Natural Language Processing For Medical Text Analysis

Priyanka Gupta[1], Poorti Sharma[2], Nidhi Mathur[3]
[1]Assistant Professor, Department of Chemistry, Kalinga University, Raipur, India.
[2]Assistant Professor, Department of Pharmacy, Kalinga University, Raipur, India.
ku.poortisharma@kalingauniversity.ac.in, 0009-0005-0442-9650
[3]Professor, New Delhi Institute of Management, New Delhi, India, nidhi.mathur@ndimdelhi.org,
https://orcid.org/0000-0003-0650-4667

Abstract
A lot of healthcare data is generated on a regular basis. This can be utilized to extract information required for disease occurrence prediction in a patient. For decision making and disease prediction, it is necessary to leverage the treatment history and health data present in the patient data most of which are 'buried' in patient data like EHR/EMR. The volume of data generated on a daily basis is extremely massive which requires leveraging data mining or machine learning methods. The hope of utilizing the analytical and ML techniques is to forecast clinical outcomes in advance so that supports the medical professionals for early diagnosis of disease and chronical diseases so that treatment can be initiated early or minimize risk of life to the patient. Early diagnosis and treatment can minimize the treatment cost to a major extent. Probabilistic modelling and deep learning method will train a Long Short-Term Memory recurrent neural network and a convolutional neural network to forecast the occurrence of the disease. The specific combination of deep learning methodologies and an abundance of data in the EHR is extremely beneficial in enhancing the understanding of human health.

Keywords: *publications, NLP, recommendations, performance*

INTRODUCTION

In the recent years, disease prediction based on patient history and health prediction based on data mining and machine learning has been ongoing with partial success. There are a number of strategies that have utilized data mining on pathology reports, medical profiles for the prediction of certain diseases [1]. There are also a number of strategies that have attempted disease re-occurrence prediction. Though others have attempted to predict the progression and control of the disease. Since with the advancements in deep learning, prediction of the disease also has also advanced much considering they are capable of learning rich hierarchal representations of raw data with minimal or no preprocessing [9]. There is substantial literature on data mining algorithms in terms of Decision Tree, Naive Bayes, neural network, support vector machine etc., employed for diagnosing heart diseases which have displayed varied accuracy results while predicting disease[2]. Most commonly, these works are implemented through WEKA. Now, for many years the Electronic Health Record (EHR) in the field of predicting the disease appeals the medical domain. EHR will be profitable both for providers as well as the patients [3]. The use of EHR enhances patient treatment with adequate care through access to patient health records, essentially low-cost, high-patient participation, and transparency. Electronic Health Record (EHR) implementation in healthcare practice delivers adequate patient care and the accessible patient's information is correct in EHR treatment and diagnosis [4] . The technology lessens the chance of medical malpractice by generating a clinical warning to doctors. Heterogeneous Recurrent Convolutional Neural Networks (HRCNN) facilitates the relationship between different heterogeneous medical incidents for Convolutional Neural Networks (CNN) model. The suggested model is efficient in finding the patient's entire health details and numerous hospitals are anticipating the application of Electronic Health records [10].

MATERIALS AND METHODS

CNN produces a pre-determined output by obtaining predetermined input. A new deep learning framework is derived through the conjunction of Heterogeneous Recurrent Convolutional Neural Networks (HRCNN) with Electronic Health Record (EHR) to forecast comorbid disease risk factors. The envisioned research is aimed in constructing a learning system that can connect the sparse convolutional layer and local pooling of heterogeneous attributes and thereby, it can model the correlations between different heterogeneous attributes. The envisioned model is aimed in forecasting the evolution of patient conditions [5]. Heterogeneous Recurrent Convolutional Neural Networks (HRCNN) achieves improved performance on risk predictions. Health Care and Life Sciences have been one of the most developed fields of research. The development of newer computing paradigms has given rise to a lot of new and untapped areas. These were initially introduced by the newer instruments that further led to newer paradigms that utilize the data that are generated through the use of these instruments [6]. With paradigms of High-Performance Computing being made increasingly available, we are now capable of making use of various machine learning / deep learning methodologies to utilize the data by digitizing it in the form of EHR, EMR [11].

