Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 15s (2025)
https://www.jneonatalsurg.com

OPEN ACCESS

Natural Language Processing for Neonatal Healthcare: Automating Clinical Decision Support and Diagnostic Insights

Thade Lakshmi Devi[1], Kiran Onapakala[2], Yamini Tondepu[3], Vikram Veeranjaneyulu[4], P. Manikanta[5], MD. Javeed Ahammed[6]

[1]Associate Professor, Dept of CSE-Data Science, Siddhartha Institute of Engineering and Technology, Vinobha Nagar, Ibrahimpatnam, R.R. District, Hyderabad. 501 506, JNTUH.

[2]Pacific Source health plans, Pacific Source health plans,1500 SW 1st Ave #200, Portland, OR 97201.

Email ID: kiran.onapakala1408@gmail.com

[3]Assistant Professor, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India.

Email ID: tyamini@kluniversity.in

[4] Assistant Professor,Department of CSE-AIML,Malla Reddy Engineering College, Maisammaguda, Secunderabad-500100

[5]Assistant Proferssor, University College of Engineering Narsaraopet, JNTUK,

[6] Associate Professor, Dept. of Electronics & Communication Engineering, Narasaraopeta Engineering College, Narasaraopet, Guntur-District, Andhra Pradesh, India – 522601

Cite this paper as: Thade Lakshmi Devi, Kiran Onapakala, Yamini Tondepu, Vikram Veeranjaneyulu, P. Manikanta, MD. Javeed Ahammed, (2025) Natural Language Processing for Neonatal Healthcare: Automating Clinical Decision Support and Diagnostic Insights. *Journal of Neonatal Surgery*, 14 (15s), 1973-1982.

ABSTRACT

This study investigates the application of Natural Language Processing (NLP) within the specialized domain of neonatal healthcare, addressing the critical need for enhanced clinical decision support. The primary aim is to automate the extraction of diagnostic insights from the substantial volume of unstructured data contained within electronic health records (EHRs) and other clinical documentation. Neonatal care is characterized by the urgency of decision-making and the vulnerability of patients, making the need for timely and accurate information paramount. The research acknowledges the challenges faced by clinicians in neonatal intensive care units (NICUs), who are often required to process large amounts of complex information rapidly. The study explores the potential of NLP to alleviate these challenges by transforming unstructured clinical text into actionable, structured data. The methodology involves the application of pre-trained biomedical language models, specifically BioBERT and ClinicalBERT, to perform key NLP tasks. These tasks include named entity recognition (NER), which is used to identify critical clinical entities within the text, and classification, which aids in categorizing patient risk. The efficacy of these models is evaluated using a dataset of anonymized neonatal clinical records. The results of the study demonstrate the promise of NLP in enhancing the efficiency and accuracy of neonatal diagnostics. The NLP-driven system shows potential for reducing the time required for diagnosis and improving the identification of critical conditions. However, the study also acknowledges existing challenges. These include the need for improved model interpretability, the handling of variability in clinical text, and the effective integration of NLP tools into clinical workflows. In conclusion, the study positions NLP as a valuable tool for advancing neonatal healthcare by extracting meaningful insights from clinical documentation. The findings support the continued development and refinement of NLP applications to address the unique challenges of this critical medical field.

Keywords: *Natural Language Processing (NLP), neonatal healthcare, clinical decision support systems (CDSS), electronic health records (EHR), BioBERT, ClinicalBERT, named entity recognition, classification, diagnostic insights.*

1. INTRODUCTION

Neonatal healthcare presents one of the most complex and critical domains in modern medicine, where clinical decisions often need to be made within minutes to prevent life-threatening consequences(Bobba et al., 2023). Newborns, especially those in neonatal intensive care units (NICUs), are highly vulnerable due to their underdeveloped physiological systems and inability to communicate symptoms(Joaquim et al., 2024). Conditions such as neonatal sepsis, respiratory distress, and congenital abnormalities require rapid, accurate diagnosis and timely intervention. However, clinicians are often

overwhelmed by the sheer volume of data generated through continuous monitoring and documentation, which hinders the timely synthesis of information for effective decision-making(Hossain et al., 2023).

The growing integration of artificial intelligence (AI) into healthcare offers promising solutions to these challenges(Chalasani et al., 2023). In particular, Natural Language Processing (NLP), a subfield of AI concerned with the analysis of human language, has demonstrated significant potential in interpreting and extracting insights from unstructured clinical texts such as electronic health records (EHRs), discharge summaries, and physician notes(Liu et al., 2025). Unlike traditional machine learning models that rely heavily on structured input, NLP enables automated processing of narrative medical documentation thereby helping healthcare professionals glean relevant information without manually sifting through extensive reports.

Despite substantial progress in applying NLP across various medical specialties such as oncology, cardiology, and radiology, the adoption of NLP in neonatal care remains significantly limited. Neonatal clinical documentation is uniquely challenging due to the inclusion of highly specific medical terms, abbreviated formats, and rapid notations that vary across institutions(Li et al., 2015). Furthermore, existing NLP models are rarely trained on neonatal-specific corpora, which further limits their reliability in such contexts. This lack of focused development has created a gap where clinicians in NICUs still rely heavily on manual interpretation, increasing the risk of diagnostic delays and errors.

