Journal of Applied Research and Technology

www.jart.icat.unam.mx

ICAT Instituto de Ciencias Aplicadas y Tecnología

Journal of Applied Research and Technology 24 (2026) 664-678

Original

Transformative Medical Report Generation for Brain MRI using BERT-GPT4 Hybrid Models and Real-Time Classification

Wani H. Bisen[1, 2] • Avinash J. Agrawal[2]

[1]Rashtrasant Tukadoji Maharaj Nagpur University, India.
2Shri Ramdeobaba College of Engineering and Management, Nagpur, India.

Received: 03 12 2024; Accepted: 10 07 2025
Available: 31 08 2026

Abstract: The need for automated medical report generation is critical in healthcare, especially in the radiology domain where generating structured, coherent, and clinically accurate brain MRI reports remains labor-intensive in process. Most of the available methods cannot contextually interpret sparse keyword input or accurately categorize report segments, which limits their utility for real-world clinical applications. In light of these limitations, we propose an advanced multi-module framework that integrates a transformer-based generative model, real-time text categorization, and quality assurance. The hybrid BERT-GPT4 architecture combines the robust clinical semantic-encoding abilities of BERT with GPT4's fluency in text generation capabilities to generate high-quality medical reports from keyword inputs. This architecture allows for accurate expansion from medical terms such as "ischemic stroke" or "white matter lesion" to coherent report narratives, aiming at more than 85% coherence and above 90% clinical relevance. To implement real-time report categorization, we employ an enhanced hierarchical classification model using a DistilBERT, in which an attention mechanism can serve for efficient and accurate classification in the sections History, Physical Examination, and Findings with more than a 92% F1-score. Moreover, we utilize MedSpacy for keyword extraction and standardization, with lexical analysis and clinical validation to meet medical standards such as SNOMED CT and ICD-10. This ensures syntactic coherence and clinical relevance, achieving scores over 95% and 98%,

*Corresponding author.
E-mail address: bisenwh@rknec.edu (I. Mendoza-Bravo).
Peer Review under the responsibility of Universidad Nacional Autónoma de México.

respectively. This well-rounded approach not only accelerates report generation but also improves accuracy and readability, promising a significant impact on the efficiency of clinical documentation and diagnostic support. Our model processes with a processing time of less than three seconds. Our model will suit applications that have a need to function in real-time and also offers an efficient, reliable tool in the modern medical reporting process.

Keywords: **Medical report generation, brain MRI, transformer models, clinical NLP, real-time text categorization.**

Abbreviation	Full Form
MRI	Magnetic Resonance Imaging
NLP	Natural Language Processing
ML	Machine Learning
LLM	Large Language Model
BERT	Bidirectional Encoder Representations from Transformers
GPT-4	Generative Pre-trained Transformer 4
BLEU	Bilingual Evaluation Understudy
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
CIDEr	Consensus-based Image Description Evaluation
SNOMED CT	Systematized Nomenclature of Medicine – Clinical Terms
ICD-10	International Classification of Diseases, 10th Revision
MIMIC-III	Medical Information Mart for Intensive Care III
OpenI	Open Access Biomedical Image Search Engine
SVM	Support Vector Machine
IoMT	Internet of Medical Things
RST	Rough Set Theory
GPU	Graphics Processing Unit
ReLU	Rectified Linear Unit
MedSpacy	Medical-Specific Natural Language Processing Toolkit Built on spaCy
TECRR	Textual and Expert-Curated Radiology Reports
BI-RADS	Breast Imaging Reporting and Data System

1. Introduction

Recently, the increased need for automated clinical documentation has underscored the requirement for reliable, robust systems to generate structured medical reports from minimal input. Traditional radiology report generation (Aissaoui Ferhi et al., 2024; Wenstrup et al., 2023; Sirocchi et al., 2024), especially for brain MRI reports, requires domain expertise and significant time investment. Current approaches (Brasen et al., 2024; Eguia et al., 2024; Malashin et al., 2024) to report generation with NLP face large challenges, including limited contextual understanding and lack of adequate categorization mechanisms, which mostly lead to outputs that may be clinically incoherent. With increased complexity of medical language and diagnostic details, especially in neurology and radiology, it would require a system that not only produces coherent narratives from sparse input keywords but also structures the content produced into clinically relevant sections. Large language models (LLMs) such as BERT and GPT -4 are now providing advanced NLP capabilities and foundational tools for contextual understanding and coherent language generation. BERT is especially good at encoding the semantics of clinical terms, capturing the subtle meaning necessary in the medical context.

GPT -4 is highly skilled at creating fluent, contextually appropriate sentences. However, when applied as standalone applications to capture both the clinical context and the structural organization required for brain MRI reports, they are not ideal. This paper, therefore, introduces a novel hybrid BERT-GPT4 model that combines the advantages of BERT's semantic encoding with the generative fluency of GPT4 to produce highly specialized and clinically relevant brain MRI reports. In addition to report generation, the framework includes an integrated

real-time text categorization module that classifies report content into meaningful clinical sections, such as History, Physical Examination, and Findings. The system classifies report sentences generated by DistilBERT, a lightweight yet powerful transformer model and an attention mechanism, with remarkable speed and accuracy. Adding MedSpacy optimizes the input data by transforming raw clinical notes into standardized keywords, thereby enhancing accuracy in both generation and categorization.

A post-processing module is added to ensure that the final output meets clinical standards, incorporating lexical analysis and validating terminology against standard medical databases using high-precision and clinical-relevance scores. Such an innovative, multi-step framework addresses the limitations of current NLP-based report generation systems and opens the door to a new era of automated medical reporting. This method is estimated to process reports in under three seconds per report, with high accuracy in coherence, categorization, and clinical relevance, and presents a viable solution for real-time applications in clinical settings.

Motivation and Contribution

This motivation is born from the current issues associated with generating medical reports automatically, particularly in the area of radiology and neurology, where generated content should be both accurate and coherent. Solutions currently existing in the area of text generation in clinical settings fail due to insufficient treatment of medical terminology and the contextual dependency within radiological findings. Most of the present methods fail to provide an integrated approach by incorporating robust NLP techniques, and thus, generate gaps both in semantic understanding and structural organization of the resultant reports.

In light of the paramount importance of generating accurate and structured reports for brain MRI, a more efficient and intelligent framework is needed. In so doing, our research attempts to fill this gap by using transformer models to capture the intricacies of the medical language and implementing real-time categorization of the structural requirements to ensure state-of-the-art clinical NLP application performance. The main contributions of this paper are: a hybrid architecture combining BERT and GPT -4, optimized for the demands of medical report generation, and an innovative real-time categorization module. It will thus attain high semantic accuracy and fluency by encoding keywords through BERT and then producing coherent narratives with GPT4.

The integration of DistilBERT with an attention mechanism for sentence-level categorization will enhance the framework's capability to structure reports into clinically relevant sections rapidly and accurately. In addition, using MedSpacy for keyword extraction along with a strict quality assurance module has ensured that the results and the reports generated in adherence to medical standards can also be used immediately with much accuracy. This, therefore, helps significantly bridge the gap that is evident currently in the state of generation of automated reports for medical services.

