IBIMA Publishing
Communications of International Proceedings
https://ibimapublishing.com/p-articles/47AI/2026/4714926/
Vol. 2026 (12), Article ID 4714926

Comparative Evaluation of Deep Learning and Transformer-Based Models for Depression Detection Using Clinical Interview Transcripts*

Sara ALMARZOOQI and Saadat ALHASHMI

University of Sharjah, Sharjah, UAE

Correspondence should be addressed to: Sara ALMARZOOQI, U23107199@sharjah.ac.ae

*Presented at the 47th IBIMA International Conference, 29-30 June 2026, Madrid, Spain

Abstract

The article is a comparative essay on the traditional deep learning structures, i.e., Convolutional Neural Network (CNN), Bidirectional Long Short-Term Memory (BiLSTM), and Bidirectional Gated Recurrent Unit (BiGRU) as opposed to a transformer-based model (DistilBERT + Logistic Regression) regarding the automated detection of depression on the clinical interview transcripts. The controlled and standardized experiments are conducted using the Distress Analysis Interview Corpus- Wizard of Oz (DAIC-WOZ) data to be able to fairly compare the results. Findings indicate that DistilBERT + Logistic Regression is better than all deep learning baselines (F1 = 0.9506), whereas classical models are also competitive and less costly to run. The comparison of the transformer advantage with a significance test (McNemar test) shows that the transformer advantage cannot be explained by the chance ($p < 0.05$). The article describes the trade-offs between predictive accuracy and computational efficiency, ethical issues, and implications of the use of AI-based decision-support in digital mental health information systems.

Keywords: Depression detection, deep learning, transformer models, DistilBERT, DAIC-WOZ, clinical NLP, digital mental health.

Introduction

Depression is one of the most widespread mental health issues globally, where its serious consequences are directed at the health of individuals, their performance, and spending on health (WHO, 2023). Its classical diagnosis is informed by structured clinical interviews and self-report scales that are valid (e.g., PHQ-8/PHQ-9), which is resource-intensive, is subjective in nature, and cannot be conducted in low-resource settings of remote areas. The emergence of artificial intelligence (AI) and natural language processing (NLP) creates a possibility to automate aspects of the screening process, therefore, making scalable and cost-effective early detection possible.

It has been proven in clinical research that depressive symptomatology is manifested in the language patterns, conversational dynamics, and the tone of emotions (Gratch et al., 2014). Text-based depression classifiers can be trained and benchmarked on large text-based annotated corpora such as DAIC-WOZ using validated labels in terms of PHQ-8 scores. Text-only systems offer a viable low resource alternative to multimodal systems and can be integrated with electronic health records (EHR), telehealth and intelligent decision-support systems.

None of the previous studies have conducted a comprehensive and rigorous comparison between classical-deep learning structures and transformer-based models when applied to text transcripts of the text set of the DAIC-WOZ, which is subject to the same experimental conditions. Majority of literature uses a single model family or integrates modalities (audio, video, text) and it is not easy to identify the role played by textual features. The gap which is bridged in this paper is that of the research.

Cite this Article as: Sara ALMARZOOQI and Saadat ALHASHMI Vol. 2026 (12) "Comparative Evaluation of Deep Learning and Transformer-Based Models for Depression Detection Using Clinical Interview Transcripts " Communications of International Proceedings, Vol. 2026 (12), Article ID 4714926, https://doi.org/10.5171/2026.4714926

Research Questions

- RQ1: How effectively can deep learning models detect depression from clinical interview transcripts?
- RQ2: Do transformer-based models outperform classical deep learning models on this task?
- RQ3: What are the computational trade-offs between model families?
- RQ4: How do results compare with prior DAIC-WOZ studies?

Related Work

AI-Based Depression Detection

Mental health analytics uses language, sound and behavioral cues to detect depression trends. Supervised models are divided into depressed/non-depressed, and text-based models are increasingly popular because they are scalable and can be used in digital health applications (Burdisso et al., 2019; Ma et al., 2016).

