Original research article

DIGITAL HEALTH

DIGITAL HEALTH
Volume 11: 1–21

DOI: 10.1177/20552076251393338
journals.sagepub.com/home/dhj
Sage

Ensemble learning for improved sentiment analysis in doctor–patient communication

Yufan Ge[1], Lingling Dai[1], Bingding Huang[2] and Rashid Khan[2,3]

Abstract

Objective: To fill the benchmarking gap in clinician–patient sentiment analysis, we compare deep learning, transformer, and ensemble models for three-class (low/medium/high) sentiment classification in doctor–patient consultations.

Methods: We used a publicly available dataset of 3325 anonymized doctor–patient consultations from the Hugging Face repository (mahfoos/Patient-Doctor-Conversation) labeled as low, medium, or high severity. Preprocessing included text cleaning, tokenization, and padding; class balancing was applied only within the training split of each fold. Models evaluated were long short-term memory (LSTM), bidirectional LSTM (BiLSTM), convolutional neural networks (CNN), CNN–LSTM, and bidirectional encoder representations from transformers (BERT); an ensemble (hard voting over Logistic Regression, Random Forest, and Support Vector Classifier (SVC)) was also tested. Evaluation used stratified five-fold cross-validation, with metrics reported as mean ± SD across outer test folds (accuracy; macro-averaged precision/recall/F1). Interpretability was examined via BERT attention and feature attributions.

Results: The ensemble achieved the highest accuracy (75.5 ± 0.5), outperforming BERT (66.98 ± 0.6), CNN–LSTM (65.68 ± 0.9), CNN (64.17 ± 0.8), BiLSTM (64.82 ± 0.7), and LSTM (58.66 ± 0.19). Class-wise analysis showed robust detection of high-severity interactions (e.g. ensemble F1 = 90.8 ± 1.3), while low-severity remained most challenging; the ensemble improved class 0 recall (58.7 ± 1.0), and BERT provided the highest class 0 precision (65.5 ± 1.0).

Conclusion: Under stratified five-fold cross-validation, ensemble learning delivered the strongest and most balanced performance for three-class sentiment classification of clinician–patient dialogue, while transformers offered complementary precision on difficult cases. Attention- and feature-attribution analyses improved transparency, supporting clinical interpretability. Future work should scale to larger, multimodal (text/audio/vision) and multilingual datasets, and develop privacy-preserving, lightweight models for real-time deployment in clinical settings.

Keywords

Sentiment analysis, healthcare, machine learning, deep learning, BERT, ensemble learning

Received: 19 May 2025; accepted: 17 October 2025

Introduction

The advent of artificial intelligence (AI) has transformed many fields, including healthcare, with the potential to improve the effectiveness and quality of interactions between doctor and patient.[1] Efficient communication between physicians and patients is critical, as it directly influences patient satisfaction, adherence to medical recommendations, and overall health outcomes.[2] For instance, studies have shown that empathetic communication can improve patient trust and compliance by up to 20%. However, the nuanced nature of human sentiments and communication makes it challenging to understand and assess these interactions fully. This has galvanized interest in using AI-driven sentiment analysis (SA) to address the gap between human empathy and data-driven insights in healthcare settings.[3] While prior work, such as Huang

[1]International College, Anhui Medical University, Hefei, China
[2]College of Big Data and Internet, Shenzhen Technology University, Shenzhen, China
[3]College of Engineering Physics, Shenzhen Technology University, Shenzhen, China

Corresponding author:
Rashid Khan, College of Big Data and Internet, Shenzhen Technology University, Shenzhen, 518188, China.
Email: Rashidkhan@sztu.edu.cn

et al.,[4] achieved promising results in emotion recognition using video-based doctor–patient interactions. However, its reliance on multimodal inputs limits its applicability to text-based settings common in telemedicine.

In the rapidly developing field of SA, traditional machine learning (ML) and advanced deep learning (DL) methods have become the main focus.[5] Sentiment analysis is a powerful tool for extracting insight from unstructured data, supporting decision making in e-commerce, education, social networks, and healthcare. For example, Mujahid et al.[6] achieved 95% accuracy in analyzing sentiments in e-learning tweets, but their focus on simpler, nonmedical datasets highlights the challenge of applying SA to complex healthcare dialogs. Similarly, Khan et al.[7] reported 96% accuracy in sentiment classification of COVID-19-related tweets using lexicon-based and ML techniques, yet their binary classification approach does not address the multiclass emotional intensity required in doctor–patient interactions.

The continued development of digital platforms, especially social media, has generated a wealth of data reflecting public sentiment.[8] Platforms such as Twitter and professional services (e.g. online consulting, academic networks, and portals) provide rich, if noisy, data sources. Applying robust methods to evaluate this data can significantly improve engagement strategies. By using AI methods to evaluate conversations, clinicians, and managers can gain insight into patient emotions,[4] identify potential problems,[9] and improve care.[10] Although lexicon-based techniques provide initial understanding by utilizing predefined emotion dictionaries, their limitations, such as oversimplification of linguistic complexity, require additional sophisticated methods. For instance, Aljedaani et al.[11] achieved 97% accuracy in Twitter sentiment analysis using hybrid long short-term memory (LSTM), gated recurrent unit (GRU) models, but their approach was not tailored to healthcare's nuanced, domain-specific language. Hybrid models that combine these techniques with deep learning architectures such as LSTM networks, bidirectional LSTM (BiLSTM), and convolutional neural networks (CNNs) offer an encouraging way forward. These approaches exploit the temporal and local properties of sequence data. In particular, recurrent models such as LSTM and BiLSTM are well-suited for capturing sequential dependencies in dialogue, while CNNs excel at identifying local textual patterns. Hybrid CNN-LSTM architectures better combine these strengths to represent temporal and local language features. Recently, transformer-based models like BERT have advanced sentiment analysis by providing deep contextual representations, making them especially effective in addressing the complexity and variability of healthcare communication.

Despite progress in SA, significant challenges persist. Current techniques often suffer from application-specific linguistic complexity,[12] human expression variability,[13] potential biases in automated analysis.[14] For example, studies like reference [15] achieved 96.49% accuracy using BERT for ChatGPT-related tweet analysis, but their reliance on single-platform data limits generalizability to diverse healthcare datasets. To address these limitations, this work emphasizes application-specific datasets and incorporates robust preprocessing methods to reduce inconsistencies and noise within the data. Furthermore, integrating explanation mechanisms can provide a more interpretable and transparent model to ensure that results are actionable and accessible to healthcare professionals.

By advancing SA methods in the context of doctor–patient interactions, this study contributes to developing a body of knowledge about the application of AI in healthcare. The results of this research could reshape how physicians approach patient engagement, providing them with tools to understand and address patients' sentiments, expectations, and concerns. This aligns with the broader vision of creating a compassionate, patient-focused healthcare system that utilizes artificial intelligence to deliver efficient and empathetic care.

In this regard, the objectives of this study are: (i) to compare the effectiveness of different deep learning and machine learning approaches for multiclass sentiment classification in doctor–patient interactions, (ii) to address domain-specific challenges such as noisy language and class imbalance, and (iii) to investigate ways of improving interpretability to support clinical trust. This research not only aims to advance the theoretical understanding of SA frameworks,[16] but also provides a practical roadmap for applying these models in real-world healthcare scenarios.

The main contributions of this study are as follows:

1. A comparative framework that systematically evaluates multiple deep learning architectures (LSTM, BiLSTM, CNN, CNN-LSTM, and BERT) alongside a hard-voting ensemble of traditional machine learning classifiers (Logistic Regression, Random Forest, and Support Vector Classifier (SVC)), providing one of the first direct comparisons of these approaches for sentiment analysis in doctor–patient interactions.
2. Domain-specific preprocessing and training-only class-balancing strategies integrated into the pipeline to effectively handle noisy language and sentiment imbalance, ensuring robust predictions across low, medium, and high-severity emotional classes.
3. Enhanced model interpretability by fine-tuning BERT and visualizing attention maps, providing transparency into the linguistic cues that drive predictions and supporting clinical trust in AI-driven sentiment analysis (Figure 1).

Related work

SA has been extensively explored across various domains, including healthcare, education, and social media, with techniques ranging from lexicon-based methods to advanced DL models. Table 1 summarizes key studies relevant to this

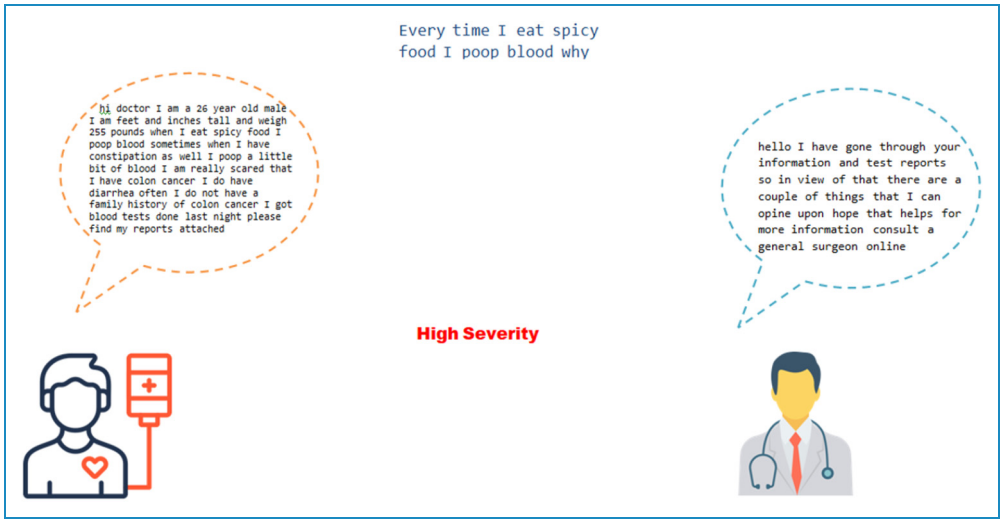

Figure 1. Medical consultation emphasizes doctor–patient communication.

work, highlighting their aims, techniques, datasets, results, and limitations.

The following subsections elaborate on hybrid, lexicon-based, and machine-learning models, building on the insights from Table 1.

Hybrid models

The proposed hybrid model integrates lexicon-based techniques like TextBlob with DL models such as LSTM–GRU, achieving high accuracy (97%) and an F1-score of 0.96.[11] ML algorithms such as SVC and the Extra Trees Classifier (ETC) also performed well with Term Frequency–Inverse Document Frequency (TF-IDF) and Bag-of-Words (BoW) features, reaching accuracies up to 0.92. While TextBlob-based annotations can assist with weak supervision and label refinement, they do not replace human annotators. Study[17] similarly aims to improve emotion classification by reducing label inconsistencies through hybrid pipelines that combine VADER/TextBlob with CNN, GRU, CNN–LSTM, and LSTM. Results show strong performance (accuracy ≈0.97, F1 ≈0.96 for LSTM–GRU), but limitations include reliance on automated interpretation without human adjudication and potential bias. In,[2] the transformative role of AI in addressing healthcare challenges (limited access, rising costs, and personalized care) is discussed, emphasizing cross-disciplinary collaboration and patient-oriented, data-driven solutions. Overall, hybrid approaches, combining lexicon cues with ML/DL architectures, leverage complementary strengths for improved performance.

Lexicon-based models

In the education domain, a Twitter dataset of 17,155 tweets was analyzed using lexicon-based techniques (TextBlob, SentiWordNet, and VADER) to address sentiment subjectivity and polarity. ML classifiers (e.g. Random Forest and SVM) achieved up to 0.95 accuracy with BoW features, and DL models (CNN–LSTM, CNN, BiLSTM, and LSTM) were also evaluated.[6] T Topic modeling highlighted challenges such as uncertainty around campus reopening, children's difficulty understanding online learning, and network limitations. Similarly,[7] examined sentiment analysis of 11,858 U.S. COVID-19 tweets (January–May 2020) using lexicon-based labeling (TextBlob) with TF-IDF/BoW features. Classifiers including RF, gradient boosting, SVM, LR, and ETC categorized emotions (neutral/positive/negative), with gradient boosting + TF-IDF achieving the best accuracy (96%). Future directions include scaling to larger datasets and enabling near-real-time analytics.

