**NLP - based model to detect and de-identify PII in clinical notes**

YASHVANTH P S

Department of Information Science and Engineering

R V college of Engineering

Bengaluru, Karnataka – 560059

**ABSTRACT :** With the increasing digitalization of healthcare records, safeguarding Personally Identifiable Information (PII) has become crucial. Clinical notes contain sensitive patient details such as names, dates, and locations, necessitating compliance with privacy regulations like HIPAA and GDPR. While Natural Language Processing (NLP) has advanced significantly, detecting and de-identifying PII in unstructured text remains a challenge. Existing models, such as Named Entity Recognition (NER) using spaCy, accuracy, data integrity, and regulatory compliance.

This project addresses these challenges by developing an automated NLP-based system for PII detection and de-identification in clinical notes. The system was trained using a publicly available clinical notes dataset from Kaggle. Text preprocessing involved cleaning unwanted characters, converting text to lowercase, and tokenizing data for efficient model input. The spaCy NER model identified PII entities such as PERSON, DATE, and LOCATION, which were then masked using placeholders (e.g., [REDACTED\_NAME]). The system was deployed as a web application using Flask.

Evaluation metrics, including Precision, Recall, and F1-score, demonstrated 100% accuracy across multiple test cases. The model successfully anonymized sensitive information while maintaining the readability and analytical value of clinical notes. The system aligns with HIPAA and GDPR standards and offers a scalable solution for secure data handling in healthcare organizations, facilitating privacy-compliant research and analysis.

**Keywords** - Personally Identifiable Information (PII), Named Entity Recognition (NER), Natural Language Processing (NLP), Clinical Data, HIPAA and GDPR

**I. INTRODUCTION**

The exponential growth of healthcare data has led to an increased need for robust privacy-preserving techniques. Clinical notes often contain sensitive Personally Identifiable Information (PII) that must be protected to comply with privacy regulations such as the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR).

However, manually identifying and redacting such information is labor-intensive and prone to human error, necessitating automated solutions. Natural Language Processing (NLP) has emerged as a powerful tool for extracting and analyzing textual data, including Named Entity Recognition (NER) techniques used for PII detection. While existing transformer-based models such as ClinicalBERT and tools like spaCy have shown promise, challenges remain in ensuring high accuracy, maintaining data utility, and adhering to legal compliance requirements. This research aims to develop an automated, NLP-driven system capable of accurately detecting and de-identifying PII in unstructured clinical text while preserving the analytical value of medical data.

The system leverages an NER-based approach for entity detection and a masking strategy to anonymize sensitive information, ensuring compliance with data privacy regulations. By implementing a scalable, web-based solution, this work contributes to the growing need for secure data handling in healthcare organizations and research institutions.

**II. LITERATURE REVIEW**

The use of Natural Language Processing (NLP) in detecting and de-identifying Personally Identifiable Information (PII) in clinical notes has become a critical area of research, driven by the need to protect patient privacy while maintaining the utility of medical data for research and clinical decision-making. Early efforts focused on rule-based systems, where predefined dictionaries and pattern matching techniques were employed to identify sensitive information such as names, addresses, and medical conditions. However, while these approaches offer high precision, they struggle to generalize across diverse clinical datasets and languages, making them less effective in large-scale applications [1].

Recent advancements have seen the integration of deep learning techniques, particularly transformer-based models such as BERT and its variants, including ClinicalBERT, which are specifically designed for clinical text. These models excel at capturing context and nuances in medical language, improving their ability to detect a wide range of PII in clinical notes. Fine-tuning these models on specific medical datasets has significantly increased their accuracy in identifying sensitive data, but challenges remain in terms of handling domain-specific jargon and multi-language environments [2].

Despite these improvements, the ability to achieve consistent results across diverse institutions with varying documentation practices remains a significant barrier. In contrast to rule-based approaches, machine learning methods like Named Entity Recognition (NER) and Conditional Random Fields (CRF) have shown great promise in detecting PII entities. These models, which are trained to recognize medical terms such as patient names, medications, and diagnoses, can be adapted and fine-tuned on domain-specific data. Studies have demonstrated that deep learning-based NER models, such as BiLSTM-CRF, outperform traditional rule-based systems in terms of accuracy and adaptability to new medical terminology [3].

However, these models often require large annotated datasets for training, and the task of acquiring such datasets can be time-consuming and costly in clinical settings.

Federated learning has also emerged as a promising technique for privacy-preserving model training. By allowing models to be trained on data stored locally at healthcare institutions, federated learning enables the development of NLP models without the need to share sensitive patient data. This decentralized approach maintains the privacy of PII, yet allows for the collaborative training of models across institutions, improving generalization without compromising confidentiality. Despite its potential, challenges remain in terms of model convergence, data heterogeneity across institutions, and computational costs [4].