Deep Learning for Text Classification

CNNs are also effective in capturing the local n-gram patterns hence suitable in sentence-level classification. BiLSTM and BiGRU are recurrent models, and dependencies are captured in both directions, and the vanishing gradient effect is alleviated, and LSTM cells are simpler than GRU, just a simplified version of the gating mechanism (Hochreiter and Schmidhuber, 1997; Goodfellow et al., 2016). Such architectures are effective in identifying emotional context in clinical transcripts and are not effective in identifying long-range global dependencies.

Transformer Models

Transformers employ self-attention to acquire global contextual relations. Bert and the smaller variant, DistilBERT, generate quality contextual representations, which can be mapped to downstream tasks (Devlin et al., 2019; Sanh et al., 2020). Transformers have a special affinity to negation and sarcasm as well as to contextual semantics that are significant in a clinical interview, and to higher computational demands.

DAIC-WOZ and Research Gap

DAIC-WOZ is the most popular depression detection benchmark (Gratch et al., 2014). The other systems that are used in the past, like LSTM + audio hybrids (Al Hanai et al., 2018), CNN-RNN models (Ma et al., 2016), and graph convolutional networks (Burdisso et al., 2019) either accept multimodal input or have completely dissimilar preprocessing, which complicates cross-study comparisons. This piece bridges that gap with a text based and multi-model test that is standardized.

Methodology

Dataset

DAIC-WOZ is semi-structured PHQ-8 annotated clinical interviews. Depression (PHQ-8 > 10) resulted in 61 depressed (32%) and 128 non-depressed (68) individuals of 189 participants, which is an imbalance in the classes that can be handled with weighted loss functions during training.

Table 1: DAIC-WOZ Dataset Summary

Attribute	Value
Total Participants	189
Depressed (PHQ-8 > 10)	61 (32%)
Non-Depressed	128 (68%)
Labeling	PHQ-8 Score
Modality Used	Text Transcripts Only

Attribute	Value
Interview Type	Semi-Structured Clinical

Preprocessing

CNN, BiLSTM, BiGRU pipelines: lowercasing, special characters removal, stop-words removal, sequence tokenization, sequence padding to a constant length. DistilBERT: DistilBERT is an auto-regressive no-preprocessing tokenizer, i.e. it does not decontextualize to create input IDs and attention mask. This offers a reasonable comparison to all the models which have their optimum input conditions.

Model Architectures

- CNN: Multi-filter convolutional layers so as to extract local n-gram patterns, max-pooling and fully connected binary classification head.
- Bidirectional LSTM learning Vanishing gradients LSTMs Vanishing gradients LSTM Attend to long range sequence dependencies with BiLSTM.
- GRU (bidirectional) with gating reduction BiLSTM It is not as costly to learn and train as the BiGRU is.
- DistilBERT + LR: The contextual embeddings of the pre-trained DistilBERT are fed into a Logistic Regression classifier- trade-offs between performance, interpretability and computation cost.

Experimental Setup

• Train/test split: 80/20; hyperparameter tuning on a held-out validation set.

• Class imbalance: Weighted loss functions to avoid biased predictions toward the majority class.

Model	Train Time (min)	Inference (ms/sample)	Memory (GB)
CNN	12	2	1.2
BiLSTM	25	3	1.5
BiGRU	20	2.5	1.4
DistilBERT + LR	75	5	4.0

• Cross-validation: 5-fold to ensure statistical robustness.

• Reproducibility: Fixed random seeds across all experiments.

Table 2: Computational Requirements per Model

Evaluation Metrics

Accuracy, precision, recall, and F1-score are all evaluated in all models. In a clinical screening setting, recall is favored due to its greater cost incurred due to a false negative (overlooking a depressed instance) as compared to a false positive. The test developed by McNemar is used to determine the difference between model pairs at the 0.05 level of statistical significance.