Machine learning models

Using LDA topic modeling,[15] identified frequently discussed themes, while sentiment classification employed a BERT model with three dense layers on hashtag-tagged tweets about ChatGPT, achieving 96.49% accuracy. A key limitation is the single-platform dataset, motivating cross-platform analyses to improve generalizability. In reference [18], a two-stage BERT-based pipeline extracted service attributes from text reviews to help patients choose appropriate doctors for online consultations. It combined public and personal preferences via a 2-increment fuzzy metric to generate comprehensive physician ratings; a case study (dxy.com) demonstrated improved rationality over traditional MADM. As a transformer language model, BERT continues to enhance NLP performance across tasks. Table 2 summarizes the main research gaps and future directions identified in prior work.

Table 1. Summary of previous studies and their key findings.

Study	Aim	Technique	Dataset	Results	Future work/limitation
11	Sentiment analysis customers' satisfaction levels	LSTM-GRU	Twitter data	97% accuracy	Different datasets will be used
6	Sentiment analysis about online Education during COVID-19	Lexicon-based techniques ML DL	Twitter dataset with 17155 Tweets	95% of highest accuracy using BoW features	Children's difficulty understanding online education
17	COVID-19 Vaccination-related sentiment analysis	Hybrid models	Labeled tweets	97% accuracy and 96% F1-score for LSTM-GRU	Integrate more powerful explanatory mechanisms
15	Analyzing Sentiments Regarding ChatGPT	BERT	Tweets tagged through hashtags	96.49% accuracy	Cross-platform data analysis work to increase applicability
18	Help patients choose the most appropriate doctor for an online consultation	Two-stage classification model based on BERT	Text reviews	Achieve comprehensive physician ratings	Big datasets and real-time decision making
7	Examined sentiment analysis from tweets in the COVID-19 pandemic	Lexicon-based techniques and machine learning	11,858 Tweets of dataset from January 2020–May 2020	96% of the highest accuracy achieved by the gradient boosting	In the future, larger datasets and analysis capabilities of additional real-time can be explored
2	AI's transformative role in handling the challenges of healthcare	Medical imaging through wearable devices	Cross-disciplinary collaboration to stabilize the benefits of AI with moral standards	Emphasized patient-oriented and healthcare data-driven solutions	Future work will explore the potential of AI for personalized medicine and universal health disaster management

AI: artificial intelligence; BERT: bidirectional encoder representations from transformers; DL: deep learning; GRU: gated recurrent unit; LSTM: long short-term memory; ML: traditional machine learning.

Beyond domain-general sentiment analysis, several works target clinician–patient conversations directly. For instance, Chatzimina et al.[19] applied topic modeling and sentiment analysis to Greek clinician–patient conversations in hematologic malignancies, finding more negative patient sentiments around pain, uncertainty, and loss, while clinicians remained more neutral. Sen et al.[20] developed an affective text analysis method to model doctor–patient communication in late-stage cancer consultations, linking linguistic/speech features to improved patient-reported outcomes. Le-Duc et al.[21] proposed a multimodal framework for sentiment reasoning in real-world doctor–patient conversations across various medical topics, incorporating annotations for sentiments and rationales to enhance AI transparency and trust. Rui et al.[22] analyzed online patient-provider communication transcripts to predict patient satisfaction, integrating sentiment metrics with provider strategies. Wang et al.[23] introduced CommSense, a wearable sensing framework for evaluating patient–clinician interactions, including sentiment-related assessments to promote health equity. Greaves et al.[24] utilized sentiment analysis to capture patient experiences from free-text online comments, categorizing them as positive or negative to inform healthcare improvements. Gohil et al.[25] applied sentiment analysis techniques to patient feedback narratives, identifying key emotional drivers in clinical interactions. Alemi et al.[26] explored sentiment in patient–doctor dialogs through natural language processing, focusing on emotional variance in mental health contexts. Khanbhai et al.[27] performed a comparative sentiment analysis on clinical conversation records, highlighting domain-specific challenges like class imbalance in emotional data. At the corpus scale, new resources such as PDCH (a multimodal depression consultation dataset with emotion annotations and clinician ratings)[28] and EHDChat (a knowledge-grounded, empathy-enhanced medical dialogue dataset)[29] further enable emotion and empathy-aware modeling in medical conversations.

Table 2. Identification of research gaps and proposed future directions.

Study	Research gap
11	This research focuses on hybrid lexicon-based models with DL techniques, but a comprehensive assessment of ensemble learning is lacking
6	This research uses lexicon-based techniques and ML algorithms to examine sentiment in e-learning, but deep learning or ensemble learning models have not been explored for better sentiment classification
2	This research highlights the role of artificial intelligence in healthcare struggles, but it does not provide specific insights into emotions in doctor-patient interactions
13	The variability of human expression is considered a challenge, but emotion-related DL models are not integrated to handle different emotions
14	Biases are discussed in the analysis of automated sentiment, but the interpretation applications for AI-driven SA in healthcare are not provided
16	Exploring theoretical advances in SA but not filling the gap between clinical AI theory and real-world performance

AI: artificial intelligence; DL: deep learning; ML: traditional machine learning; SA: sentiment analysis.

Additionally, the Distress Analysis Interview Corpus—Wizard-of-Oz (DAIC-WOZ) and its extended version E-DAIC are clinical interview corpora in which participants are interviewed by a virtual agent; ongoing work on these resources advances affect and depression detection, underscoring the broader applicability of sentiment/emotion signals in clinician–patient dialogue.[30,31] Collectively, these studies underscore the value of tailored sentiment analysis in healthcare dialogs; most emphasize binary or emotion-specific labels, leaving room for multiclass severity modeling as pursued here.

Methodology

This study is a retrospective computational analysis conducted from June 2024 to April 2025 at Anhui Medical University (Hefei, China) and Shenzhen Technology University (Shenzhen, China). It was based entirely on secondary, anonymized data from doctor–patient consultations without any direct patient recruitment or clinical intervention. All analyses used a publicly available dataset and were performed per relevant research ethics and data-governance standards. Building on this foundation, the study applies an AI-driven SA framework to model and interpret doctor–patient interactions using natural language processing (NLP).

We employed ML and DL methods to classify interaction severity into three classes (low, medium, and high). LSTM networks address the vanishing-gradient problem that can hinder vanilla RNNs from learning long-term dependencies.[32] BiLSTM processes sequences in both forward and backward directions and often outperforms unidirectional LSTM.[33] CNNs specialize in recognizing local patterns of sequential data, like n-grams in the text.[34] Hybrid CNN–LSTM outperforms traditional ML and DL models regarding accuracy, precision, recall, and f-measure.[35] Ensemble learning combines multiple models and can improve predictive accuracy, generalizability, and robustness.[36] Figure 2 illustrates the overall framework of the proposed study.

Data collection

The dataset used in this study was obtained from a publicly available Hugging Face repository (https://huggingface.co/datasets/mahfoos/Patient-Doctor-Conversation/viewer/default/train?p=4). It consists of 3325 anonymized doctor–patient consultation records. Each record includes key fields description, doctor, patient, and status (low/medium/high severity), which we concatenated into a single input text for modeling. Figure 3 illustrates the dataset structure and field organization.

The dataset supports multiclass sentiment classification with three severity levels: low (class 0), medium (class 1), and high (class 2). Before balancing, the classes were imbalanced (class 0 = 1780; class 1 = 1072; class 2 = 473), motivating the class-balancing procedures described in the Dataset balancing section. Under the cross-validation protocol, oversampling was applied only within the training split of each fold to prevent information leakage; the validation and test splits retained their natural label distributions.

For model development and evaluation, we used stratified five-fold cross-validation to preserve label proportions; the whole protocol is provided in the Train-validation-test split, and stratified cross-validation section.

Dataset preprocessing

The dataset was carefully preprocessed to ensure it was suitable for training the models. First, we merged relevant columns (description, doctor, and patient) into a single input field to capture the context of doctor–patient interaction. This concatenation provides a comprehensive representation of an interaction, $T_i = concat(description_i,\ \ doctor_i,\ patient_i)$.

Text cleaning. All text was lowercased and punctuation removed. Extraneous symbols and extra whitespace were eliminated; numerals were retained to preserve clinical context. Stop words were retained, as they can contribute to semantic meaning in conversational healthcare data.

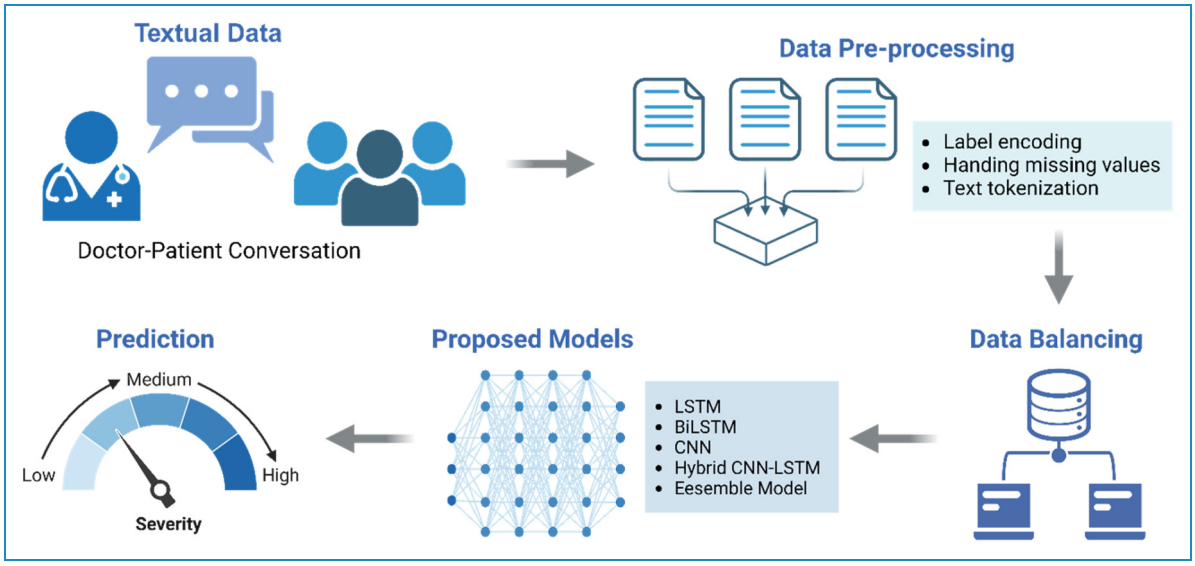

Figure 2. Structure of the proposed AI-driven framework for sentiment analysis in doctor–patient interactions.

Description	Doctor	Patient	Status
foot why does he get such pain	growing children are usually benign normal blood…	of shin pain in the left foot once in every mont…	low severity
applying conut oil for dry skin on penis shaft cause redness and burning in urethra is it…	hello normally the tip of urethra is slightly more reddish compared to the rest of the penis…	hello doctor I have dry masturbated for years and it created dry itchy skin on my penis shaft I…	medium severity
how to find that I am allergic to hydroxychloroquine	hello it is less likely you would be allergic to hydroxychloroquine however allergy is not the…	hello doctor I am allergic to quinolones metronidazole ibuprofen diclofenac and aspirin…	low severity
I get stomach pain if I do not eat for hours why	hello welcome you to icliniq com I can imagine how upset you have been since the start of your…	hello doctor I have regular pain in the stomach while not having food in hours or so I have…	low severity
has metformin 500 been recalled due to cancer scare	hello greetings happy to give consultation to with the above given history you sugar levels ar…	i have diabetes type I am a female my age is 27 and my weight is 43 kegs I am currently been…	low severity
my grade sprained ankle is paining extremely and I am unable to bear weight why	dear thank you for your query as per the information provided by you I would suggest you…	i have a grade sprained ankle im male 9 160lbs I had severe swelling to the size of a softball th…	medium severity
my fiance is couid positive how can I protect myself	hi I understand you are worried about couid infection you need to understand that if couid…	hello doctor I have a question on tuesday I went to go get tested with my fiance for couid becaus…	high severity

Figure 3. Details of the dataset used in the study.

