Journal of Computer Science

Research Article

Deep Learning Analysis of Drug Reviews: Identifying Patient Conditions, Predicting Drug Ratings, and Sentiment Analysis

Rafat Hammad, Mohammad Ashraf Ottom and Ayah Bani Baker

Department of Information Systems, Faculty of Information Technology and Computer Sciences, Yarmouk University, Irbid, Jordan

Article history
Received: 15-03-2026
Revised: 10-07-2026
Accepted: 07-08-2026

Corresponding Author:
Rafat Hammad
Department of Information Systems, Faculty of Information Technology and Computer Sciences, Yarmouk University, Irbid, Jordan
Email: rafat.hammad@yu.edu.jo

Abstract: Understanding patients' experiences with medications is essential for improving drug effectiveness, detecting potential side effects, and enhancing overall patient satisfaction. This research investigates applying machine learning to patient drug reviews across three main tasks: Identifying patient conditions, predicting drug ratings, and performing sentiment analysis. We implemented and evaluated Convolutional Neural Networks (CNNs), Artificial Neural Networks (ANNs), and Recurrent Neural Networks (RNNs) using a large dataset of patient reviews collected from reputable medical platforms. Model performance was assessed using accuracy, precision, recall, and F1-score for classification tasks, Mean Squared Error (MSE), and coefficient of determination (R^2) for rating prediction. For patient condition identification, the RNN achieved the highest accuracy of 96.50%, followed by the ANN (96.30%) and CNN (95.80%). In the drug rating prediction task, the CNN outperformed the other models, achieving an MSE of 0.0296 and an R^2 of 0.9692. For sentiment classification, the RNN achieved the best performance with an accuracy of 94.82%. Cross-validation results demonstrated the robustness and stability of all models, with standard deviations below 0.3%. The findings indicate that deep learning models can effectively extract meaningful insights from unstructured patient-generated reviews. Moreover, the results suggest that model effectiveness is task-dependent: RNNs excel at tasks that require sequential text understanding, such as condition identification and sentiment classification, while CNNs perform better at predicting drug ratings through efficient local feature extraction.

Keywords: Deep Learning, Natural Language Processing, Drug Review Analysis, Sentiment Analysis, Patient Experience, CNN, RNN, ANN

Introduction

Online health platforms have emerged as the primary source of health advice information, with patients visiting websites such as WebMD to learn about drugs and discuss their uses. Usually, many people write reviews about the drugs, discussing the side effects and their overall satisfaction with the medicines. Studying the reviews of these drugs could offer insights to healthcare practitioners, researchers, and pharmaceutical firms on what to expect from a drug's performance in the real world (Mulkalwar et al., 2025).

The rapid growth of user-generated content has created new opportunities in the healthcare sector for advanced analytics using Natural Language Processing (NLP) and deep learning. These techniques provide valuable insights from unstructured data and support tasks such as classifying medical conditions, predicting patient drug ratings, and determining patient sentiment (Sorayaie Azar et al., 2024; Liu et al., 2020). Such analysis could also be applied to support evidence-based reasoning in the medical field or to further personalize treatments for individual patients based on a particular medication (Omranian et al., 2023).

Deep learning offers several benefits over traditional statistical approaches because it can handle large-scale computations that other methods can't, and provides rapid progress through the entire workflow. Models like CNNs, RNNs, and ANNs became notable for processing large amounts of data and recognizing subtle linguistic signals-potentially symptoms, treatment response, and emotional tone (Wei et al., 2020; Devlin et al., 2019). Patient-

reported symptoms such as "severe joint pain" and "persistent headache" may indicate underlying diseases and provide valuable signals for automated disease classification. Text-based sentiment analysis can be utilized to predict users' drug ratings, providing insights into their acceptance of drug treatments as well as their perceptions of the drug's perceived effectiveness.

Apart from the prediction model, understanding of drug–condition relationship, rating distributions, and sentiment trends will enable clinicians and researchers to discover trends that are not clear from the text summary. These perspectives provide new evidence for evaluating the effectiveness of drug therapies across diverse demographic groups, for identifying common side effects, and for integrating these insights into clinical or policy decisions (Vijayaraj et al., 2024).

The incorporation of deep learning into drug reviews is anticipated to be one of the most significant advancements in health analytics, given the opportunities it would bring. This is a great way to understand what patients are actually feeling, so that practical tools could be developed for use in making clinical decisions in drug development as well as patient-centered treatment.