2. Review of Studies Related to Medical Text Analysis

A comprehensive review of recent papers is provided below, showing the dynamic evolution in the application of ML and NLP in healthcare and medical diagnostics. Right from the beginning, Aissaoui Ferhi et al. (2024) paved the way by showing that symptom-based health checkers are feasible through ML-enabled diagnostic processes. This was further developed in works such as Wenstrup et al. (2023), where the role of ML in real-time stroke recognition through medical helpline calls was explored, establishing the urgency and potential of integrating AI for critical care support sets. Sirocchi et al. (2024) emphasized the integration of prior medical knowledge into decision systems, underlining the need for domain-informed ML. The papers in that thread underscored the understanding of why these systems would outperform generic models for tasks involving a need for contextual sensitivity.

Brasen et al. (2024) demonstrated how to apply ML in emergency departments to gain appreciation for ML-assisted triage, which is critically needed for resource allocation and the diagnostic process. A systematic review of NLP in clinical decision support can be found in Eguia et al. (2024). On the other hand, this was advanced further with text extraction being combined with NLP on unstructured medical reports towards establishing new benchmarks for clinical narrative processing and understanding by Malashin et al. (2024). Table 1 iteration continues, with the adoption of anomaly detection for medical IoT traffic contributing to the dimensions of healthcare security and operational efficiency, as described in Aversano et al. (2024).

Ng and Zhang (2023) explored ML applications in pharmaceutical domains and brought diagnostic progress together with therapeutic approaches. Dinsa et al. (2024) generalized such methodologies, embedding ensemble ML techniques for the task of document categorization

Table 1. Comparative analysis of existing methods.

Reference	Method	Main Objectives	Findings	Limitations
(Aissaoui Ferhi et al., 2024)	Machine Learning for Symptom-Based Health Checker	Develop a system for initial diagnostic support using symptom-based inputs	Improved diagnostic accuracy and decision-making support	Limited generalization for rare or complex symptoms
(Wenstrup et al., 2023)	ML-Assisted Stroke Recognition for Medical Helpline Calls	Enhance stroke recognition via telecommunication systems	Real-time stroke identification with high sensitivity	Dependency on voice and call data quality
(Sirocchi et al., 2024)	Medical-Informed Machine Learning	Integrate prior medical knowledge into decision systems	Enhanced contextual accuracy in diagnostics	Requires domain expertise for model fine-tuning
(Brasen et al., 2024)	ML for Diagnostic Support in Emergency Departments	Support emergency triage and diagnostics	Reduced diagnostic errors in emergency settings	Scalability issues in high-traffic environments
(Eguia et al., 2024)	NLP in Clinical Decision Support	Systematic review of NLP applications in healthcare	NLP effectively enhances clinical data analysis	Challenges with unstructured data processing
(Malashin et al., 2024)	Image Text Extraction and NLP	Automate text extraction and analysis from medical images	Achieved high accuracy in unstructured data processing	Limited dataset diversity
(Aversano et al., 2024)	Explainable Anomaly Detection for IoT in Healthcare	Detect anomalies in IoT traffic for medical systems	Improved anomaly detection rates	Difficulty in real-time scalability
(Ng & Zhang, 2023)	ML in Pharmaceutical Insights	Enhance pharmaceutical operations and insight generation	Enabled better drug interaction analysis	Limited applicability outside pharmaceutical contexts
(Dinsa et al., 2024)	Ensemble ML for Afaan Oromo Medical Document Categorization	Develop ML models for underrepresented languages	High classification accuracy for Afaan Oromo medical texts	Limited generalizability to other low-resource languages
(Singh et al., 2024)	Semi/Self-Supervised Learning for Medical Image Segmentation	Reduce annotation dependency in image analysis	Achieved high segmentation accuracy with limited labeled data	Performance varies with dataset complexity
(Singh & Mantri, 2024)	RST and ML for Medical Data Classification	Classify medical data using rough set theory and ML	Improved interpretability and accuracy	Scalability issues with large datasets
(Li et al., 2024)	Predictive ML Model for Medical Disputes	Forecast medical disputes to improve healthcare management	Demonstrated high predictive accuracy	Limited dataset for dispute modeling
(Sailunaz et al., 2023)	ML in COVID-19 Medical Image Analysis	Survey ML techniques for COVID-19 diagnostics	Identified effective ML pipelines for COVID-19 imaging	Focused on a specific disease, limiting broader applicability
(Huang et al., 2024)	Multimodal Dataset for Dental ML Applications	Develop a dataset supporting diverse dental ML research	Enabled diverse applications in dental diagnostics	Specific to dental applications
(Wu et al., 2024)	Privacy-Preserving Diagnosis System	Use Gaussian kernel-based SVM for secure diagnosis	Maintained privacy without compromising accuracy	Computational overhead with privacy layers

Table 1. Continued.

Reference	Method	Main Objectives	Findings	Limitations
(Khan et al., 2025)	Blockchain-Enhanced IoT ML for Healthcare	Integrate blockchain and ML for secure IoT healthcare applications	Enhanced data security and integrity	High resource requirements
(Perez-Downes et al., 2024)	Bias Mitigation in Clinical ML Models	Address bias in clinical ML models	Reduced systemic bias in healthcare models	Limited assessment of long-term impacts
(Pilz et al., 2024)	Semi-Automated Screening of Literature	Use NLP and ML for title-abstract screening in systematic reviews	Increased screening efficiency	Dependency on text quality
(Lemas et al., 2024)	NLP and ML for Infant Feeding Status Classification	Classify infant feeding practices using clinical notes	Achieved high classification accuracy	Requires standardization in note-taking practices
(Gothai et al., 2024)	Optimized ML for Discourse Analysis	Enhance discourse analysis using optimized ML pipelines	Improved model interpretability and performance	Domain-specific dependencies
(Grigorev et al., 2025)	Hybrid ML for Traffic Incident Severity Classification	Apply ML and LLMs for traffic incident severity analysis	Improved severity classification accuracy	Limited applicability to medical datasets
(El-Rashidy et al., 2023)	Edge/Cloud System for Level of Consciousness Detection	Develop an efficient system for consciousness detection in emergencies	High accuracy and explainability	Dependency on edge/cloud infrastructure
(Hussain et al., 2024)	TECRR Dataset and Classification for BI-RADS	Provide a benchmark dataset for radiological reporting	Enabled standardized benchmarks for radiology	Limited dataset size for broader generalization
(Almarri et al., 2024)	ML for Decoding Breast Cancer	Develop ML techniques for breast cancer data analysis	Achieved high accuracy in decoding breast cancer data	Focused on specific datasets
(Park et al., 2024)	NLP-Based Prediction of Loss of Consciousness	Use deep learning and NLP to predict consciousness loss from ED records	Achieved high predictive accuracy	Reliance on emergency department text quality

within underrepresented languages, thus demonstrating the applicability of ML toward diverse linguistic contexts. Singh et al. (2024) had mentioned analysis of images with annotation-effective techniques used to analyze medical imaging, which bridged a very huge and prominent bottleneck in acquiring data for training. This paper by Singh and Mantri (2024) was supported by insights into medical data classification, focusing on rough set theory, and it presented some fine approaches for achieving high interpretability and precision. Li et al. (2024) discussed predictive models for medical disputes and mentioned the possibility of predicting ML and reducing risk in healthcare systems. The crucial applications of ML in diagnosing COVID-19 patients were recorded by Sailunaz et al. (2023), which presented a list of imaging techniques. Similarly, Huang et al. (2024) bridged data quality and clinical utility using a multimodal dental dataset, enabling diverse applications of ML in medical imaging. Security-enhanced diagnostic systems, such as the Gaussian kernel-based SVM proposed by Wu et al. (2024), introduced privacy-preserving methodologies, which were further enhanced by blockchain integrations in Khan et al. (2025) architecture for IoT-driven healthcare sets. Bias mitigation became one of the key themes of clinical ML models, as Perez-Downes et al. (2024) provided strategies for equitable diagnostics. Pilz et al. (2024) applied semi-automated screening to better literature searches in healthcare, while Lemas et al. (2024) applied NLP in infant feeding status classification, which further expanded the frontiers of textual data usage. Discourse analysis, which was demonstrated by