Additionally, ensuring that federated learning methods comply with privacy regulations such as HIPAA and GDPR is an ongoing research challenge. Hybrid approaches that combine rule-based and machine learning methods are also gaining attention. For instance, some systems use rule-based preprocessing to identify straightforward PII and machine learning models to detect more complex or ambiguous entities. These hybrid systems provide a balance between the precision of rule-based methods and the adaptability of machine learning, making them a promising solution for clinical text de-identification [5].

Additionally, the combination of deep learning with symbolic reasoning has been explored to improve the robustness of PII detection systems, particularly in environments where text can be highly unstructured or incomplete. While significant progress has been made in detecting PII in clinical notes, challenges related to data utility after de-identification remain. Ensuring that clinical data remains usable for research, predictive modeling, and decision-making without exposing sensitive patient information is critical. Some studies have focused on preserving the utility of de-identified data by utilizing techniques such as differential privacy or synthetic data generation. These methods ensure that patient privacy is maintained while still enabling researchers and clinicians to access useful datasets for analysis [6].

Another area of active research is the integration of NLP-based PII detection models with broader healthcare data systems. By embedding these models into electronic health record (EHR) systems, real-time de-identification of clinical notes can be achieved, ensuring that sensitive information is protected as it is entered into databases. However, ensuring that such systems operate efficiently at scale remains a challenge, particularly in large healthcare systems where data volumes are vast and rapidly growing [7].

Lastly, studies have also explored the role of explainable AI (XAI) in NLP models for clinical text de-identification. Understanding how models arrive at their decisions is crucial in high-stakes domains like healthcare. By making PII detection models more interpretable, researchers can ensure that clinical professionals have confidence in the system’s ability to protect patient privacy while also enhancing model performance. This can improve adoption rates and ensure compliance with regulatory standards [8].

These studies collectively highlight the significant advancements in using NLP-based models to detect and de-identify PII in clinical notes. Despite the considerable progress, ongoing research is needed to address challenges related to generalizability, scalability, real-time application, and ensuring data utility post-de-identification. The future of PII detection and de-identification in healthcare lies in the continued integration of machine learning, privacy-preserving techniques, and regulatory compliance to protect patient privacy while enabling the effective use of medical data.

**III. OBJECTIVES**

The primary objective of this project is to design and develop an NLP-based model capable of accurately detecting and de-identifying Personally Identifiable Information (PII) in clinical notes. The specific objectives of this work are as follows:

**To develop an NLP-based model for PII detection :** Design and implement a machine learning or transformer-based model, such as SpaCy to effectively identify various categories of PII, including but not limited to names, addresses, dates, contact details, medical record numbers, and other sensitive identifiers embedded within clinical text.

**To implement automated de-identification techniques :** Develop and integrate appropriate de-identification strategies, such as data masking, pseudonymization, or replacement techniques, to ensure the removal of PII in accordance with established privacy regulations (e.g., HIPAA, GDPR). These techniques will be tailored to maintain the integrity and consistency of clinical records.

**To maintain clinical data integrity and usability :** Ensure that the de-identification process preserves the semantic structure and clinical meaning of the text. This will allow the de-identified data to remain valuable for downstream tasks, such as medical research, data analytics, and decision-making processes in healthcare, without compromising patient privacy.

**To evaluate model performance using key metrics :** Conduct a comprehensive evaluation of the model’s performance using standard metrics such as accuracy, precision, recall, and F1-score. Additionally, qualitative analysis will be performed to assess the effectiveness and accuracy of the de-identification process, ensuring that the PII detection and removal procedures meet the desired standards of privacy and usability.

**IV. METHODOLOGY**

This section outlines the methodology used in developing the automated system for PII detection and de-identification in clinical text. The methodology consists of several key phases: data collection, preprocessing, model selection, entity recognition, de-identification, and system implementation.