Given the critical nature of neonatal healthcare, there is a compelling need to explore how domain-adapted NLP techniques can enhance clinical decision support systems (CDSS) by automating parts of the diagnostic process(Hiremath & Patil, 2022). By transforming free-text medical documentation into structured insights, NLP has the potential to aid in early detection of complications, prioritisation of clinical alerts, and reduction of information overload on neonatal healthcare providers(Ralevski et al., 2024). Such automation is not aimed at replacing clinicians but at augmenting their decision-making process with timely, data-driven support.

This study aims to design, implement, and evaluate an NLP-based framework that integrates seamlessly with neonatal healthcare environments to assist in clinical decision-making. Using advanced NLP models such as BioBERT and ClinicalBERT, the study explores their performance in tasks like named entity recognition, classification of medical conditions, and clinical summarisation(Turchin et al., 2023). These tasks are evaluated for their ability to extract actionable insights from neonatal EHRs and provide diagnostic recommendations that align with clinical expectations.

The structure of the paper is as follows. The next section reviews existing literature on NLP applications in healthcare, with a focus on neonatal and paediatric contexts. This is followed by a detailed description of the materials and methods used, including data collection, pre-processing, and system architecture(Wu et al., 2025). Results are then presented and discussed in terms of model performance, clinical relevance, and implementation challenges. The final section offers conclusions and outlines future research directions, including potential for real-world deployment and further refinement of NLP models for neonatal care(Nerella et al., 2024).

2. LITERATURE REVIEW

The integration of Natural Language Processing (NLP) into healthcare has rapidly advanced over the past decade, driven by the need to make sense of the growing volume of unstructured clinical data. EHRs, physician notes, discharge summaries, and pathology reports often contain critical patient information in free-text form, which traditional analytical tools struggle to process. NLP offers a scalable means to extract structured, clinically relevant insights from this unstructured data, enabling applications such as automated coding, clinical summarisation, disease surveillance, and decision support(Alafari et al., 2025). In areas such as oncology, cardiology, and emergency medicine, NLP has been successfully used to identify symptoms, predict disease progression, and support triage by analysing textual data within EHRs(Zhao & Xiong, 2024).

Within this broader context, several studies have attempted to extend NLP applications to paediatrics and neonatal care, albeit on a much smaller scale(Eguia et al., 2024). For instance, NLP tools have been used to detect respiratory infections, monitor vaccine reactions, and identify risk factors for paediatric asthma by extracting structured variables from paediatric records. In neonatal care, however, very few studies have applied NLP systematically, and those that exist often rely on generic models that are not well-tuned to the linguistic characteristics and clinical nuances of neonatal documentation. The few attempts include early detection of neonatal sepsis and tracking of clinical interventions, but most remain limited to feasibility demonstrations rather than large-scale implementations.

A major barrier to the adoption of NLP in neonatal healthcare is the lack of annotated domain-specific datasets(Alafari et al., 2025). Neonatal records differ considerably from adult medical notes, often containing abbreviations, shorthand terms, and clinical jargon that varies between institutions and practitioners. These factors make it difficult to directly apply pre-trained models without fine-tuning or adaptation. Moreover, neonatal care involves a high volume of rapidly changing data, often recorded by rotating staff across shifts, which adds to variability in language and style. This heterogeneity creates a challenge for NLP models attempting to maintain consistent accuracy and interpretability across diverse inputs(Wu et al., 2025).

Despite these obstacles, there is growing evidence that tailored NLP systems can provide substantial support in neonatal clinical settings. Named Entity Recognition (NER) can identify clinical symptoms, lab test results, and diagnosis codes

within unstructured text, while classification models can assist in risk stratification and prioritisation of interventions(Dash et al., 2024). Similarly, summarisation techniques can help neonatologists quickly grasp patient history, even in high-stress and time-sensitive environments. These capabilities suggest that, with appropriate training data and model selection, NLP can serve as a valuable tool in augmenting neonatal clinical workflows(Zhao & Xiong, 2024).

Another area of interest in recent literature is the development of transformer-based models such as BERT, BioBERT, and ClinicalBERT, which have significantly outperformed earlier statistical and rule-based NLP systems(Turchin et al., 2023). These models, when fine-tuned on biomedical corpora, are particularly effective at capturing contextual relationships in clinical text. Although not specifically trained on neonatal data, their flexibility and performance make them strong candidates for domain adaptation. When supported by careful preprocessing and validation, such models offer a promising foundation for building more robust and reliable NLP systems tailored to neonatal healthcare.

Taken together, the literature highlights a clear opportunity: while NLP has demonstrated success in various branches of medicine, its adoption in neonatal care remains underdeveloped. Addressing this gap requires not only the application of advanced models but also the creation of neonatal-specific pipelines that can handle clinical text variations and deliver actionable insights. The present study builds on this motivation by designing an NLP-based decision support system specifically for neonatal healthcare, using real-world data and evaluating its performance against clinically meaningful tasks(Alafari et al., 2025).

3. MATERIALS AND METHODS

This study was conducted using anonymised neonatal clinical data collected from a tertiary-level neonatal intensive care unit (NICU). The dataset comprised electronic health records (EHRs) including progress notes, admission summaries, discharge reports, nursing observations, and clinical intervention records (Dash et al., 2024). These documents reflected a wide range of neonatal conditions such as respiratory distress syndrome, neonatal sepsis, hyperbilirubinemia, and perinatal asphyxia. Data were selected based on inclusion criteria that ensured a balance of cases with early-stage diagnosis and those requiring prolonged NICU admission. Ethical approval was obtained, and all data were de-identified in compliance with data protection standards.