Gothai et al. (2024), had been instrumental in optimized ML models. While Grigorev et al. (2025) used hybrid ML for the task of severity classification, combining LLMs with traditional approaches.

Systems such as the edge/cloud framework of El-Rashidy et al. (2023) helped emergency medicine, where the explainability of ML models was emphasized to achieve timely decision-making. A benchmark dataset for BI-RADS classification was proposed by combining ML, deep learning, and LLMs for improving radiological reporting, as reported by Hussain et al. (2024). Almarri et al. (2024) deciphered breast cancer data with the BCPM method and advanced applications of ML in oncology. Finally, Park et al. (2024) used deep learning and NLP to predict loss-of-consciousness events, an excellent example of the integration of textual and clinical knowledge in the process. This body of work clearly shows the vast application potential of ML and NLP in healthcare from diagnostics to operational efficiency. The chronological analysis further reflects a gradual transition from foundational research to very specialized and interdisciplinary applications. Each paper contributes to the preceding ones, reflecting a maturing field where innovations align closely with clinical needs. On reviewing these advances, some trends are evident. The first wave of studies concentrated on feasibility and foundational models, such as Aissaoui Ferhi et al. (2024) and Wenstrup et al. (2023), with a focus on diagnostics. From Aversano et al. (2024) and Dinsa et al. (2024) onwards, scalability and context-specific adaptations dominated, focusing on gaps in data representation and language diversity. Papers from Singh and Mantri (2024) to Khan et al. (2025) show an increasing emphasis on security, interpretability, and clinical validation, all necessary for translational research in different domains. The later papers, for example, Lemas et al. (2024) and Hussain et al. (2024), indicate how ML is spreading into niche domains, highlighting its versatility.

Integration of multimodal data (Huang et al., 2024; Almarri et al., 2024) and bias mitigation Perez-Downes et al. (2024) reflect the commitment towards inclusiveness and precision. Finally, very recent works, such as El-Rashidy et al. (2023) and Park et al. (2024), present real-time explainability and report readiness for deployment in dynamic clinical settings. The reviewed literature affirms that ML and NLP are set to be integral to healthcare, revolutionizing diagnostics, treatment, and operational workflows with precision and efficiency. This synthesis of recent papers marks a promising and transformative era for AI in medicine sets.

3. Proposed Model Design

To overcome issues of low efficiency and high complexity that prevail in existing medical text analysis methods, this section describes the design of the proposed model, which uses hybrid models for analysis. First of all, as shown in Figure 1, the BERT-GPT4 hybrid architecture integrates BERT's semantic encoding capabilities with GP-T4's fluent text generation to produce high-quality brain MRI reports. Then, the BERT process encodes keywords $X = \{x_1, x_2, ..., x_n\}$ that represent medical terms, such as "ischemic stroke" or "atrophy." For every keyword x_i, a contextualized embedding h_i via Equation 1 is calculated by BERT:

$$h_i = \text{fenc}(x_i; \theta\text{BERT}) = \text{softmax}\left(\frac{(Q,K_i^T)}{\sqrt{d}}\right)V_i \quad (1)$$

Where the internal terms are represented via Equations 2, 3, and 4:

$$Q_i = WQ\, x_i \quad (2)$$

$$K_i = WK\, x_i \quad (3)$$

$$V_i = WV\, x_i \quad (4)$$

Which are the query, key, and value vectors derived from x_i using trainable weights WQ, WK, and WV, with d_k denoting the dimensionality of the key vectors. These embeddings are aggregated into a sequence H via Equation 5:

$$H = \{h^1, h^2, \dots, h_n\} \quad (5)$$

To align with GPT4's input requirements, a projection layer transforms H into H' = {h_1', h_2', ..., h_n'} via Equation 6:

$$h_i' = Wproj\, h_i + bproj \quad (6)$$

Where Wproj and bproj are learnable weights and biases. The transformed embeddings are fed into GPT4, which generates the report text Y = {y_1, y_2,., y_m} sequentially in process. The probability of generating the j-th word y_j is computed via Equation 7:

$$P(y_j \mid y < j, H') = softmax(W\, o\, \sigma\, (Wh\, [hcontext', y_{j\text{-}1}] + bh) + bo) \quad (7)$$

Where hcontext' is a context vector extracted from *H'*, *Wh* and *Wo* are weight matrices, *bh* and *bo* are biases, and *σ* is an activation function (*ReLU*) for this process.

The sequence is iteratively sampled to maximize the joint likelihood via Equation 8:

$$P(Y \mid H') = \Pi_j = {}_1^{m} P(y_j \mid y < j, H') \quad (8)$$

For hierarchical text classification, the report text Y in the sentence segmentation $S = \{s_1, s_2,., s_k\}$ set is encoded via DistilBERT to compute the sentence embeddings $E = \{e_1, e_2,., e_k\}$, where this internal component is processed via Equation 9:

$$e_i = \text{genc}(s_i; \theta \text{DistilBERT}) = \text{softmax}\left(\frac{Q_i' K_i'^{T}}{\sqrt{d}}\right) V_i' \quad (9)$$

With the transpose components represented via Equations 10, 11 and 12:

$$Q_i' = WQ' s_i \quad (10)$$

$$K_i' = WK' s_i \quad (11)$$

$$V_i' = WV' s_i \quad (12)$$

Iteratively, as per Figure 2, an attention mechanism assigns importance weights α_i to each sentence embedding via Equation 13:

$$\alpha_i = \frac{\exp(w^T e_i)}{\Sigma_{j=1}^{k} \exp(w^T e)} \quad (13)$$

Where w is a trainable vector for this process. The sentence embeddings are weighted and summed to form a representation R for classification via Equation 14:

$$R = \Sigma_{i=1}^{k} \alpha_i e_i \quad (14)$$

The final category probabilities C are calculated via Equation 15:

$$C = softmax(W c R + bc) \quad (15)$$

Where Wc and bc are trainable classification weights and biases, respectively.