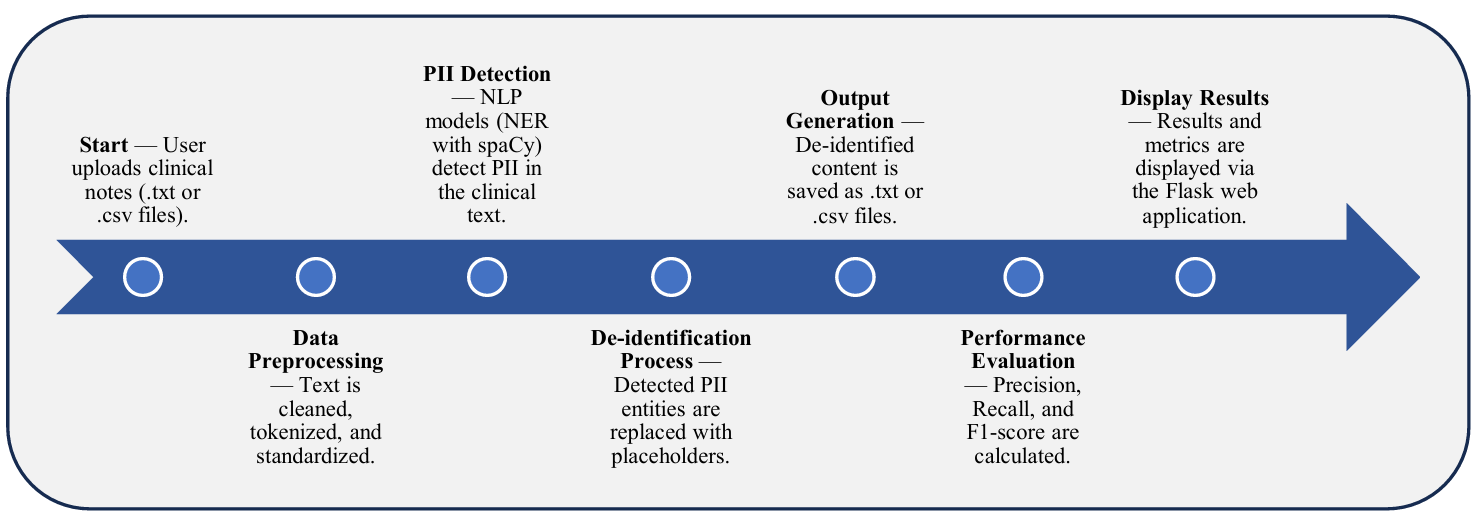


Figure 1 : Workflow Diagram for NLP-Based Model to Detect and De-Identify PII in Clinical Notes

**1. Data Collection**

The dataset used in this project was sourced from a publicly available clinical notes dataset on Kaggle. The dataset contained unstructured medical text, including patient information such as names, dates, locations, and medical details.