Label encoding. Labels in the status column representing interaction severity levels were normalized (lowercased; punctuation removed) and then integer-encoded, $L_i = encode(Status_i)$.[37] Table 3 shows the mapping from severity level to encoded label.

Handling missing values. We addressed missing data by replacing any missing or null entries with blank strings, ensuring completeness and avoiding tokenization errors.[38] $T_i = if\ T_i\ is\ missing,\ then\ replace\ with\ .$

Text tokenization. Tokenization was applied to transform the merged text into integer sequences, converting each word to its corresponding index in the vocabulary. To prevent information leakage, we fitted the Keras Tokenizer on the training portion within each cross-validation fold only, with a vocabulary size of 5000 and an out-of-vocabulary token (<OOV>>).[39] Formally,

$$T_i \rightarrow \{w_1, w_2, \ldots, w_n\} \tag{1}$$

Padding sequences. To ensure that all sequences input to the model has the same length,[40] we padded or truncated sequences to a maximum of 100 tokens. Mathematically,

$$T_i = padding\left(T_i, \max_{length} = 100\right) \tag{2}$$

Train-validation-test split, and stratified cross-validation. We evaluated models using stratified five-fold cross-validation (scikit-learn's *StratifiedKFol*, k = 5, shuffle = True, and random_state = 42) at the conversation level. In each outer fold, 80% of the data served as the training portion and 20% as the test set; within the training portion, a stratified 20% was reserved as a validation set for early stopping/model

Table 3. Mapping of severity levels to encoded labels.

Severity level	Encoded label
Low severity	0
Medium severity	1
High severity	2

selection. All data-dependent preprocessing (e.g. tokenizer fitting and vocabulary/statistics) and any oversampling were fit only on the training partition of each fold to prevent information leakage; the validation and test partitions remained unchanged.

Through these steps, the dataset was prepared for sentiment classification. The combination of cleaning, missing-value handling, tokenization, padding, and stratified cross-validation ensured effective and unbiased model training.

Dataset balancing

Considering class imbalance within the training portion of each cross-validation fold, we applied random oversampling to the minority classes (classes 1 and 2) to reduce bias.[41] Imbalanced class distributions can lead to systematic performance shifts toward the majority class.[42] Accordingly, oversampling was performed only within the training split of each fold; the corresponding validation and test splits remained unchanged. For per-fold balancing, minority-class counts were increased to match the per-fold majority count:

$$N_{train,\,fold,\,after}(C) = N_{train,\,fold,\,majority} \text{ for } c \in \{Class\ 1,\ Class\ 2\} \tag{3}$$

We used scikit-learn's *sklearn.utils.resample* with replacement ($replace = True$, $random_{state} = 42$) to generate additional minority-class instances within the training split of each fold.[43,44] The resulting per-fold training set is (Table 4):

$$Train_{balanced} = Train_{Class\ 0} \cup Train_{Class\ 1,\,oversampled} \cup Train_{Class\ 2,\,oversampled} \tag{4}$$

Overall, the dataset counts are N = 3,325, with class 0 = 1,780, class 1 = 1,072, and class 2 = 473 (Data collection section). Minor fold-to-fold deviations may occur due to stratification on finite counts.

Methods

This research compares multiple model architectures, combining traditional ML and DL techniques, to classify doctor–patient interactions by severity. LSTM, BiLSTM, CNN, a hybrid CNN–LSTM, an ensemble classifier, and BERT were applied. Each model's architecture was designed to handle sequence-based input.

Table 4. Per-fold training class distribution before and after balancing (stratified 5-fold CV).

Labels	Training proportion (before) per fold	Training proportion (after) per fold
Class 0	~53.5%	~33.3%
Class 1	~32.2%	~33.3%
Class 2	~14.2%	~33.3%

LSTM. LSTM is a DL-based technique designed for modeling languages, speech analysis, text data prediction, and sentiment analysis.[45] This model was built using the Keras Sequential API to take advantage of its ability to model sequential data.

Embedding layer: Input sequences have been converted to a fixed-dimensional dense vector representation (see equation (5))

$$X_{embed} = f_{embed}(\mathrm{X}) \tag{5}$$

where, X shows the tokenized of input sequence and, f_{embed} shows the embedding function.

Stacked LSTM layers: The stacked LSTM model is an extension of LSTM with many hidden LSTM layers; every layer consists of numerous memory cells.[46] Two LSTM layers were stacked to capture long-term dependencies (equation (6)):

$$h_t = f_{LSTM}(x_t, h_{t-1}, c_{t-1}) \tag{6}$$

where x_t shows input at t time, h_t the hidden state, and c_{t-1} the cell state of the last timestep.

Dropout regularization: After every layer of LSTM, dropout is used to decrease overfitting[47] (see equation (7))

$$h_{t'} = f_{dropout}(h_t) \tag{7}$$

Dense output layer: A dense layer maps hidden states to class probabilities via softmax[48] (equation (8)):

$$\hat{y} = softmax(Wh' + b) \tag{8}$$

where $\hat{y}$ is the probability distribution prediction, W the weights, and b the biases. The L2 regularization term with λ=0.01 helps reduce overfitting, while a dropout rate of p = 0.5 further mitigates this risk. As detailed in equation (9), the class probabilities are computed by a dense layer with the Softmax activation function.

$$p(y = k|x) = \frac{\exp(z_k)}{\sum_{j=1}^{3} \exp(z_j)} \tag{9}$$

where z_k represents the score for the class k. The Adam optimizer, with a learning rate of $\partial = 0.0001$ is used to train the model along with sparse categorical cross-entropy loss. This can be represented in equation (10),

$$\tau = -\frac{1}{N}\sum_{i=1}^{N}\log P(y_i|x_i) \tag{10}$$

BiLSTM (bidirectional LSTM). The BiLSTM layer improves the model's ability to understand sequential dependencies through processing the sequence input forward and backward, leveraging prior and subsequent context.[49] In the embedding layer, each word is mapped t_i index with dense vector $e_i \epsilon R^d$, here $d = 64$. In this BiLSTM layer, both forward h_t and backward $\overleftarrow{} \rightarrow h_t$ hidden states are combined like equation (11),

$$h_t = concat\,(h_t\,, \overleftarrow{} \rightarrow h_t) \tag{11}$$

where $h_t \in R^{128}$ the dropout layer: A $p = 0.5$ dropout for regularization has been applied, and the sense layer is calculated using the class probabilities. The Softmax activation function is used.

CNN. CNNs specialize in recognizing local patterns of sequential data (n-grams) in text, making them suitable for this task.[34] Integer-encoded tokens were embedded into dense vectors $E = embedding(w; W_e)$, where $w \in Z^N$ is an input word sequence, $E \in R^{N\times d}$ is the embedding matrix with $d = 64$, and W_e the embedding lookup matrix. A one-dimensional (1D) convolution extracts n-gram features $f_i = ReLU\left(\sum_{j=0}^{k-1} x_{i+j}.\ W_c + b_c\right)$, where f_i is the feature at position i, x_{i+j} the input window of size k, W_c the convolutional filter, and b_c the bias. We used two Conv1D layers (128 filters, kernel $k_1 = 5$; then 64 filters, kernel $k_2 = 5$). MaxPooling1D reduces the temporal dimension while retaining salient features: $p_i = \max(x_{i;i+p})$ with pool size $p = 2$. The multidimensional tensor is flattened for dense layers $Flatten\,(X) = reshape(X,\ [-1])$; dense activations are computed as $h = ReLU\,(W_d.x + b_d)$. A Softmax output produces class probabilities $y_i = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}}$, $C = 3$. We optimized with Adam (equation (12)):

$$\theta_{t+1} = \theta_t - \partial.\ \frac{m_t}{\sqrt{v_t + \epsilon}} \tag{12}$$

where, ∂ represents the learning rate, m_t the first time estimate, and v_t the second. The loss was sparse categorical cross-entropy:

$\tau = -\frac{1}{N}\sum_{i=1}^{N}\log(p_{yi})$, where the true label of i^{th} sample is yi, and the predicted probability is p_{yi} .

Hybrid CNN-LSTM. For multiclass classification, we combine CNN and LSTM models. CNN layers perform local feature extraction,[50] while LSTM layers model long-term temporal dependencies within the data.[51] In the embedding layer, integer-encoded tokens are mapped to dense vectors in a shared vector space. A 1D convolutional layer extracts local n-gram features, and a MaxPooling1D layer downsamples the convolutional outputs by retaining the maximum within each pool. The LSTM layer then captures long-term sequential structure. Dropout reduces overfitting by randomly dropping units during training. Finally, a dense Softmax layer outputs three-class probabilities (low/medium/high).

Ensemble model. ML ensembles are an effective way to enhance predictive performance.[36] By aggregating diverse base classifiers, they reduce individual-model weaknesses and yield more stable predictions. We used a hard-voting classifier that combines Logistic Regression (LR), Random Forest (RF), and SVC. In hard voting, each base learner outputs a class label $\hat{y}_m$; the ensemble prediction is the majority class:

$$\hat{y}_{ensemble} = mode\{\hat{y}_{LR,}\hat{y}_{RF}\hat{y}_{SVC}\} \tag{13}$$

The LR classifier uses a logistic function to express the relationship among the X input features and the y target variable. $P(y = 1|X) = \sigma(wX + b)$. The decision limit is donated by $\hat{y} = \begin{cases} P(y = 1|X) \geq 0.5 \text{ then } 1 \\ otherwise \qquad\qquad 0 \end{cases}$. SVC attempts to expand the M margin between classes within high dimensional space of feature. The function of decision is $f(X) = sign\left(\sum_{i=1}^{N} a_i y_i K(X_i,\ X) + b\right)$. The SVC predicts $\hat{y} = sign(f(X))$. RF is a technique of decision trees ensemble. Each T_k tree within the forest generates $\hat{y}_k$. prediction and the result is aggregated by the forest $P(y|X) = \frac{1}{K}\sum_{k=1}^{k} P_k(y|X)$, and the last prediction is computed by $\widehat{y} = mode\{\hat{y}_1,\ \hat{y}_2\,,\ \ldots, \hat{y}_k\}$. Finally, the prediction for the voting classifier is $\widehat{y} = mode\{\hat{y}_{LogReg},\ \hat{y}_{RF}\,,\ \ldots, \hat{y}_{SVC}\}$.

BERT. Bidirectional Encoder Representation from Transformers (BERT) is an ML transformer-based model developed for a variety of NLP tasks, containing sequence classification.[52] The model is pretrained for the goal of masked language modeling and fine-tuned for particular downstream tasks. We also implemented BERT to classify text data. Text data is converted into numeric form through tokenization suited input for BERT model. Let the input sequence of text is T. Tokenizer mapping T in token ID $x = \{x_1,\ x_2,\ \ldots\ , x_N\}$, where N shows the length sequence and $x_i \in Z$. $x = Tokenizer\big(T,\ padding = True,\ truncation = True,$ $\max_{length} = 128\big)$. BERT model can be represented for sequence classification is $y = Softmax(W_c.\ h_{[cls]} + b_c)$, where $y = R^C$ and C represents classes' counts. $h_{[cls]}$ shows hidden representation. The loss function cross entropy is used $\tau = -\frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{C} y_{ij}\log(\hat{y}_{ij})$. Adam optimizer is used with the learning rate $\eta = 2 \times 10^{-5}$, loss function is

Table 5. Architecture, components, and hyperparameters of each model used in the study.

