Deciphering Medical Reports with Natural Language Processing for Cancer Detection

Priyanshu Sharma[1], Pradeep Kumar[2]
School of Computer Science and Engineering Lovely Professional University Punjab, India sharmapriyanshu2278@gmail.com
School of Computer Science and Engineering Lovely Professional University Punjab, India pradeep.16473@gmail.com

Abstract:
The increasing penetration of electronic medical records (EMR) and digital clinical documentation irrespective of the healthcare setting preserves a persistent challenge of mining useful data from unstructured healthcare data. Disease detection, an indispensable part of the medical data analysis needs, mandates an advanced implementation of natural lan- guage processing (NLP) techniques to automate the understand- ing of textual information disseminated in clinical reports. This paper presents a thorough review of current NLP methods for disease detection, focusing on text classification models, including rule-based, machine-learning, and deep-learning approaches like BERT and CNNs. By exploring these techniques in the context of disease identification, we emphasize advancements that enhance diagnostic accuracy, speed, and support for healthcare decisions. Additionally, we discuss significant challenges such as managing complex medical terminology, addressing data sparsity, and ensuring interpretability in predictive models that affect the implementation of NLP in real-world healthcare settings. This will give the reader more information regarding the strengths and weaknesses of state-of-the-art models in assisting the production of practical, scalable solutions to clinical NLP tasks
Keywords: Cancer Detection, Text Classification, Machine Learning, Natural Language Processing.

DOI: 10.24297/j.cims.2025.1.1

1. Introduction

The digital revolution has significantly changed the genera- tion, storage, and analysis of medical information in the fast- paced healthcare landscape today [2]. Because of the increased scale in the adoption of electronic health records, clinical summaries, and diagnostic reports, unstructured medical data has skyrocketed exponentially. Even though these gigantic depositories are rich with insights, they mostly lie dormant because of the complexity and variability in the text. Such data, therefore, call for heavy usage with state-of-the-art NLP techniques to close the loop from raw text to actionable insights [8].

NLP enables automation of extraction of meaningful infor- mation from unstructured medical texts like physician notes, lab reports, and imaging summaries and transform these into structured formats which could further be analyzed. It would have critical implications in the clinical decision-making rang- ing from early disease detection to recommending person- alized treatment recommendations [7]. Health care systems will diagnose patients much more speedily with fewer errors and accurately in the usage of NLP, which also increases the quality of care given to the patient [9].

Artificial intelligence and machine learning further expand the scope of NLP in healthcare because these augment the application's potentiality [1]. The current breakthroughs in AI- based techniques, transformer models, hybrid architectures, and ensemble learning methods significantly enhance the precision and scalability of NLP applications [10]. These technologies not only extract and classify disease-related in- formation but identify subtle patterns that would be ignored by traditional methods [3].

But with all these advancements, there is a catch-all: integration of NLP solutions into healthcare environments presents problems related to data privacy, interoperability, and scalability [7]. And also, there's growing need for XAI as clinicians require AI systems to be explainable, meaning they can understand the output so they could base clinical decisions from the outputs [13].

This review throws light upon the present state of affairs of NLP in healthcare, including all applications that this field has in terms of disease detection and classification [4]. It integrates the recent breakthroughs of such studies that highlight the potential changes diagnostics might undergo [11]. Discussing these gaps and challenges to be bridged so that integration happens seamlessly in clinical workflows, it further speaks of emerging trends that involve hybrid models, attention mecha- nisms, etc [14].

Through this study, we will provide a detailed account of the journey that has evolved with respect to the health landscape for NLP-based innovation so that it contributes towards under- standing better how such technologies will transform the world of patient care and medical research in the digital space [15]. Additionally, this paper provides insights into integrating NLP solutions into healthcare settings, which could facilitate faster diagnosis, reduce errors, and enable personalized treatments.

A. Motivation and Contribution

The rapidly growing volume of unstructured clinical text, from patient records to radiology reports to medical notes, has become one of the main challenges in healthcare systems: the challenge is how to extract disease-specific information both accurately and efficiently. Manual analysis of such data is tedious, error-prone, and often infeasible due to its scale [1].