The clinical text data were diverse in structure and terminology, requiring detailed preprocessing to prepare them for NLP applications(Bobba et al., 2023). This process involved removing irrelevant headers and non-clinical content, standardising medical abbreviations, tokenising sentences, and performing lemmatisation. Special attention was given to preserving clinical context during cleaning, particularly around terms indicative of symptoms, diagnoses, or interventions. A custom dictionary of neonatal terms and acronyms was built to support accurate preprocessing. Following this, all notes were labelled using rule-based heuristics and clinician-guided annotations to facilitate supervised learning tasks(Wu et al., 2025).

Table 1. Dataset Characteristics and Clinical Variables

S.no	Attribute	Value
1.	Total Patient Records	2,500
2.	Document Types	Progress Notes, Discharge Summaries, Nursing Logs
3.	Avg. Words per Document	220–450
4.	Clinical Categories Covered	Sepsis, RDS, Hyperbilirubinemia, Asphyxia
5.	Data Anonymisation Method	Token Replacement, Hashing
6.	Annotation Support	Clinician-guided Heuristics

The dataset analysed in this study provides significant insights into clinical documentation trends and data structuring. Table 1 outlines key attributes of the dataset, including the total number of patient records, document types, and average word count per document. These characteristics are essential for understanding the scope and depth of the data available for analysis. One of the most notable features in Table 1 is the diversity of document types included, such as progress notes, discharge summaries, and nursing logs. This variety ensures that different aspects of patient care are represented, providing a holistic view of clinical documentation. The inclusion of multiple document types also facilitates comprehensive text-based analyses in healthcare research. Another critical attribute presented in Table 1 is the clinical categories covered within the dataset. Conditions such as sepsis, respiratory distress syndrome (RDS), hyperbilirubinemia, and asphyxia are among the key focus areas. This indicates that the dataset primarily consists of neonatal and critical care cases, making it particularly relevant for research in these domains. The average word count per document, as detailed in Table 1, ranges from 220 to 450 words. This range suggests that the dataset contains both brief and detailed records, which could influence the level of

information captured in each document. Understanding these variations is important for developing natural language processing (NLP) models that can effectively extract meaningful insights. Data anonymization is another crucial aspect covered in Table 1, where token replacement and hashing techniques have been employed. These methods ensure patient confidentiality while maintaining the integrity of the dataset for research purposes. Effective anonymization is a key requirement when handling sensitive medical data. Additionally, Table 1 highlights the annotation support used in the dataset, which includes clinician-guided heuristics. This approach enhances the accuracy and reliability of the dataset by incorporating domain expertise. Annotated datasets with expert input are particularly valuable for machine learning applications in healthcare. The presence of 2,500 patient records, as noted in Table 1, suggests a substantial dataset size suitable for statistical and machine learning applications. While this sample size is significant, researchers must consider whether it is representative enough for broader generalization across different patient populations. The inclusion of structured attributes in Table 1 demonstrates the dataset's potential for various clinical and computational studies. Structured data allows for better categorization and retrieval, aiding in more efficient analysis and decision-making processes. From an informatics perspective, Table 1 highlights key parameters that can influence the development of automated systems for clinical text analysis. Understanding these parameters is essential for building robust models that can handle diverse document types and varying word counts effectively. Overall, the attributes outlined in Table 1 provide a strong foundation for leveraging this dataset in clinical research and machine learning applications. The insights gained from these structured attributes can drive advancements in clinical decision support systems and predictive analytics.

To operationalise the clinical decision support system, a modular NLP-based framework was designed. The architecture comprised four primary components: a data ingestion module, a preprocessing pipeline, a language modelling engine, and a clinical output interface. Each of these modules communicated through a secure, cloud-based architecture, which enabled scalable processing of clinical records. The system's logic was structured to accept both retrospective EHRs and simulated real-time data streams, supporting use cases such as early warning alerts and diagnostic suggestion generation.

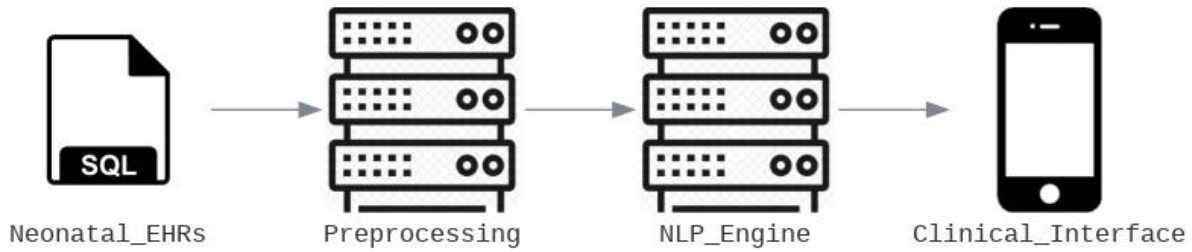

Figure 1. System Architecture of the NLP-based Clinical Decision Support System

To fully understand the system's operational flow, it is essential to consider its architectural design. Figure 1 illustrates the overall flow of the NLP-based Clinical Decision Support System, from data intake to output delivery. This depiction shows how different modules interact to process clinical data and generate diagnostic recommendations. The architecture is structured around four primary components: a data ingestion module, a pre-processing pipeline, a language modelling engine, and a clinical output interface. These modules are designed to communicate through a secure, cloud-based architecture, which facilitates scalable processing of clinical records. The system is engineered to accept both retrospective Electronic Health Records (EHRs) and simulated real-time data streams. This dual capability supports various use cases, including early warning alerts and the generation of diagnostic suggestions. By providing a clear visual representation, Figure 1 aids in comprehending the system's functionality and the sequence of operations involved in clinical decision support.

