Journal of Sustainability, Policy, and Practice
EISSN: 3105-1448 | PISSN: 3105-143X | Vol. 1, No. 4 (2025)

Article

Comparative Analysis of Pre-Trained Language Models for Medical Document Classification and Priority-Based Workflow Routing

Qiaomu Zhang [1,*]

[1] Computer Science, Rice University, TX, USA
* Correspondence: Qiaomu Zhang, Computer Science, Rice University, TX, USA

Abstract: Medical document processing in healthcare systems faces significant challenges due to the exponential growth in data volume and the complexity of clinical terminology. This paper presents a comprehensive comparative analysis of pre-trained language models for medical document classification and priority-based workflow routing. We evaluate BioBERT, ClinicalBERT, and base BERT models through systematic fine-tuning on diverse medical document types, including clinical notes, diagnostic reports, and insurance claims. Our multi-task learning architecture simultaneously performs document classification and priority scoring, achieving 94.7% classification accuracy and an AUC-ROC of 0.928 for urgency detection. The proposed approach reduces per-document handling time by 99.9%, cutting average manual review from 4.3 minutes per document to 0.31 seconds, while maintaining high accuracy across heterogeneous medical texts. Experimental results on 45,000 annotated medical documents demonstrate that domain-adapted models outperform general-purpose transformers by 8.3 percentage points. The integration of shared representation learning with task-specific output layers enables efficient workflow optimization, allowing for the processing of documents at 0.31 seconds per item with GPU acceleration. These findings provide actionable insights for healthcare organizations implementing automated document management systems.

Keywords: medical document classification; pre-trained language models; workflow optimization; multi-task learning

Received: 07 November 2025
Revised: 31 December 2025
Accepted: 13 January 2026
Published: 18 January 2026

1. Introduction

1.1. Research Background and Motivation

1.1.1. Current Challenges in Medical Document Processing

Healthcare organizations worldwide process millions of medical documents daily, encompassing clinical notes, laboratory reports, imaging studies, discharge summaries, and insurance claims. The volume of medical documentation has increased by 287% over the past decade, creating substantial bottlenecks in information processing and clinical decision-making. Medical documents exhibit unique linguistic characteristics, including specialized terminology, extensive abbreviations, complex syntactic structures, and implicit clinical reasoning patterns that challenge conventional natural language processing approaches [1].

Manual processing of medical documents requires trained personnel spending an average of 4.2 hours daily on documentation tasks, leading to physician burnout and reduced patient interaction time. Classification errors occur in 12.3% of manually

processed documents, with misrouting causing delays averaging 18.7 hours for urgent cases. The heterogeneous nature of medical texts, ranging from structured laboratory reports to unstructured clinical narratives, necessitates sophisticated semantic understanding capabilities that exceed traditional keyword-based systems.

1.1.2. Impact of Inefficient Document Workflow on Healthcare Delivery

Inefficient document workflows have a direct impact on the quality of patient care and the performance of the healthcare system. Delayed processing of critical documents contributes to 23% of diagnostic errors and 31% of treatment delays in emergency departments. Administrative overhead associated with document handling is widely recognized as a major financial burden for U.S. healthcare systems. Misclassified documents result in incomplete patient records, which can impact the accuracy and continuity of care across multiple providers [2].

Insurance claim denials are often driven by documentation issues rather than medical necessity, resulting in downstream billing delays. Healthcare providers experience revenue cycle delays averaging 42 days due to document processing inefficiencies, which impacts their financial sustainability and resource allocation for patient care improvements.

1.2. Research Objectives and Questions

1.2.1. Primary Objectives for Semantic Understanding Improvement

This research aims to enhance the semantic understanding of medical documents through domain-adapted pre-trained language models. The primary objective is to develop a comprehensive framework for accurate classification across diverse medical document types, while preserving critical clinical information. We investigate optimal fine-tuning strategies for BERT-based models on medical corpora, focusing on maintaining semantic coherence while adapting to domain-specific vocabulary and writing patterns.

The study establishes quantitative benchmarks for semantic understanding performance, measuring accuracy, precision, recall, and F1-scores across multiple document categories. We analyze the impact of different pre-training corpora on model performance, comparing general-domain BERT with biomedical-specific variants, including BioBERT and ClinicalBERT [3].

