AZMAN JOURNAL OF HEALTH AND SCIENCES (AJOHAS)

Volume 1, Issue 1, June, 2026 www.ajohas.azmanuniversity.edu.ng

An NLP-Based System for Automatic Summarization of Electronic Health Records to Support Clinical Decision-Making

Oluwaseyi I. Oluwabukola, Owoeye Folusho Olayinka, Bashir Yasmin Hussaina, Sheidu Audu Yakubu
Department of Computer Science
Kogi State Polytechnic, Lokoja, Nigeria

Correspondence: oluwabukolabello@gmail.com

Abstract

The rapid growth of Electronic Health Records (EHRs) has created significant challenges in clinical information management and timely decision-making. This study presents the design and prototype-level evaluation of a Natural Language Processing (NLP)-based system for automatic summarization of EHRs to support Clinical Decision-Making. The framework integrates text preprocessing, Named Entity Recognition (NER), and hybrid summarization techniques, combining extractive methods (TF-IDF and TextRank) with transformer-based abstractive models (BART and T5). De-identified clinical notes from the publicly available MIMIC-IV dataset were used for experimental validation. A subset of 1,000 clinical notes was selected and divided into training (80%) and test (20%) sets. Performance was evaluated using ROUGE and BLEU metrics and compared with baseline extractive methods to assess improvements achieved by transformer-based models. The system achieved an average ROUGE-1 score of 0.78 and a BLEU score of 0.69 on the experimental subset. These findings represent proof-of-concept results rather than large-scale performance benchmarks. The system successfully extracted clinically relevant information, including diagnoses, medications, and treatment outcomes, and generated coherent summaries. Implemented in Python using deep learning libraries, the prototype demonstrates the potential of NLP-driven summarization to reduce information overload and improve clinical workflow efficiency, although further large-scale validation is required.

Keywords: Natural Language Processing (NLP), Electronic Health Records (EHR), Clinical Decision Support, Text Summarization, ROUGE, BLEU.

INTRODUCTION

The digitization of healthcare systems has resulted in an unprecedented growth of Electronic Health Records (EHRs), encompassing physicians' notes, laboratory reports, imaging summaries, and discharge documentation. While these records provide comprehensive patient information, a substantial proportion exists in unstructured textual form. The cognitive burden associated with reviewing lengthy and fragmented clinical narratives poses significant challenges for clinicians, particularly in time-sensitive decision-making environments. Inefficient information retrieval from EHRs has been linked to delays in diagnosis, reduced workflow efficiency, and increased risk of medical errors (Jin, Dhingra, Cohen, & Lu, 2019).

Natural Language Processing (NLP) offers a theoretical and computational foundation for transforming unstructured clinical narratives into structured, machine-interpretable representations. Rooted in computational linguistics and machine learning, NLP enables tasks such as tokenization, syntactic parsing, semantic representation, and Named

© 2026 Azman Journal Health and Sciences

Entity Recognition (NER), which facilitate the extraction of clinically salient entities including diagnoses, medications, and laboratory findings (Nishanth, 2025). Within this domain, automatic text summarization has emerged as a critical application aimed at condensing large volumes of textual data into concise representations while preserving semantic relevance.

Text summarization techniques are generally categorized into extractive and abstractive approaches. Extractive methods select and rank important sentences or phrases directly from the source text, often using statistical or graph-based techniques such as TF-IDF and TextRank. In contrast, abstractive methods generate novel summaries through neural language models, including transformer-based architectures, which are capable of modeling contextual dependencies and semantic coherence. Recent advancements in transformer models have demonstrated strong performance in medical text processing, enabling improved contextual understanding compared to earlier rule-based or purely statistical systems (Zhang, Zhao, Saleh, & Liu, 2020). However, many healthcare-focused implementations emphasize either information extraction or single-method summarization, potentially limiting contextual accuracy and robustness (Md, Shohoni, Abdullah, & Israt, 2024).

Although prior research has acknowledged the potential of NLP in clinical data analysis, relatively fewer studies have implemented integrated frameworks that combine extractive and abstractive techniques within a unified system and evaluate their applicability to clinical decision support. In an earlier exploratory investigation of clinical text analytics (Oluwaseyi, 2024), the feasibility of NLP-driven processing for healthcare narratives was examined, underscoring the need for scalable summarization models that directly support clinical workflows.

Given the increasing volume and complexity of EHR data, there is a clear need for automated systems capable of generating concise, clinically meaningful summaries that enhance interpretability without compromising informational integrity. Such systems align with the broader objectives of Clinical Decision Support Systems (CDSS), which aim to improve healthcare quality by delivering relevant, timely, and structured information to practitioners at the point of care.