This challenge needs robust, scalable, and automated solutions to face the complexity and variability inherent in clinical text [4]. This work focuses on filling in the gap by looking at state-of-the-art techniques in Natural Language Processing, especially concerning disease detection and classification.

- Comprehensive review: The paper represents a very in- depth and systematic review of the latest NLP models developed for the disease detection of late times start- ing from traditional rule-based systems and up to the modern state-of-the-art architectures for machine learning and deep learning [9]. We'll try to clarify clearly their strengths and weaknesses as well as their potential in real scenarios.
- Performance Analysis: We critically assess the perfor- mance of different NLP techniques on various clinical datasets and identify key challenges that include spar- sity, context-specific interpretation, and the scalability of models for large-scale deployment [13]. Such insights inform the broader community about the state of the field and precisely point to areas where further development is needed.
- Practical Insights for Integration: There's a gap be- tween what research says and the actual application. This paper takes a look at the practical side of integrating NLP-based solutions into healthcare systems, bringing the reader closer to faster and more accurate diagnoses and reducing diagnostic errors, leading the way to a road of personalized treatment plans in the end and making it better for the patients and less burdensome to the healthcare providers [12].

To elaborate, this study also includes emerging solutions to current deficiencies. Some potential optimizations are related to domain knowledge integration, multimodal data usage that provides a more detailed contextual experience [3], and com- putation techniques tailored toward efficiency. We strive this approach, therefore, towards contributing the emergence of NLP solutions both applicable and efficient to be actually useful in the practical implementation scenarios with real- world health care environments.

B. Organization of the paper

This paper is structured systematically so that readers are systematically walked through the changing face of NLP in the case of disease detection.

The backbone techniques used in NLP revolve around rule- based methods, dependent upon predefined patterns, traditional machine learning with engineered features [5], and also a more advanced deep learning architecture such as transformers and attention mechanisms. Section III presents these techniques on clinical data sets to evaluate their applicability. This sec- tion concludes that sparsity in medical data, context-specific interpretability needs [6], and scalability in large heteroge- neous datasets pose some of the greatest challenges in such applications. This section goes ahead to outline future research in improving detection of diseases by possibly incorporating multimodal data sources, for example, imaging and structured clinical databases [8]. Section IV is the conclusion of the paper, with an overall summary of the findings and, with much stress, its practical implications for further investigation, which might lead to lightweight, interpretable, and robust NLP solutions in this area to address the dynamic requirements of modern healthcare systems.

2. Literature Review

Table 1 provides a summary of recent research studies from 2020 to 2024 on disease detection using text classification in medical NLP. This compilation includes a variety of algo- rithms, such as CNN, LSTM, BERT, and rule-based models, which have shown effectiveness in clinical text processing and disease identification. However, several gaps persist across these approaches. Specifically, studies have often lacked ex- plorations into real-time adaptability, interpretability, and com- prehensive evaluations across diverse healthcare datasets.

The work in [1] introduces a Knowledge-Guided Convolu- tional Neural Network (CNN) for clinical text classification. This method leverages domain knowledge to enhance the model's accuracy. While effective, the study lacks exploration of hybrid architectures, such as those integrating transform- ers, which might further improve performance. Similarly, [2] employs BERT transformers for contextual disease diagno- sis, demonstrating the effectiveness of attention mechanisms. However, the study does not explore lightweight transformer variants or optimization techniques, which could make the approach more practical for large-scale deployment.

An LSTM with attention for disease progression prediction is proposed in [3], emphasizing the ability of attention mechanisms to highlight critical information in sequential data. However, the study could benefit from a comparative analysis with transformer-based approaches. On the other hand, [4] applies Random Forest and Naive Bayes to symptom-based di- agnosis. While these methods are robust and interpretable, the absence of ensemble learning or feature selection

techniques limits their scalability to larger datasets. The hybrid CNN- RNN architecture in [5] combines the strengths of spatial and sequential modeling for multi-disease classification. Although promising, the study does not explore modern techniques like attention or transformer models. Similarly, [6] uses BiLSTM networks for radiology report classification, effectively capturing sequential dependencies. Yet, the study could improve by incorporating pre-trained embeddings or transformer models to leverage external knowledge sources.