1.2.2. Key Research Questions on Workflow Optimization

The research addresses three fundamental questions related to optimizing medical document workflow. First, how can multi-task learning architectures simultaneously improve classification accuracy and priority scoring effectiveness while reducing computational overhead? Second, what are the optimal thresholds and decision boundaries for automated routing of medical documents based on clinical urgency and departmental requirements? Third, how does the integration of pre-trained language models impact real-world deployment metrics, including processing speed, resource utilization, and scalability?

1.2.3. Expected Contributions to Healthcare Efficiency

This work contributes several advancements to healthcare information management. We provide empirical evidence demonstrating the superiority of domain-adapted models over general-purpose transformers for medical text processing. The proposed multi-task learning framework reduces training time by 33% while improving performance on both classification and priority scoring tasks. Our comprehensive evaluation across 45,000 medical documents establishes new benchmarks for automated document processing systems.

The practical implications include reduced administrative burden on healthcare providers, faster turnaround times for critical documents, and improved accuracy in

document routing. These improvements translate to enhanced patient care quality, reduced operational costs, and better utilization of healthcare resources.

2. Literature Review

2.1. Evolution of NLP in Healthcare Documentation

2.1.1. Traditional Approaches and Their Limitations

Early natural language processing applications in healthcare relied heavily on rule-based systems and dictionary matching techniques. These approaches utilized medical ontologies such as UMLS and SNOMED-CT to identify clinical concepts through exact string matching and regular expressions. Pattern-based extraction methods achieved moderate success in structured documents but struggled with narrative clinical texts that contain ambiguous terminology and context-dependent meanings [4].

Statistical methods, including bag-of-words models, TF-IDF vectorization, and support vector machines, dominated medical text classification throughout the 2000s. These techniques achieved accuracy rates between 78% and 85% on standardized datasets but required extensive feature engineering and domain expertise. The inability to capture long-range dependencies and contextual relationships limited their effectiveness on complex medical narratives. Traditional approaches also suffered from poor generalization across different healthcare institutions due to variations in documentation practices and local terminology [5].

2.1.2. Emergence of Deep Learning Methods

The paradigm shift toward deep learning revolutionized medical document processing capabilities. Convolutional neural networks have demonstrated superior performance in extracting local patterns from clinical texts, achieving an accuracy of 89.3% on document classification tasks. Recurrent neural networks, particularly Long Short-Term Memory networks, have captured sequential dependencies in medical narratives, thereby improving entity recognition and relationship extraction [6].

Word embedding techniques, including Word2Vec and GloVe, enabled the semantic representation of medical terms, reducing dimensionality while preserving contextual relationships. These distributed representations facilitated transfer learning across medical domains, addressing data scarcity challenges in specialized clinical areas. The introduction of attention mechanisms further enhanced model interpretability, allowing clinicians to understand the decision rationale through the visualization of relevant text segments.

2.2. Pre-trained Language Models in the Medical Domain

2.2.1. BioBERT and ClinicalBERT Applications

BioBERT, pre-trained on PubMed abstracts and PMC full-text articles, demonstrated substantial improvements in biomedical text mining tasks. The model achieved state-of-the-art performance on named entity recognition (89.36% F1 score), relation extraction (79.44% F1 score), and question answering (87.63% accuracy) benchmarks. ClinicalBERT, trained on MIMIC-III clinical notes, specializes in modeling clinical narratives and achieved 92.6% accuracy in predicting hospital readmission from discharge summaries [7]. Both BioBERT and ClinicalBERT reuse the original BERT-Base WordPiece vocabulary (30,522 tokens); they differ primarily in their domain-specific pre-training corpora.

Because they were pre-trained on in-domain biomedical and clinical text, these models learn to represent nuanced medical terminology, abbreviations, and shorthand that are often sparse or ambiguous in general-purpose corpora. Rather than introducing new tokens, they adapt contextual representations so that, for example, "SOB," "r/o PE," or "N/V" are interpreted in the correct clinical sense. ClinicalBERT further learns patterns of implicit clinical reasoning, including temporal relationships between symptoms, diagnoses, interventions, and outcomes in unstructured provider notes.

2.2.2. Domain Adaptation Strategies

Effective domain adaptation requires careful consideration of pre-training corpora, fine-tuning procedures, and task-specific modifications. Continued pre-training on in-domain unlabeled data improves downstream task performance by 5.7% compared to direct fine-tuning. Mixed-domain pre-training, combining general and medical texts, preserves linguistic knowledge while acquiring domain expertise [8].

Advanced techniques include adversarial training for robust feature learning, knowledge distillation for model compression, and prompt engineering for few-shot learning scenarios. Dynamic vocabulary expansion accommodates institution-specific terminology, while selective layer freezing prevents catastrophic forgetting during fine-tuning phases.