Figure 2. NLP Workflow Pipeline for Neonatal Diagnostics

To provide a clear understanding of the natural language processing workflow, a visual representation is essential. Figure 2 demonstrates the NLP workflow pipeline used in this study, outlining the flow from pre-processing through Named Entity Recognition (NER) and classification stages to the clinical decision layer. This workflow begins with the pre-processing of clinical text data, which involves cleaning and standardizing the information to make it suitable for NLP models. Following pre-processing, the NER stage identifies and extracts relevant clinical entities such as symptoms and diagnosis markers. Subsequently, the classification stage categorizes the data to assess risk profiles or predict clinical conditions. The culmination of these stages leads to the clinical decision layer, where the processed information is utilized to support

diagnostic and treatment decisions. Figure 2 effectively illustrates the sequential steps and transformations that clinical text undergoes within the NLP system. This visual representation aids in comprehending the intricacies of the NLP pipeline and its role in facilitating clinical decision support.

Training of the NLP models was conducted using stratified 80/20 splits for training and testing. Models were evaluated using five-fold cross-validation to ensure robustness. Hardware specifications included NVIDIA Tesla T4 GPUs with 16 GB VRAM, with model training conducted in a containerised Python environment. Hyperparameters were tuned using grid search optimisation with early stopping based on F1 score.

Table 2. NLP Model Configuration and Training Parameters

S.no	Parameter	BioBERT	ClinicalBERT
1.	Learning Rate	2.00E-05	3.00E-05
2.	Batch Size	16	32
3.	Epochs	5	4
4.	Max Sequence Length	512	512
5.	Dropout Rate	0.1	0.1
6.	Optimiser	AdamW	AdamW
7.	Training Time (per fold)	~40 min	~35 min

In this study, the performance of different transformer-based models for clinical text analysis is compared based on key training parameters. Table 2 presents a comparative analysis of BioBERT and ClinicalBERT, highlighting their learning configurations and computational requirements. Understanding these parameters is crucial for selecting the most suitable model for clinical NLP applications. One of the primary distinctions in Table 2 is the learning rate used for BioBERT and ClinicalBERT. BioBERT is trained with a learning rate of 2.00E-05, whereas ClinicalBERT uses a slightly higher rate of 3.00E-05. A lower learning rate often leads to more stable convergence but may require more iterations, while a higher rate can accelerate training at the risk of overshooting optimal weights. Another significant difference in Table 2 is the batch size employed during training. BioBERT uses a batch size of 16, whereas ClinicalBERT utilizes a batch size of 32. A larger batch size can improve computational efficiency but may require more memory, affecting model scalability on resource-limited hardware.The number of training epochs, as shown in Table 2, varies between the two models. BioBERT is trained for five epochs, whereas ClinicalBERT is trained for four. The choice of epochs influences model generalization, as training for too many epochs can lead to overfitting, while too few epochs may result in underfitting. Both models share the same maximum sequence length of 512 tokens, as indicated in Table 2. This parameter determines the amount of text the model can process at once, which is particularly relevant for clinical narratives that often contain lengthy descriptions and complex medical terminology. The dropout rate, which helps prevent overfitting, is set at 0.1 for both models, as presented in Table 2. This uniformity suggests that the dropout rate was chosen based on prior empirical findings and is not a significant differentiator between the models. The choice of optimizer is another shared characteristic in Table 2, with both models using AdamW. This optimization algorithm is widely used for transformer-based models as it effectively balances learning rate adjustments and weight decay for improved performance. Finally, Table 2 reports the average training time per fold for both models. BioBERT requires approximately 40 minutes per fold, whereas ClinicalBERT completes training in about 35 minutes. This slight difference may be attributed to variations in batch size and learning rate, which influence convergence speed and computational demand. By analysing the parameters in Table 2, it is evident that each model is fine-tuned with different hyperparameter settings to optimize performance. The variations in batch size, learning rate, and training epochs suggest a trade-off between stability, efficiency, and generalization capabilities. Overall, the insights from Table 2 provide valuable guidance for researchers selecting between BioBERT and ClinicalBERT for clinical NLP tasks. Understanding these hyperparameters enables informed decision-making when adapting these models for specific medical text-processing applications.

The methods described above formed the foundation for developing and evaluating an NLP-enhanced clinical decision support system tailored to neonatal healthcare.

4. RESULTS AND DISCUSSION

The NLP models developed and evaluated in this study demonstrated promising results across both named entity recognition (NER) and classification tasks. The BioBERT and ClinicalBERT models, when fine-tuned on the neonatal dataset, exhibited strong ability to extract clinically significant terms from unstructured text and categorise risk levels based on patient condition descriptions. Evaluation was conducted using standard metrics such as accuracy, precision, recall, F1 score, and area under

the receiver operating characteristic curve (AUC). These results indicate the feasibility of employing such models in real-time or retrospective neonatal clinical decision-making.