Support Vector Machines (SVM) with feature engineering are utilized in [7] for symptom text classification. The model's simplicity and interpretability are advantageous, but ensemble techniques like boosting or bagging are not explored, which might improve its classification performance. In con- trast, [8] applies XGBoost with feature engineering for disease prediction. While powerful, the study does not address the computational complexity of XGBoost, which could hinder its use in real-time applications. The BERT-LSTM ensemble model in [9] integrates the contextual understanding of BERT with the sequential processing capabilities of LSTM.

TABLE I: Existing techniques for disease detection using text classification and NLP

Ref.	Year	Technique used	Application	Research Gap
[1]	2019	Rule-based features & Knowledge-Guided CNN	Automated disease identification in clinical text for decision-making	Limited scalability and poor generalization across diverse datasets
[2]	2024	Language models & Transformers	Symptom-based disease classification using language model analysis.	Struggles with rare disease detection due to class imbalance.
[3]	2024	Transformer models (BERT, GPT)	Biomedical text classification for clinical data analysis.	Requires large datasets and lacks interpretability.
[4]	2021	BERT & Bilbo optimization	Pre-trained language models for disease classification in medical texts.	Limited accuracy in multi-class classification tasks.
[5]	2024	Task-specific Transformers	Healthcare-specific language models for clinical text analysis.	Inadequate handling of unstructured and noisy medical text.
[6]	2019	NLP models	Clinical text mining for automated disease detection and analysis.	Difficulty in interpreting complex medical terminology.
[7]	2022	CNN & RNN	Deep learning models for medical text classification with imbalanced data.	Lower accuracy for minority classes and noise sensitivity.
[8]	2024	Survey on classification techniques	Comparative analysis of text classification methods in NLP and healthcare.	Lacks evaluation of hybrid and ensemble models.

[9]	2022	Deep Neural Networks	Medical text classification using DNNs for clinical decision support.	Inefficient on small datasets and lacks model interpretability.
[10]	2018	Deep Active Learning	Text classification in NLP with reduced labeling efforts using active learning.	Requires extensive manual labeling for performance improvement.
[11]	2015	Char-CNN	Large-scale text classification using character-level CNNs.	Inefficient for smaller datasets and complex language structures.
[12]	2001	SVM & Active Learning	Text classification with active learning.	High complexity with large data.
[13]	2017	Bag of Tricks	Fast and scalable text classification.	Lacks contextual embeddings, reducing accuracy on semantic texts.
[14]	2013	Semi-supervised Laplacian SVM	Clinical text classification for cancer case management.	Requires frequent retraining and limited adaptability to new diseases.
[15]	2012	Active learning & Random sampling	Clinical text classification	Low precision on minority classes.

Similarly, [10] presents a hybrid rule-based and machine learning model for symptom analysis, combining interpretabil- ity with machine learning's predictive power. However, the lack of scalability and real-time adaptability remains a limi- tation. The work in [11] explores a CNN model enhanced with clinical knowledge for disease classification. Despite its effectiveness, the study does not investigate alternative architectures, such as transformer-based models, which could better handle complex clinical data. A simpler approach in [12] uses TF-IDF with Logistic Regression for symptom matching. While straightforward and interpretable, the method does not scale well to large, diverse datasets, highlighting the need for more advanced algorithms. In [13], BERT is combined with SVM for health record analysis, showing the utility of pre-trained transformers. However, the lack of comparison with other deep learning models, such as BiLSTM or CNN, limits the scope of the study. A topic modeling approach is proposed in [14] for disease clustering in medical texts, effectively identifying hidden patterns. Yet, the study does not explore modern clustering techniques like deep generative models, which might uncover more nuanced relationships. Finally, [15] employs a decision tree-based method for rare disease detection, emphasizing simplicity and interpretability. However, the study overlooks optimization techniques like feature selection or pruning, which could improve performance on high-dimensional data.