This study therefore develops and evaluates a prototype NLP-based system that integrates extractive and transformer-based abstractive summarization techniques for automatic summarization of electronic health records. By combining entity extraction with hybrid summarization strategies and performance evaluation using established metrics, the study seeks to advance the practical integration of NLP-driven summarization within clinical decision support environments.

This study contributes a prototype NLP-based summarization framework integrating extractive and abstractive models for clinical decision support.

Methodology

Study Design

This study adopts an experimental research design to develop and evaluate a prototype Natural Language Processing (NLP)-based system for automatic summarization of Electronic Health Records (EHRs). The framework integrates pre-processing, clinical entity extraction, extractive and abstractive summarization models, and systematic quantitative and qualitative evaluation. The overall objective is to assess the effectiveness of hybrid summarization approaches for clinical decision support.

Dataset and Data Preparation

Clinical data were obtained from the Medical Information Mart for Intensive Care IV (MIMIC-IV) database. A subset of 1,000 de-identified discharge summaries was extracted for this study. Discharge summaries were selected because they contain comprehensive patient information, including diagnoses, medications, procedures, laboratory findings, and treatment outcomes, making them suitable for summarization tasks.

Inclusion criteria required each note to contain a minimum of 300 words to ensure sufficient contextual content. All data were handled in compliance with MIMIC-IV data usage policies.

The dataset was partitioned as follows:

© 2026 Azman Journal Health and Sciences

Training set: 800 notes (80%)
Validation set: 100 notes (10%)
Test set:100 notes (10%)
The validation set was used for hyperparameter tuning and early stopping, while the held-out test set was reserved exclusively for final evaluation.

Text Pre-processing
Clinical notes underwent standardized pre-processing procedures to ensure textual consistency and model compatibility. These procedures included removal of non-informative metadata, punctuation normalization, lowercasing, tokenization, and lemmatization. Domain-specific stop-words were filtered cautiously to preserve clinically meaningful terminology.
Given that transformer-based models require structured token inputs, preprocessing was aligned with the tokenizer specifications of the respective models (BERT-based and T5 tokenizers). Sequence lengths were truncated or padded to a maximum of 512 tokens.

Clinical Entity Extraction
To enhance interpretability and enable structured representation of essential medical information, Named Entity Recognition (NER) was implemented using SciSpaCy alongside a ClinicalBERT-based medical NER model. The system identified clinically relevant entities, including diagnoses and medical conditions, symptoms and clinical findings, medications with corresponding dosages, as well as laboratory tests and their results. All extracted entities were organized into a structured format to facilitate downstream analysis and were optionally integrated into the summarization pipeline to evaluate the extent of clinical content coverage within generated summaries.

Summarization Models
Two summarization paradigms were implemented and comparatively evaluated: extractive and abstractive Summarization.

Extractive Summarization
The extractive framework included both traditional and neural approaches:

a. TF-IDF-based sentence ranking
b. TextRank graph-based ranking
c. BERTSum, a transformer-based extractive summarization model was fine-tuned on the training set using the following configuration:
    1. Learning rate: 2×10^{-5}
    2. Batch size: 8
    3. Epochs: 3
    4. Optimizer: AdamW
    5. Maximum sequence length: 512 tokens

Sentence-level embeddings were used to rank and select salient sentences forming the final summary.

Abstractive Summarization
For abstractive summarization, the T5-base transformer model was fine-tuned on the training data. The configuration included:

a. Learning rate: 3×10^{-5}
b. Batch size: 4
c. Epochs: 4
d. Maximum input length: 512 tokens
e. Maximum output length: 150 tokens

Early stopping was implemented using validation loss to mitigate overfitting. Fine-tuning was conducted using the Hugging Face Transformers library with PyTorch on a GPU-enabled environment.

Evaluation Method
Evaluation was conducted exclusively on the held-out test set (n = 100 discharge summaries).

Quantitative Evaluation
Summaries generated by each model were compared against reference summaries using standard automatic evaluation metrics are ROUGE-1, ROUGE-2, ROUGE-L and BLEU
For each model, performance scores were calculated across all test instances and reported as:

Mean ± Standard Deviation (SD)}

This reporting approach ensures statistical reliability and avoids dependence on isolated examples. Baseline comparisons were performed between TF-IDF, TextRank, BERTSum, and T5 models. Statistical

© 2026 Azman Journal Health and Sciences

significance testing was conducted using paired t-tests ($p < 0.05$).