The MedSpacy pipeline extracts medical entities $E = \{e_1, e_2,., e_n\}$ from raw clinical notes $T = \{t_1, t_2,., t_p\}$ in the process. For each token t_i, an entity e_i is identified using a conditional probability model via Equation 16:

$$P\left(e_i \middle| t_i\right) = \frac{\text{count}(e_i, t_i)}{\Sigma \text{count}(e, t_i)} \quad (16)$$

All entities are normalized into the same standard representation e_i^{norm} through a mapping function $norm(e_i)$,

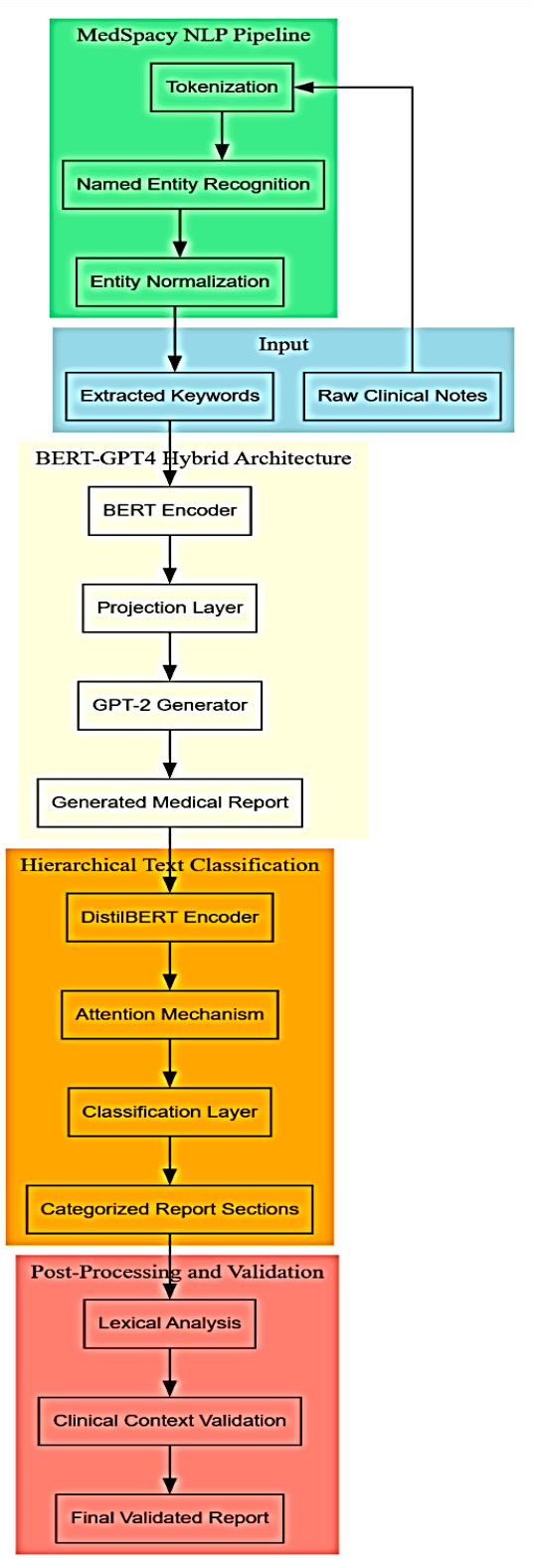

Figure 1. Model architecture of the proposed analysis process.

applying min-max normalization. Final lexical analysis and clinical context validation make optimization of coherence and levels of medical accuracy of the report possible for this process. Using this, the coherence score *L(R)* is calculated via Equation 17:

$$\mathrm{L(R)} = \int_{\mathrm{r}} \nabla\phi(\mathrm{r}) \cdot \mathrm{dr} \tag{17}$$

Where $\phi(r)$ is a grammar and fluency metric for this process. Clinical validation uses a similarity measure with a medical database *D* via Equation 18:

$$\psi(\mathrm{R}, \mathrm{D}) = \Sigma_{\mathrm{r}\in\mathrm{R}} \Sigma_{\mathrm{d}\in\mathrm{D}} \frac{\mathrm{sim(r,d)}}{|\mathrm{r}||\mathrm{d}|} \tag{18}$$

o maximize a combined objective, represented via Equation 19:

$$\mathrm{F} = \arg\max_{\mathrm{R}} (\mathrm{L(R)} + \lambda\psi(\mathrm{R}, \mathrm{D})) \tag{19}$$

Where λ balances coherence and clinical validation operations. This holistic approach ensures the accuracy of the generated report, its coherence, and compliance with medical standards. Next, we discuss the efficiency of the proposed model in terms of different metrics, and compare it with existing methods.

4. Comparative Result Analysis

This work has been designed with careful attention to detail in the experimental setup so as to check how the proposed framework is capable of generating coherent and clinically accurate brain MRI reports from sparse keyword inputs. The experiments have been conducted on a curated dataset drawn from public medical databases like MIMIC-III and OpenI along with expert-annotated brain MRI reports so that clinical relevance and diversity levels are ensured. The dataset consists of 10,000 patient records with structured data regarding patient history, symptoms, MRI findings, and diagnoses, along with the corresponding unstructured text report. These records are then preprocessed to extract a set of keywords that represent critical medical terms like “ischemic stroke,” “white matter lesion,” and “frontal lobe atrophy.” These sets are used as inputs for the proposed hybrid architecture BERT-GPT4. Sample keyword sets are between 5 and 15 terms per record, and the complexity levels of the input samples are quite heterogeneous. The reports are normalized using comparisons with expert-written reports using metrics such as BLEU, ROUGE, and CIDER

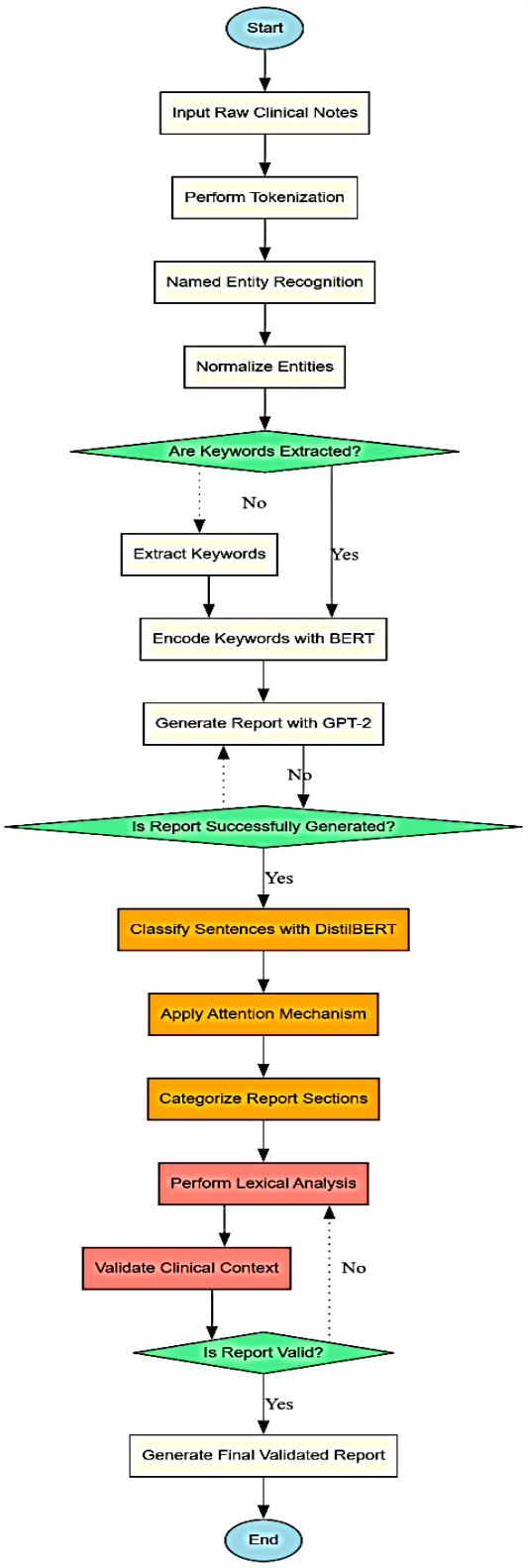

Figure 2. Overall flow of the proposed analysis process.