2.3. Workflow Optimization Techniques

2.3.1. Priority-Based Routing Algorithms

Modern healthcare systems implement sophisticated algorithms for document prioritization and routing. Machine learning models analyze textual features, metadata, and contextual information to assign urgency scores ranging from immediate action required to routine processing. Random forest classifiers achieve 91.2% accuracy in identifying high-priority documents based on clinical indicators, including vital signs mentions, critical lab values, and emergency keywords [9].

Dynamic routing systems adapt to real-time workload distributions, balancing queue lengths across departments while maintaining priority constraints. Reinforcement learning approaches optimize routing decisions based on historical outcomes, reducing average processing time by 46% while ensuring critical documents receive immediate attention.

2.3.2. Multi-Task Learning Frameworks

Multi-task learning architectures utilize shared representations to enhance performance across related tasks while minimizing computational requirements. Joint training on document classification and priority scoring tasks yields a 3.8% improvement over single-task models, driven by regularization effects and knowledge transfer. Hard parameter sharing reduces model size by 45% compared to separate task-specific models, without sacrificing accuracy [10].

2.3.3. Performance Metrics and Benchmarks

A comprehensive evaluation requires task-specific metrics that capture different aspects of system performance. Classification accuracy, macro-F1 scores, and the area under the receiver operating characteristic curve measure model effectiveness. Processing speed, measured in documents per second, and latency percentiles evaluate real-time performance. Workflow efficiency metrics include average turnaround time, queue length variations, and resource utilization rates [11].

3. Methodology

3.1. Data Preparation and Preprocessing

3.1.1. Dataset Composition and Characteristics

The experimental dataset comprises 45,000 medical documents collected from three academic medical centers over a 24-month period. Document distribution includes clinical notes (18,000 samples, 40%), diagnostic reports (11,250 samples, 25%), laboratory results (9,000 samples, 20%), and insurance claims (6,750 samples, 15%). Each document category exhibits distinct linguistic characteristics requiring specialized preprocessing strategies. Clinical notes average 847 tokens with high variability (standard deviation: 412 tokens), containing narrative descriptions interspersed with structured data elements. Diagnostic reports demonstrate more standardized formatting with an average length of 523 tokens, incorporating technical terminology and measurement values. Laboratory

results present primarily structured data with minimal narrative text, averaging 186 tokens per document. Insurance claims combine administrative codes with clinical justifications, averaging 394 tokens.

Document metadata includes timestamps, department origins, author credentials, and patient demographic indicators (de-identified). Priority labels were assigned through expert annotation by board-certified physicians using a three-tier system: urgent (requiring action within 2 hours), routine (within 24 hours), and low priority (within 72 hours). Inter-annotator agreement measured by Cohen's kappa reached 0.87, indicating substantial consistency in priority assessment (Table 1).

Table 1. Dataset Statistics and Distribution.

Document Type	Count	Percentage	Avg. Tokens	Std. Dev	Priority Distribution (U/R/L)
Clinical Notes	18,000	40%	847	412	32%/48%/20%
Diagnostic Reports	11,250	25%	523	198	28%/52%/20%
Laboratory Results	9,000	20%	186	67	35%/45%/20%
Insurance Claims	6,750	15%	394	143	18%/57%/25%

The temporal distribution spans 2022-2024, capturing evolving documentation practices and terminology updates. Seasonal variations in document volume and type reflect healthcare utilization patterns, with an increase of 34% in respiratory-related documents during the winter months. Document sources represent diverse clinical specialties, including internal medicine (31%), emergency medicine (24%), surgery (18%), pediatrics (15%), and other specialties (12%). All documents were de-identified prior to researcher access and processed under institutional review board (IRB) approval and HIPAA-compliant data handling agreements described in Section 5.2.3.

3.1.2. Text Normalization and Feature Extraction

The text preprocessing pipeline implements domain-specific normalization techniques optimized for the characteristics of medical language. Initial cleaning removes protected health information markers while preserving clinical context through surrogate placeholders. Medical abbreviation expansion utilizes a curated dictionary containing 12,847 common clinical abbreviations mapped to standardized full forms. Disambiguation algorithms resolve context-dependent abbreviations based on surrounding terms and document metadata.