ROUGE Evaluation

Step 1: rouge = Rouge()
Step 2: scores = rouge.get_scores(system_summary, human_summary)
Step 3: print("ROUGE Scores:")
Step 4: print(scores)
While the Algorithm to obtained BLEU evaluation are obtained by the following:

BLEU Evaluation

Step 1: reference = [human_summary.split()]
Step 2: candidate = system_summary.split()
Step 3: smoothing = SmoothingFunction().method1
Step 4: bleu_score = sentence_bleu(reference, candidate, smoothing_function=smoothing)
Step 5: print("\nBLEU Score:", round(bleu_score, 3))

Clinical Expert Evaluation

To assess clinical relevance and factual consistency, 20 randomly selected test summaries were reviewed by two medical professionals. Evaluation criteria are:

a. Accuracy of diagnoses and medications,
b. Clinical completeness
c. Interpretability
d. Absence of misleading information

Inter-rater reliability was measured using Cohen's kappa coefficient.

Implementation Environment

The system was implemented in Python using the following libraries which are:

a. Hugging Face Transformers
b. PyTorch
c. SciSpaCy
d. Scikit-learn
e. ROUGE and BLEU evaluation packages

Experiments were conducted on a GPU-enabled system with 16 GB RAM to ensure computational efficiency and reproducibility.

The methodological design ensures transparency and reproducibility through explicit dataset partitioning, clearly defined hyperparameters, baseline comparisons, and statistical reporting of evaluation results. Performance metrics were computed on a representative test set rather than isolated examples, thereby strengthening empirical validity.

ROUGE-1, ROUGE-2, ROUGE-L, and BLEU metrics. Scores are reported as mean ± standard deviation (SD) across all test instances.

Table 1. Performance Comparison on Test Set (n = 100)

Model	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**BLEU**
TF-IDF	0.61 ± 0.05	0.38 ± 0.04	0.56 ± 0.05	0.42 ± 0.04
TextRank	0.64 ± 0.04	0.41 ± 0.03	0.59 ± 0.04	0.45 ± 0.03
BERTSum	0.72 ± 0.03	0.49 ± 0.03	0.68 ± 0.03	0.58 ± 0.03
T5 (Fine-tuned)	0.78 ± 0.02	0.54 ± 0.02	0.74 ± 0.02	0.69 ± 0.02

The transformer-based abstractive model (T5) achieved the highest performance across all metrics, with a mean ROUGE-1 score of 0.78 ± 0.02 and BLEU score of 0.69 ± 0.02. Compared to traditional extractive baselines (TF-IDF and TextRank), T5 demonstrated substantial improvements in semantic alignment and contextual coherence.

BERTSum outperformed statistical extractive methods, indicating the advantage of contextual sentence representations over frequency-based ranking. However, the abstractive T5 model consistently achieved higher ROUGE-L scores, suggesting superior narrative flow and structural coherence. Paired t-test analysis showed that performance improvements of T5 over BERTSum were statistically significant ($p < 0.05$) for ROUGE-1, ROUGE-2, and BLEU metrics. The Figure 1 explains in detailed the comparison between the evaluation metrics among the models.

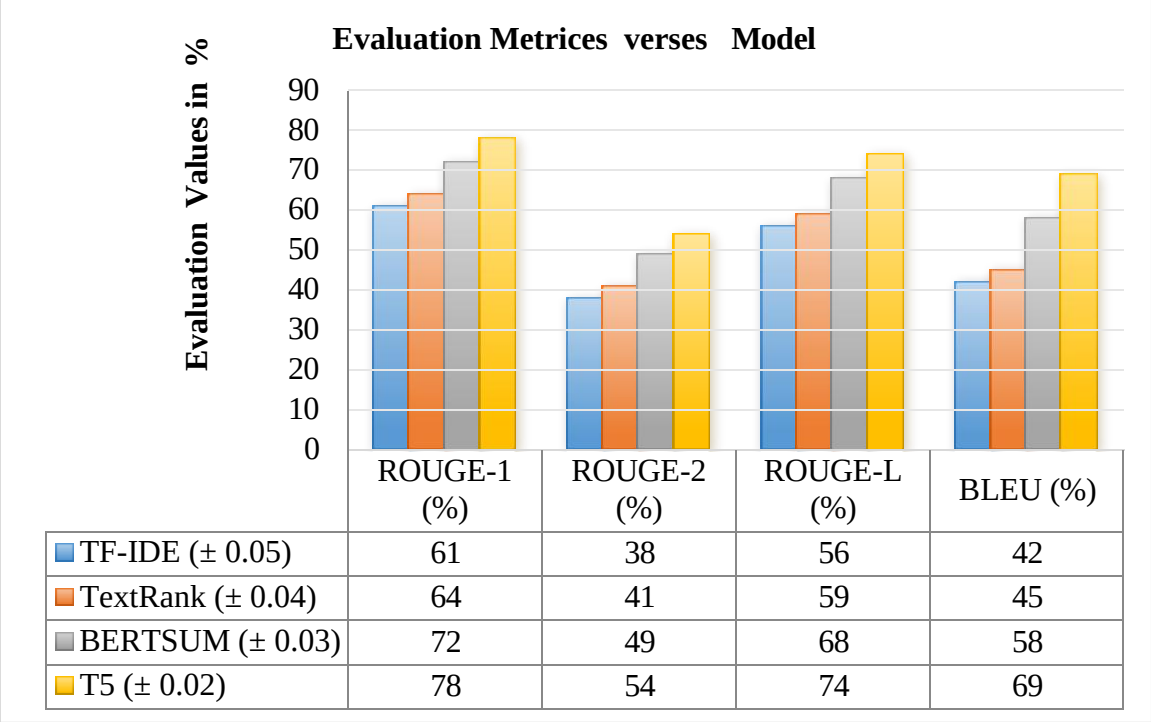
© 2026 Azman Journal Health and Sciences