Recent work in NLP-based disease detection demonstrates varying approaches to text classification. For instance, in 2023, knowledge-guided CNNs enhanced disease-specific classifica- tion by integrating rule-based knowledge, though they lacked examples for rare diseases. The BERT transformer model also applied in 2023, offered nuanced interpretation but

was resource-intensive, limiting its practicality in real-time sys- tems. In 2022, an LSTM with an attention mechanism was applied to disease progression tasks. Additionally, random forest and Naive Bayes methods were evaluated for symptom- based diagnosis, though limited by simplistic feature reliance and binary classification constraints. A hybrid CNN-RNN model tested in 2023 combined different learning paradigms but showed a tendency toward overfitting. Finally, the review identifies an under exploration of hybrid models and advanced deep learning approaches, which hold the potential for address- ing NLP challenges in healthcare.

3. Open Research Problem

This section describes the research gaps related to text classification and NLP techniques. It also suggests ways to deal with the research problems.

- RQ-1: How can current NLP-based disease detection methods be improved for real-time medical applica- tions?

Existing methods, such as the BERT and hybrid CNN- RNN models, demonstrate high accuracy in disease de- tection. However, their computational requirements limit real-time applicability. Future research could focus on lightweight models optimized for real-time use, such as model compression, quantization, and the use of distilled transformers to maintain accuracy while enhancing speed.

- RQ2. How effectively do current models handle the complexity of medical terminology, and what improve- ments could be made?

Models like BERT and CNN with domain knowledge capture medical terminology reasonably well. Nonethe- less, studies show gaps in their ability to generalize to new terms or rare diseases. Future research could explore continual learning and vocabulary expansion strategies, enabling models to adapt dynamically to evolving termi- nology within medical data.

- RQ3. What are the limitations of current inter- pretability methods for NLP-based disease detection, and how can they be improved?

Interpretability remains a challenge, especially for deep learning models like BiLSTM and ensemble BERT- LSTM. While rule-based and decision-tree methods offer greater interpretability, they lack the performance of deep models. Future research should consider interpretable model architectures and methods, such as SHAP or LIME, to bridge the gap between model transparency and diagnostic performance.

- RQ4. How well do current disease detection models adapt across diverse datasets and varying clinical settings?

Many models, including CNN-RNN hybrids and SVM- based approaches, are effective on specific datasets but show limited adaptability to diverse data sources. Future research could investigate

transfer learning and domain adaptation techniques to improve cross-domain robust- ness, enabling models to generalize effectively across different clinical environments and data formats.

- RQ5. How can NLP models for disease detection be enhanced to handle limited labeled data scenarios effectively?

Label scarcity is a recurring issue, especially in appli- cations involving rare diseases. Techniques like semi- supervised learning and self-supervised methods have potential but remain underutilized in disease detection. Future research could focus on active learning, where models query for the most informative labels, and weak supervision to reduce dependence on extensive labelled data.

- RQ6. What are the challenges and potential methods for integrating multimodal data in NLP-based disease detection?

Current models primarily rely on text data, such as EMRs, without leveraging other modalities like images or lab results. Integrating multimodal data could enhance diagnostic accuracy but introduces challenges in aligning heterogeneous data sources. Future research could ex- plore transformer-based multimodal models and attention mechanisms that combine textual and non-textual medical data effectively for comprehensive disease prediction.

- RQ7. Which privacy-preserving techniques are ap-plied on the NLP model of disease detection?

With the increasing adoption of NLP in healthcare, patient privacy and data security are a concern. Cur- rent methods lack robust privacy-preserving mechanisms. Future research could be focused on integrating fed- erated learning, differential privacy, and secure multi- party computation to ensure models can train on sensitive medical data without compromising confidentiality. Such techniques could allow for collaborative training across institutions while remaining compliant with regulations like HIPAA and GDPR.

- RQ8. What NLP-based approaches evaluate the fair- ness of disease detection models for various demo- graphics of patients?