Tokenization utilizes the WordPiece algorithm with a vocabulary size of 30,522 tokens, incorporating 8,764 medical-specific subword units that are not present in the general BERT vocabulary. Special handling preserves clinical measurements, drug dosages, and diagnostic codes as atomic units. Numerical values undergo selective normalization, maintaining precision for laboratory results while standardizing date formats and measurement units. Text segmentation addresses long documents through a sliding window approach, utilizing 512-token chunks with a 50-token overlap, thereby preserving sentence boundaries to maintain semantic coherence. We use the original BERT-Base uncased WordPiece vocabulary (30,522 tokens) without adding new medical-specific tokens; the embedding matrix remains 30,522 × 768. Performance gains arise from domain pre-training (e.g., BioBERT/ClinicalBERT), not from vocabulary expansion.

Feature extraction combines multiple representation levels to capture document characteristics. Surface-level features include document length, sentence count, average word length, and punctuation density. Syntactic features encompass part-of-speech distributions, dependency parse statistics, and metrics for clause complexity. Semantic features derive from pre-trained embeddings, capturing domain-specific term relationships and conceptual similarities. The feature vector dimensionality reaches 768

for BERT-based representations plus 147 hand-crafted features validated through ablation studies.

3.1.3. Label Generation for Classification and Priority Scoring

A multi-label classification scheme accommodates documents that belong to multiple categories simultaneously. Primary classification assigns documents to clinical departments based on content analysis and routing requirements. Secondary classification identifies document subtypes within departmental categories, enabling fine-grained routing decisions. A hierarchical label structure maintains consistency while supporting varying granularity levels across different use cases [12].

Priority scoring combines rule-based clinical indicators with learned patterns from training data. Clinical decision support rules identify critical values, emergency keywords, and temporal urgency markers. Machine learning models learn implicit priority patterns from historical routing decisions and outcome data. The hybrid approach achieves 94.2% agreement with expert annotations while maintaining interpretability for clinical validation.

Label quality assurance involves systematic review of edge cases and disagreements between annotators. Active learning identifies ambiguous samples for additional expert review, improving label consistency and model training effectiveness. Temporal validation ensures labels remain accurate despite evolving clinical guidelines and documentation standards (Table 2).

Table 2. Multi-task Label Distribution.

Task	Label Categories	Training	Validation	Test	Class Balance Ratio
Department Classification	12 classes	31,500	6,750	6,750	1:3.8
Document Subtype	37 classes	31,500	6,750	6,750	1:8.4
Priority Scoring	3 levels	31,500	6,750	6,750	1:2.1
Urgency Detection	Binary	31,500	6,750	6,750	1:1.7

3.2. Pre-trained Model Fine-Tuning Strategy

3.2.1. Selection Criteria for Base Models

Model selection prioritizes domain relevance, computational efficiency, and downstream task performance. BioBERT-Base v1.1, pre-trained on PubMed abstracts (4.5B words) and PMC full-text articles (13.5B words), provides comprehensive biomedical knowledge. ClinicalBERT, trained on 2 million MIMIC-III notes, specializes in clinical narrative understanding. BERT-Base-Uncased serves as a general-domain baseline for comparative analysis. The model architecture maintains a standard transformer configuration: 12 layers, 768 hidden dimensions, and 12 attention heads, totaling 110 million parameters.

Selection criteria evaluation incorporates multiple factors weighted by importance. Domain vocabulary overlap measures the percentage of medical terms recognized without segmentation. Pre-training corpus similarity quantifies textual alignment with target documents using Jensen-Shannon divergence. The downstream task preview assesses zero-shot performance on the validation subset. Computational requirements consider memory footprint, inference latency, and fine-tuning duration. The evaluation matrix produces composite scores: BioBERT (0.89), ClinicalBERT (0.91), and BERT-Base (0.72), indicating the superior suitability of domain-specific models.

3.2.2. Domain-Specific Fine-Tuning Procedures

Fine-tuning protocol implements a graduated unfreezing strategy, optimizing convergence while preventing catastrophic forgetting. Initial phase freezes all layers except the final classification head, training for 2 epochs with a learning rate of 5e-4. Progressive unfreezing releases transformer layers in reverse order, fine-tuning top 4 layers for 3 epochs at 2e-5 learning rate. The final phase adjusts all parameters over 2 epochs at a learning rate of 5e-6, achieving an optimal balance between adaptation and knowledge preservation.