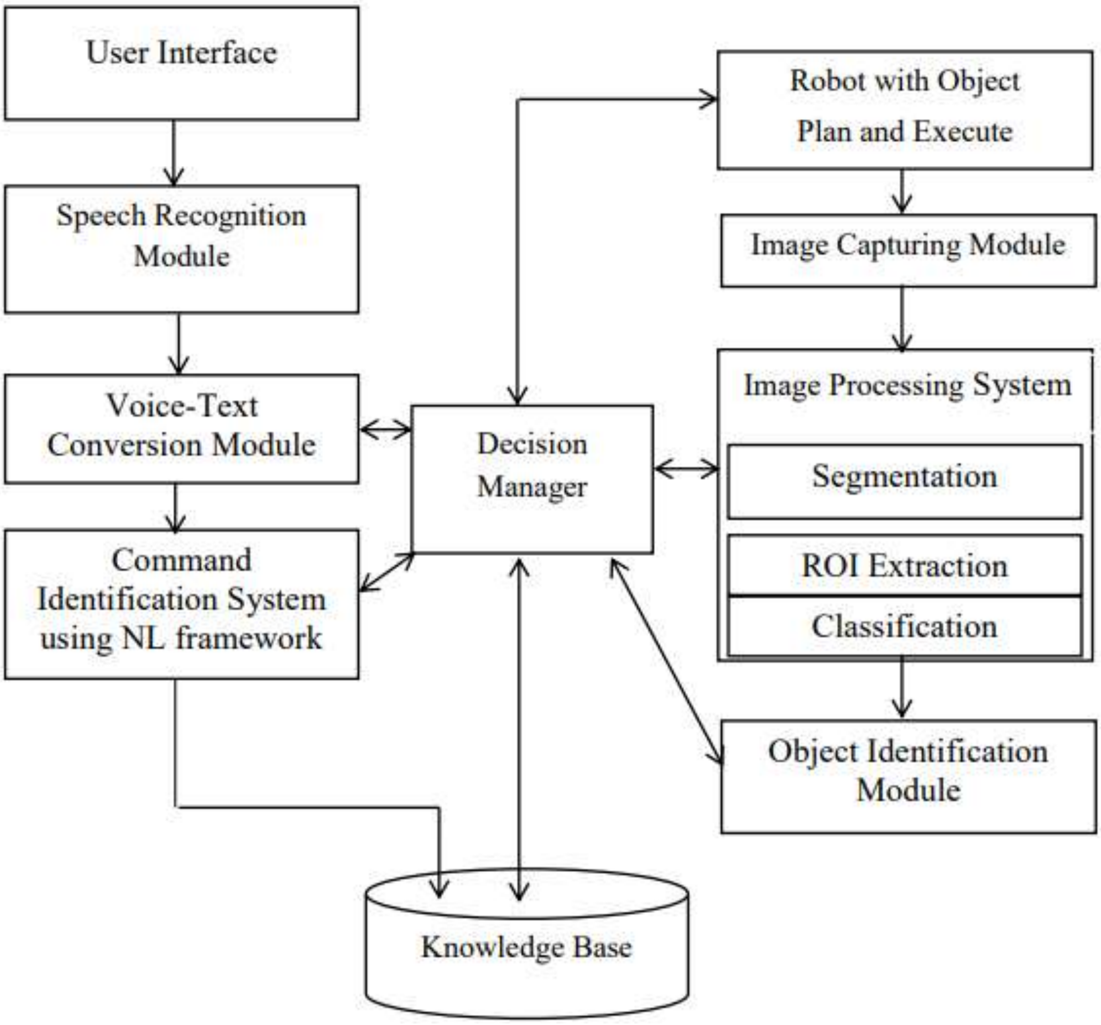
User Interface
Speech Recognition Module
Voice-Text Conversion Module
Command Identification System using NL framework
Decision Manager
Robot with Object Plan and Execute
Image Capturing Module
Image Processing System
Segmentation
ROI Extraction
Classification
Object Identification Module
Knowledge Base

Figure 1: System Architecture

It offered insights into the text's emotional tone by identifying keywords, classifying them, and assigning scores for sentiment (positive, negative, neutral) and emotions (anger, fear, contempt, joy, and sadness) [7]. According to NLP analysis, conversations about administration were linked to wrath, whereas references to body parts were linked to fear and melancholy[8]. This illustrates how NLP may be used to analyse unstructured text data from online forums in order to find patient feelings and emotions around particular subjects, like medications, symptoms, and complications.