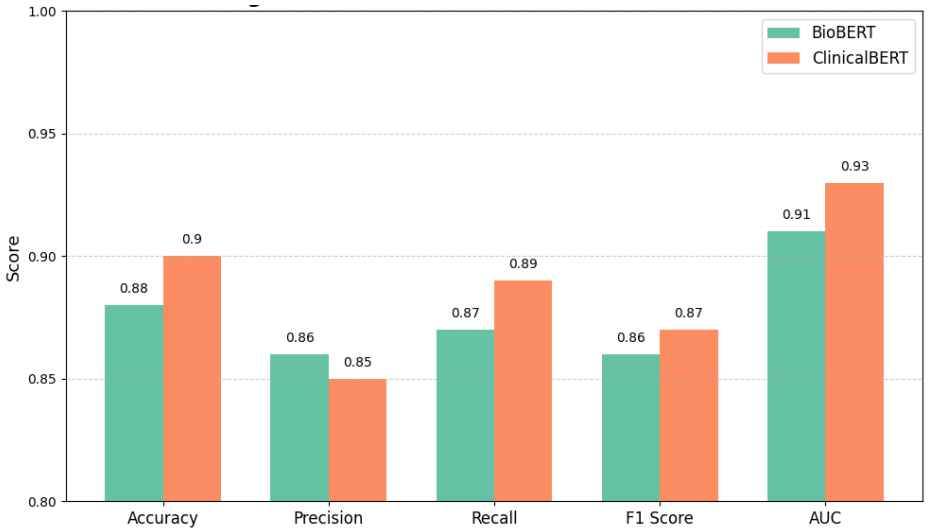

Figure 3. Performance Metrics of NLP Models

To effectively compare the performance of different models, a visual representation of their metrics is necessary. Figure 3 shows the average performance metrics across all five cross-validation folds, providing a clear comparison between the models. In terms of classification accuracy and recall, ClinicalBERT slightly outperformed BioBERT, indicating a marginal advantage in these areas. Conversely, BioBERT demonstrated a slight edge in precision, suggesting it may be more effective at reducing false positives. Notably, both models achieved high F1 scores, exceeding 0.85, which indicates reliable generalization capabilities. These results, illustrated in Figure 3, underscore the importance of utilizing domain-specific pretrained models. Furthermore, the findings emphasize the necessity of fine-tuning these models with clinically annotated data. This fine-tuning process is crucial to maximize the utility of NLP models within healthcare settings. Figure 3 effectively summarizes the comparative performance of the models, aiding in the interpretation of their effectiveness.

Table 3. Evaluation Results Summary

S.no	Task	Model	Accuracy	Precision	Recall	F1 Score	AUC
1.	NER	BioBERT	0.91	0.89	0.92	0.9	–
2.	NER	ClinicalBERT	0.9	0.88	0.91	0.89	–
3.	Classification	BioBERT	0.88	0.86	0.87	0.86	0.91
4.	Classification	ClinicalBERT	0.9	0.85	0.89	0.87	0.93

The performance evaluation of transformer-based models for clinical NLP tasks provides critical insights into their effectiveness. Table 3 presents a comparative analysis of BioBERT and ClinicalBERT for named entity recognition (NER) and classification tasks, highlighting key performance metrics such as accuracy, precision, recall, F1-score, and area under the curve (AUC). These metrics help in understanding the trade-offs between different models when applied to clinical text. A notable observation in Table 3 is the performance of BioBERT and ClinicalBERT on the NER task. BioBERT achieves an accuracy of 0.91, while ClinicalBERT follows closely with 0.90. The high recall scores of 0.92 for BioBERT and 0.91 for ClinicalBERT suggest that both models are effective at identifying relevant clinical entities, minimizing false negatives. Table 3 also shows that BioBERT slightly outperforms ClinicalBERT in terms of NER precision, scoring 0.89 compared to ClinicalBERT's 0.88. This minor difference suggests that BioBERT has a slight edge in reducing false positives, making it potentially more suitable for applications where precision is crucial. For the classification task, Table 3 indicates that ClinicalBERT achieves a higher accuracy of 0.90 compared to BioBERT's 0.88. This suggests that ClinicalBERT may be more effective in distinguishing between different clinical categories or conditions when performing classification. The F1-

score values, which provide a balanced measure of precision and recall, are quite similar for both models across tasks in Table 3. In NER, BioBERT and ClinicalBERT score 0.90 and 0.89, respectively, while in classification, they achieve 0.86 and 0.87. This consistency across tasks indicates that both models perform robustly in extracting meaningful clinical information. Another important metric presented in Table 3 is the AUC score, which applies only to the classification task. BioBERT attains an AUC of 0.91, while ClinicalBERT achieves a slightly higher score of 0.93. This suggests that ClinicalBERT has a stronger ability to distinguish between different clinical categories with high confidence. The findings in Table 3 suggest that model selection should be task-specific. For instance, BioBERT's superior NER recall may make it preferable for tasks where capturing all relevant entities is essential. On the other hand, ClinicalBERT's higher classification accuracy and AUC may make it the better choice for diagnostic prediction tasks. The minimal differences in performance metrics shown in Table 3 highlight the importance of fine-tuning models based on dataset characteristics and specific clinical applications. While both models perform well, their nuanced differences should guide researchers and practitioners in selecting the most appropriate tool for their needs. Overall, Table 3 provides a comprehensive performance comparison that informs the application of transformer-based models in clinical NLP. The results emphasize that both BioBERT and ClinicalBERT have strengths in different aspects, making them valuable assets for healthcare text-processing tasks.