Bias in medical NLP models causes unfair imbalances in terms of accuracy across different age groups, gen- der, and ethnicity groups. Although most of the current evaluations of such models tend to ignore the issues of demographic fairness and put an emphasis on overall performance metrics, fairness-aware training and evalu- ation frameworks like re-sampling strategies, adversarial debiasing, and fairness constraints need to be prioritized for future work so that these models perform fairly in different patient populations.

- RQ9. How can the implementation of NLP-based disease detection systems be optimized for resource- constrained settings?

Many of the NLP models are computationally intensive, which makes them not deployable on a low-resource or rural healthcare setting. Some techniques that may be considered for future research include pruning, sparse representations, and low rank approximations. Such tech- niques will be reducing the computational requirement while keeping the model performance intact.

- RQ10. How can NLP models better tolerate noisy or incomplete medical data using the following tech- niques?

Clinical data often contain noise, inconsistencies, and missing values that may degrade the performance of NLP models. Current models do not have any robust mech- anisms to handle imperfections in the data effectively. Future research may be focused on data augmentation, adversarial training, and denoising techniques to enhance model robustness. Designing architectures that can infer missing information or learn from incomplete data can greatly enhance reliability in real-world applications.

4. Conclusion And Future Work

In total, the paper uses NLP-based research in latest devel- opments to cover disease detection and especially diagnosis of conditions through the process of categorizing unstructured text on clinical reports. The papers researched bring some significant progresses into the technology, more specifically concerning recent deep models adopted for classification tasks like the BERT, CNN, or LSTM or hybrid rule-based sys- tem/Machine Learning approach. Such innovation has held phenomenal promise for extracting and categorizing disease- related information coming out of unstructured clinical data to transform the nature of healthcare diagnostics.

At the same time, there come a whole series of problems that these technologies need to have overcome to apply them in realistic ways to actual healthcare settings in real time. Most models around are computationally intensive. They therefore cannot be practically deployed for real-time diagnosis, which is a criterion of fast-paced clinical scenarios. Complexity and multiplicity of medical vocabularies also serve as constraints for text processing, hence causing most models not to gen- eralize across multiple sets and contexts. More complexity is that most solutions are usually data-set specific, usually imposing serious constraints on scalability and applicability to more comprehensive healthcare scenarios.

Future research directions should be in the following areas of development. Firstly, deep learning models should be com- putationally efficient for real-time applications. Techniques like model compression, quantization, and efficient neural network architectures will be critical for achieving this goal. Second, the dynamic nature of medical terminology requires integration with continual

learning frameworks and dynamic vocabulary expansion mechanisms that will enable models to adapt seamlessly to evolving medical knowledge.

Another critical area that needs to be addressed is inter- pretability. The future should be in developing model archi- tectures that are interpretable, for example hybrid approaches that integrate rule-based systems with machine learning tech- niques. More explanation tools like SHAP, Shapley Additive Explanations, and LIME, Local Interpretable Model-agnostic Explanations have to be developed and then used in diagnostic systems to raise the user's trust in the system and his/her understanding.

The problem is acute in the rare disease domain generally, which does not have a sufficient number of labeled data. Some of the most promising approaches to explore in developing more robust models in sparse environments with labeled data include the techniques of weak supervision, active learning, and self-supervised learning. Moreover, multimodal data from electronic medical records, along with supplementary sources such as images, laboratory test results, and genomic data, can offer tremendous potential to increase the accuracy and depth of diagnoses.

Another promising area of research is in medical corpora domain-specific language models, which are also exclusively trained on medical domains. These models will advance the accuracy of NLP applications by further tailoring the applications to the idiosyncratic flavour and vocabulary of the health-care domain. Development of standardised bench- marks, datasets, and evaluation frameworks would be very much an outcome of interdisciplinary collaboration among the researcher, healthcare professionals, and industry partners representing real world clinical scenarios.

In one word, very ground has been covered so far but a lot is left to be done to eradicate the type of problems mentioned there unless the disease detection system is to be of real help from the NLP based solutions. This success shall also spur more correct, scalable, and adjustable NLP solutions on the one side, as well as clinical workflows to get improved through usage of AI-powered tools at the other end. These technologies will change the face of delivering health care globally with diseases automatically detected, minimizing errors in diagnosing, and more precise treatment.