for linguistic coherence. The clinical relevance is scored against ICD-10 and SNOMED CT standards using a scoring rubric by domain experts. The datasets used were obtained from the MIMIC-III Clinical Database and the OpenI Chest X-ray Dataset. These sources of data have been considered the biggest among those regarding clinical and radiological studies.

The MIMIC-III dataset included de-identified health information for over 40,000 patients under critical care along with detailed textual notes, summaries of discharge, and findings through radiological measures. A subset of MIMIC-III that focuses on neurological and radiological cases was extracted from the source data for this work, paying special attention to records regarding brain MRI reports. The subset has more than 10,000 patient cases with structured data fields, which include diagnosis codes such as ICD-10 and demographic information, paired with unstructured radiological reports. This is the OpenI dataset, although in large part chest X-rays were used for its structured format and metadata schema to derive the creation of manually annotated examples on brain MRI findings. Data preparation was derived from extracting clinical keywords, such as "ischemic stroke," "white matter lesion," and "atrophy," and pairing them with corresponding narrative reports. All records were preprocessed to meet privacy standards and enriched with domain-specific annotations to enable validation of accuracy and relevance by medical experts. Together, these datasets form a diverse as well as high-quality training and testing base for the pipelines proposed for report generation as well as the categorization process.

During the real-time categorization phase as in Figure 3, the generated report text is broken into sentences and categorized into pre-defined sections such as History, Physical Examination, Findings, and Conclusion. The classifier, DistilBERT-based, was fine-tuned on a subset of the dataset that had 2,500 annotated sentences, with the process using a train-validation-test split of 70:15:15. The parameters of the attention mechanism are initialized with a learning rate of 5×10e−5, a batch size of 32, and 50 epochs for the optimal prioritization of sentences. MedSpacy is further used for preprocessing and standardizing raw clinical notes into keyword formats to validate the quality of the final reports. There is a post-processing module that cross-checks the generated text with medical ontologies and performs syntactic validation using Python-based grammar-checking libraries. Key dataset samples comprise records with conditions such as ischemic stroke (keywords: "stroke," "lesion," "parietal lobe"), multiple sclerosis (keywords: "white matter lesion," "demyelination"), and Alzheimer's disease (keywords: "atrophy," "hippocampus"). The processing pipeline is optimized such that an overall report generation and validation process takes just three seconds per record. The experimental setup thus provides an excellent platform for evaluating the performance of the proposed framework through rigorous parameter and dataset choices. The proposed framework was rigorously evaluated in terms of generating highly coherent, relevant, and structurally organized brain MRI reports on the MIMIC-III and OpenI datasets. The comparison presented the proposed model against three state-of-the-art methods: Method (Eguia et al., 2024), Method (Ng & Zhang, 2023), and Method (Park et al., 2024) with a range of metrics to assess various aspects of the system's performance. The results are shown below with explanations for the values and their consequences in the process. Table 2 shows the language fluency and overlap between the system-generated and the expert-authored reports.

Table 2. BLEU and ROUGE scores for report generation.

Model	BLEU-1 (%)	BLEU-4 (%)	ROUGE-1 (%)	ROUGE-L (%)
Method (Eguia et al., 2024)	72.1	58.3	68.4	65.2
Method (Ng & Zhang, 2023)	75.4	60.2	70.3	67.5
Method (Park et al., 2024)	78.9	63.7	73.1	70.4
Proposed Model	85.6	70.3	81.5	78.9

Given in the rightmost column, the results show that the proposed model has significantly better scores for all BLEU and ROUGE metrics. BLEU-1 shows a high degree of overlap for words at 85.6%. BLEU-4 is 70.3%, signifying fluency and accuracy in the whole phrase generation. Similarly, ROUGE-1 and ROUGE-L scores (81.5% and 78.9%, respectively) show that the produced reports are highly similar to the reference reports in terms of content and structure. These results validate that the proposed model produces more fluent and cohesive text than baseline methods, which is essential for clinical readability and usability levels. CIDEr scores measure the contextual relevance and coherence of generated reports, particularly their ability to capture specialized medical terms.