Patient reviews are an excellent source of information about the outcomes of the patients' treatment and their experience with healthcare. The volume and diversity of these reviews, however, make them difficult to analyze. Traditional analytical methods are often ineffective at handling and understanding this textual data and fail to provide meaningful patterns and insights that can guide action. Therefore, advanced machine learning techniques are required to actually support operations, including medical condition classification, drug rating prediction, and sentiment analysis of patient reviews. The main objective of this study was to address the above concerns by leveraging the potential of deep learning models for drug review analysis, focusing on three core tasks: Identifying patient conditions, predicting drug ratings from review content, and performing sentiment analysis of patient reviews.

Literature Survey

Using deep learning has now become crucial for sifting through and analyzing patients' own drug reviews to generate automated representations of patient conditions, predict drug ratings, and classify sentiment features. Identifying diseases, symptoms, and medical concepts in reviews is typically handled using Named Entity Recognition (NER). NER is an NLP technique that automatically identifies and classifies key information (entities) in unstructured text into predefined categories, such as names, organizations, locations, dates, and monetary values. Transformer models such as BioBERT (Lee et al., 2020), SciBERT (Beltagy et al., 2020), BlueBERT (Kormilitzin et al., 2021), Med7 (Huang et al., 2020), and ClinicalXLNet (Ling, 2023) consistently achieve state-of-the-art F1 scores of 91%-93%. These models are pre-trained separately on domain-specific corpora, such as PubMed abstracts, clinical notes, and scientific literature, thereby acquiring deep knowledge of biomedical terminology and informal patient language. Their strong contextual representations make them particularly effective for extracting medical entities from unstructured drug reviews.

Drug rating prediction is considered a text-based regression problem. Previous work has considered different approaches based on CNNs and RNNs for analysis. Transformer models have shown superior efficacy, as evidenced by recent work, with fine-tuned BERT achieving an MAE of 0.43 in star category prediction (Tran et al., 2021). Coupling CNNs, RNNs, and BERT has improved progress toward a Mean Absolute Error (MAE) of 0.38 (Raffel et al., 2020). They improve the performance of more recent language models, such as T5 (Brown et al., 2020) and GPT-3 (Sun et al., 2019), further enhancing MAE's accuracy to 0.35-0.37, thereby promoting large-scale text-to-text and generative transformers.

Sentiment analysis is a valuable tool for estimating patient satisfaction and the drug's effectiveness. BERT-based models have shown outstanding performance compared to classical machine learning and any earlier deep learning methods. According to Lan et al. (2020), the accuracy scores were 90.5% for BERT and 92.3% for ROBERTa. Efficient versions such as ALBERT and DistilBERT maintain model performance in the 90%-92% range while reducing model size and computational cost, making them deployable at scale.

The increasing availability of advanced computational techniques has enhanced the significance of deep learning techniques in drug discovery and has made them enormously impactful on clinical applications and analytic approaches in healthcare. Experiments with deep learning and knowledge-graph enhanced models have shown promising results in predicting drug–target interactions, drug–drug interactions, and identification of potential drugs for repurposing (Askr et al., 2023). Together, these works are really indicative of how deep learning has spread in biomedicine. Still, many aspects need to be addressed in terms of data quality, interpretability of the model, and generalization in the domain.

The existing works are concentrated in particular applications such as disease diagnosis, drug rating, sentiment analysis, etc. In this work, these three tasks are intertwined in a single framework by proposing advanced transformer-based models, preprocessing standardization, and hyperparameter optimization at the task level, along with cross-validation procedures. The fundamental idea of the study is to develop a deep learning model that uses large-scale data input to make

predictions, derive insights, and conduct sentiment analysis on patients' feedback regarding drugs.

Materials and Methods

Dataset Description

This study explores patient reviews using deep learning methods, including sentiment analysis, drug ratings projections, and medical condition identification. The raw data came from public data made available by Kaggle in three such datasets that had been collected from the reviews of patients available on WebMD. There are three data sets on WebMD relating to the medications for diabetes, hypertension, and psychiatric disorders. These are WebMD Reviews for Diabetes Drugs (2023), WebMD Reviews for Hypertension Medications (2023), and WebMD Reviews for Psychiatric Medications (2023). Each patient review includes information about the drug name, the treated medical condition, and several evaluation metrics such as overall rating, effectiveness, ease of use, and satisfaction. These are large-scale and diverse datasets of textual and numerical information that reflect real-world patient experiences, treatment success or failure, and views of the treatment as perceived across a wide variety of diseases. Because of its richness and variety, the datasets are well-suited for the development and evaluation of deep learning models for sentiment analysis, drug rating prediction, or medical condition classification. Table 1 contains detailed source information and references for each dataset used for this study.