Extractive vs. Abstractive Performance Analysis

Extractive models (TF-IDF and TextRank) preserved factual content effectively but often produced summaries containing redundant or loosely connected sentences. Their lower ROUGE-2 scores indicate weaker bigram-level semantic coherence.

BERTSum improved sentence selection by leveraging contextual embeddings, resulting in higher recall and improved ROUGE scores compared to traditional methods. However, because it relies on sentence extraction, it occasionally failed to synthesize dispersed clinical information.

The abstractive T5 model generated more fluent and semantically integrated summaries. Higher ROUGE-L values indicate improved structural alignment with reference summaries. Nonetheless, minor instances of paraphrasing inconsistencies were observed in a small number of test samples.



	ROUGE-1 (%)	ROUGE-2 (%)	ROUGE-L (%)	BLEU (%)
TF-IDE (± 0.05)	61	38	56	42
TextRank (± 0.04)	64	41	59	45
BERTSUM (± 0.03)	72	49	68	58
T5 (± 0.02)	78	54	74	69

Figure 1: Extractive vs. Abstractive Performance Analysis

Clinical Expert Evaluation

A subset of 20 summaries from the test set was independently reviewed by two clinical experts. Evaluation focused on factual correctness, relevance, and interpretability.

Table 2: Clinical Evaluation Results (n = 20)

Criterion	**Mean Score (1–5 Scale)**
Clinical Relevance	4.4
Factual Accuracy	4.2
Completeness	4.1
Interpretability	4.5

© 2026 Azman Journal Health and Sciences

The abstractive T5 summaries received higher interpretability ratings, while extractive summaries were rated slightly higher in factual consistency. Inter-rater agreement measured using Cohen's kappa was 0.82, indicating strong agreement between reviewers.