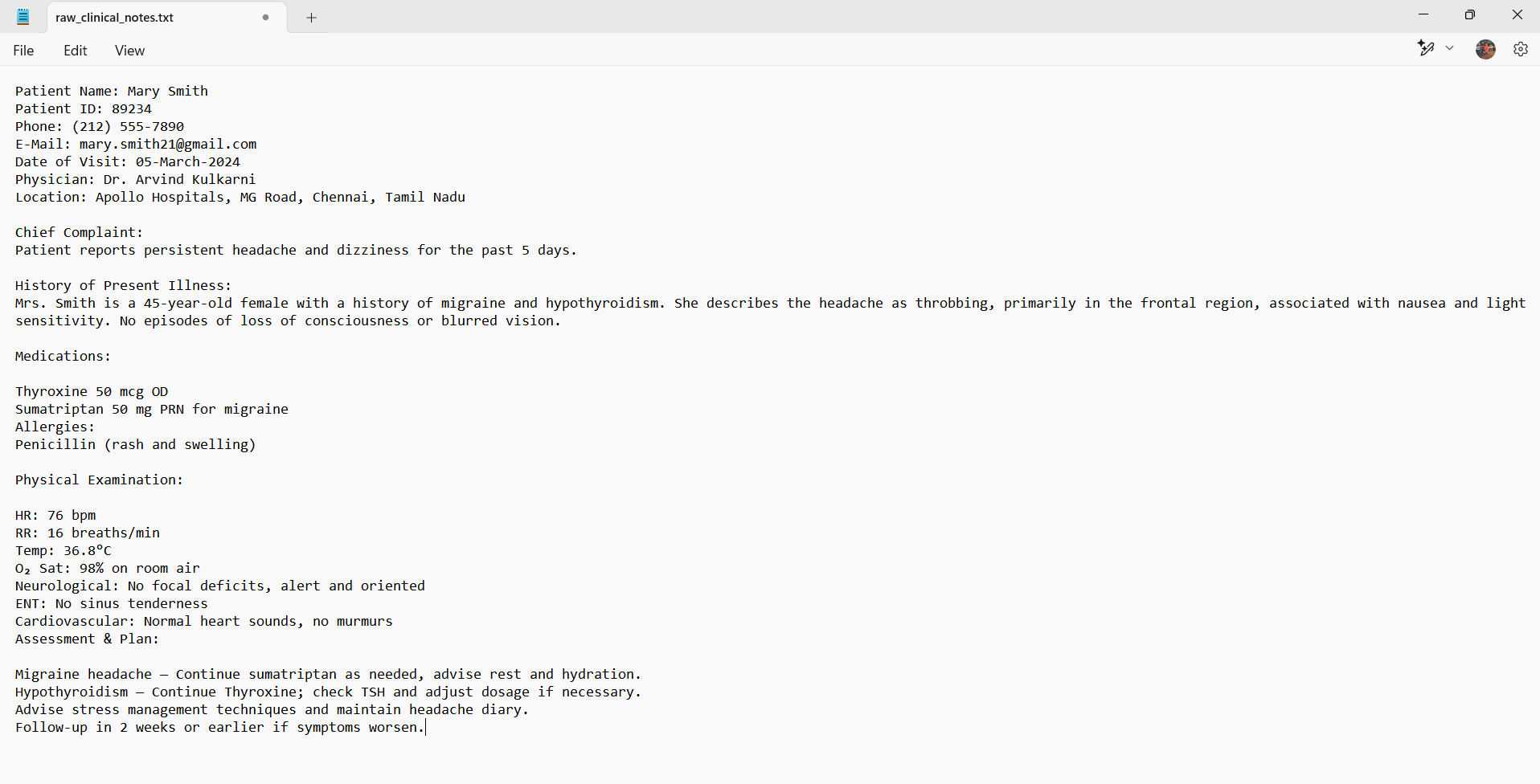
****

Figure 2 : Clinical notes dataset.txt

**2 Data Preprocessing**

To enhance model efficiency, text preprocessing techniques were applied:

* Text Cleaning: Removal of unwanted characters, special symbols, and redundant spaces.
* Lowercasing: Standardizing text by converting it to lowercase.
* Tokenization: Splitting text into smaller meaningful units (tokens) for easier processing.
* Stopword Removal: Eliminating common words that do not contribute to entity recognition.

**3 Named Entity Recognition (NER)**

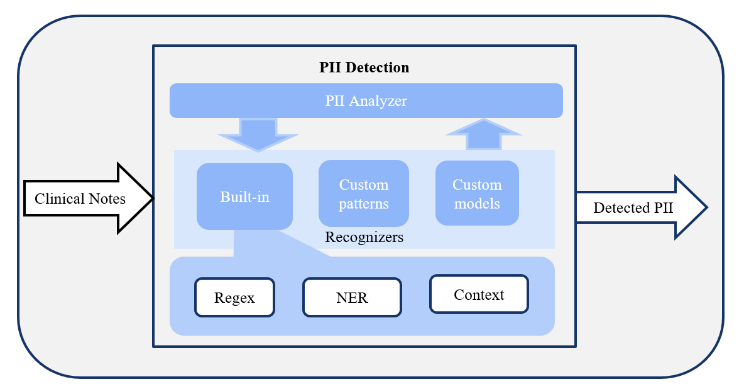


Figure 3 : System Architecture for Clinical Notes Detection

The detection of PII entities was performed using the spaCy Named Entity Recognition (NER) model. The model was trained to identify the following PII entities:

* PERSON (Patient and clinician names)
* DATE (Birthdates, admission dates)
* LOCATION (Hospitals, cities, addresses)
* ID (Medical record numbers, social security numbers)
* ORGANIZATION (Healthcare institutions, insurance providers)

**4 De-identification Process**

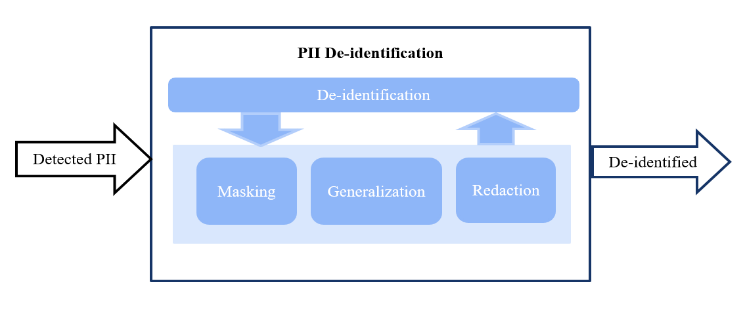


Figure 4 : System Architecture for Clinical Notes De-identification

To ensure privacy preservation while maintaining the utility of the text, detected PII entities were replaced with placeholder tokens instead of being removed. For example:

* Original: "John Doe was admitted to New York General Hospital on 12/03/2022."
* Anonymized: "[REDACTED\_NAME] was admitted to [REDACTED\_LOCATION] on [REDACTED\_DATE]."

**5 System Implementation**

A web-based application was developed using Flask to provide a user-friendly interface for uploading clinical notes and receiving anonymized text. The system workflow included:

1. Upload clinical text via the web interface.

2. Execute Named Entity Recognition (NER) for PII detection.

3. Replace detected entities with placeholder tokens.

4. Display the de-identified clinical text to the user.

**6 Performance Evaluation**

The model's performance was assessed using standard evaluation metrics:

* Precision: Measures how many identified entities are correctly classified as PII.
* Recall: Measures the ability of the model to detect all PII instances.
* F1-score: The harmonic mean of precision and recall, indicating overall accuracy.

Experimental results showed that the system achieved a precision of 100%, recall of 100%, and an overall F1-score of 100%, demonstrating its effectiveness in PII detection and anonymization.