Results

Model Performance

Table 3: Model Performance on DAIC-WOZ Test Set (Best result in bold shading)

Model	Accuracy	Precision	Recall	F1-Score
CNN	0.9274	0.9510	0.9274	0.9391
BiGRU	0.9254	0.9509	0.9254	0.9380
BiLSTM	0.9254	0.9509	0.9254	0.9380
DistilBERT + LR	0.9496	0.9516	0.9496	0.9506

DistilBERT + LR scores the best on all the metrics. Contextual embeddings enable the model to acquire delicate language cues of structured clinical language. Despite being the quickest model to train, CNN has the second-best F1-score. BiLSTM and BiGRU are equal in terms of score, and this implies that they have similar sequential modeling capabilities in this problem.

Confusion Matrix Analysis

Clinically significant differences are found in the analysis of confusion matrices. DistilBERT + LR produces only 3 false negative (false depressed cases) compared to 1011 false negative with deep learning models. False negativity is a serious issue in clinical screening where a misdiagnosis is a risk to the safety of the patient. Class-weighted training has the benefit of reducing the biasness of prediction to most non-depressed classes.

Table 4: Confusion Matrix Summary (Depressed-class predictions)

Model	True Positive	False Negative	False Positive
CNN	45	10	12
BiLSTM	44	11	14
BiGRU	44	11	13
DistilBERT + LR	~58	~3	~5

Comparison with Prior Work

Table 5: Comparison with Prior DAIC-WOZ Studies (NR = Not Reported)

Study	Model	Acc.	Prec.	Rec.	F1
Al Hanai et al. (2018)	LSTM + Audio/Text	NR	NR	NR	~0.84
Ma et al. (2016)	CNN-RNN	NR	NR	NR	~0.87
Burdisso et al. (2019)	GCN (Social Media)	NR	NR	NR	~0.90
This Study	DistilBERT + LR	0.9496	0.9516	0.9496	0.9506

Direct numerical comparisons are not to be taken seriously: the preceding literature is inconsistent in the input modality, preprocessing and datasets division. It would be worth adding that the multifaceted systems (audio + text) are already at an advantage when it comes to the information. Our text-only DistilBERT + LR system with F1 = 0.9506 is a good starting point of text-based and repeatable text screening of depression.

Statistical Significance

Model Pair	McNemar's p-value	Significant (α=0.05)?
DistilBERT + LR vs. BiLSTM	0.032	Yes
DistilBERT + LR vs. BiGRU	0.028	Yes
BiLSTM vs. BiGRU	0.842	No

Table 6: McNemar's Statistical Significance Tests

The statistical difference between the performance of the transformer model and the recurrent architectures is statistically significant ($p < 0.05$). The statistical significance of BiLSTM and BiGRU ($p = 0.842$) shows that the same F1 scores are not due to the coincidence of the same architecture.

Discussion

Model Trade-offs

The combination of DistilBERT and LR is always the most effective in predictive tasks, especially when remembering the most clinically important one. Its contextualizations assimilate negation, semantic subtlety and long-range interactions of structured clinical speech. Nonetheless, its training duration of 75-minute and memory footprint of 4 GB are 6 times higher in the resource consumption of CNN. Classical models are a possible alternative in resource-constrained applications: CNN converges in 12 minutes at 1.2 GB and has F1 = 0.9391- a fine of just 0.012 as compared to DistilBERT + LR.

Information Systems Implications

Concerning the IS design, the architecture should be determined by the context of the deployment: DistilBERT + LR when it is deployed to a cloud-hosted EHR and telehealth platforms and can access the computational resources; CNN or BiGRU when deployed to a mobile or low-bandwidth environment and requires fast inference. The hybrid of the DistilBERT and LR (logistic regression on top of contextual embeddings) is more interpretable compared to full-trained transformers, both in clinical trust and regulatory compliance.