The data preprocessing is quite an important step during the preparation of textual data for the modeling step. It is done to eliminate the irrelevant data and to standardize the structure of the text; only the terms and expressions with medical relevance will be included. The initial process in the pipeline was to tokenize the text and then convert all words in the text to lowercase and remove stop words. Following this, the rest of the words were lemmatized to produce their stem form in order to ensure consistency in the language as well as to alleviate the sparsity of words in the vocabulary. The preprocessed and cleaned text was then mapped to numerical vectors through a known word-embedding model, Word2Vec, to preserve the semantic information of words and medical terms. The cleaned and preprocessed text was then translated into numerical representations, which allowed us to capture semantic relationships between words and medical concepts.

The preprocessing process filtered the reviews to 82,250 from 107,000, as shown in Figure 1. The obtained data set was then randomly split into two distinct groups a training set (80%, 65800 reviews) and a test set (20%, 16450 reviews). This data set was then randomly split into a training set (80% - 65,800 reviews) and a testing set (20% - 16,450 reviews). All experiments were run with the same random seed of 42, to ensure reproducibility. Furthermore, the division between the training and validation sets remained constant throughout the modeling and testing process, allowing for an impartial and fair comparison of all the approaches explored.

Model Selection Rationale

For the purpose of solving the pointed out research objectives, three different architectures of deep learning techniques have been developed and tested as follows: Convolutional Neural Network (CNN), Artificial Neural Network (ANN), and Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM) units. They are selected for their supplementing role in representing and learning textual data:

- CNNs can be particularly useful for learning more localized semantic or syntactic properties, such as a phrase like "not effective", which may be excellent indicators of the sentiment of the patients or their medical condition. They possess solid pattern sense and are able to recognize patterns, although words may not be in the same order in each text
- RNNs with LSTMs are designed to capture the temporal structure and long-distance relationships in sequential data, such as text. This feature is particularly useful for patient quality statements as experiences and feelings can change throughout many clauses/sentences (such as: "e.g., "I started feeling better after two weeks, but then the side effects began")

Table 1: Raw data sources

Disease Category	Kaggle Dataset Name	URL	# reviews
Diabetes	WebMD Reviews for Diabetes Drugs	https://www.kaggle.com/datasets/sepidehparhami/webmd-reviews-for-diabetes-drugs	14,000
Hypertension	WebMD Reviews for Hypertension Drugs	https://www.kaggle.com/datasets/sepidehparhami/webmd-reviews-for-hypertension-drugs	32,000
Psychiatric	Psychiatric Drug WebMD Review	https://www.kaggle.com/datasets/sepidehparhami/psychiatric-drug-webmd-reviews	61,000
Total			107,000