Model	Components	Hyperparameters
LSTM	Embedding: Vocabulary size (input) = 5000, embedding size (output) = 64, input_length = 100 LSTM 1st layer: Units = 64, L2 regularization = 0.01, return sequence = true Dropout: 0.5 LSTM 2nd layer: L2 regularization = 0.01, Units = 64, Output layer (dense): Units = 3 (three classes), L2 regularization = 0.01, activation = softmax	Optimizer: Adam optimizer, learning rate = 0.0001 Loss_function: Cross entropy loss function Batch_size = 32 Epochs = 50
BiLSTM	Embedding: Vocabulary size (input) = 5000, embedding size (output) = 64, input_length = 100 BiLSTM: Units = 64 Dropout: 0.5 Output layer (dense): Units = 3 (three classes), activation = softmax	Optimizer: Adam optimizer, Loss_function: Cross entropy loss function Batch_size = 32 Epochs = 50
CNN	Embedding: Vocabulary size (input) = 5000, embedding size (output) = 64, input_length = 100 Conv1D: ReLU, filter = 128, kernel = 5 MaxPooling1D: Pool = 2 Flatten Dense: Units = 64, ReLU Dropout: 0.5 Output layer (dense): Units = 3, activation = softmax	Optimizer: Adam optimizer, Loss_function: Cross entropy loss function Batch_size = 32 Epochs = 50
CNN-LSTM (hybrid)	Embedding: Vocabulary size (input) = 5000, embedding size (output) = 64, input_length = 100 Conv1D: ReLU, filter = 128, kernel = 5 MaxPooling1D: Pool = 2 LSTM: Units = 64 Dropout: 0.5 Output layer (Dense): Units = 3, activation = softmax	Optimizer: Adam optimizer, Loss_function: Cross entropy loss function Batch_size:=32 Epochs = 50
Ensemble (voting classifier)	Base Models: log_reg, SVC, RF Type: Hard voting (predicted probabilities used to make better decisions)	-
BERT	Pre-trained: *bert-base-uncased* model Tokenizer: bert tokenizer	Optimizer: Adam optimizer, *lr* = *2e-5* *lr* = *2e-5* Epochs = 50 Batch_size = 8 Loss: Implicit

LSTM: long short-term memory; RF: Random Forest; SVC: Support Vector Classifier.

minimized by the model, and the total counts of training attempts are calculated as $train_{steps_{No}} = \frac{sample_{No}}{batch_{Size}} \times epochs_{No}$.

Models' architecture

Table 5 shows the architecture of each model, detailing the components and hyperparameters used for LSTM, BiLSTM, CNN, CNN-LSTM hybrid, Ensemble, and BERT models. Table 5 highlights the specific configurations and settings applied to each model to achieve optimal performance in sentiment analysis tasks. The hyperparameters reported in Table 5 were determined through empirical tuning rather than systematic grid or random search. Initial ranges for parameters such as learning rate, batch size, and dropout probability were selected based on values commonly reported in related literature, and multiple pilot experiments were conducted to refine them. The final choices were made by monitoring validation accuracy and F1-scores while balancing computational efficiency. For the ensemble classifier, standard scikit-learn implementations of LR, RF, and SVC were used

with default settings, as the primary focus was on evaluating the combined voting strategy rather than fine-tuning each base learner.

Evaluation metrics

To comprehensively evaluate the performance of sentiment classification models in doctor–patient interactions, multiple metrics were employed. Overall accuracy was used to measure the proportion of correctly classified consultations across all severity levels:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{14}$$

Since accuracy alone can be misleading in imbalanced datasets, precision, and recall were also calculated for each sentiment severity class (low = 0, medium = 1, and high = 2). Precision indicates how many of the predicted interactions for a given class were correct:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{15}$$

whereas recall reflects the proportion of true class samples that were successfully identified:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{16}$$

The F1-score, defined as the harmonic mean of precision and recall, provided a balanced measure of performance, which is particularly important in this study given the class imbalance present in the doctor–patient dataset:

$$F1 = 2.\frac{Precision \cdot Recall}{Precision + Recall} \tag{17}$$

Finally, the area under the receiver operating characteristic curve (AUC-ROC) was computed, and ROC curves were plotted for all models. This metric captures the ability of the classifiers to separate low, medium, and high-severity emotional interactions at varying thresholds, offering an additional perspective on robustness beyond class-specific metrics. Together, these metrics ensured that the evaluation not only measured overall accuracy but also captured the models' capability to detect minority classes, which is critical for identifying high-severity emotional interactions in clinical communication.

Results

All metrics are reported on the outer test folds of the stratified five-fold cross-validation; values are reported as mean ± standard deviation across folds. For the LSTM model, the cross-validated accuracy was 58.66 ± 0.19 (indicating that, on average, the model correctly classified ~58.7% of test samples across folds). Table 6 shows the class-wise precision, recall, and F1.

Table 6. Classification report for the LSTM model (%, mean ± SD across 5 folds), including precision, recall, and F1-score for each severity class.

Class	Precision	Recall	F1-score
0	49 ± 1.1	40 ± 1.3	44 ± 1.2
1	66 ± 0.8	47 ± 1.0	55 ± 0.9
2	59 ± 1.5	87 ± 2.1	70 ± 1.8

LSTM: long short-term memory.

Table 7. Classification report for the BiLSTM model (%, mean ± SD across 5 folds), including precision, recall, and F1-score for each severity class.

Class	Precision	Recall	F1-score
0	45 ± 1.2	40 ± 1.4	42 ± 1.3
1	61 ± 0.9	58 ± 1.1	60 ± 1.4
2	79 ± 1.5	91 ± 2.1	84 ± 1.7

BiLSTM: bidirectional long short-term.

For the LSTM model, precision for each category measures the accuracy of positive predictions. Class 0 has the lowest precision at 49.0 ± 1.1, which indicates a higher false-positive rate. In contrast, class 1 has the highest precision at 66.0 ± 0.8, reflecting a good positive prediction for this class. Recall quantifies the model's ability to identify true positives. Class 2 achieved the highest recall at 87.0 ± 2.1, demonstrating that the model successfully identifies most samples from this class. However, class 0 has the lowest recall at 40.0 ± 1.3, indicating a high false-negative rate. The F1-score balances precision and recall. Class 2 achieved the highest F1-score of 70.0 ± 1.8, indicating strong overall performance, while class 0 has the lowest F1-score at 44.0 ± 1.2, signaling the need for improvement in this class.

For the BiLSTM model, the cross-validated accuracy was 64.82 ± 0.70. Compared with LSTM (Table 6), BiLSTM shows higher mean accuracy and stronger detection of high-severity cases (class 2), while class 0 remains comparatively more challenging. The class-wise report (%, mean ± SD across 5 folds) is shown in Table 7.

Class 0 has the lowest precision (45 ± 1.2) and recall (40 ± 1.4), indicating higher false-positive and false-negative rates, respectively. In contrast, class 2 attains the highest precision (79 ± 1.5) and recall (91 ± 2.1), yielding the strongest F1-score (84 ± 1.7). Class 1 shows balanced precision and recall (≈61%/58%), resulting in a mid-range F1-score (60 ± 1.4).

Table 8. Classification report for the CNN model (%, mean ± SD across 5 folds), including precision, recall, and F1-score for each severity class.

Class	Precision	Recall	F1-score
0	49.2 ± 1.5	33.7 ± 1.2	40.0 ± 1.1
1	61.4 ± 0.9	61.8 ± 1.0	61.1 ± 0.9
2	73.2 ± 1.4	90.4 ± 1.8	81.2 ± 1.5

CNN: convolutional neural networks.

Table 9. Classification report for the CNN–LSTM model (%, mean ± SD across 5 folds), including precision, recall, and F1-score for each severity class.

Class	Precision	Recall	F1-score
0	47.6 ± 1.1	37.2 ± 1.3	41.3 ± 1.2
1	63.2 ± 0.9	62.6 ± 1.0	63.8 ± 0.9
2	77.7 ± 1.4	90.4 ± 1.8	83.2 ± 1.5

CNN: convolutional neural networks; LSTM: long short-term memory.

For the CNN model, the cross-validated accuracy was 64.17 ± 0.8. Relative to BiLSTM (Table 7), CNN attains comparable mean accuracy but exhibits lower sensitivity to class 0 and similarly strong performance on class 2. The class-wise report (%, mean ± SD across 5 folds) is shown in Table 8.

Class 0 has the lowest precision (49.2 ± 1.5) and the lowest recall (33.7 ± 1.2), reflecting higher false-positive and false-negative rates, respectively. In contrast, class 2 shows the highest precision (73.2 ± 1.4) and recall (90.4 ± 1.8), yielding the strongest F1-score (81.2 ± 1.5). Class 1 exhibits balanced precision/recall (≈61.4%/61.8%), producing an F1-score of 61.1 ± 0.9.

For the CNN–LSTM (hybrid) model, the cross-validated accuracy was 65.68 ± 0.90. Relative to CNN (Table 8), CNN–LSTM yields a modest gain in mean accuracy and macro-F1, with similarly strong detection of class 2 and persistent difficulty on class 0. The class-wise report (%, mean ± SD across 5 folds) is shown in Table 9.

For the hybrid CNN–LSTM model, class 0 shows the lowest mean precision (47.6 ± 1.1) and recall (37.2 ± 1.3), indicating higher false-positive and false-negative rates, respectively. Class 2 maintains the highest precision (77.7 ± 1.4) and recall (90.4 ± 1.8), yielding the strongest F1-score (83.2 ± 1.5). Class 1 exhibits balanced precision and recall (≈63.2%/62.6%), resulting in a mid-range F1 (63.8 ± 0.9).

For the ensemble model, the cross-validated accuracy was 75.5 ± 0.5 across the outer test folds. Relative to CNN–LSTM (Table 9), the ensemble achieves the highest mean accuracy and macro-F1 across folds, with consistently strong detection of class 2 and improved performance on class 0. The class-wise report (%, mean ± SD across 5 folds) is shown in Table 10.

For the ensemble model, class 0 shows the lowest mean precision (62.4 ± 0.8) and recall (58.7 ± 1.0), indicating higher false-positive and false-negative rates than the other classes. Class 2 attains the highest precision (89.7 ± 1.2) and recall (92.6 ± 1.5), yielding the strongest F1-score (90.8 ± 1.3). Class 1 remains balanced (~71–72%), producing a stable F1 (72.6 ± 0.7).

Table 10. Classification report for the ensemble model (%, mean ± SD across 5 folds), including precision, recall, and F1-score for each severity class.

Class	Precision	Recall	F1-score
0	62.4 ± 0.8	58.7 ± 1.0	60.2 ± 0.9
1	71.3 ± 0.7	72.4 ± 0.8	72.6 ± 0.7
2	89.7 ± 1.2	92.6 ± 1.5	90.8 ± 1.3

Table 11. Classification report for the BERT model (%, mean ± SD across 5 folds), including precision, recall, and F1-score for each severity class.

Class	Precision	Recall	F1-score
0	65.5 ± 1.0	44.7 ± 1.2	53.3 ± 1.1
1	62.4 ± 0.8	61.2 ± 0.9	62.7 ± 0.8
2	72.5 ± 1.3	90.1 ± 1.6	80.3 ± 1.4

BERT: bidirectional encoder representations from transformers.

For the BERT model, the cross-validated accuracy was 66.98 ± 0.6. Relative to the ensemble (Table 10), BERT yields lower mean accuracy but competitive performance on class 2; compared with recurrent baselines (Tables 6 and 7), it shows markedly higher precision on class 0. The class-wise report (%, mean ± SD across 5 folds) is shown in Table 11.

Class 0 shows the lowest mean recall (44.7 ± 1.2), indicating more false negatives, despite relatively strong precision (65.5 ± 1.0). Class 2 again attains the highest recall (90.1 ± 1.6) and F1 (80.3 ± 1.4). Class 1 remains balanced around 61% to 62% across metrics. Loss over epochs and accuracy over epochs have been visualized to display the learning process of each model, which is shown in Figure 4 to 8, respectively.