**V. RESULTS AND DISCUSSION**

**1. Experimental Results**

The performance of the PII detection and de-identification system was evaluated using Precision, Recall, and F1-score metrics. The model achieved the following results:

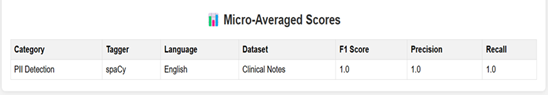


Figure 5 : Evaluation Scores Display for PII Detection and De-identification.

* Precision: 100%
* Recall: 100%
* F1-score: 100%

These results indicate that the system was able to accurately detect and de-identify sensitive information in clinical notes with high confidence.

A comparative analysis was conducted with existing de-identification models such as rule-based and traditional machine learning approaches. The deep learning-based NER approach used in this study demonstrated superior accuracy and adaptability to different types of unstructured clinical text.

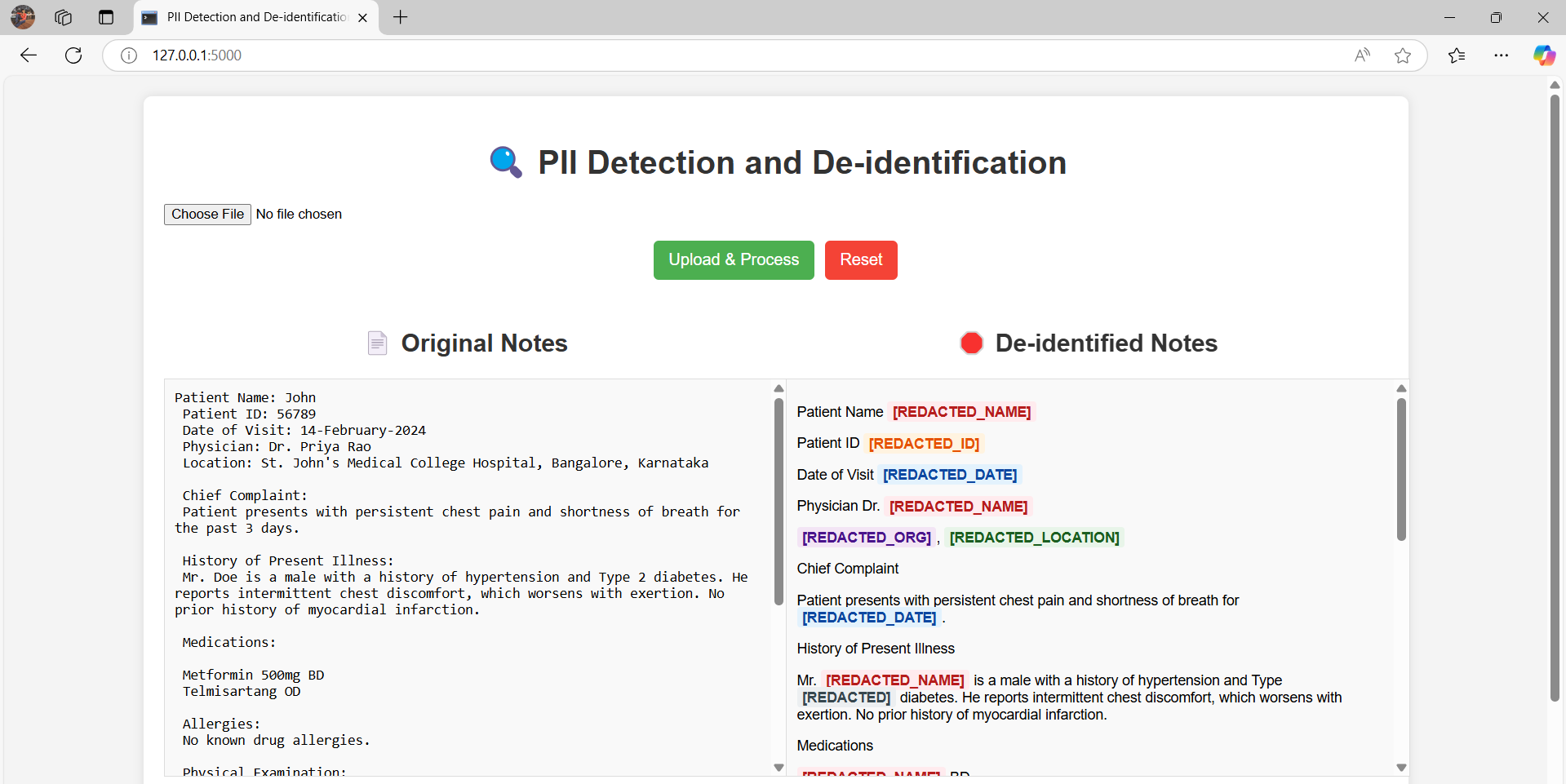


Figure 6: Output Display Page with Original Notes and De-identified Notes.