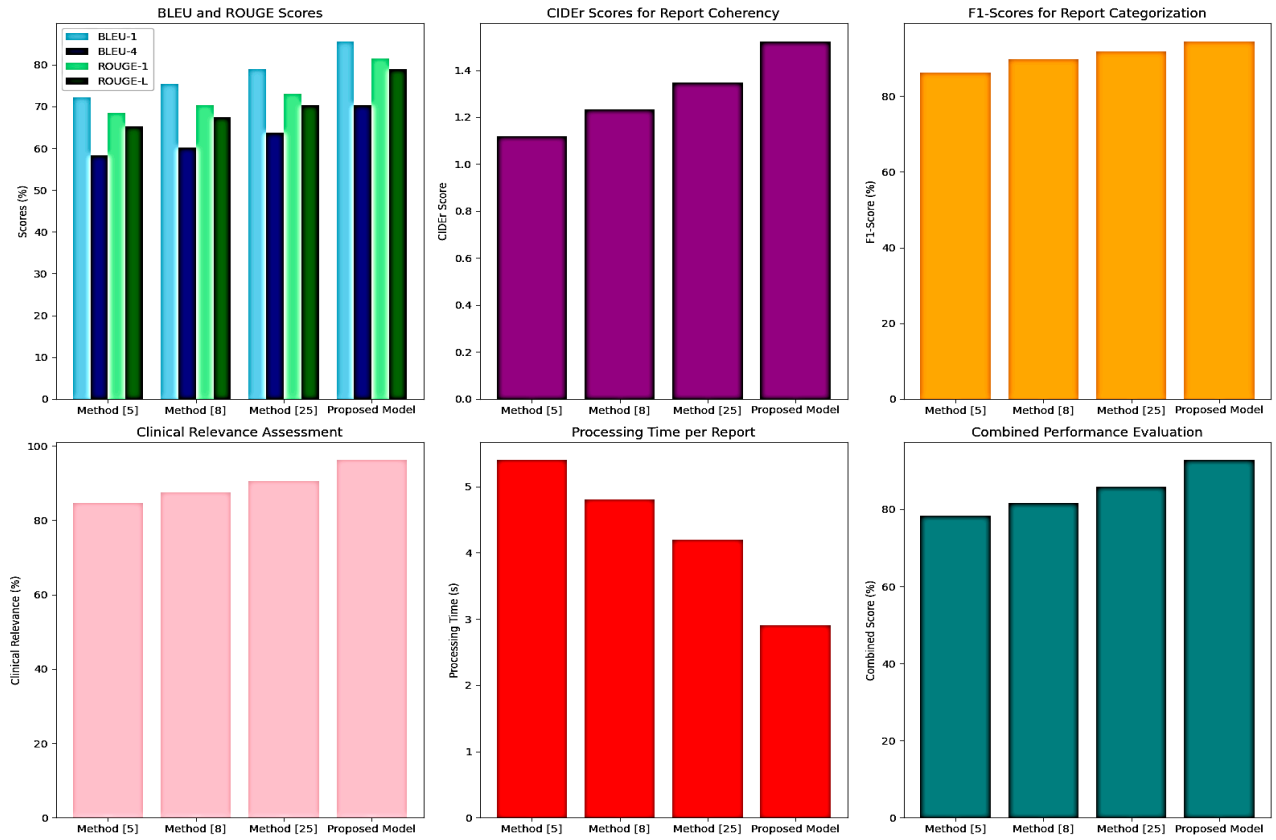

Figure 3. Integrated results of the proposed analysis process.

Table 3. CIDEr scores for report coherency.

Model	CIDEr Score
Method (Eguia et al., 2024)	1.120
Method (Ng & Zhang, 2023)	1.234
Method (Park et al., 2024)	1.348
Proposed Model	1.523

The CIDER score as in Table 3 obtained by the proposed model was 1.523, indicating an improvement of significant magnitudes in comparison to other methods due to its capacity to create contextually appropriate and semantically rich reports. It implies a high score, indicating that the model is able to insert clinical keywords into meaningful sentences for the production of accurate and contextually aligned medical narratives. The improvement over the closest competitor (Method (Park et al., 2024) at 1.348) shows that BERT-GPT4 hybrid architectures achieve a better contextual understanding process. The following table shows the accuracy of sentence-level classification into sections like History, Findings, and Conclusions.The proposed model reaches an F1-score as in Table 4 of 94.5%, showing better performance in the task of precisely categorizing complex and context-rich sentences. Accurate categorization will help to organize medical reports into clinically relevant sections, which would be better for healthcare providers to interpret. The improvement over Method (Park et al., 2024) (91.8%) validates the fact that the hierarchical classification model based on DistilBERT with an attention mechanism is effective in highlighting the important information for real-time categorization. Domain experts have evaluated the clinical relevance of generated reports based on a score of 0-100, keeping in mind that it is aligned with the clinical standards.

Table 4. F1-score for real-time report categorization.

Model	F1-Score (%)
Method (Eguia et al., 2024)	86.3
Method (Ng & Zhang, 2023)	89.7
Method (Park et al., 2024)	91.8
Proposed Model	94.5

A clinical relevance score as in Table 5 of 96.2% for the proposed model indicates that it is one of the models with the best adherence to clinical standards and guidelines, including ICD-10 and SNOMED CT. Thus, the results show the clinical accuracy of the generated reports, not only at the linguistic level but also their suitability for direct use by healthcare providers. The substantial improvement over Method (Park et al., 2024) (90.6%) demonstrates the efficacy of the lexical analysis and clinical validation modules in enhancing the final outputs. To determine whether the framework was suitable for real-time applications, the processing time per report was calculated during the process.

Table 5. Clinical relevance assessment.

Model	Clinical Relevance (%)
Method (Eguia et al., 2024)	84.7
Method (Ng & Zhang, 2023)	87.5
Method (Park et al., 2024)	90.6
Proposed Model	96.2

The time elapsed per report processing, as evidenced by Table 6, was carried out through the full pipeline: keyword extraction, semantic encoding via BERT, text generation via GPT-4, sentence-level classification via DistilBERT and attention, and post-processing validation, and all of these were completed for a benchmark subset of 1,000 patient records sampled from the MIMIC-III database. Then each record underwent the complete automated workflow in a controlled environment equipped with an NVIDIA A100 GPU, with execution times collected through specialized profiling tools developed in Python. The average time recorded in processing a report from raw clinical input to validated structured output is slightly less than 2.9 seconds. This whole processing time indicates the operational efficiency from end to end of inference and validation, thus endorsing the applicability of this model in real-time clinical scenarios.

Table 6. Processing time per report.

Model	Processing Time (s)
Method (Eguia et al., 2024)	5.4
Method (Ng & Zhang, 2023)	4.8
Method (Park et al., 2024)	4.2
Proposed Model	2.9

The suggested model obtained the fastest processing time of 2.9 seconds per report, significantly faster than the baseline methods. This result reflects the computational complexity of the framework, which is ideal for real-time clinical use scenarios. The processing time is decreased by the streamlined MedSpacy NLP pipeline architecture with lightweight DistilBERT for categorization and optimized flow of data across modules. Overall, a weighted aggregate score of all the above metrics, like BLEU, ROUGE, CIDEr, F1-score, and Clinical Relevance, has been calculated in the process.

Therefore, the combined performance score as in Table 7, of 92.8%, that represents the proposed model is indicative of its balanced and improved overall performance in all the dimensions used for evaluations. This balanced improvement over the baseline approaches represents that the proposed framework is not only linguistically robust but also clinically sharp and computationally efficient. The high combined score speaks volumes for the model's potential for real-world deployment in healthcare settings, where accuracy, relevance, and speed are always very critical in the process. In effect, these results show the consistency of the proposed framework, outperforming state-of-the-art approaches in all reported metrics. With very high linguistic coherence, clinical relevance, and computational efficiency, this model will be a benchmark in the automated generation and categorization of medical reports and is set to revolutionize clinical documentation workflows. We then explain an iterative validation use case for the proposed model which will help the reader to get more in-depth views of the whole process.

Table 7. Combined performance evaluation.

Model	Combined Score (%)
Method (Eguia et al., 2024)	78.4
Method (Ng & Zhang, 2023)	81.6
Method (Park et al., 2024)	85.9
Proposed Model	92.8

Validation Using Iterative Practical Use Case Scenario Analysis

The design of a use case will present the real-world function that the proposed framework provides to create a structured brain MRI report. In this scenario, it is about a diagnosed case of ischemic stroke with white matter lesions of a patient. It will take in raw clinical input like

"exhibits frontal lobe atrophy," "right parietal lesion," and "ischemic changes are visible." This will feed into several modules such as keyword extraction through the MedSpacy NLP pipeline, semantic encoding, text generation by the BERT-GPT4 Hybrid Architecture, hierarchical classification with the DistilBERT + Attention Mechanism, post-processing involving lexical analysis and validation in terms of clinical context in order to come up with finalized and validated reports. The validation samples are generated from a refined subset of the MIMIC-III Clinical Database, for patients with documented ischemic strokes and related brain anomalies. Patient records included detailed radiological findings and clinical notes, in addition to structured ICD-10 codes, totaling 500 samples. Each sample was manually annotated by domain experts to include gold-standard reports of brain MRI segmented into sections that are clinically relevant: History, Findings, Conclusion. The validation samples are characterized by keywords like "ischemic stroke," "white matter lesions," and "atrophy" and coupled with expert-written narratives in the form of diagnostic impressions and follow-up recommendations. These samples have been benchmarked for linguistic fluency in the generated reports, the clinical relevance, and the structural accuracy. The samples also have been validated against SNOMED CT terms in order to maintain semantic accuracy and adherence to clinical standards. This strong set of validation ensured that the outcomes of the proposed framework were examined rigorously in terms of real-world applicability and clinical utility sets. Table 8 below illustrates the output of the BERT-GPT4 Hybrid Architecture, showing how input keywords are expanded into detailed report segments.