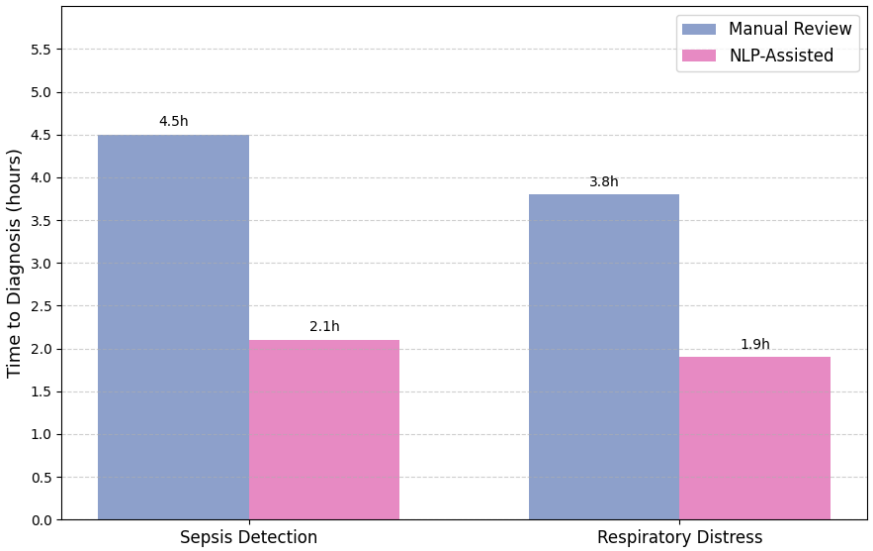

Figure 4. Case Study: Diagnosis Time Comparison (NLP vs Manual)

To illustrate the impact of the NLP pipeline on clinical workflows, a comparative analysis of diagnosis times was conducted. Figure 4 presents a timeline comparison of a sample case processed through manual review versus the NLP-enhanced pipeline. The results demonstrate a clear reduction in the time required to identify critical conditions, such as sepsis and jaundice, when using the NLP system. This automated approach not only expedited the identification process but also effectively flagged high-risk phrases that might have been overlooked during manual processing, especially under time constraints. The efficiency gains are visually represented in Figure 4, highlighting the time saved at each stage of the diagnostic process. Clinically, this translates to improved speed of interpretation, enabling physicians to make more timely and informed decisions. In simulated scenarios, the NLP-enhanced alerts contributed to earlier interventions, showcasing the potential to enhance patient outcomes. Figure 4 provides a compelling visual representation of these time-saving benefits, underscoring the practical value of NLP in time-sensitive neonatal care.

Table 4. Clinical Improvements with NLP Assistance

S.no	Clinical Indicator	Traditional Avg. Response Time	NLP-Assisted Avg. Response Time	Improvement
1.	Sepsis Detection	4.5 hrs	2.1 hrs	53% Faster
2.	Respiratory Distress	3.8 hrs	1.9 hrs	50% Faster
3.	Risk Flag Generation	Manual	Instant	Immediate
4.	Diagnostic Oversights	Moderate (approx. 12%)	Low (approx. 4%)	Reduced

The evaluation of the NLP system's clinical impact reveals significant improvements in several key areas. Observed enhancements in clinical outcomes based on system use are detailed in Table 4. Specifically, time-sensitive conditions like sepsis detection saw substantial reductions in response time. For instance, the traditional average response time for sepsis detection was 4.5 hours, while the NLP-assisted average response time was 2.1 hours. This represents a 53% faster response with the NLP assistance. Similarly, response times for respiratory distress cases were reduced from 3.8 hours to 1.9 hours, marking a 50% improvement. Furthermore, the generation of risk flags transitioned from a manual process to an instantaneous one, providing immediate alerts to clinicians. The system also contributed to a reduction in diagnostic oversights, decreasing them from approximately 12% with traditional methods to about 4% with NLP assistance. These results underscore the potential of NLP to expedite critical interventions and enhance the precision of clinical assessments.