DOI: 10.3844/jcssp.2026.2692.2699

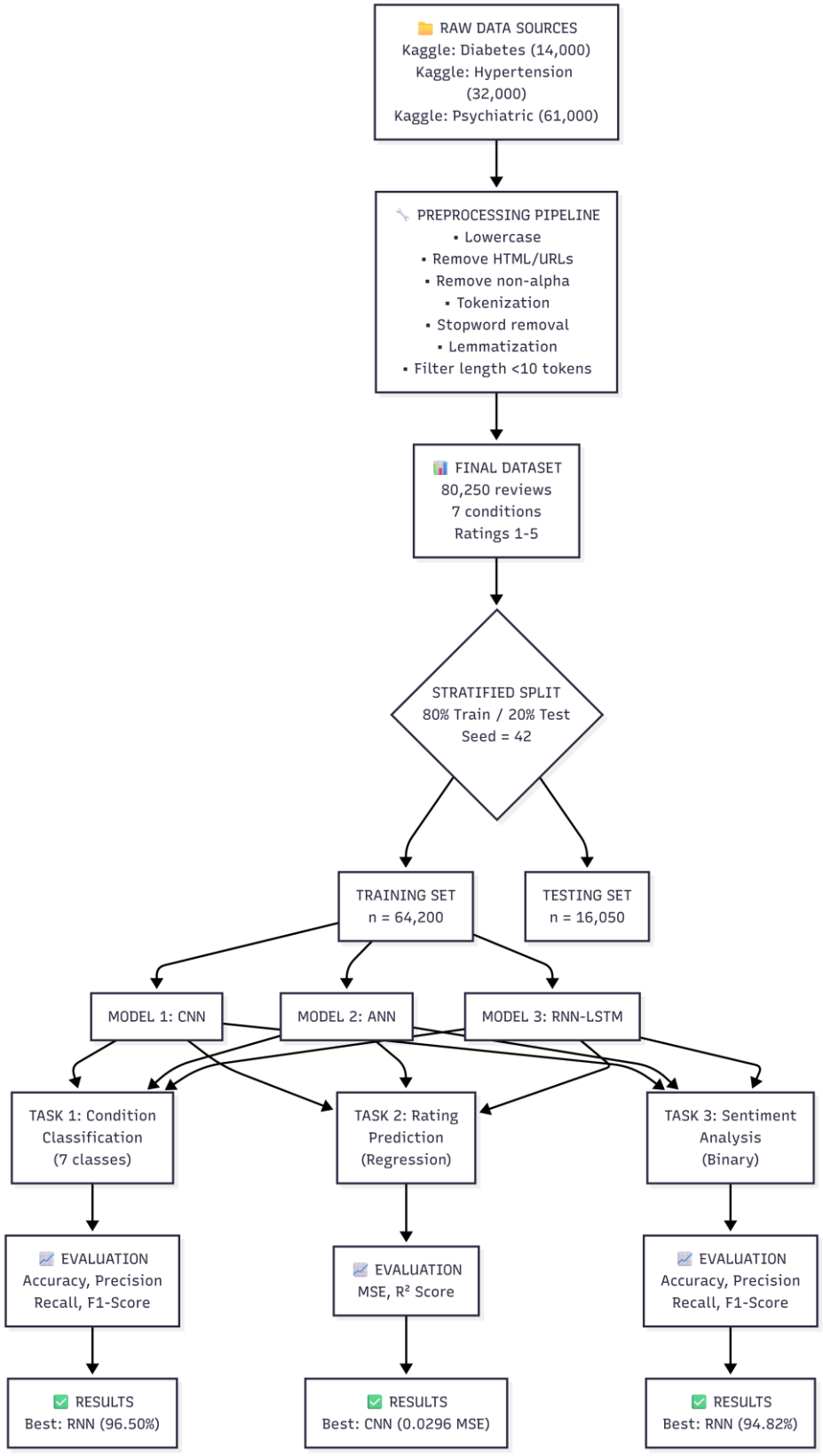

Fig. 1: Overall Methodology Framework

- ANNs, in contrast, do not explicitly model sequential information. Rather, they are trained on a set of nonlinear relationships between the input features, as a good benchmark for evaluating the benefits of local pattern extraction provided by CNNs and sequential context modeling offered by RNNs

The comparison of these three architectures provides a systematic analysis of the importance of the local features extracted by CNN, its sequential context by RNN-LSTM, and its interactions with other nonlinear features by ANN, for each of the prediction problems studied in this article.

Hyperparameter Sensitivity and Regularization

In order to increase the model robustness and ability to generalize, several major hyperparameters were tuned in a systematic manner, which include learning rates (1×10^{-5}, 1×10^{-4}, and 1×10^{-3}); batch sizes (16, 32, and 64); embedding dimensions (100, 200, and 300); and LSTM unit size (64, 128, and 256). Early stopping based on validation loss with a patience value of 10 epochs was applied to prevent overfitting. Dropout was used at rates of 0.3 and 0.5 across neural network layers to boost generalization performance. The best setting of all the hyperparameters was reached by further determining the maximum achieved validation performance during the training period.

Experimental Environment

All experiments were done on Python in Google Colab with its GPU acceleration (NVIDIA T4). The latest versions of TensorFlow and Keras were used for the implementation of the deep learning models, whereas scikit-learn was used for the evaluation support. Text processing was done with NLTK and spaCy; Matplotlib displayed the visual output.