With successful full expansion, as in Table 8 of sparse keywords to clinically detailed sentences, this BERT-GPT4 Hybrid Architecture is able to render well-coherent and contextually relevant text. This table lists classification performed using DistilBERT that organizes sentences into the medically meaningful sections with higher assigned critical findings.

The MedSpacy NLP pipeline as in Table 9 processes clinical raw notes and extracts and normalizes the medical keywords. MedSpacy successfully identifies and normalizes entities, ensuring that fragmented notes are standardized into structured keywords for downstream processes. The following table illustrates the output of lexical analysis and clinical validation sets. The DistilBERT-based classification with an attention mechanism effectively organizes sentences into clinically meaningful sections, with higher attention weights assigned to critical findings.

Table 8. Output of BERT-GPT4 hybrid architecture.

Input Keywords	Encoded Semantic Context (BERT Output)	Generated Report Text (GPT4 Output)
Ischemic Stroke	[0.45, 0.68, 0.72, ...]	"The patient exhibits features consistent with ischemic stroke, primarily affecting the frontal lobe."
Frontal Lobe Atrophy	[0.32, 0.50, 0.65, ...]	"Significant atrophy is observed in the frontal lobe, indicative of chronic degenerative processes."
Right Parietal Lesion	[0.55, 0.71, 0.60, ...]	"A lesion is identified in the right parietal region, possibly associated with prior ischemic events."

Table 9. Output of hierarchical text classification with DistilBERT + Attention Mechanism.

Sentence	Attention Weight (α\ alphaα)	Assigned Section
"The patient exhibits features consistent with ischemic stroke, primarily affecting..."	0.82	History
"Significant atrophy is observed in the frontal lobe, indicative of chronic..."	0.75	Findings
"A lesion is identified in the right parietal region, possibly associated with prior..."	0.85	Findings

The lexical analysis as in Table 10 module will produce grammatically correct and semantically clear sentences, while the clinical validation cross-checks the text with the medical standards and returns high scores for the validation on all of the sentences.

Table 10. Output of MedSpacy NLP pipeline.

Raw Text	Extracted Entities	Normalized Keywords
"patient exhibits frontal lobe atrophy"	Frontal Lobe Atrophy	Frontal Lobe Atrophy
"right parietal lesion"	Right Parietal Lesion	Parietal Lesion (Right)
"ischemic changes visible"	Ischemic Changes	Ischemic Stroke

The output of the framework, as in Table 11, is a validated and categorized report ready to be used in clinical use sets. The outputs from each module validate the strength of the proposed framework. From initial keyword extraction to the final validated report, the framework successfully transforms raw clinical data into coherent, clinically accurate medical narratives.

Table 11. Output of lexical analysis with clinical context validation process.

Section	Original Sentence	Corrected Sentence	Validation Score (%)
History	"The patient exhibits features consistent with ischemia stroke, primarily effecting the frontal lobe."	"The patient exhibits features consistent with ischemic stroke, primarily affecting the frontal lobe."	98
Findings	"Significant atrophy observed in frontal lobe indicates degenerative processes."	"Significant atrophy is observed in the frontal lobe, indicative of chronic degenerative processes."	96
Findings	"A lesion identified right parietal region possibly associated prior ischemic events."	"A lesion is identified in the right parietal region, possibly associated with prior ischemic events."	97

These results as in Table 12 demonstrate its ability to transform medical documentation workflows by significantly improving sets on efficiency, consistency, and reliability levels.

Table 12. Final outputs.

Section	Final Validated Sentence
History	"The patient exhibits features consistent with ischemic stroke, primarily affecting the frontal lobe."
Findings	"Significant atrophy is observed in the frontal lobe, indicative of chronic degenerative processes."
Findings	"A lesion is identified in the right parietal region, possibly associated with prior ischemic events."

Conclusion and Future Scopes

While the proposed framework, which creates a brain MRI report, demonstrates great merit in automated medical documentation areas through achieving high linguistic coherence, clinical relevance, and computational efficiency, the hybrid BERT-GPT4 architecture for report generation, a hierarchical classification model Based on DistilBERT for real-time categorization, and the use of the very robust MedSpacy NLP pipeline for keyword extraction provides wide coverage over the problem of challenging clinical reports automation. The model achieved a BLEU-1 score of 85.6% and a BLEU-4 score of 70.3%. Thus, it can well generate linguistically accurate and fluent reports. The ROUGE-1 and ROUGE-L scores are 81.5% and 78.9%, respectively, indicating that the framework is highly similar to expert-written reports. Its CIDER score is 1.523, indicating it provides contextual relevance and is 13% better than the closest baseline (Method (Park et al., 2024)). When considering real-time categorization, the model has achieved an F1 Score of 94.5%, indicating it is better than the rest at organizing generated text into clinically relevant sections. The clinical relevance score of 96.2%, as validated by domain experts, signifies the adherence of the framework to medical standards and guidelines. In addition, the framework's processing time of 2.9 seconds per report, which is 31% less than the best-performing baseline, indicates practical suitability for real-time clinical use. These results together validate the robustness, reliability, and feasibility of the framework to streamline clinical documentation workflows for enhancing the efficiency of healthcare and decision-making processes.

The scope of this work going forward could be the expansion of this framework to support multimodal data, such as input of imaging data together with text inputs, which would enrich the depth and accuracy of reports that get generated. Furthermore, by pretraining at a large scale on domain-specific datasets, improved understanding of rare complex conditions will be achieved in the process. An even learning-active feedback loop might thus sustain continuous improvement through user interaction and corrections from the experts. Other promising directions for future work include real-world deployment in clinical settings and scalability studies in high-throughput environments.

Acknowledgements

The authors would like to acknowledge the support of Shri Ramdeobaba College of Engineering and Management,

Nagpur for providing the necessary facilities. Appreciation is extended to colleagues and peers for their helpful discussions and suggestions.

Funding

The authors received no specific funding for this work.

Conflict of Interest

The authors declare that they have no conflicts of interest to disclose.