RESULT AND DISCUSSION

By demythologizing technical terms, NLP-based summary can increase the accessibility of difficult publications. However, just 3% of NLP summarization efforts are applied to clinical EHR notes, indicating a significant gap in this technology's application, underscoring the need for further attention in this field [12].

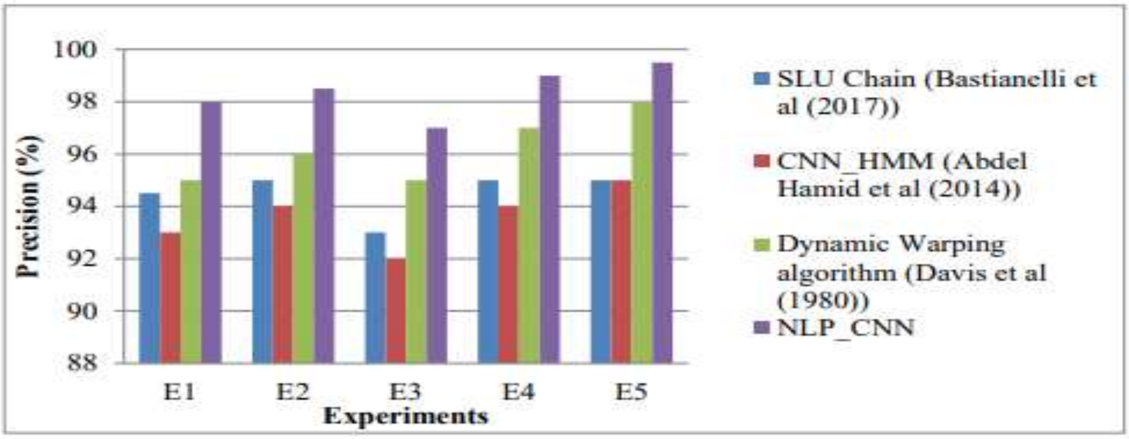

Figure 2: Performance Analysis

Electronic scribes with NLP capabilities can relieve doctors of some of the paperwork so they can concentrate more on patient care. These scribes can increase efficiency, boost patient happiness, and fortify patient-physician interactions by automating documentation and interacting with text extraction pipelines[13]. Background noise, technical jargon, fillers, disfluencies, and non-linear talks can all make it difficult for NLP-powered scribes to capture reliable clinical documentation. It is essential to overcome these obstacles in order to create dependable and efficient electronic scribes.

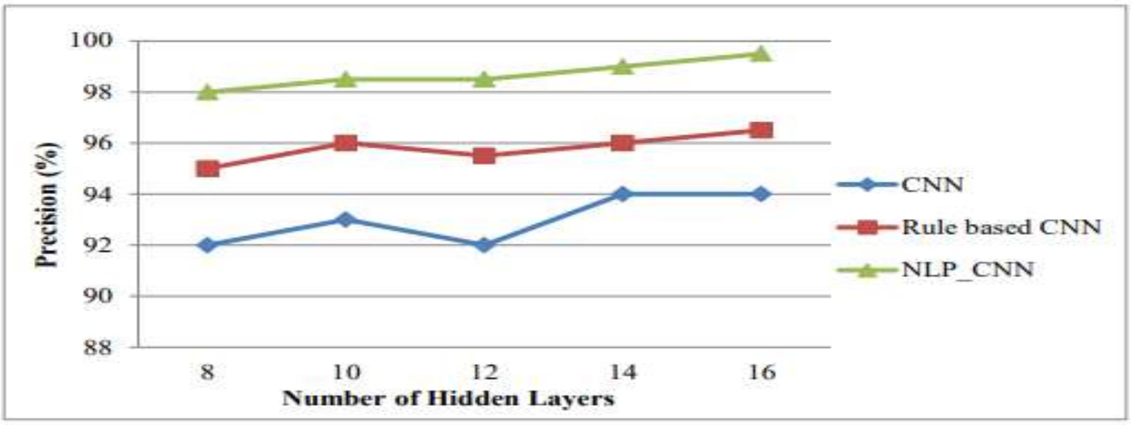

Figure 3: Comparisons with Precision Values

By anticipating the next word in a sentence, large language models (LLMs) provide responses that are human-like in chatbot applications [14]. With trillions of criteria, they can generate content that is contextually appropriate and coherent, frequently creating the appearance of comprehension.