Results and Discussion

A performance benchmark in this study compares the models over three main aspects: patient conditions, drug rating prediction, and sentiment analysis. From our learning curves and other results on the three selected deep learning algorithms - CNNs, ANNs, and RNNs - we show that each model has some benefits and limitations, which are certainly task-oriented and related to language detail.

Patient Condition Identification

Notably, all models have shown continuous improvement in the accurate classification of disease conditions in patient reviews during five rounds of training. The CNN model emerged as a powerful learner, with an accuracy of 95.80%, high precision measured at 95.10%, and recall (95.60%), which resulted in an F-score of 95.35%. As illustrated in Fig. 2, the growth in the training and test curves increased accuracy as convergence approached, minimizing overfitting.

The RNN is seen as performing best among the models tested, with estimated 96.50, 96.31, 95.21, and 95.76% for accuracy, precision, recall, and F1-score, respectively, as shown in Fig. 3.

A model for ANN, although with slightly low accuracy, has brought an accuracy of 96.30% along with a precision of most (91.20%) and a recall of 90.60%, reaching an F1-score of 94.50%. Furthermore, the accuracy curves in Fig. 4 highlight the model's ability to capture and learn complex nonlinear relationships in the data.

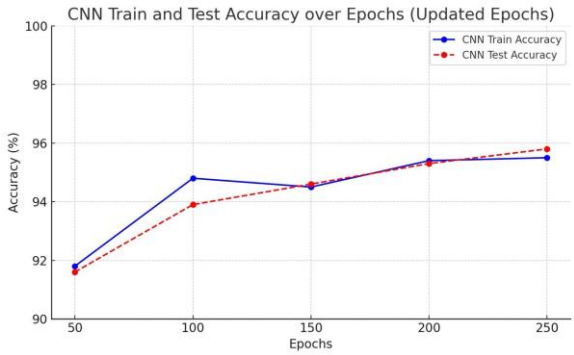

Fig. 2: The accuracy of the CNN model over 250 epochs

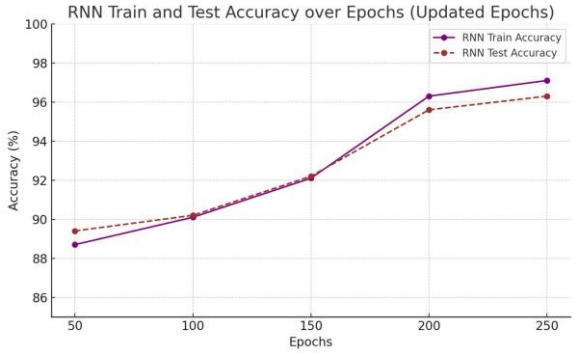

Fig. 3: The accuracy of the RNN model over 250 epochs

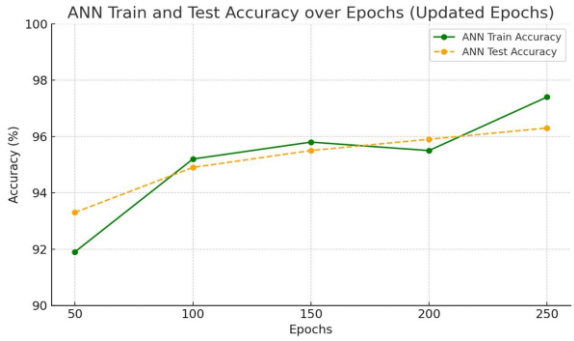

Fig. 4: The accuracy of the ANN model over 250 epochs

In general, all three models are effective at identifying patient condition, but the RNN was found to offer the best balance and robustness across the evaluation metrics.

Drug Rating Prediction

The CNN model's ability to predict drug ratings improves substantially after 250 rounds, each with 50 epochs. This model outperformed others, with the lowest mean error of 0.0296 and the highest R^2 score of 0.9692, which correspond to the highest accuracy and variance, as indicated in Fig. 5.

Despite its high predictive accuracy, the ANN model performed relatively poorly, with an MSE of 0.0359 and an R^2 of 0.9563 (Fig. 6).

The RNN model showed good performance, with an MSE of 0.01328 and an R^2 of 0.9625, indicating its ability to capture specific patterns in sequential data (Fig. 7).

Overall, CNN and RNN models outperformed ANN models in predicting drug ratings, with CNN emerging as the most precise predictor across the evaluations.