The loss plot describes how the model's prediction error decreases during training and validation. The accuracy plot

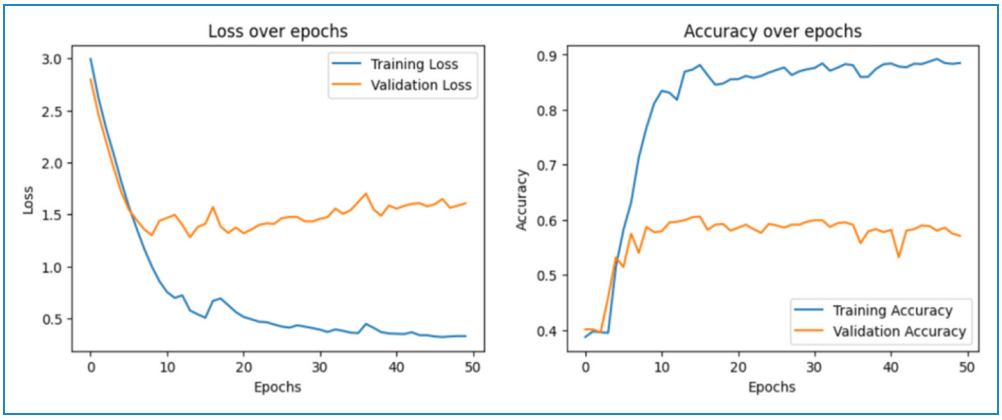

Figure 4. Loss over epochs and accuracy over epochs of the long short-term memory (LSTM) model.

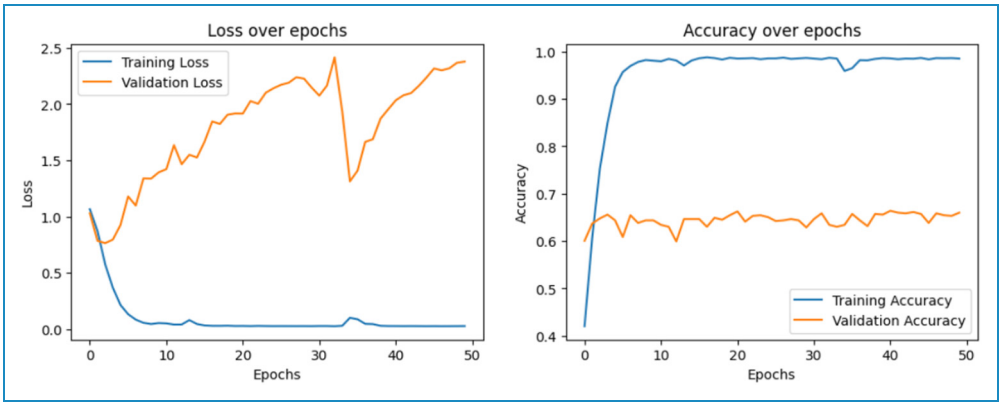

Figure 5. Loss over epochs and accuracy over epochs of the bidirectional long short-term memory (BiLSTM) model.

explains how well our model predicts results during training and validation. A confusion matrix is also visualized to analyze the model's ability to classify samples correctly. Figures 9 and 10 show the confusion matrix of each model.

This heat map highlights the counts of TP (true positives), FP (false positives), FN (false negatives), and TN (true negatives) of each class.

ROC curve

An ROC curve shows how well a model performs across different threshold values.

Above, Figures 11(a)–(f) plots ROC curve for each category to display the exchange between sensitivity actual positive rate (TPR) and specificity false positive rate (FPR). The AUC value represents the ability of the models to differentiate between classes.

Comparative analysis

This section presents a comparative analysis of the implemented models, explaining their selection and relative performance. LSTM and BiLSTM capture sequential dependencies in conversational text, CNN extracts local features, CNN–LSTM combines spatial and temporal cues, and BERT provides deep contextual representations. The ensemble integrates LR, RF, and SVC via hard voting to leverage complementary decision boundaries. All comparative results are reported under stratified five-fold cross-validation as mean $\pm$ SD across the outer test folds. A detailed per-class comparison appears in Table 12; model-specific tables (Tables 6–11) provide the same metrics per model.

The ensemble achieves the strongest overall accuracy (75.5 ± 0.5), followed by BERT (66.98 ± 0.6), CNN–LSTM (65.68 ± 0.9), BiLSTM (64.82 ± 0.7), CNN (64.17 ± 0.8), and LSTM (58.66 ± 0.19). These margins exceed fold-to-fold variability (SDs $\lesssim$1%), indicating robust differences across folds. Class-wise patterns mirror this ranking: class 2 is generally easiest (high recall across models), whereas class 0 remains most challenging. BERT offers higher precision on class 0 (65.5 ± 1.0) than the ensemble (62.4 ± 0.8), but the ensemble substantially better class 0 recall (58.7 ± 1.0 vs. 44.7 ± 1.2) yields stronger class 0 F1

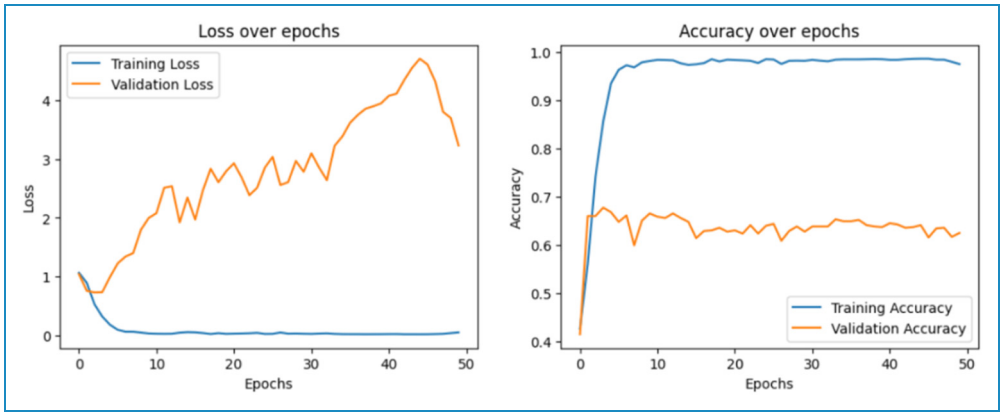

Figure 6. Loss over epochs and accuracy over epochs of the convolutional neural networks (CNN) model.

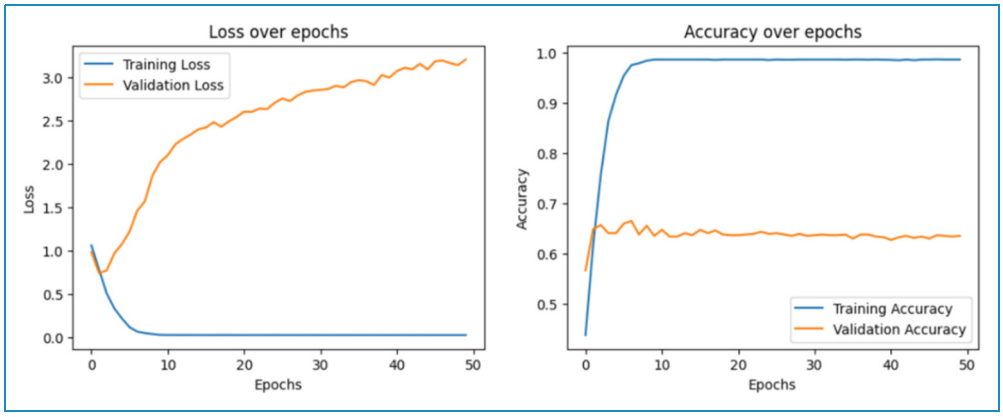

Figure 7. Loss over epochs and accuracy over epochs of the hybrid convolutional neural networks long short-term memory (CNN–LSTM) model.

overall. Comparison results of different models in this study are shown in Figure 12.

Sentiment trends are explored by Mujahid et al.[6] in e-learning, and the traditional ML techniques, such as SVM and RF, performed with 95% accuracy. The strength of the transformer model was demonstrated by Wu et al.[18] and Mujahid et al.[15]; the BERT-based framework outperformed traditional models and achieved 96.49% accuracy for sentiment classification. At the same time, Khan et al. [7] exploited lexicon-based and an ML technique for sentiment classification in the COVID-19 tweets, using TF-IDF and gradient boosting by achieving 96% accuracy. In comparison, our work raises an important point by showing that DL techniques achieved considerable accuracy scores, such as 58.7 ± 0.2 and 65.7 ± 0.9. Although our ensemble model positively improves performance with an accuracy of up to $75.5 \pm 0.5\%$, it still lags behind advanced transformer models. Noticeably, in our research, the BERT approach obtained 67.0 ± 0.6 accuracy, demonstrating that larger datasets and more fine-tuning could improve its performance. Unlike existing studies that mainly focused on the binary classification and the simple dataset, our multi-classification approach, such as low, medium, and high, and a complex dataset introduces further complexity, thus troubling performance. Generally, our findings emphasize the need to leverage ensemble learning and transformer-based approaches to fill the execution gap and enhance classification in the complex datasets.

Interpretability analysis

Interpretability is crucial for clinical sentiment analysis systems, as healthcare practitioners must trust both the predictions and their underlying rationale. To enhance transparency, we investigated BERT's attention mechanisms to uncover contextual patterns and applied SHapley Additive exPlanations (SHAP) values to a TF-IDF + LR model to identify key textual features driving classification decisions.

Figure 13 presents the attention maps of 10 heads from the third layer of BERT for a representative doctor–patient dialogue. Each subplot corresponds to a different attention head, with tokens on both axes and brighter cells indicating

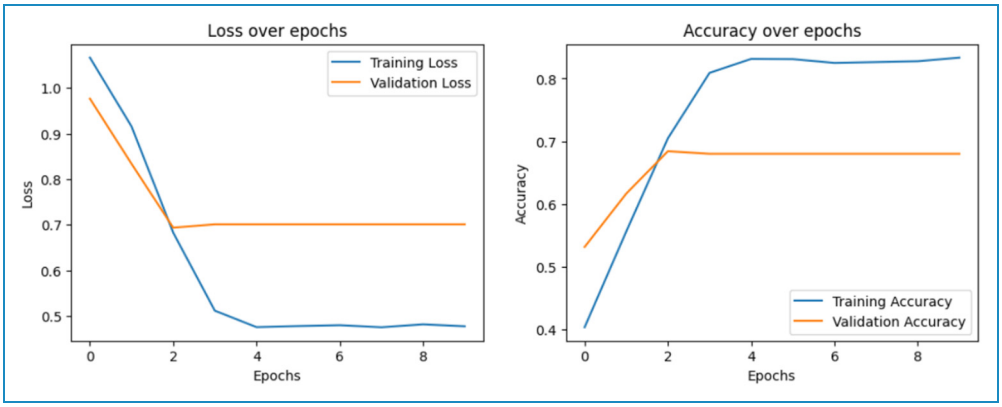

Figure 8. Loss over epochs and accuracy over epochs of the bidirectional encoder representations from transformers (BERT) model.