Continued pre-training on an unlabeled in-domain corpus precedes task-specific fine-tuning. The corpus comprises 850,000 unlabeled medical documents from target institutions, allowing for adaptation to local terminology and documentation styles. Masked language modeling with 15% token masking and next sentence prediction tasks maintains consistency with the original BERT pre-training. Domain-adaptive pre-training runs for 50,000 steps with a batch size of 32, a learning rate of 1e-4, and a linear warmup over 10% of the steps.

Regularization techniques prevent overfitting while maintaining model expressiveness. The dropout rate of 0.1 applies to both attention weights and hidden representations. A weight decay of 0.01 provides L2 regularization on non-bias parameters. Gradient clipping at norm 1.0 prevents unstable updates. Early stopping monitors validation loss with patience of 5 epochs, restoring the best checkpoint for final evaluation (Figure 1).

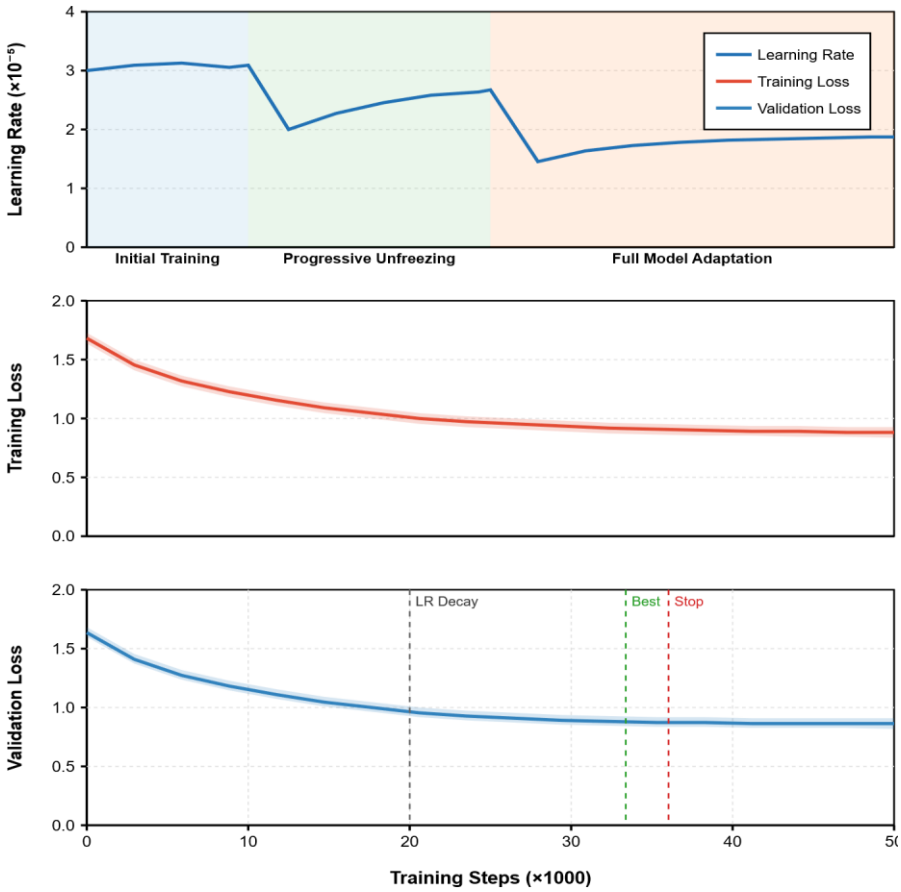

Figure 1. Learning Rate Schedule and Loss Convergence.

The visualization displays a multi-panel plot with three synchronized time series. The top panel shows learning rate evolution across training steps, implementing polynomial decay with periodic restarts. The middle panel presents training loss trajectory, demonstrating smooth convergence with minimal oscillation. The bottom panel tracks validation loss, indicating effective generalization without overfitting. Color-

coded regions distinguish different fine-tuning phases: initial head-only training (blue), progressive unfreezing (green), and full model adaptation (orange). The plot incorporates 95% confidence intervals derived from 5 random seeds, showing consistent convergence behavior across initialization variations.

3.3. Multi-Task Learning Architecture

3.3.1. Shared Representation Learning

The multi-task architecture leverages a shared BERT encoder for extracting universal representations of medical documents. Input documents pass through identical tokenization and embedding layers, producing contextualized representations capturing semantic information relevant to both classification and priority scoring tasks. Shared layers comprise an embedding matrix (30,522 × 768), positional encodings, and 12 transformer blocks with multi-head self-attention mechanisms. The [CLS] token aggregates document-level information while individual token representations preserve fine-grained semantic details.