Sentiment Classification

In sentiment analysis, all models achieved strong performance, with accuracies greater than 93%. The best-performing model was the CNN, with 94.70% accuracy and a maximum recall of 92.76% across the rounds, as shown in Fig. 8.

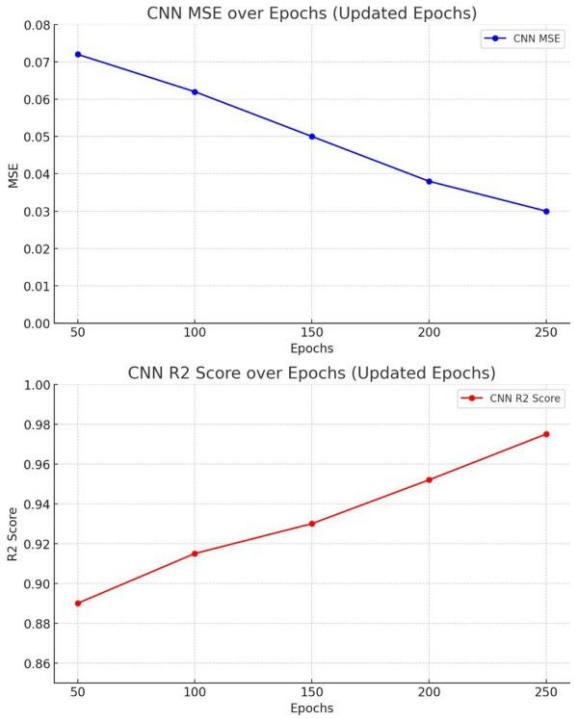

Fig. 5: The Mean Squared Error (MSE) and R^2 Score for test sets across 250 epochs (CNN)

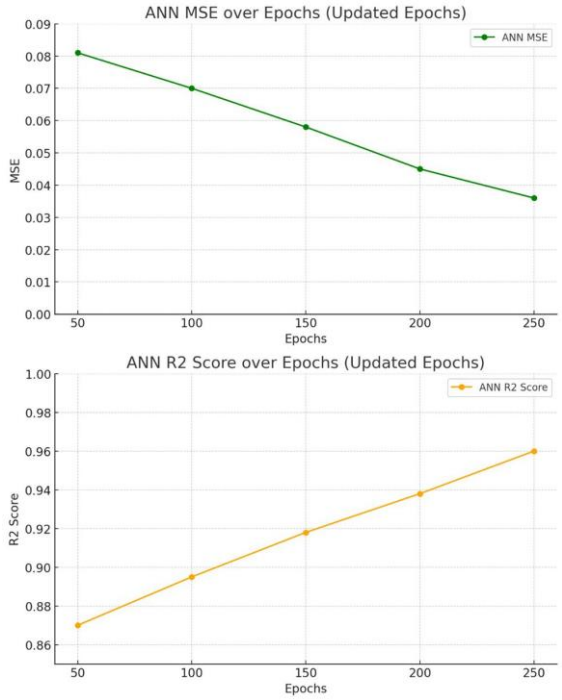

Fig. 6: The Mean Squared Error (MSE) and R^2 Score for test sets across 250 epochs (ANN)

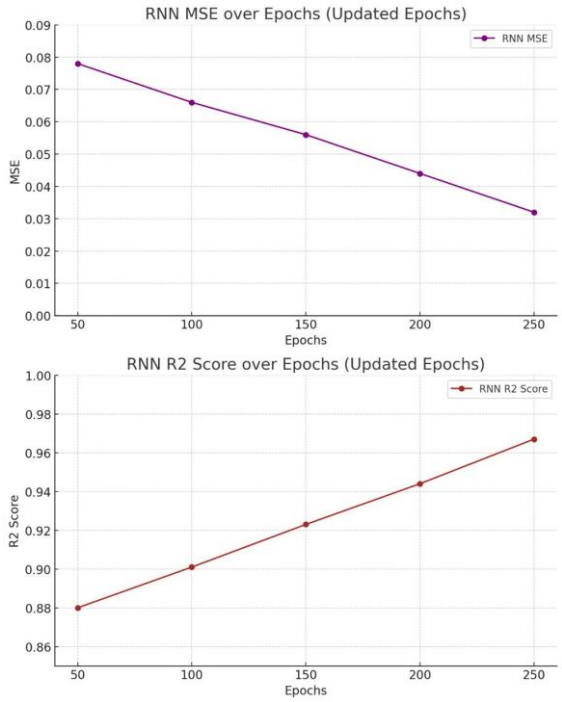

Fig. 7: The Mean Squared Error (MSE) and R^2 Score for test sets across 250 epochs (RNN)

The ANN model demonstrated balanced performance across all metrics, as shown in Fig. 9. Accuracy was 94.20%, precision 87.91%, recall 88.45%, and F1 89.12%.