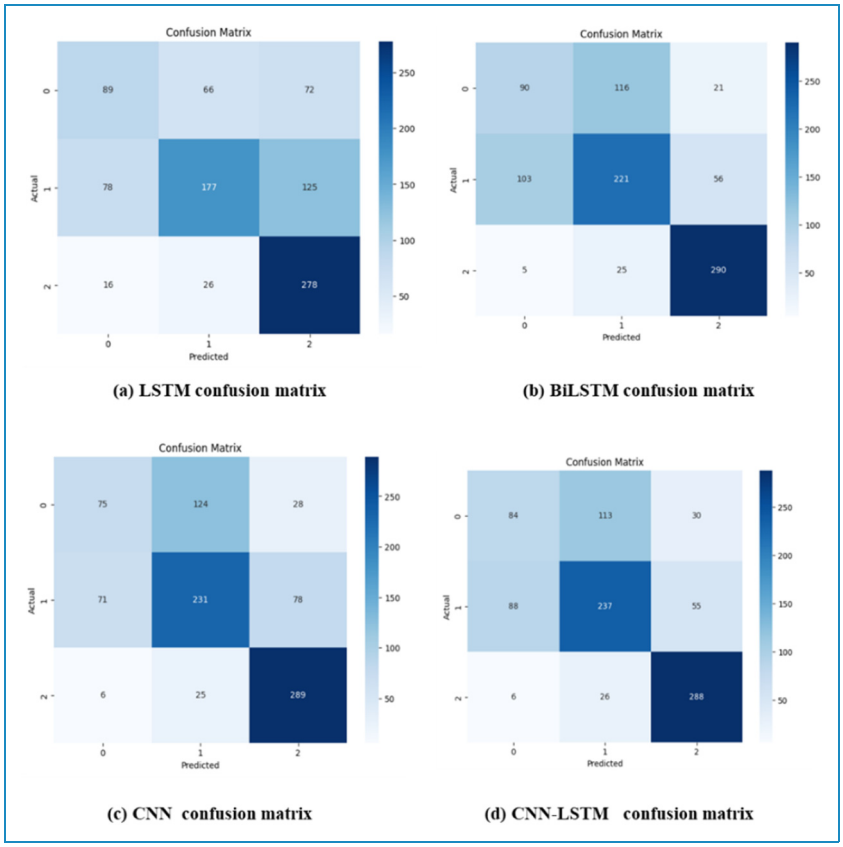

Figure 9. Confusion matrix reports of the evaluated models: (a) LSTM model, (b) BiLSTM model, (c) CNN model, and (d) hybrid CNN-LSTM model. BiLSTM: bidirectional long short-term; CNN: convolutional neural networks.

stronger weights. The maps reveal diverse interpretive strategies: some heads focus narrowly along the diagonal, reflecting word-level consistency, while others attend to medically salient expressions such as "worried" and "pain." Certain heads also differentiate between conversational roles (doctor vs. patient), indicating that the model captures dialogue structure and word meaning. These patterns confirm that BERT distributes attention across multiple complementary mechanisms rather than relying on a single pathway, which helps explain its robust performance in sentiment classification.

To complement this deep-model perspective, Figure 14 shows SHAP values for a TF-IDF + LR classifier, highlighting the n-grams most influential in predicting high emotional severity (class 2). Terms such as "blood," "cancer," "brain," "chest," and "scan" exert substantial positive

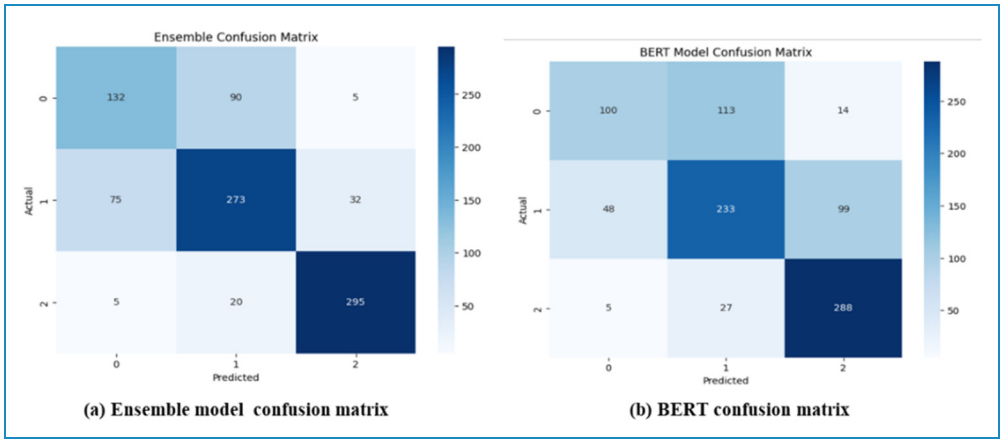

Figure 10. Confusion matrix reports of the advanced models: (a) ensemble model and (b) bidirectional encoder representations from transformers (BERT) model.

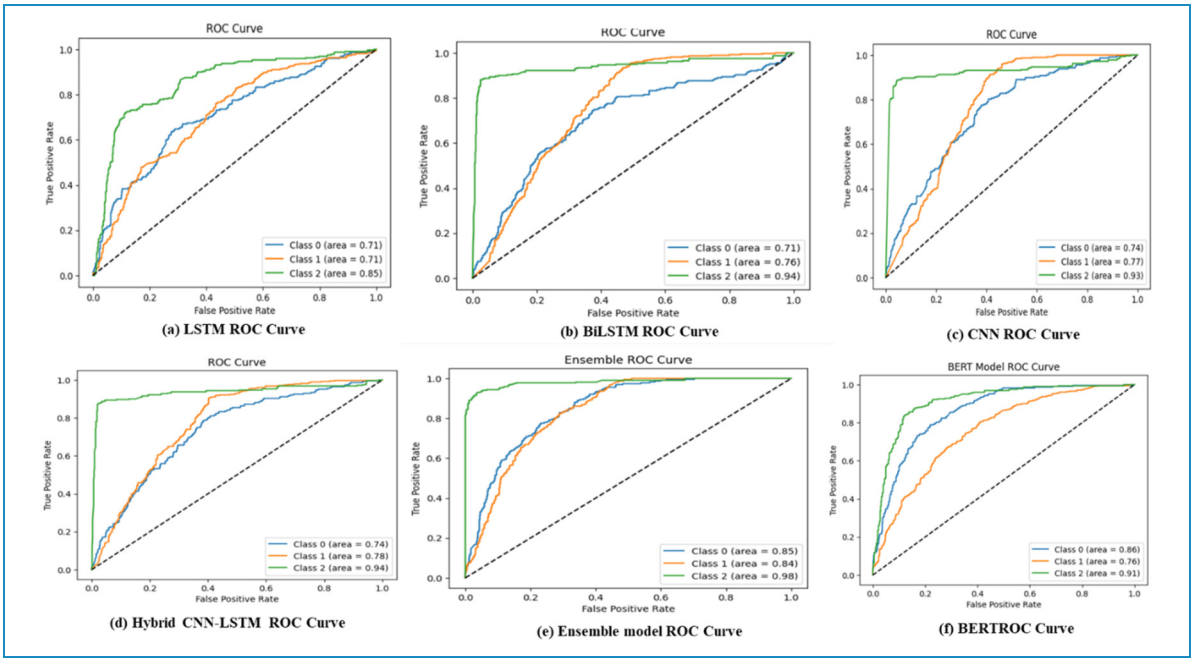

Figure 11. ROC curves of all models evaluated in this study: (a) LSTM model, (b) BiLSTM model, (c) CNN model, (d) hybrid CNN–LSTM model, (e) ensemble model, and (f) BERT model. BERT: bidirectional encoder representations from transformers; BiLSTM: bidirectional long short-term; CNN: convolutional neural networks; ROC: receiver operating characteristic.

contributions, while phrases like "feel fine" reduce the probability of a high-severity classification. Although our emphasis is on class 2, for completeness, it is worth noting that SHAP can also clarify the drivers of lower-severity categories: reassuring expressions typically contribute toward class 0 (low severity), while ambiguous or mildly negative terms influence class 1 (medium severity). This reinforces that different lexical cues consistently align with each severity level, providing a transparent link between model predictions and the underlying language. Together, these interpretability techniques provide complementary insights. Attention maps show how deep contextual dependencies are captured in BERT, whereas SHAP visualizations expose which explicit n-grams are most responsible for classification outcomes in a transparent linear model. By grounding predictions in medically meaningful tokens and dialogue structure, both methods support clinician trust and strengthen the case for using sentiment analysis in sensitive healthcare contexts.

Discussion

This study demonstrates the potential of AI-driven SA to improve understanding of doctor–patient interactions, highlighting the complementary strengths of different modeling

Table 12. Comparative evaluation across models (%, mean ± SD across 5 folds).

Model	Accuracy	Precision	Recall	F1-score
LSTM	58.66 ± 0.19	Class 0 = 49 ± 1.1	Class 0 = 40 ± 1.3	Class 0 = 44 ± 1.2
		Class 1 = 66 ± 0.8	Class 1 = 47 ± 1.0	Class 1 = 55 ± 0.9
		Class 2 = 59 ± 1.5	Class 2 = 87 ± 2.1	Class 2 = 70 ± 1.8
BiLSTM	64.82 ± 0.70	Class 0 = 45 ± 1.2	Class 0 = 40 ± 1.4	Class 0 = 42 ± 1.3
		Class 1 = 61 ± 0.9	Class 1 = 58 ± 1.1	Class 1 = 60 ± 1.4
		Class 2 = 79 ± 1.5	Class 2 = 91 ± 2.1	Class 2 = 84 ± 1.7
CNN	64.17 ± 0.80	Class 0 = 49.2 ± 1.5	Class 0 = 33.7 ± 1.2	Class 0 = 40.1 ± 1.1
		Class 1 = 61.4 ± 0.9	Class 1 = 61.8 ± 1.0	Class 1 = 61.1 ± 0.9
		Class 2 = 73.2 ± 1.4	Class 2 = 90.4 ± 1.8	Class 2 = 81.2 ± 1.5
Hybrid (CNN–LSTM)	65.68 ± 0.90	Class 0 = 47.6 ± 1.1	Class 0 = 37.2 ± 1.3	Class 0 = 41.3 ± 1.2
		Class 1 = 63.2 ± 0.9	Class 1 = 62.6 ± 1.0	Class 1 = 63.8 ± 0.9
		Class 2 = 77.7 ± 1.4	Class 2 = 90.4 ± 1.8	Class 2 = 83.2 ± 1.5
BERT	66.98 ± 0.60	Class 0 = 65.5 ± 1.0	Class 0 = 44.7 ± 1.2	Class 0 = 53.3 ± 1.1
		Class 1 = 62.4 ± 0.8	Class 1 = 61.2 ± 0.9	Class 1 = 62.7 ± 0.8
		Class 2 = 72.5 ± 1.3	Class 2 = 90.1 ± 1.6	Class 2 = 80.3 ± 1.4
Ensemble model	75.5 ± 0.50	Class 0 = 62.4 ± 0.8	Class 0 = 58.7 ± 1.0	Class 0 = 60.2 ± 0.9
		Class 1 = 71.3 ± 0.7	Class 1 = 72.4 ± 0.8	Class 1 = 72.6 ± 0.7
		Class 2 = 89.7 ± 1.2	Class 2 = 92.6 ± 1.5	Class 2 = 90.8 ± 1.3

BERT: bidirectional encoder representations from transformers; BiLSTM: bidirectional long short-term; CNN: convolutional neural networks.

strategies. Unless otherwise stated, performance figures below are reported as mean ± SD across the outer test folds of the stratified five-fold cross-validation. Among the tested models, the ensemble classifier achieved the highest overall accuracy (75.5 ± 0.5) and delivered particularly strong results for the medium sentiment class, with an F1-score of 72.6 ± 0.7. This superior performance can be attributed to the ensemble's ability to combine diverse decision boundaries: LR captured linear separations, RF modeled nonlinear patterns, and SVC handled complex margins, which collectively balanced precision and recall across sentiment categories. In contrast, individual deep learning models showed uneven performance: BiLSTM and CNN–LSTM captured long-range dependencies and local textual features effectively, excelling in high-severity interactions (mean F1 ≈ 84.0 ± 1.7 and 83.2 ± 1.5, respectively), while BERT achieved strong contextual understanding but was comparatively weaker on class 0 recall. Notably, BERT attained the highest precision on class 0 (65.5 ± 1.0), whereas the ensemble achieved markedly higher class 0 recall (58.7 ± 1.0), yielding a better overall balance for this difficult class. Fold-to-fold variability was modest (typically ≤ ~1% SD across metrics), indicating stable estimates. These patterns underscore that model performance depends not only on architecture but also on the balance of class distributions and the subtlety of sentiment expression in clinical dialogue.