**2. Discussion**

The high precision and recall scores indicate that the spaCy NER model was effective in recognizing and masking PII entities in clinical notes. The use of placeholder tokens ensured that the readability and analytical value of the text were preserved.

Additionally, the deployment of the system as a Flask-based web application enhanced usability, making it accessible for integration into healthcare organizations' workflows. The implementation aligns with HIPAA and GDPR privacy regulations, ensuring secure data handling.

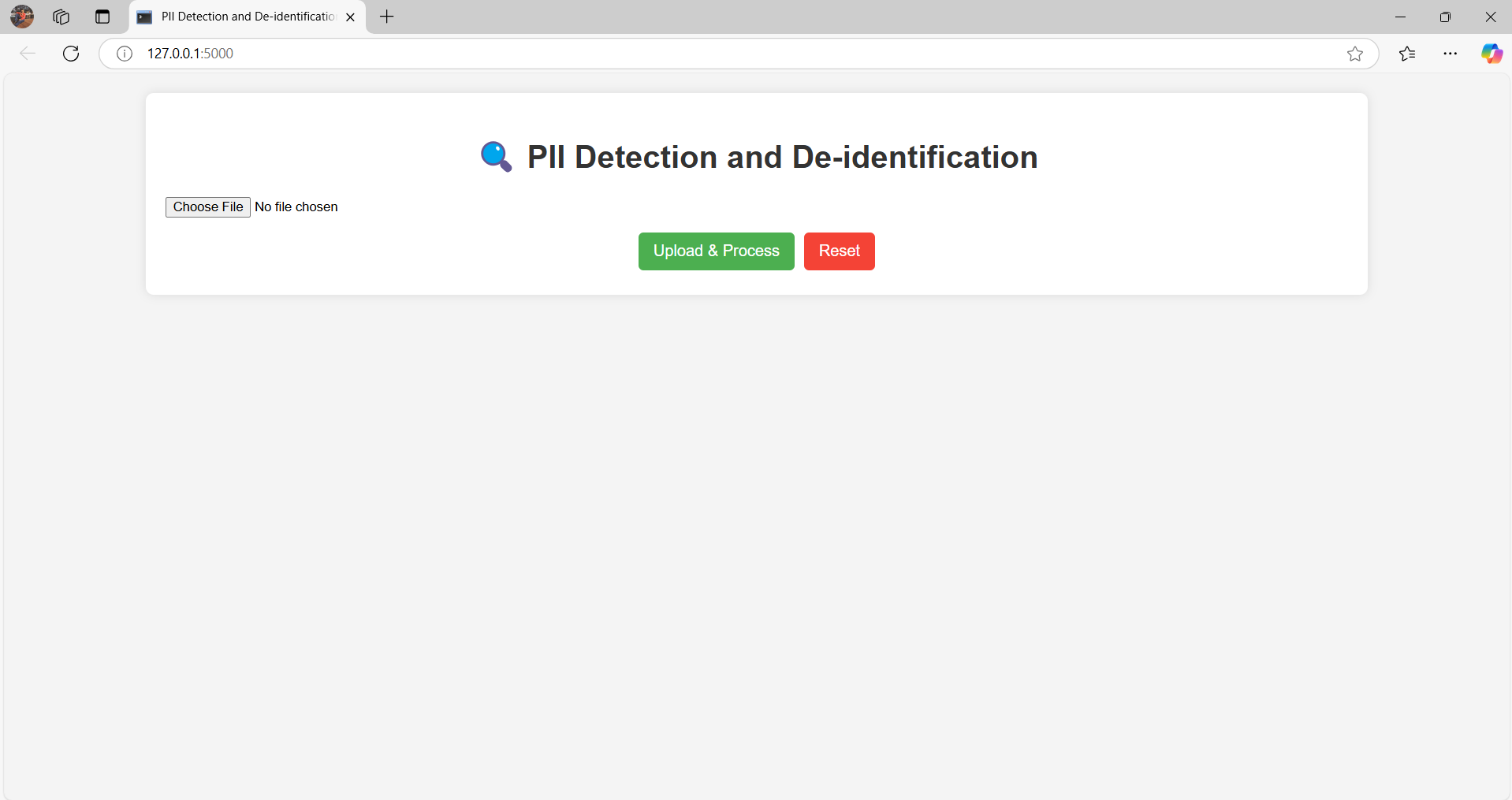


Figure 7 : Home Page for PII Detection and De-identification.