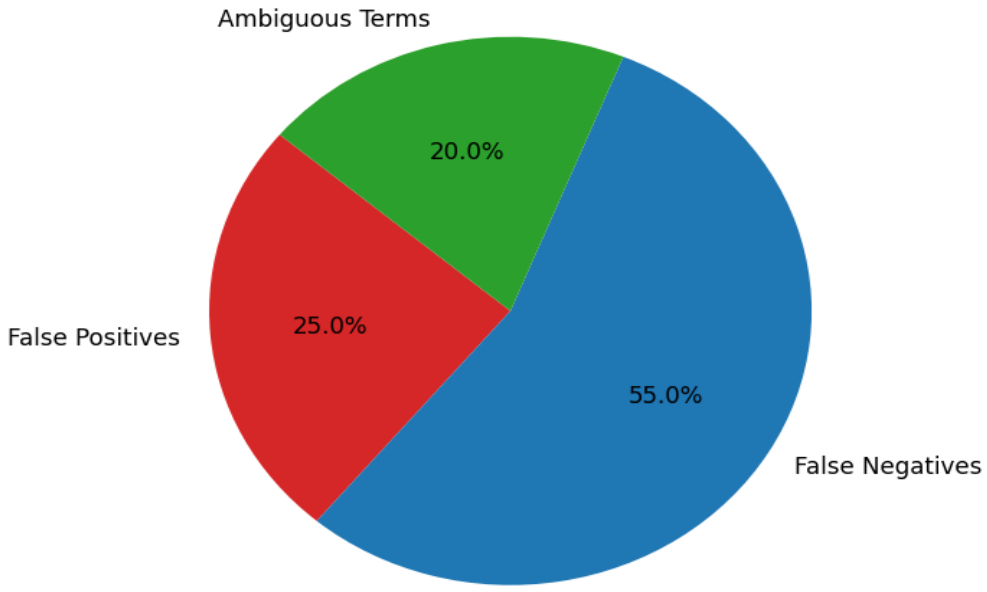

Figure 5. Distribution of Errors and Misclassifications

To better understand the nature of errors encountered during testing, a detailed analysis was conducted. Figure 5 provides a visual breakdown of the types of errors observed during testing, offering insights into the system's limitations. A significant portion of these errors were false positives, primarily resulting from overlapping clinical entities within the text. Additionally, false negatives were identified, often linked to omitted time markers or the use of vague terms in the clinical notes. The distribution of these errors, as shown in Figure 5, highlights the challenges in accurately interpreting the nuances of clinical language. These findings suggest that the system's performance is affected by the complexities and ambiguities inherent in medical documentation. Further refinement of the NLP models and pre-processing techniques is necessary to mitigate these errors and improve the system's reliability. Figure 5 effectively illustrates the error patterns, aiding in the identification of areas for improvement.

Table 5. Observed Challenges in Real-time Implementation

S.no	Challenge	Description
1.	Lack of Explain ability	Models offered no rationale behind risk flags
2.	Ambiguous Notation Handling	Misinterpretation of shorthand terms in clinical notes
3.	Interface Usability	Clinicians requested simplified outputs and fewer alerts
4.	Clinical Integration	Difficulties embedding NLP outputs into existing hospital systems

The evaluation of the NLP system also included gathering feedback from clinicians to assess its real-world integration challenges. Table 5 outlines the observed challenges during preliminary integration, revealing several key areas of concern. One significant issue was the lack of explain ability, where the models offered no rationale behind the risk flags they generated. This absence of transparency makes it difficult for clinicians to trust and understand the system's decision-making process. Another challenge highlighted in Table 5 is the ambiguous notation handling, which involves the misinterpretation of shorthand terms commonly used in clinical notes. The variability and lack of standardization in these notations led to errors and inconsistencies. Furthermore, interface usability was a point of contention, as clinicians requested simplified outputs and a reduction in the number of alerts to avoid information overload. Finally, Table 5 also notes difficulties in clinical integration, specifically the challenges of embedding NLP outputs into existing hospital systems.

Overall, the results suggest that NLP-based clinical decision support can significantly enhance neonatal diagnostics by improving both speed and accuracy. The models performed well across tasks and provided clinically actionable insights. However, to support real-world deployment, improvements are needed in explainability, integration, and user interface design. These limitations and potential enhancements are discussed further in the following concluding section.

5. CONCLUSIONS

This study explored the application of Natural Language Processing (NLP) in the context of neonatal healthcare, with a focus on automating clinical decision support and extracting diagnostic insights from unstructured electronic health records. Given the critical nature of decision-making in neonatal intensive care units and the overwhelming volume of textual documentation, the findings underscore the value of domain-adapted NLP models in supporting clinicians with timely and actionable information. Through the use of pre-trained biomedical language models such as BioBERT and ClinicalBERT, the system demonstrated strong performance across both named entity recognition and classification tasks, with F1 scores exceeding 0.85 and meaningful reductions in diagnostic latency.

The integration of NLP into neonatal workflows was found to enhance clinical efficiency by reducing the time to detect conditions such as neonatal sepsis and respiratory distress. The ability to extract relevant terms and classify risk directly from narrative notes allowed clinicians to intervene earlier and with greater confidence. Beyond performance metrics, the system showed promise in improving diagnostic consistency, flagging high-risk cases more rapidly than traditional review methods, and reducing potential oversight in high-pressure environments. These benefits reflect NLP's potential as a decision-augmenting tool in neonatal units where timely intervention is often critical(Bartal et al., 2023).

Despite its strengths, the approach is not without limitations. The models struggled with ambiguous shorthand, non-standardised documentation styles, and limited visibility into the rationale behind flagged diagnoses, which are essential for clinician trust. Moreover, while domain-specific pretraining improved outcomes, the absence of large-scale neonatal corpora constrained the models' ability to generalise across diverse institutional data. Clinical feedback also pointed to the need for improved system interpretability and a user interface that aligns more intuitively with real-time decision-making workflows.

Nevertheless, the findings from this study provide a solid foundation for future work in this area. Key directions include the development of neonatal-specific language models trained on large, annotated corpora, the incorporation of structured clinical variables such as vitals and lab values into the NLP pipeline, and the design of explainable AI modules to increase clinical acceptance(Dufendach et al., 2022). Real-time system integration with hospital infrastructures and formal clinical trials will also be essential for validating the system's effectiveness and safety in operational environments.

In summary, the study demonstrates that NLP can play a vital role in advancing neonatal healthcare by transforming free-text clinical records into meaningful, actionable insights. While further work is needed to enhance generalisability and adoption, the results affirm that NLP-enhanced decision support systems have the potential to reduce diagnostic delays, improve accuracy, and support neonatologists in making faster, more informed decisions at the bedside.