The findings align with and extend prior work in AI-driven SA. Studies on domain-general datasets have reported very high accuracies using BERT-based models, on online doctor reviews, often exceeding 95%.[6] These results, however, were largely based on binary classification tasks and large nonclinical corpora. In contrast, our study addresses a more demanding three-class sentiment

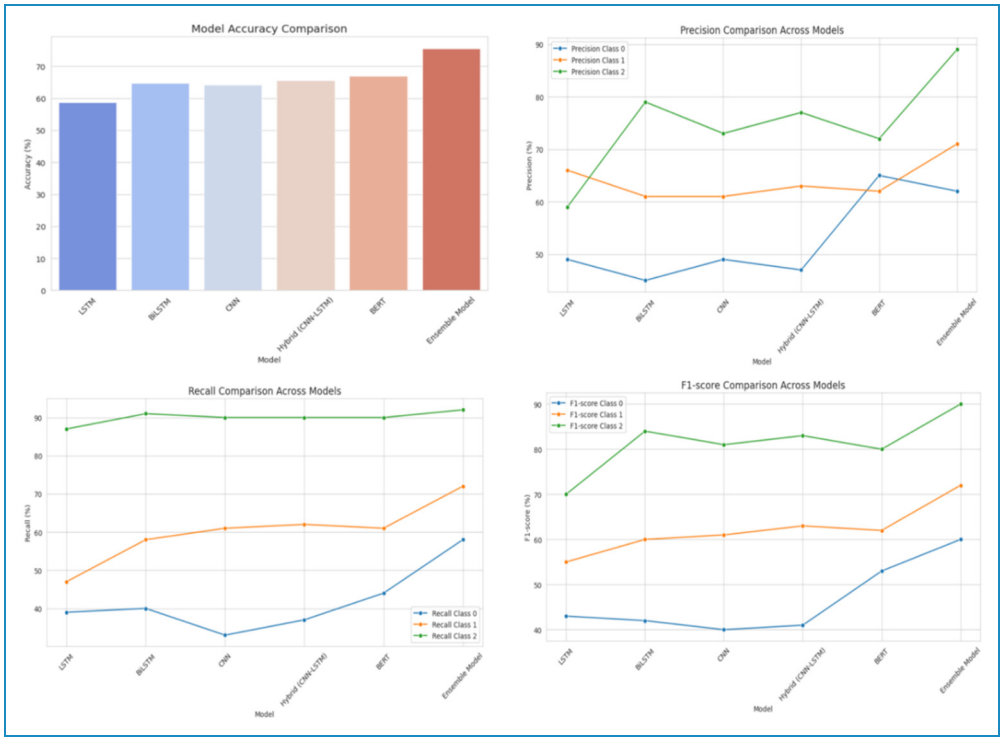

Figure 12. Comparison of performance across different models, highlighting accuracy, precision, recall, and F1 scores.

classification problem in healthcare dialogs, where emotional intensity spans from mild reassurance to high-severity distress. This naturally results in lower absolute accuracies but highlights the novelty and difficulty of our setting. Our findings are consistent with those of Hsu et al.,[53] who similarly emphasized the promise of transformer-based models for emotion recognition, pointing to their potential to enhance therapeutic communication and support early detection of mental health concerns. Our study demonstrates that contextual models like BERT capture nuanced emotional cues in doctor–patient interactions. Significantly, unlike the conceptual focus of many domain-general studies, our work contributes empirical evidence grounded in real clinical dialogue. It advances interpretability through attention maps and SHAP analyses, providing transparency into the linguistic cues driving predictions.

The clinical implications of these findings are significant. By reliably detecting both high-severity distress and moderate levels of patient concern, the proposed models could assist clinicians in tailoring communication strategies, improving empathy, and identifying when additional psychological support may be warranted. Ensemble methods, in particular, show promise for integration into real-time systems because of their robustness across sentiment classes. Moreover, attention maps revealed that BERT focuses on medically salient words (e.g. "pain" and "worried") and speaker roles, while SHAP analysis identified domain-relevant n-grams (e.g. "cancer," "blood," and "scan") as critical signals. These insights strengthen trust by demonstrating that models attend to clinically meaningful patterns rather than spurious correlations.

At the same time, several limitations must be acknowledged. This study analyzed only text transcripts, whereas real consultations involve multimodal cues such as prosody, facial expressions, and gestures that can convey sentiment more richly. The dataset, while valuable, is relatively small and imbalanced, particularly for high-severity classes, and it may not fully reflect the diversity of real clinical conversations, which limits generalizability across healthcare contexts. Furthermore, linguistic diversity poses challenges: the dataset is English-only, and results may not extend to other languages or cultural communication styles. Additionally, doctor–patient interactions may include indirect or nuanced expressions, such as sarcasm or understatement, which current models do not capture. Computationally, resource-intensive models like BERT remain difficult to deploy in resource-constrained healthcare environments without optimization. Methodologically, we mitigated some risks by using stratified five-fold cross-validation and applying oversampling only within each training partition, leaving validation/test portions unchanged; nonetheless, residual class imbalance and limited data remain important constraints. These constraints highlight the need for caution in directly translating findings into practice and pointing toward specific improvement directions.

Beyond technical issues, ethical considerations are central to applying sentiment analysis in healthcare. Although the dataset was anonymized, real-world deployment must

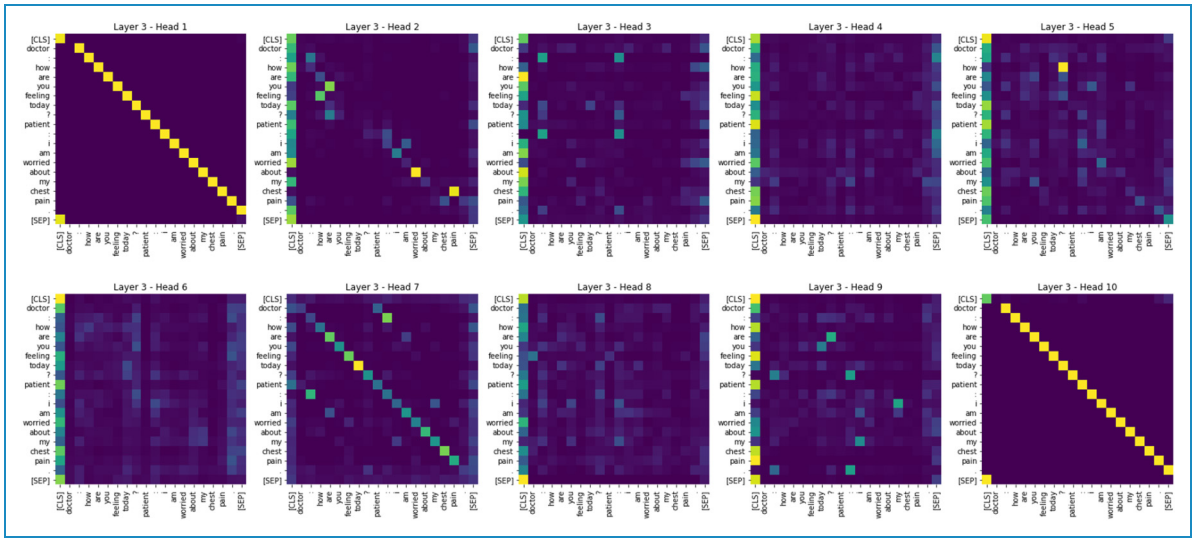

Figure 13. Multihead attention maps from the third layer of BERT for a doctor–patient dialogue. Each heatmap shows one head's attention distribution across tokens, with brighter values indicating stronger focus. BERT: bidirectional encoder representations from transformers.

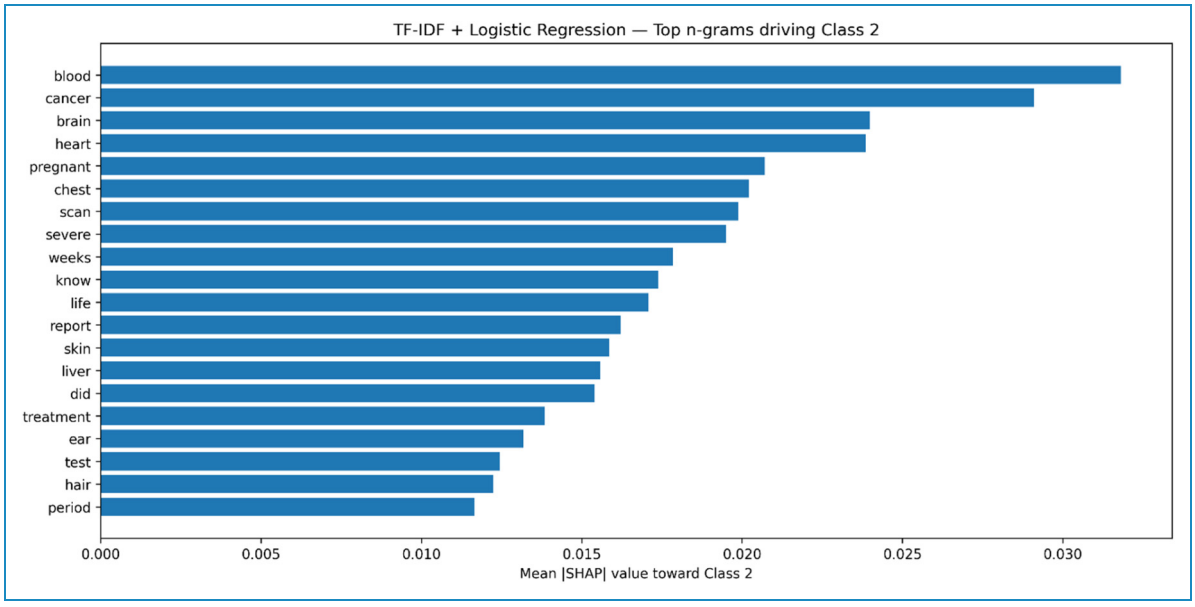

Figure 14. SHAP values for the TF-IDF + LR model, showing the top n-grams driving the classification of high severity (class 2). Higher values indicate stronger contributions toward predicting high severity.

ensure strict privacy protection and compliance with data governance standards. Automated sentiment predictions must not replace human clinical judgment, as misclassifications could seriously affect patient trust and care quality. Addressing these issues requires technical refinements, clear ethical guidelines, and stakeholder engagement.

Future work should expand datasets to include larger, more balanced, and multilingual samples, while also applying multimodal fusion techniques (e.g. speech and visual cues) and advanced data augmentation methods (e.g. back-translation and paraphrasing) to mitigate class imbalance. Exploring oversampling strategies, adaptive loss functions, and regularization techniques will further address uneven performance and reduce overfitting in complex models. Lightweight model compression methods such as pruning, quantization, or knowledge distillation will be critical to enable deployment in clinical environments. Given our results, an ensemble–transformer hybrid that preserves BERT's contextual precision on difficult classes while improving recall (as seen in the ensemble) is a promising

direction. Additionally, comparative studies across diverse healthcare settings would clarify generalizability. Finally, embedding interpretability at the core of model design will remain essential for fostering clinician trust and ensuring safe adoption. Overall, this study provides one of the first systematic comparisons of deep learning, transformer, and ensemble models for sentiment analysis in doctor–patient interactions. By highlighting both the potential and the limitations of these approaches, it contributes to advancing AI methods that are not only accurate but also interpretable, ethically grounded, and clinically relevant.