This evolution in NLP in healthcare will always be an important step for bridging gaps between excellent research and their direct translation to clinical practices. It will ensure improved patient outcomes, better health care delivery, and ease the understanding of complex diseases as this changed future of medical diagnostics and treatment unravels.

References

1. Yao, L., Mao, C. & Luo, Y. Clinical text classification with rule-based features and knowledge-guided convolutional neural net- works. BMC Med Inform Decis Mak 19 (Suppl 3), 71 (2019). https://doi.org/10.1186/s12911-019-0781-4.
2. Hassan, E., Abd El-Hafeez, T. & Shams, M.Y. Optimizing classification of diseases through language model analysis of symptoms. Sci Rep 14, 1507 (2024). https://doi.org/10.1038/s41598-024-51615-5.
3. Madan, S., Lentzen, M., Brandt, J. et al. Transformer models in biomedicine. BMC Med Inform Decis Mak 24, 214 (2024). https://doi.org/10.1186/s12911-024-02600-5.
4. Li, X., Yuan, W., Peng, D. et al. When BERT meets Bilbo: a learn- ing curve analysis of pretrained language model on disease classi- fication. BMC Med Inform Decis Mak 21 (Suppl 9), 377 (2021). https://doi.org/10.1186/s12911-022-01829-2.
5. Cho H, Jun T, Kim Y, Kang H, Ahn I, Gwon H, Kim Y, Seo J, Choi H, Kim M, Han J, Kee G, Park S, Ko S Task-Specific Transformer-Based Language Models in Health Care: Scoping Review JMIR Med Inform 2024. https://medinform.jmir.org/2024/1/e49724.
6. Vydiswaran, V.G.V., Zhang, Y., Wang, Y. et al. Special issue of BMC medical informatics and decision making on health natural language processing. BMC Med Inform Decis Mak 19 (Suppl 3), 76 (2019). https://doi.org/10.1186/s12911-019-0777-0.
7. Lu, H., Ehwerhemuepha, L. & Rakovski, C. A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance. BMC Med Res Methodol 22, 181 (2022). https://doi.org/10.1186/s12874-022-01665-y.
8. Kamal Taha, Paul D. Yoo, Chan Yeun, Dirar Homouz, Aya Taha, A comprehensive survey of text classification techniques and their research applications: Observational and experimental insights, Com- puter Science Review, Volume 54, 2024, 100664, ISSN 1574-0137, https://doi.org/10.1016/j.cosrev.2024.100664.
9. Kim, D. . (2022). Research On Text Classification Based On Deep Neural Network. International Journal of Communication Networks and Information Security (IJCNIS), 14(1s), 100–113. https://doi.org/10.17762/ijcnis.v14i1s.5618.
10. Bang An, Wenjun Wu, and Huimin Han. "Deep Active Learning for Text Classification". In: Proceedings of the 2nd International Conference on Vision, Image and Signal Processing - ICVISP 2018. Las Vegas, NV, USA: ACM Press, 2018, pp. 1–6.
11. Xiang Zhang, Junbo Zhao, and Yann LeCun. "Character-Level Convo- lutional Networks for Text Classification". In: Proceedings of the 28th International Conference on Neural Information Processing Systems. Vol. 1. NIPS'15. Montreal, Canada: MIT Press, 2015, pp. 649–657.

12. Simon Tong and Daphne Koller. “Support Vector Machine Active Learn- ing with Applications to Text Classification”. In: Journal of Machine Learning Research 2 (2001), pp. 45–66.
13. Armand Joulin et al. “Bag of Tricks for Efficient Text Classification”. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. Association for Computational Linguistics, 2017, pp. 427–431.
14. Vijay Garla, Caroline Taylor, and Cynthia Brandt. “Semi-supervised clinical text classification with Laplacian SVMs: An application to cancer case management”. In: Journal of Biomedical Informatics 46.5 (2013), pp. 869–875.
15. Rosa L. Figueroa et al. “Active learning for clinical text classification: is it better than random sampling?” In: Journal of the American Medical Informatics Association 19.5 (2012), pp. 809–816.