Ethical Considerations

The problem of the depression screening assisted by the AI is quite grave on the issue of the ethics. Privacy: mental health information privacy should be highly ensured when it comes to confidentiality. Interpretability: opaque black-box models cannot be applied to clinical practice; the LR classifier based on DistilBERT embeddings is partially handling this problem. Bias: demographic underrepresentation in DAIC-WOZ cannot be cross-culturally generalized; intra-dataset imbalance can be alleviated by class-weighted training, but not population bias. Possibility of false-negative: false-negatives of depression are clinically detrimental; the transformer is cheaper but has lower false-negative. During the deployment, accountability should always be taken by adhering, reporting and monitoring model constraints to the clinicians.

limitations and future work

Certainly, there exist several limitations that are to be mentioned. DAIC-WOZ is an interview on a professional level, and can only be applied in the instance of informal care (social media, free-form telehealth chat). There is low population diversity as it impacts on cross-cultural validity. There were no audio and visual modalities, and multimodal fusion may be applied to increase the accuracy even more. Future research should: (1) include multimodal inputs; (2) investigate explainable AI (XAI) techniques to uncover linguistically-motivated clinical observations; (3) investigate domain adaptation and cross-lingual transfer to increase the scope of generalizability; and (4) conduct formal fairness audits to demographic subgroups.

Conclusion

This paper reported a controlled comparative analysis of CNN, BiLSTM, BiGRU, and DistilBERT + Logistic Regression on the detection of depression based on transcripts of clinical interviews. DistilBERT + LR had the best performance (Accuracy = 0.9496, F1 = 0.9506) and had much lower numbers of false negatives (vital clinical safety factor). Deep learning models based on classical methods can still be used especially in situations where there are restriction of computational resources. The paper shows that text-based NLP systems may produce high-quality depression screening performance, which adds a replicable benchmark and evidence-based recommendations to AI incorporation into digital mental health information systems. Multimodality, explainability, and cross-cultural validation should be discussed in future work and maximize the impact in the real world.

References

- Al Hanai, T., Ghassemi, M. and Glass, J. (2018) 'Detecting depression with audio/text sequence modeling of interviews', *Proceedings of Interspeech 2018*, pp. 1716–1720. Available at: https://doi.org/10.21437/Interspeech.2018-2522
- Burdisso, S.G., Errecalde, M. and Montes-y-Gómez, M. (2019) 'A text classification framework for simple and effective early depression detection over social media streams', *Expert Systems with Applications*, 133, pp. 182–197. Available at: https://doi.org/10.1016/j.eswa.2019.05.023
- Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2019) 'BERT: Pre-training of deep bidirectional transformers for language understanding', *Proceedings of NAACL-HLT 2019*, pp. 4171–4186. Available at: https://doi.org/10.18653/v1/N19-1423
- Goodfellow, I., Bengio, Y. and Courville, A. (2016) *Deep Learning*. Cambridge, MA: MIT Press.
- Gratch, J., Artstein, R., Lucas, G.M. *et al.* (2014) 'The Distress Analysis Interview Corpus of human and computer interviews', *Proceedings of LREC 2014*, pp. 3123–3128.
- Hochreiter, S. and Schmidhuber, J. (1997) 'Long short-term memory', *Neural Computation*, 9(8), pp. 1735–1780.
- Kroenke, K., Steer, R.A. and Lowe, B. (2009) 'The Patient Health Questionnaire somatic, anxiety, and depressive symptom scales: A systematic review', *General Hospital Psychiatry*, 31(4), pp. 309–325.
- LeCun, Y., Bengio, Y. and Hinton, G. (2015) 'Deep learning', *Nature*, 521(7553), pp. 436–444.
- Ma, X., Yang, H., Chen, Q. and Huang, J. (2016) 'Depression detection based on deep learning using text data', *Proceedings of ICCI*, pp. 1–8.
- Sanh, V., Debut, L., Chaumond, J. and Wolf, T. (2020) 'DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter', *Proceedings of Energy Efficient ML and Cognitive Computing Workshop*, pp. 1–5.
- Vaswani, A., Shazeer, N., Parmar, N. *et al.* (2017) 'Attention is all you need', *Advances in Neural Information Processing Systems*, pp. 5998–6008.
- World Health Organization (2023) *Depression*. Available at: https://www.who.int/news-room/fact-sheets/detail/depression (Accessed: 7 April 2026).