Conclusion

This study provides one of the first systematic evaluations of DL, transformer, and ensemble models for SA in doctor–patient interactions. Under stratified five-fold cross-validation, the ensemble achieved the strongest overall accuracy (75.5 ± 0.5), outperforming the transformer (BERT, 66.9 ± 0.6), hybrid CNN–LSTM (65.68 ± 0.9), CNN (64.17 ± 0.8), BiLSTM (64.82 ± 0.7), and LSTM (58.66 ± 0.19). Fold-to-fold variability was modest (SDs $\lesssim$1%), supporting the robustness of these findings. The ensemble approach also delivered balanced performance across classes, with improved recall on the difficult low-severity class, while BERT offered higher precision on that class, reflecting complementary strengths. At the same time, BiLSTM and CNN-LSTM proved effective for high-severity cases, and BERT demonstrated strong contextual understanding despite data limitations. By combining rigorous model comparison with interpretability analyses through attention maps and SHAP values, the work advances both methodological transparency and practical relevance. Beyond improving sentiment classification accuracy, the findings show how AI can help clinicians recognize patient distress and tailor communication strategies more effectively. Future research should scale to larger, multilingual, and multimodal datasets; explore ensemble–transformer hybrids to couple BERT's precision with improved recall; and develop lightweight, interpretable models for reliable real-time deployment in clinical practice.

ORCID iD

Rashid Khan https://orcid.org/0000-0002-2410-044X

Informed consent

This study is a computational analysis of an anonymized, publicly available dataset of doctor–patient conversations from *Hugging Face*. It does not involve human participants, animal subjects, or experimental protocols requiring institutional approval or informed consent. All methods comply with relevant guidelines for secondary data analysis.

Contributorship

YG was involved in conceptualization, methodology, writing—review & editing, funding acquisition, and data curation; LD in writing—review & editing, formal analysis, and data curation; BH in writing—review & editing, resources, and supervision; and RK in writing—review & editing, supervision, and project administration.

Funding

The author disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was funded by the Philosophy and Social Sciences in Colleges and Universities of Anhui, China (project number 2023AH050544), the philosophy and Social Science of Anhui, China (project number AHSKQ2022D185), the philosophy and Social Science of Anhui, China, the Philosophy and Social Sciences in Colleges and Universities of Anhui, China (grant number AHSKQ2022D185, 2023AH050544).

Declaration of competing interest

The authors state that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper. Furthermore, no AI tools were used to develop or edit this manuscript.

Data availability

The dataset used in this study is publicly available from the Hugging Face repository at https://huggingface.co/datasets/mahfoos/Patient-Doctor-Conversation.

Permissions for figures

All figures (Figures 1–14) were created by the authors from data analyzed in this study; no third-party materials were used, and no permissions are required.

Guarantor

The guarantors for this manuscript are Rashid Khan, (rashidkhan@sztu.edu.cn), and Yufan Ge, (geyufan@ahmu.edu.cn).

References

1. Alowais SA, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Med Educ* 2023; 23: 689.
2. Maleki Varnosfaderani S and Forouzanfar M. The role of AI in hospitals and clinics: transforming healthcare in the 21st century. *Bioengineering* 2024; 11: 337.
3. Knowles SE, et al. Participatory codesign of patient involvement in a learning health system: how can data-driven care be patient-driven care? *Health Expect* 2022; 25: 103–115.
4. Huang C-W, et al. Emotion recognition in doctor-patient interactions from real-world clinical video database: initial development of artificial empathy. *Comput Methods Programs Biomed* 2023; 233: 107480.

5. Gadri S, et al. Sentiment analysis: developing an efficient model based on machine learning and deep learning approaches. In: *International conference on intelligent computing & optimization*. Cham, Switzerland: Springer, 2021, pp.237–247.
6. Mujahid M, et al. Sentiment analysis and topic modeling on tweets about online education during COVID-19. *Appl Sci* 2021; 11: 8438.
7. Khan R, et al. US based COVID-19 tweets sentiment analysis using textblob and supervised machine learning algorithms. In: *2021 international conference on artificial intelligence (ICAI)*. Islamabad, Pakistan: IEEE, 2021, pp.1–8.
8. Chauhan P, Sharma N and Sikka G. The emergence of social media data and sentiment analysis in election prediction. *J Ambient Intell Humaniz Comput* 2021; 12: 2601–2627.
9. Sauerbrei A, et al. The impact of artificial intelligence on the person-centred, doctor-patient relationship: some problems and solutions. *BMC Med Inform Decis Mak* 2023; 23: 73.
10. Bates DW, et al. The potential of artificial intelligence to improve patient safety: a scoping review. *NPJ Digital Med* 2021; 4: 54.
11. Aljedaani W, et al. Sentiment analysis on Twitter data integrating TextBlob and deep learning models: the case of US airline industry. *Knowl Based Syst* 2022; 255: 109780.
12. Almeida C, et al. Optimizing sentiment analysis models for customer support: methodology and case study in the Portuguese retail sector. *J Theoretical Appl Electr Commerce Res* 2024; 19: 1493–1516.
13. Tan L, et al. Emotional variance analysis: a new sentiment analysis feature set for artificial intelligence and machine learning applications. *PLoS ONE* 2023; 18: e0274299.
14. Ungless EL, Ross B and Belle V. Potential pitfalls with automatic sentiment analysis: the example of queerphobic bias. *Soc Sci Comput Rev* 2023; 41: 2211–2229.
15. Mujahid M, et al. Analyzing sentiments regarding ChatGPT using novel BERT: a machine learning approach. *Information* 2023; 14: 474.
16. Osório L and Fachada N. Patient-centered healthcare: a framework for analyzing patient feedback through sentiment analysis and topic modeling. In: *Doctoral conference on computing, electrical and industrial systems*. Cham, Switzerland: Springer, 2024, pp.152–163.
17. Reshi AA, et al. COVID-19 vaccination-related sentiments analysis: a case study using worldwide twitter dataset. *Healthcare* 2022; MDPI10,3: 411–438.
18. Wu J, et al. A sentiment analysis driven method based on public and personal preferences with correlated attributes to select online doctors. *Appl Intel* 2023; 53: 19093–19114.
19. Chatzimina ME, et al. Topic modeling and sentiment analysis of Greek clinician–patient conversations in hematologic malignancies. *Int J Med Inf* 2025; 204: 106071.
20. Sen T, et al. Modeling doctor-patient communication with affective text analysis. In: *2017 seventh international conference on affective computing and intelligent interaction (ACII)*. San Antonio, TX: IEEE, 2017, pp.2156–8111.
21. Nguyen K-N, et al. Sentiment reasoning for healthcare. *arXiv Preprint* 2024; arXiv:2407.21054.
22. Rui JR, Guo J and Yang K. How do provider communication strategies predict online patient satisfaction? A content analysis of online patient-provider communication transcripts. *Digital Health* 2024; 10: 20552076241255617.
23. Wang Z, et al. Commsense: a wearable sensing computational framework for evaluating patient-clinician interactions. *Proc ACM Human-Computer Interaction* 2024; 8: 1–31.
24. Greaves F, et al. Use of sentiment analysis for capturing patient experience from free-text comments posted online. *J Med Internet Res* 2013; 15: e2721.
25. Gohil S, Vuik S and Darzi A. Sentiment analysis of health care tweets: review of the methods used. *JMIR Public Health Surveillance* 2018; 4: e5789.
26. Alemi F, et al. Feasibility of real-time satisfaction surveys through automated analysis of patients' unstructured comments and sentiments. *Qual Manag Healthcare* 2012; 21: 9–19.
27. Khanbhai M, et al. Applying natural language processing and machine learning techniques to patient experience feedback: a systematic review. *BMJ Health Care Inform* 2021; 28: e100262.
28. Cao P, et al. A multimodal depression consultation dataset of speech and text with hamd-17 assessments. *Sci Data* 2025; 12: 1577.
29. Wu S, Hsu W and Lee M-L. EHDChat: a knowledge-grounded, empathy-enhanced language model for healthcare interactions. In: *Proceedings of the second workshop on social influence in conversations (SICon 2024)* Miami, FL: Association for Computational Linguistics, 2024, pp.141–151.
30. Sadeghi M, et al. Harnessing multimodal approaches for depression detection using large language models and facial expressions. *NPJ Mental Health Res* 2024; 3: 66.
31. Zhang X, et al. Optimizing depression detection in clinical doctor-patient interviews using a multi-instance learning framework. *Sci Rep* 2025; 15: 6637.
32. Al-Selwi SM, et al. RNN-LSTM: from applications to modeling techniques and beyond—systematic review. *J King Saud Univ-Computer Inform Sci* 2024; 36: 102068.
33. Siami-Namini S, Tavakoli N and Namin AS. The performance of LSTM and BiLSTM in forecasting time series. In: *2019 IEEE international conference on big data (Big Data)*. Los Angeles, CA: IEEE, 2019, pp.3285–3292.
34. Tassopoulou V, Retsinas G and Maragos P. Enhancing handwritten text recognition with n-gram sequence decomposition and multitask learning. In: *2020 25th International Conference on Pattern Recognition (ICPR)*. Milan, Italy: IEEE, 2021, pp.10555–10560.
35. Rehman AU, et al. A hybrid CNN-LSTM model for improving accuracy of movie reviews sentiment analysis. *Multimed Tools Appl* 2019; 78: 26597–26613.
36. Rane N, Choudhary SP and Rane J. Ensemble deep learning and machine learning: applications, opportunities, challenges, and future directions. *Stud Medical Health Sci* 2024; 1: 18–41.

37. Chai CP. Comparison of text preprocessing methods. *Nat Lang Eng* 2023; 29: 509–553.
38. Fritz M. Decision tree classification with missing values. Technische Universität Wien 2023.
39. Dotan E, et al. Effect of tokenization on transformers for biological sequences. *Bioinformatics* 2024; 40: btae196.
40. Dang Y, et al. Repeated padding for sequential recommendation. Proceedings of the 18th ACM Conference on Recommender Systems 2024.
41. Bej S, et al. LoRAS: an oversampling approach for imbalanced datasets. *Mach Learn* 2021; 110: 279–301.
42. Thabtah F, et al. Data imbalance in classification: experimental evaluation. *Inf Sci (Ny)* 2020; 513: 429–441.
43. Khandelwal R and Deshmukh S. Towards addressing bias and fairness in machine learning. *PIJET J* 2024; 3: 100–109.
44. Pedregosa F, et al. Scikit-learn: machine learning in Python. *J Machine Learning Res* 2011; 12: 2825–2830.
45. Gandhi UD, et al. Sentiment analysis on twitter data by using convolutional neural network (CNN) and long short term memory (LSTM). *Wirel Pers Commun* 2021; 122: 1–10.
46. Ma M, et al. Predicting machine's performance record using the stacked long short-term memory (LSTM) neural networks. *J Appl Clin Med Phys* 2022; 23: e13558.
47. Anh DT, et al. Effect of gradient descent optimizers and dropout technique on deep learning LSTM performance in rainfall-runoff modeling. *Water Resour Manage* 2023; 37: 639–657.
48. Nabil A, Seyam M and Abou-Elfetouh A. Prediction of students' academic performance based on courses' grades using deep neural networks. *IEEE Access* 2021; 9: 140731–140746.
49. Hameed Z and Garcia-Zapirain B. Sentiment classification using a single-layered BiLSTM model. *IEEE Access* 2020; 8: 73992–74001.
50. Liu Y, Pu H and Sun D-W. Efficient extraction of deep image features using convolutional neural network (CNN) for applications in detecting and analysing complex food matrices. *Trends Food Sci Technol* 2021; 113: 193–204.
51. Al-Selwi SM, et al. LSTM inefficiency in long-term dependencies regression problems. *J Adv Res Appl Sci Eng Technol* 2023; 30: 16–31.
52. Mohammed AH and Ali AH. Survey of BERT (bidirectional encoder representation transformer) types. *J Phys Conf Ser* 2021; 1963: 012173–012178.
53. Hsu C-Y, et al. The impact of AI-driven sentiment analysis on patient outcomes in psychiatric care: a narrative review. *Asian J Psychiatr* 2025; 107: 104443.