. *J. Couns. Dev.* **2025**, *103*, 110–125. [\[CrossRef\]](#)
25. Schlicher, M.; Li, Y.; Murthy, S.M.K.; Sun, Q.; Schuller, B.W. Emotionally adaptive support: A narrative review of affective computing for mental health. *Front. Digit. Health* **2025**, *7*, 1657031. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Farsadaki, V.; Griy-Brown, C. AI affective computing and behavioral health. *Front. Comput. Sci.* **2026**, *7*, 1692728. [\[CrossRef\]](#)
27. Liu, Y.; Palacio, M.-I.; Bikki, T.; Toledo, C.; Ouyang, Y.; Li, Z.; Wang, Z.; Toledo, F.; Zeng, H.; Herrero, M.-T. Machine Learning, Physiological Signals, and Emotional Stress/Anxiety: Pitfalls and Challenges. *Appl. Sci.* **2025**, *15*, 11777. [\[CrossRef\]](#)
28. Han, E.-G.; Kang, T.-K.; Lim, M.-T. Physiological Signal-Based Real-Time Emotion Recognition Based on Exploiting Mutual Information with Physiologically Common Features. *Electronics* **2023**, *12*, 2933. [\[CrossRef\]](#)
29. Mosolova, E.S.; Alfimov, A.E.; Kostyukova, E.G.; Mosolov, S.N. Prospects of machine learning applications in affective disorders. *Digit. Diagn.* **2025**, *6*, 97–115. [\[CrossRef\]](#)
30. Sansyzbay, K.M.; Bakhtiyarova, Y.A.; Iliev, T.; Patokin, G.S.; Tasbolatova, L.T.; Sagmedinov, D.B. Development of an Algorithm for a National Microprocessor-Based Centralization System with a Modular Architecture KZ-MPC-MA Featuring Advanced Intelligent Control Functions. *IEEE Access* **2024**, *12*, 193229–193240. [\[CrossRef\]](#)
31. Babenko, T.; Kolesnikova, K.; Bakhtiyarova, Y.; Yeskendirova, D.; Sansyzbay, K.; Sysoyev, A.; Kruchinin, O. Hybrid GNN-LSTM Architecture for Probabilistic IoT Botnet Detection with Calibrated Risk Assessment. *Computers* **2026**, *15*, 26. [\[CrossRef\]](#)
32. Islam, M.M.; Hassan, S.; Akter, S.; Jibon, F.A.; Sahidullah, M. A comprehensive review of predictive analytics models for mental illness using machine learning algorithms. *Healthc. Anal.* **2024**, *6*, 100350. [\[CrossRef\]](#)
33. Singh, G.; Phukan, O.C.; Kumar, R. Stress recognition with multi-modal sensing using bootstrapped ensemble deep learning model. *Expert. Syst.* **2023**, *40*, e13239. [\[CrossRef\]](#)
34. Uria-Rivas, R.; Rodriguez-Sanchez, M.C.; Santos, O.C.; Vaquero, J.; Boticario, J.G. Impact of Physiological Signals Acquisition in the Emotional Support Provided in Learning Scenarios. *Sensors* **2019**, *19*, 4520. [\[CrossRef\]](#) [\[PubMed\]](#)
35. Palko, D.; Babenko, T.; Bigdan, A.; Kiktev, N.; Hutsol, T.; Kuboń, M.; Hnatienko, H.; Tabor, S.; Gorbovy, O.; Borusiewicz, A. Cyber Security Risk Modeling in Distributed Information Systems. *Appl. Sci.* **2023**, *13*, 2393. [\[CrossRef\]](#)
36. Swain, V.D.; Gao, L.; Mondal, A.; Abowd, G.D.; De Choudhury, M. Sensible and Sensitive AI for Worker Wellbeing: Factors that Inform Adoption and Resistance for Information Workers. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24), Honolulu, HI, USA, 11–16 May 2024*; ACM: New York, NY, USA, 2024; 30p. [\[CrossRef\]](#)
37. Grawemeyer, B.; Mavrikis, M.; Holmes, W.; Gutiérrez-Santos, S.; Wiedmann, M.; Rumme, N. Affective learning: Improving engagement and enhancing learning with affect-aware feedback. *User Model. User-Adapt. Interact.* **2017**, *27*, 119–158. [\[CrossRef\]](#)
38. Lieskovská, E.; Jakubec, M.; Jarina, R.; Chmulk, M. A Review on Speech Emotion Recognition Using Deep Learning and Attention Mechanism. *Electronics* **2021**, *10*, 1163. [\[CrossRef\]](#)
39. Poria, S.; Cambria, E.; Winterstein, G.; Huang, G.-B. Sentic patterns: Dependency-based rules for concept-level sentiment analysis. *Knowl.-Based Syst.* **2014**, *69*, 45–63. [\[CrossRef\]](#)
40. Gunes, H.; Schuller, B. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image Vis. Comput.* **2013**, *31*, 120–136. [\[CrossRef\]](#)
41. Tzirakis, P.; Trigeorgis, G.; Nicolaou, M.A.; Schuller, B.W.; Zafeiriou, S. End-to-End Multimodal Emotion Recognition Using Deep Neural Networks. *IEEE J. Sel. Top. Signal Process.* **2017**, *11*, 1301–1309. [\[CrossRef\]](#)
42. Faust, O.; Hagiwara, Y.; Hong, T.J.; Lih, O.S.; Acharya, U.R. Deep learning for healthcare applications based on physiological signals: A review. *Comput. Methods Programs Biomed.* **2018**, *161*, 1–13. [\[CrossRef\]](#) [\[PubMed\]](#)
43. Zunimova, G.; Soltan, G.; Ismailova, A.; Smayl, A.; Abdikadyr, Z.; Kaipova, A. The Impact of Education Management Digitalisation on the Quality of Student Learning. *Int. J. Comput.* **2024**, *23*, 432–439. [\[CrossRef\]](#)
44. Bekmurat, O.; Akpanbetov, D.; Tursynkhan, A.; Demeubayeva, L.; Duisenbekkyzy, Z.; Sansyzbay, K.; Kadirkulov, S.; Bakhtiyarova, Y. Explainable and Computationally Efficient NLP Framework for Detecting Psycho-Emotional Risk Signals in Social Media. *Computers* **2026**, *15*, 327. [\[CrossRef\]](#)

45. Izzah, N.; Sutarto, A.P.; Hariyadi, M. Machine Learning models for the Cognitive Stress Detection Using Heart Rate Variability Signals. *J. Tek. Ind.* **2022**, *24*, 83–94. [[CrossRef](#)]
46. Sarinova, A.; Bekbayeva, A.; Dunayev, P.; Sarsikeev, Y.; Sansyrbay, K. Hyperspectral image compression algorithms for phytosanitary inspection of agricultural crops in aerospace photography. *J. Theor. Appl. Inf. Technol.* **2021**, *99*, 6280–6290.
47. Aliaskarov, S.; Bekmurat, O.; Serbin, V.; Sansyrbay, K.; Bakhtiyarova, Y.; Andekin, E. Predicting Adolescent Suicide Risk in Smart Cities: An AI-Driven, Privacy-Preserving Architecture. *HighTech Innov. J.* **2025**, *6*, 1361–1389. [[CrossRef](#)]
48. Aliaskarov, S.; Uskenbayeva, R.; Serbin, V.; Bekmurat, O.; Bazarbayeva, U.; Bakhtiyarova, Y.; Sansyrbay, K. Mathematical Approaches and Algorithms in Big Data Architecture and Hybrid System Efficiency. *HighTech Innov. J.* **2025**, *6*, 1013–1034. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.