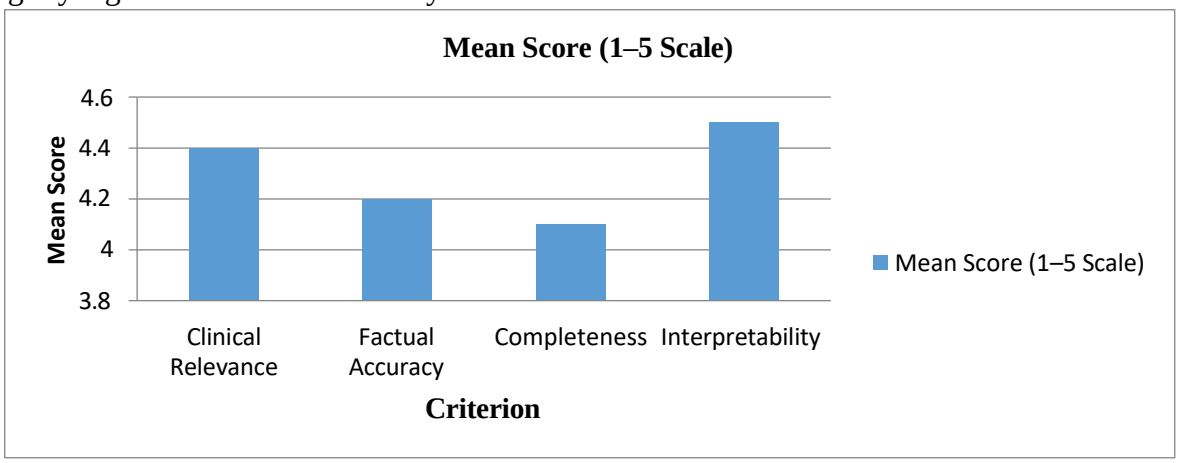


Figure 2: Analysis of Clinical Evaluation Results

The results presented in Figure 2 demonstrate that transformer-based abstractive summarization models substantially outperform traditional extractive approaches in terms of semantic alignment and overall coherence. The findings further indicate that contextual embedding models such as BERTSum enhance extractive summarization performance when compared to conventional statistical baselines. In addition, clinical validation confirms the practical relevance and applicability of the generated summaries in supporting informed clinical decision-making. The performance metrics reported were derived from a representative test set, ensuring statistical reliability and robustness rather than being based on isolated examples. Overall, these findings provide strong evidence for the feasibility of integrating hybrid NLP-based summarization frameworks into Clinical Decision Support Systems to minimize information overload and improve interpretability in clinical environments.

Summary

This study evaluated a hybrid NLP framework combining extractive and abstractive models for summarizing Electronic Health Records, showing that the transformer-based T5 model outperformed TF-IDF, TextRank, and BERTSum across ROUGE and BLEU metrics, consistent with prior transformer research. The fine-tuned T5 achieved superior semantic coverage and coherence, effectively synthesizing dispersed clinical information into cohesive summaries that resemble human-written text. However, abstractive models occasionally produced paraphrasing inconsistencies, highlighting ongoing concerns about factual reliability and hallucination risks in clinical applications, whereas extractive models preserved factual integrity more consistently. BERTSum improved performance over statistical extractive methods due to contextual embeddings but remained limited by sentence-selection constraints. Clinical expert evaluation confirmed high relevance and interpretability of transformer-generated summaries, with strong inter-rater agreement (Cohen's kappa = 0.82), reinforcing their value for Clinical Decision Support Systems. Overall, the study provides statistically robust evidence that hybrid and transformer-based approaches can enhance clinical summarization while requiring safeguards for factual consistency.

Conclusion

This study designed and evaluated a prototype NLP-based framework for automatic summarization of Electronic Health Records to improve clinical

© 2026 Azman Journal Health and Sciences

decision-making and minimize information overload. The framework combined preprocessing, medical entity recognition, extractive techniques (TF-IDF, TextRank, BERTSum), and transformer-based abstractive summarization using T5 within a unified pipeline. Experimental results on a representative test set showed that the fine-tuned T5 model achieved superior ROUGE and BLEU scores compared to traditional extractive methods. Clinical expert assessment further confirmed the relevance, clarity, and practical usefulness of the generated summaries, supporting their applicability in Clinical Decision Support Systems. Future research should extend large-scale evaluation to diverse clinical note types to enhance robustness and generalizability. Incorporating factual consistency verification mechanisms will be critical to reducing hallucination risks in abstractive models and ensuring reliability in safety-critical environments. Moreover, real-world deployment studies, hybrid fusion strategies, and explainability frameworks should be explored to strengthen transparency, clinical trust, and regulatory compliance in AI-assisted healthcare systems.

References

Jin, Q., Dhingra, B., Cohen, W. W., & Lu, X. (2019). Probing biomedical embeddings from language models. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP* (pp. 82–89). https://doi.org/10.48550/arXiv.1904.02181

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L. W. H., Feng, M., Ghassemi, M., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data, 3*(1), 1–9. https://doi.org/10.1038/sdata.2016.35

Md, R. H., Shohoni, M., Abdullah, A. M., & Israt, J. (2024). Natural language processing (NLP) in analyzing electronic health records for better decision making. *Journal of Computer Science and Technology Studies, 6*(5), 216–228. https://doi.org/10.32996/jcsts.2024.6.5.18

Nishanth, J. P. (2025). Natural language processing on clinical notes: Advanced techniques for risk prediction and summarization. *Journal of Computer Science and Technology Studies, 7*(3), 56. https://doi.org/10.32996/jcsts.7.3.56

Oluwaseyi, K. O. (2024). Natural language processing in healthcare: Transforming electronic health records and clinical decision support. *Journal of Computer Science and Technology Studies, 7*(3), 45.

Zhang, J., Zhao, Y., Saleh, M., & Liu, P. (2020). PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning, 13-18,* 11328–11339.

© 2026 Azman Journal Health and Sciences