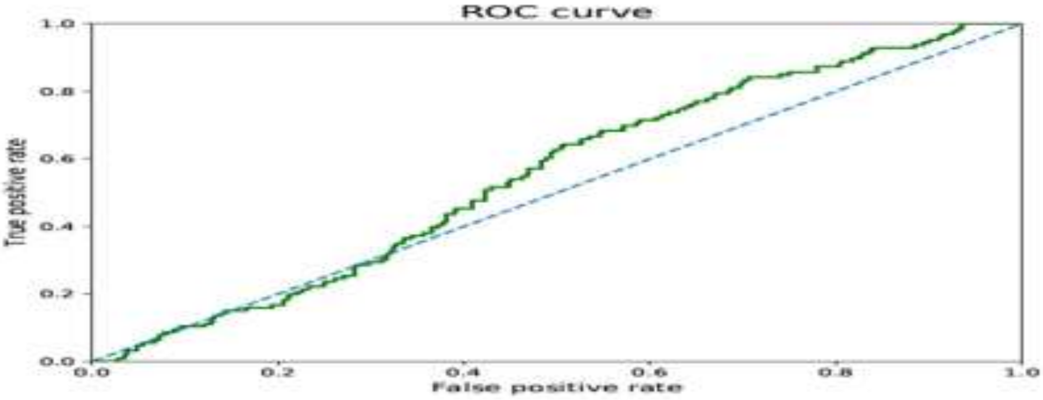

Figure 4: ROC curve

But it's important to understand that LLMs are just sophisticated statistical models and cannot fully understand meaning.

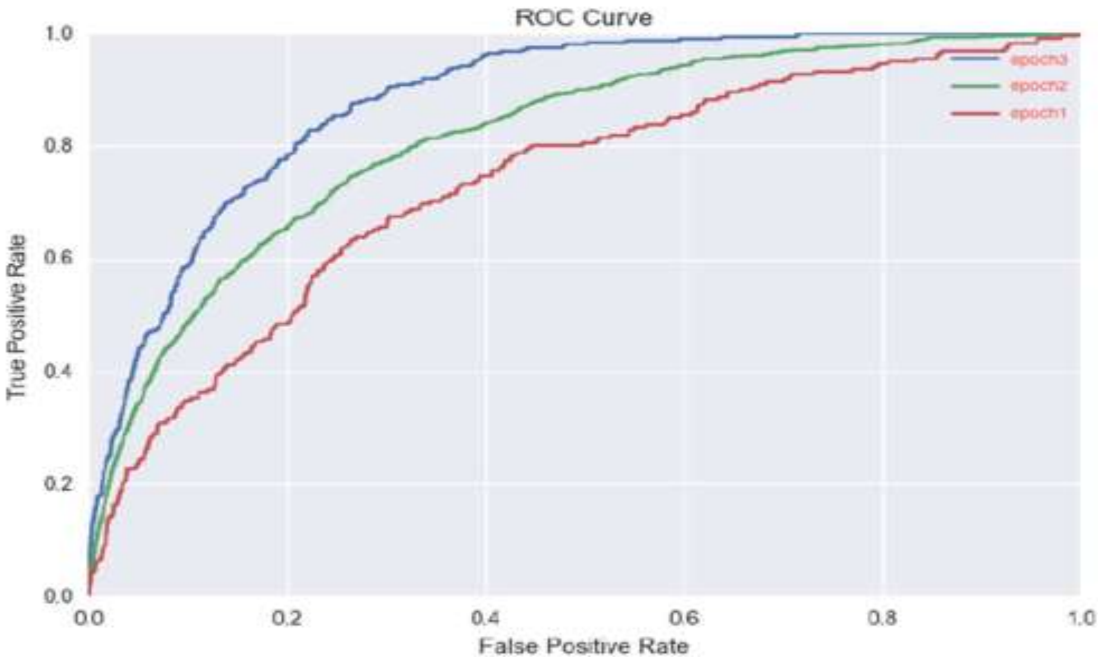

Figure 5: Additional Layer + LSTM

Applications like chatbots (such as ChatGPT and Gemini) can now process and produce language that sounds human thanks to LLMs, which are a major development in natural language processing [15].