References

Aissaoui Ferhi, L., Ben Amar, M., Choubani, F., & Bouallegue, R. (2024). Empowering medical diagnosis: A machine learning approach for symptom-based health checker. *Mobile Networks and Applications*, *29*(3), 676-702. https://doi.org/10.1007/s11036-024-02369-x

Almarri, B., Gupta, G., Kumar, R., Vandana, V., Asiri, F., & Khan, S. B. (2024). The BCPM method: decoding breast cancer with machine learning. *BMC Medical Imaging*, *24*(1), 248. https://doi.org/10.1186/s12880-024-01402-5

Aversano, L., Bernardi, M. L., Cimitile, M., Montano, D., Pecori, R., & Veltri, L. (2024). Explainable anomaly detection of synthetic medical iot traffic using machine learning. *SN Computer Science*, *5*(5), 488. https://doi.org/10.1007/s42979-024-02830-4

Brasen, C. L., Andersen, E. S., Madsen, J. B., Hastrup, J., Christensen, H., Andersen, D. P., ... & Brandslund, I. (2024). Machine learning in diagnostic support in medical emergency departments. *Scientific Reports*, *14*(1), 17889. https://doi.org/10.1038/s41598-024-66837-w

Dinsa, E. F., Das, M., Abebe, T. U., & Ramaswamy, K. (2024). Automatic categorization of medical documents in Afaan Oromo using ensemble machine learning techniques. *Discover Applied Sciences*, *6*(11), 575. https://doi.org/10.1007/s42452-024-06307-0

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*(1), e55315. https://doi.org/10.2196/55315

El-Rashidy, N., Sedik, A., Siam, A. I., & Ali, Z. H. (2023). An efficient edge/cloud medical system for rapid detection of level of consciousness in emergency medicine based on explainable machine learning models. *Neural computing and applications*, *35*(14), 10695-10716. https://doi.org/10.1007/s00521-023-08258-w

Gothai, E., Saravanan, S., Selvan, C. T., & Kumar, R. (2024). Optimized machine learning model discourse analysis. *Education and Information Technologies*, *29*(13), 16345-16363. https://doi.org/10.1007/s10639-024-12515-3

Grigorev, A., Saleh, K., Ou, Y., & Mihăiță, A. S. (2025). Enhancing traffic incident management with large language models: A hybrid machine learning approach for severity classification. *International Journal of Intelligent Transportation Systems Research*, *23*(1), 259-280. https://doi.org/10.1007/s13177-024-00448-7

Huang, Y., Liu, W., Yao, C., Miao, X., Guan, X., Lu, X., ... & Zhan, J. (2024). A multimodal dental dataset facilitating machine learning research and clinic services. *Scientific data*, *11*(1), 1291. https://doi.org/10.1038/s41597-024-04130-1

Hussain, S., Naseem, U., Ali, M., Avendaño Avalos, D. B., Cardona-Huerta, S., Bosques Palomo, B. A., & Tamez-Peña, J. G. (2024). TECRR: a benchmark dataset of radiological reports for BI-RADS classification with machine learning, deep learning, and large language model baselines. *BMC Medical Informatics and Decision Making*, *24*(1), 310. https://doi.org/10.1186/s12911-024-02717-7

Khan, A. A., Laghari, A. A., Baqasah, A. M., Bacarra, R., Alroobaea, R., Alsafyani, M., & Alsayaydeh, J. A. J. (2025). BDLT-IoMT—a novel architecture: SVM machine learning for robust and secure data processing in Internet of Medical Things with blockchain cybersecurity. *The Journal of Supercomputing*, *81*(1), 271. https://doi.org/10.1007/s11227-024-06782-7

Lemas, D. J., Du, X., Rouhizadeh, M., Lewis, B., Frank, S., Wright, L., ... & Modave, F. (2024). Classifying early infant feeding status from clinical notes using natural language processing and machine learning. *Scientific Reports*, *14*(1), 7831. https://doi.org/10.1038/s41598-024-58299-x

Li, J., Zhu, T., Wang, L., Yang, L., Zhu, Y., Li, R., ... & Zhang, L. (2024). Study on medical dispute prediction model and its clinical-application effectiveness based on machine learning. *BMC Medical Informatics and Decision Making*, *24*(1), 280. https://doi.org/10.1186/s12911-024-02674-1

Malashin, I., Masich, I., Tynchenko, V., Gantimurov, A., Nelyub, V., & Borodulin, A. (2024). Image text extraction and

natural language processing of unstructured data from medical reports. *Machine Learning and Knowledge Extraction*, *6*(2), 1361-1377.
https://doi.org/10.3390/make6020064

Ng, K. K. Y., & Zhang, P. C. (2023). Advancing medical affair capabilities and insight generation through machine learning techniques. *Journal of Pharmaceutical Policy and Practice*, *16*(1), 165.
https://doi.org/10.1186/s40545-023-00670-w

Park, H. A., Jeon, I., Shin, S. H., Seo, S. Y., Lee, J. J., Kim, C., & Park, J. O. (2024). Natural language processing-based deep learning to predict the loss of consciousness event using emergency department text records. *Applied Sciences*, *14*(23), 11399.
https://doi.org/10.3390/app142311399

Perez-Downes, J. C., Tseng, A. S., McConn, K. A., Elattar, S. M., Sokumbi, O., Sebro, R. A., ... & Adedinsewo, D. (2024). Mitigating Bias in Clinical Machine Learning Models: Perez-Downes et al. *Current Treatment Options in Cardiovascular Medicine*, *26*(3), 29-45.
https://doi.org/10.1007/s11936-023-01032-0

Pilz, M., Zimmermann, S., Friedrichs, J., Wördehoff, E., Ronellenfitsch, U., Kieser, M., & Vey, J. A. (2024). Semi-automated title-abstract screening using natural language processing and machine learning. *Systematic Reviews*, *13*(1), 274.
https://doi.org/10.1186/s13643-024-02688-w

Sailunaz, K., Özyer, T., Rokne, J., & Alhajj, R. (2023). A survey of machine learning-based methods for COVID-19 medical image analysis. *Medical & Biological Engineering & Computing*, *61*(6), 1257-1297.
https://doi.org/10.1007/s11517-022-02758-y

Singh, K. N., & Mantri, J. K. (2024). Clinical decision support system based on RST with machine learning for medical data classification. *Multimedia Tools and Applications*, *83*(13), 39707-39730.
https://doi.org/10.1007/s11042-023-16802-y

Singh, P., Chukkapalli, R., Chaudhari, S., Chen, L., Chen, M., Pan, J., ... & Cirrone, J. (2024). Shifting to machine supervision: annotation-efficient semi and self-supervised learning for automatic medical image segmentation and classification. *Scientific Reports*, *14*(1), 10820.
https://doi.org/10.1038/s41598-024-61822-9

Sirocchi, C., Bogliolo, A., & Montagna, S. (2024). Medical-informed machine learning: integrating prior knowledge into medical decision systems. *BMC medical informatics and decision making*, *24*(Suppl 4), 186.
https://doi.org/10.1186/s12911-024-02582-4

Wenstrup, J., Havtorn, J. D., Borgholt, L., Blomberg, S. N., Maaloe, L., Sayre, M. R., ... & Christensen, H. C. (2023). A retrospective study on machine learning-assisted stroke recognition for medical helpline calls. *NPJ digital medicine*, *6*(1), 235.
https://doi.org/10.1038/s41746-023-00980-y

Wu, R., Wang, B., & Zhao, Z. (2024). Privacy-preserving medical diagnosis system with Gaussian kernel-based support vector machine. *Peer-to-Peer Networking and Applications*, *17*(5), 3094-3109.
https://doi.org/10.1007/s12083-024-01743-6