The RNN model achieved the highest accuracy of 94.82% and the highest F1-score of 89.28%, as shown in Fig. 10. This suggests efficient performance overall but could also be interpreted as a lower precision of 85.42%, implying a higher rate of false-positive predictions.

These results indicate that while RNN slightly outperforms the other models in overall accuracy, CNN is more effective in maximizing recall, and ANN provides a more balanced trade-off between precision and recall. Depending on the type of application, therefore, the selection of a model might rely on the overall accuracy, the sensitivity to positive sentiments, or on the balanced performance in classification.

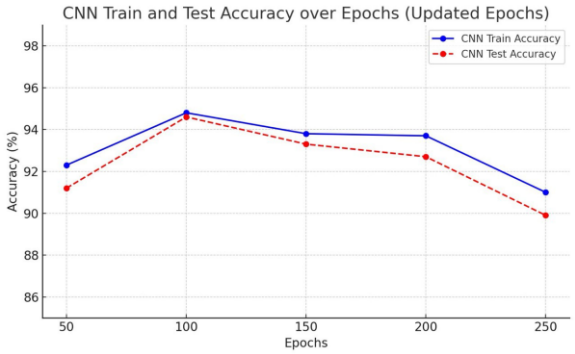

Fig. 8: The accuracy of the CNN model across 250 epochs

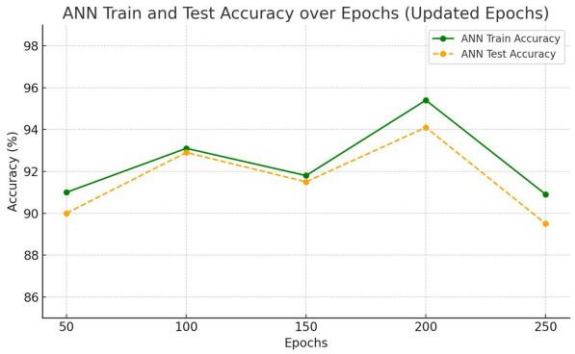

Fig. 9: The accuracy of the ANN model across 250 epochs

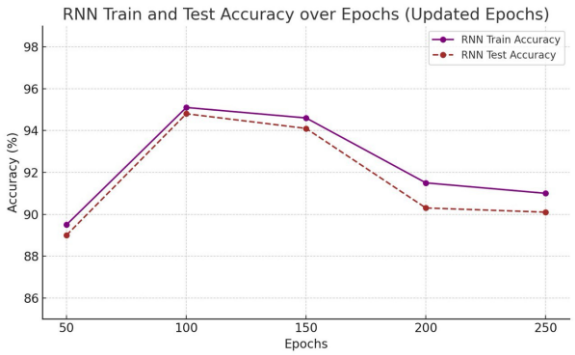

Fig. 10: The accuracy of the RNN model across 250 epochs

Conclusion

This study validates the application of three well-known algorithms, namely, CNN, ANN, and RNN, in evaluating the user health review seeking to predict drug ratings, medical condition reviews, and sentiment analysis of patients. All appear effective in these given operations and show linguistic features as their strengths. RNNs have a greater performance than CNNs and ANN when sequences of words need to be managed, such as in patient health identification and sentiment analysis. CNN emerged as an excellent option for drug rating prediction among them since it can efficiently learn and extract localized textual patterns.

Deep learning is well-suited to translating vast volumes of unstructured healthcare data into valuable insights; this translation scales to support real-world monitoring by health professionals and patient outcomes. As future work, we aim to extend the proposed models for drug reviews in Arabic. Due to the increasing amount of Arabic medical content on the internet, the application of deep learning techniques to Arabic text is likely to provide valuable insights into patient experiences and the efficacy of drugs within Arabic-speaking populations.