CONCLUSION

Clinicians' lack of familiarity with NLP technologies and their possible advantages and disadvantages may be the cause of their limited acceptance in clinical practice. A number of past studies that have arrived at feature extraction and representation of EHR information have addressed these challenges and came up with several methods. In general, all the methods initially determine the data representation followed by the adaptation of learning infrastructure and condition perdition experiments based on available data and target label. Feature representation in this case refers to vector, sequence, tensor or temporal matric to graph representations are employed. The learning infrastructure adaptation is done by employing support sector machines, tensor factorization and regression. But we are of the opinion that there is plenty of scopes to contribute even greater value to the work that has been done particularly through exploring more sophisticated models like GANs which can be utilized to derive greater insights from medical images, etc. Additionally, the above-stated architecture that assists us in removing the uncertain notations in the healthcare data can further be improvised to remove other hyper-parameter-based adjustments and better management of missing values in real-world data.

REFERENCES

1. Iroju, Olaronke G., and Janet O. Olaleke. "A systematic review of natural language processing in healthcare." International Journal of Information Technology and Computer Science 8, no. 8 (2015): 44-50.
2. Dallal, H. R. H. A. (2024). Changes to Communication Infrastructure Caused by Blockchain Technology. International Academic Journal of Science and Engineering, 11(1), 25–39. https://doi.org/10.9756/IAJSE/V11I1/IAJSE1105
3. Duch, Włodzisław, Paweł Matykiewicz, and John Pestian. "Neurolinguistic approach to natural language processing with applications to medical text analysis." Neural Networks 21, no. 10 (2008): 1500-1510.
4. Samyadevi, V., Anguraj, S., Singaravel, G., & Suganya, S. (2024). Image Based Authentication Using Zero-Knowledge Protocol. International Academic Journal of Innovative Research, 11(1), 01–05. https://doi.org/10.9756/IAJIR/V11I1/IAJIR1101
5. Rajput, Adil. "Natural language processing, sentiment analysis, and clinical analytics." In Innovation in health informatics, pp. 79-97. Academic Press, 2020.
6. Sio, A. (2025). Integration of embedded systems in healthcare monitoring: Challenges and opportunities. SCCTS Journal of Embedded Systems Design and Applications, 2(2), 9–20.

7. Baud, R. H., A-M. Rassinoux, and J-R. Scherrer. "Natural language processing and semantical representation of medical texts." Methods of information in medicine 31, no. 02 (1992): 117-125.

8. Christian, J., Paul, M., & Alexander, F. (2025). Smart traffic management using IoT and wireless sensor networks: A case study approach. Journal of Wireless Sensor Networks and IoT, 2(2), 45-57.

9. Dreisbach, Caitlin, Theresa A. Koleck, Philip E. Bourne, and Suzanne Bakken. "A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data." International journal of medical informatics 125 (2019): 37-46.

10. Fanni, Salvatore Claudio, Maria Febi, Gayane Aghakhanyan, and Emanuele Neri. "Natural language processing." In Introduction to artificial intelligence, pp. 87-99. Cham: Springer International Publishing, 2023.

11. Popowich, Fred. "Using text mining and natural language processing for health care claims processing." ACM SIGKDD Explorations Newsletter 7, no. 1 (2005): 59-66.

12. Locke, Saskia, Anthony Bashall, Sarah Al-Adely, John Moore, Anthony Wilson, and Gareth B. Kitchen. "Natural language processing in medicine: a review." Trends in Anaesthesia and Critical Care 38 (2021): 4-9.

13. Quinby, B., & Yannas, B. (2025). Future of tissue engineering in regenerative medicine: Challenges and opportunities. Innovative Reviews in Engineering and Science, 3(2), 73–80. https://doi.org/10.31838/INES/03.02.08

14. Schmidt, J., Fischer, C., & Weber, S. (2025). Autonomous systems and robotics using reconfigurable computing. SCCTS Transactions on Reconfigurable Computing, 2(2), 25–30. https://doi.org/10.31838/RCC/02.02.04

15. Ahmad, A. B., & Prabowo, S. (2025). Real time operating systems for embedded applications: Design and implementation. Journal of Integrated VLSI, Embedded and Computing Technologies, 2(1), 37–45. https://doi.org/10.31838/JIVCT/02.01.05