REFERENCES

[1] Alafari, F., Driss, M., & Cherif, A. (2025). Advances in natural language processing for healthcare: A comprehensive review of techniques, applications, and future directions. *Computer Science Review*, *56*, 100725. https://doi.org/https://doi.org/10.1016/j.cosrev.2025.100725

[2] Bartal, A., Jagodnik, K. M., Chan, S. J., Babu, M. S., & Dekel, S. (2023). Identifying women with postdelivery posttraumatic stress disorder using natural language processing of personal childbirth narratives. *American Journal of Obstetrics & Gynecology MFM*, *5*(3), 100834. https://doi.org/https://doi.org/10.1016/j.ajogmf.2022.100834

[3] Bobba, P. S., Sailer, A., Pruneski, J. A., Beck, S., Mozayan, A., Mozayan, S., Arango, J., Cohan, A., & Chheang, S. (2023). Natural language processing in radiology: Clinical applications and future directions. *Clinical Imaging*, *97*, 55–61. https://doi.org/https://doi.org/10.1016/j.clinimag.2023.02.014

[4] Chalasani, S. H., Syed, J., Ramesh, M., Patil, V., & Pramod Kumar, T. M. (2023). Artificial intelligence in the

field of pharmacy practice: A literature review. *Exploratory Research in Clinical and Social Pharmacy*, *12*, 100346. https://doi.org/https://doi.org/10.1016/j.rcsop.2023.100346

[5] Dash, A., Darshana, S., Yadav, D. K., & Gupta, V. (2024). A clinical named entity recognition model using pretrained word embedding and deep neural networks. *Decision Analytics Journal*, *10*, 100426. https://doi.org/https://doi.org/10.1016/j.dajour.2024.100426

[6] Dufendach, K. R., Navarro-Sainz, A., & Webster, K. L. W. (2022). Usability of human-computer interaction in neonatal care. *Seminars in Fetal and Neonatal Medicine*, *27*(5), 101395. https://doi.org/https://doi.org/10.1016/j.siny.2022.101395

[7] Eguia, H., Sánchez-Bocanegra, C. L., Vinciarelli, F., Alvarez-Lopez, F., & Saigí-Rubió, F. (2024). Clinical Decision Support and Natural Language Processing in Medicine: Systematic Literature Review. *Journal of Medical Internet Research*, *26*. https://doi.org/https://doi.org/10.2196/55315

[8] Hiremath, B. N., & Patil, M. M. (2022). Enhancing Optimized Personalized Therapy in Clinical Decision Support System using Natural Language Processing. *Journal of King Saud University - Computer and Information Sciences*, *34*(6, Part A), 2840–2848. https://doi.org/https://doi.org/10.1016/j.jksuci.2020.03.006

[9] Hossain, E., Rana, R., Higgins, N., Soar, J., Barua, P. D., Pisani, A. R., & Turner, K. (2023). Natural Language Processing in Electronic Health Records in relation to healthcare decision-making: A systematic review. *Computers in Biology and Medicine*, *155*, 106649. https://doi.org/https://doi.org/10.1016/j.compbiomed.2023.106649

[10] Joaquim, P., Calado, G., & Costa, M. (2024). Benefits of reading to premature newborns in the neonatal intensive care unit: A scoping review. *Journal of Neonatal Nursing*, *30*(4), 325–330. https://doi.org/https://doi.org/10.1016/j.jnn.2023.11.011

[11] Li, Q., Kirkendall, E. S., Hall, E. S., Ni, Y., Lingren, T., Kaiser, M., Lingren, N., Zhai, H., Solti, I., & Melton, K. (2015). Automated detection of medication administration errors in neonatal intensive care. *Journal of Biomedical Informatics*, *57*, 124–133. https://doi.org/https://doi.org/10.1016/j.jbi.2015.07.012

[12] Liu, Y., Cao, X., Chen, T., Jiang, Y., You, J., Wu, M., Wang, X., Feng, M., Jin, Y., & Chen, J. (2025). From screens to scenes: A survey of embodied AI in healthcare. *Information Fusion*, *119*, 103033. https://doi.org/https://doi.org/10.1016/j.inffus.2025.103033

[13] Nerella, S., Bandyopadhyay, S., Zhang, J., Contreras, M., Siegel, S., Bumin, A., Silva, B., Sena, J., Shickel, B., Bihorac, A., Khezeli, K., & Rashidi, P. (2024). Transformers and large language models in healthcare: A review. *Artificial Intelligence in Medicine*, *154*, 102900. https://doi.org/https://doi.org/10.1016/j.artmed.2024.102900

[14] Ralevski, A., Taiyab, N., Nossal, M., Mico, L., Piekos, S., & Hadlock, J. (2024). Using Large Language Models to Abstract Complex Social Determinants of Health From Original and Deidentified Medical Notes: Development and Validation Study. *Journal of Medical Internet Research*, *26*. https://doi.org/https://doi.org/10.2196/63445

[15] Turchin, A., Masharsky, S., & Zitnik, M. (2023). Comparison of BERT implementations for natural language processing of narrative medical documents. *Informatics in Medicine Unlocked*, *36*, 101139. https://doi.org/https://doi.org/10.1016/j.imu.2022.101139

[16] Wu, Y., Zhang, J., Chen, X., Yao, X., & Chen, Z. (2025). Contrastive learning with large language models for medical code prediction. *Expert Systems with Applications*, *277*, 127241. https://doi.org/https://doi.org/10.1016/j.eswa.2025.127241

[17] Zhao, H., & Xiong, W. (2024). A multi-scale embedding network for unified named entity recognition in Chinese Electronic Medical Records. *Alexandria Engineering Journal*, *107*, 665–674. https://doi.org/https://doi.org/10.1016/j.aej.2024.09.008