Acknowledgment

Thank you to the publisher for their support in the publication of this research article. We are grateful to the publisher for the resources and platform, which have enabled us to share our findings with a wider audience. We appreciate the editorial team's efforts in reviewing and editing our work, and we are thankful for the opportunity to contribute to the field of research through this publication.

Funding Information

The authors have not received any financial support or funding to report.

Author's Contributions

All authors equally contributed to this study.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all other authors have read and approved the manuscript and that no ethical issues are involved.

References

Askr, H., Elgeldawi, Enas , Aboul Ella, H., Elshaier, Yaseen A. M. M. , Gomaa, Mamdouh M., & Hassanien, Aboul Ella. (2023). Deep learning in drug discovery: An integrative review and future challenges. *Artificial Intelligence Review, 56*(7), 5975–6037. https://doi.org/10.1007/s10462-022-10306-1

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3615–3620. https://doi.org/10.18653/v1/d19-1371

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186). https://doi.org/10.18653/v1/N19-1423

Huang, K., Altosaar, J., & Ranganath, R. (2020). ClinicalXLNet: Modeling clinical notes and predicting hospital readmission. *Journal of Biomedical Informatics*, *117*, 103732.

Kormilitzin, A., Vaci, N., Liu, Q., & Nevado-Holgado, A. (2021). Med7: A transferable clinical natural language processing model for electronic health records. *Artificial Intelligence in Medicine*, *118*, 102086. https://doi.org/10.1016/j.artmed.2021.102086

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*. https://doi.org/10.48550/arXiv.1909.11942

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

Ling, Y. (2023). Bio+Clinical BERT, BERT Base, and CNN performance comparison for predicting drug-review satisfaction. *Computer Science > Computation and Language*. https://doi.org/10.48550/arXiv.2308.03782

Liu, J., Zhou, Y., Jiang, X., & Zhang, W. (2020). Consumers' satisfaction factors mining and sentiment analysis of B2C online pharmacy reviews. *BMC Medical Informatics and Decision Making*, *20*(1), 194. https://doi.org/10.1186/s12911-020-01214-x

Mulkalwar, S., Chitale, S., Rathi, B., Khan, U., Shende, V., & Shah, A. (2025). Online Patients Reviews - A Novel Approach to Evidence-based Medicine. *Medical Journal of Dr. D.Y. Patil Vidyapeeth*, *18*(Suppl 1), S71–S75. https://doi.org/10.4103/mjdrdypu.mjdrdypu_669_24

Omranian, S., Zolnoori, M., Huang, M., Campos-Castillo, C., & McRoy, S. (2023). Predicting Patient Satisfaction With Medications for Treating Opioid Use Disorder: Case Study Applying Natural Language Processing to Reviews of Methadone and Buprenorphine/Naloxone on Health-Related Social Media. *JMIR Infodemiology*, *3*, e37207. https://doi.org/10.2196/37207

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*(140), 1–67.

Sorayaie Azar, A., Babaei Rikan, S., Naemi, A., Bagherzadeh Mohasefi, J., & Wiil, U. K. (2024). Predicting patients' sentiments about medications using artificial intelligence techniques. *Scientific Reports*, *14*(1), 31928. https://doi.org/10.1038/s41598-024-83222-9

Sun, C., Huang, L., & Qiu, X. (2019). Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. *NAACL-HLT*, 380–385. https://doi.org/10.18653/v1/N19-1035

Tran, H., Tran, D., & Ngo, C. (2021). Predicting drug ratings using hybrid CNN-RNN and BERT models. *IEEE Journal of Biomedical and Health Informatics*, *25*(5), 1738–1748.

Vijayaraj, A., Murugan, V. P., Jebakumar, R., NV, G., V, S., & R, S. K. (2024). A Sentimental Analysis Approach for Personalized Drug Recommendations Using Machine Learning. *2024 International Conference on Computing and Data Science (ICCDS)*, 1–6. https://doi.org/10.1109/iccds60734.2024.10560460

Wei, Q., Ji, Z., Li, Z., Du, J., Wang, J., Xu, J., Xiang, Y., Tiryaki, F., Wu, S., Zhang, Y., Tao, C., & Xu, H. (2020). A study of deep learning approaches for medication and adverse drug event extraction from clinical text. *Journal of the American Medical Informatics Association*, *27*(1), 13–21. https://doi.org/10.1093/jamia/ocz063