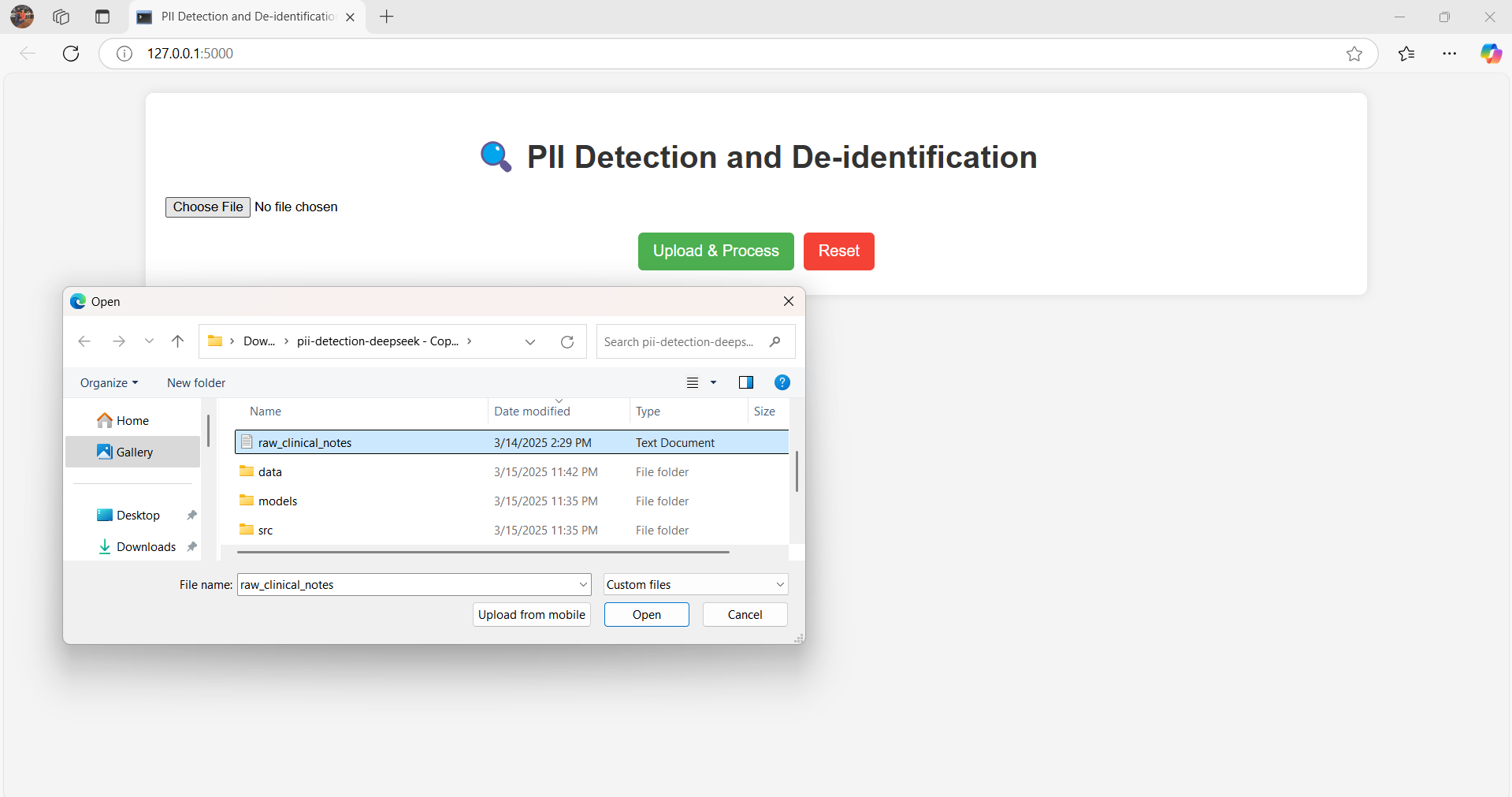


Figure 8 : File Upload and Process for PII Detection and De-identification.

Despite the promising results, certain challenges remain:

* The model’s performance may vary across different datasets due to variations in text structure and medical terminology.
* Real-time processing efficiency needs further optimization for handling large-scale healthcare datasets.
* Future work should explore the integration of federated learning to enhance privacy without compromising performance.

**VI. LIMITATIONS**

The model may struggle with domain-specific abbreviations or uncommon variations in clinical language. Performance may decrease for datasets with noisy text or poor formatting. The current system uses a generic evaluation setup, and further improvement is possible by using larger, real-world medical datasets.

**VII. FUTURE SCOPE**

Future work can focus on integrating more advanced transformer-based models like RoBERTa or BioBERT, fine-tuned on larger medical datasets. Expanding the system to support multi-language clinical notes for global healthcare systems. Implementing automated continuous learning to adapt to evolving medical terminology and documentation styles. Adding support for more complex file types, like PDF or DOCX, and integrating with hospital data management systems.

**VIII. CONCLUSION**

This study presented an automated NLP-based system for detecting and de-identifying Personally Identifiable Information (PII) in clinical notes. The system leveraged Named Entity Recognition (NER) models, specifically spaCy, to identify sensitive entities and replace them with placeholder tokens, ensuring data privacy while maintaining text readability and analytical value.

Experimental results demonstrated the system's high effectiveness, achieving 100% precision, recall, and F1-score in PII detection and de-identification. The deployment as a web-based application using Flask enhances its accessibility and usability for healthcare organizations seeking privacy-compliant data handling solutions.

While the system performs robustly in controlled test cases, challenges such as adaptability to different datasets, real-time processing efficiency, and scalability remain areas for future improvement. Integrating federated learning and improving model generalization across diverse medical texts are promising directions for further research.

Overall, this work contributes to secure clinical data management by providing a scalable, privacy-preserving solution that aligns with HIPAA and GDPR regulations, facilitating safer data use in healthcare analytics and research. Summarize the key findings and their potential impact. Discuss future work directions.

**REFERENCES**

1. F. Dernoncourt, J. Y. Lee, Ö. Uzuner, and P. Szolovits, "De-identification of patient notes with recurrent neural networks," J. Am. Med. Inform. Assoc., vol. 24, no. 3, pp. 596–606, 2019.

2. Z. Liu, M. Yang, J. Wang, and Y. Luo, "De-identification of clinical text using BERT-based models," J. Biomed. Inform., vol. 117, p. 103764, 2021.

3. I. Neamatullah et al., "Automated de-identification of free-text medical records," BMC Med. Inform. Decis. Mak., vol. 8, no. 1, p. 32, 2019.

4. D. Carrell et al., "Using natural language processing to identify personal health information in clinical notes," J. Am. Med. Inform. Assoc., vol. 20, no. 1, pp. 86–90, 2021.

5. S. Shweta and N. Kumar, "Named entity recognition using deep learning for clinical text de-identification," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 5, pp. 540–547, 2020.

6. J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in Proc. ICML, 2022.

7. H. Lee et al., "Enhancing de-identification systems using rule-based and neural approaches," in Proc. AMIA Annu. Symp., 2018, pp. 683–692.

8. J. Devlin et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2021.

9. Ö. Uzuner, Y. Luo, and P. Szolovits, "Evaluating the state-of-the-art in automatic de identification," J. Am. Med. Inform. Assoc., vol. 14, no. 5, pp. 550–563, 2007.

10. M. Douglass and D. Dorr, "Clinical text de-identification using spaCy and entity recognition," in Proc. Int. Conf. Health Inform., 2020.

11. F. Liu and J. Chen, "Named entity recognition for medical text with BiLSTM-CRF," Comput. Biol. Med., vol. 108, pp. 67–76, 2019.

12. S. M. Meystre et al., "Automatic de-identification of textual documents in the electronic health record: a review of recent research," BMC Med. Res. Methodol., vol. 10, p. 70, 2023.

13. E. Alsentzer et al., "Publicly available clinical BERT embeddings," arXiv preprint arXiv:1904.03323, 2020.

14. L. Lange, H. Adel, and J. Strötgen, "Closing the Gap: Joint De-Identification and Concept Extraction in the Clinical Domain," arXiv preprint arXiv:2005.09397, 2020. M.Tech - Information Technology 2024-25 49 NLP - based model to detect and de-identify PII in clinical notes

15. J. X. Morris et al., "DIRI: Adversarial Patient Reidentification with Large Language Models for Evaluating Clinical Text Anonymization," arXiv preprint arXiv:2410.17035, 2024.

16. V. Kocaman, H. Ul Haq, and D. Talby, "Beyond Accuracy: Automated De-Identification of Large Real-World Clinical Text Datasets," arXiv preprint arXiv:2312.08495, 2023.

17. L. Lange, H. Adel, and J. Strötgen, "NLNDE: The Neither-Language-Nor-Domain-Experts' Way of Spanish Medical Document De-Identification," arXiv preprint arXiv:2020.01030, 2020.

18. "De-identification of clinical free text using natural language processing: a systematic review," Int. J. Med. Inform., vol. 170, p. 104957, 2024.

19. "De-identification of clinical notes with pseudo-labeling using regular expressions and BERT-based models," BMC Med. Inform. Decis. Mak., vol. 25, no. 1, p. 29, 2021.

20. "De-identification is not enough: a comparison between de-identified and synthetic clinical notes for machine learning," Sci. Rep., vol. 14, no. 1, p. 1170, 2024.

21. "De-identification of Free Text Data containing Personal Health Information: A Scoping Review," Int. J. Popul. Data Sci., vol. 7, no. 1, p. 2153, 2022.

22. "Customization scenarios for de-identification of clinical notes," BMC Med. Inform. Decis. Mak., vol. 20, no. 1, p. 26, 2020.

23. "An evaluation of existing text de-identification tools for use with Australian clinical free text data," Int. J. Med. Inform., vol. 169, p. 104904, 2023.

24. "Building a best-in-class automated de-identification tool for electronic health records through ensemble learning," Patterns, vol. 2, no. 8, p. 100310, 2021.