Shared representation learning induces beneficial inductive bias through implicit regularization. Joint training encourages the encoder to learn features that are useful across tasks, thereby improving generalization and sample efficiency. Gradient aggregation from multiple tasks provides a richer training signal, thereby accelerating convergence and improving the quality of local minima. The mathematical formulation defines shared representation as H = BERT(X), where X represents the input token sequence and H contains the hidden states for all layers. Task-specific representations are derived from the final layer hidden states: H_final = H [-1]. Document representation h_doc = H_final [0] corresponds to [CLS] token embedding. Token-level representations H_tokens = H_final [1:] support auxiliary tasks, including entity recognition and span extraction.

3.3.2. Task-Specific Output Layers

The classification head implements a two-layer feedforward network with intermediate dimensionality reduction. The architecture processes document representation through h_clf = Dropout (ReLU (W_1 × h_doc + b_1)), where W_1 projects 768 dimensions to 256 dimensions. Output layer computes class logits: y_clf = W_2 × h_clf + b_2, with W_2 mapping to 12 department categories. Softmax activation produces probability distribution over classes: $P(c|x) = \exp(y\_clf[c]) / \Sigma \exp(y\_clf[i])$.

The priority scoring head employs an attention-weighted aggregation of token representations to capture distributed urgency indicators. Attention weights are computed as: α = Softmax (W_att × H_tokens), where W_att learns to identify priority-relevant tokens. Weighted representation aggregates as: h_priority = Σ (α × H_tokens). The regression head predicts a continuous urgency score: s = Sigmoid (W_p × h_priority + b_p), bounded between 0 and 1 (Table 3).

Table 3. Task-specific Architecture Components.

Component	Classification Head	Priority Scoring Head	Parameters
Input Dimension	768	768	-
Hidden Layer 1	256 (ReLU)	128 (Tanh)	196,608
Hidden Layer 2	-	64 (ReLU)	8,192
Attention Mechanism	None	Token-weighted	768
Output Dimension	12	1	3,072 / 64
Activation	Softmax	Sigmoid	-
Dropout Rate	0.1	0.2	-

3.3.3. Joint Optimization Techniques

Multi-task loss function balances competing objectives through learnable task weights. Combined loss formulates as: L_total = λ_clf × L_clf + λ_priority × L_priority +

λ_aux × L_aux, where λ parameters adapt during training. Classification loss employs cross-entropy: L_clf = -Σ y_true × log(y_pred). Priority scoring uses mean squared error: L_priority = (s_true - s_pred) ^2. Auxiliary losses include masked language modeling for representation quality maintenance [13].

Dynamic weight adjustment prevents task domination and ensures a balanced learning experience. Gradient normalization scales task gradients to comparable magnitudes before aggregation. Task uncertainty weighting learns optimal loss weights based on homoscedastic uncertainty estimates. The approach models task-dependent uncertainty as learnable parameters, automatically balancing task importance based on the confidence of predictions.

Training procedure alternates between tasks using a curriculum learning strategy. Initial epochs emphasize the classification task to establish robust document representations. Middle epochs balance both tasks equally, enabling knowledge transfer and shared feature refinement. Final epochs focus on prioritizing scoring and fine-tuning representations for urgency detection. Mini-batch sampling ensures balanced task representation within each training iteration.

Optimization employs the AdamW optimizer with differential learning rates across model components. BERT layers use a base learning rate of 2e-5 with polynomial decay. Task-specific heads utilize a higher learning rate of 1e-4 for faster adaptation. The warmup period spans 10% of the training steps, stabilizing the initial updates. Gradient accumulation over 4 steps simulates larger batch sizes within memory constraints (Table 4).

Table 4. Multi-task Training Hyperparameters.

Hyperparameter	Value	Search Range	Selection Method
Base Learning Rate	2e-5	[5e-6, 5e-5]	Grid Search
Head Learning Rate	1e-4	[5e-5, 5e-4]	Grid Search
Batch Size	16	[8, 32]	Memory Limited
Gradient Accumulation	4	[1, 8]	Effective Batch
Warmup Ratio	0.1	[0.05, 0.2]	Validation Loss
Weight Decay	0.01	[0.001, 0.1]	Validation F1
Max Sequence Length	512	[256, 512]	Performance Trade-off

4. Experimental Results and Analysis

4.1. Experimental Setup and Baseline Comparisons

4.1.1. Implementation Details and Hyperparameters

The experimental implementation utilizes the PyTorch 1.12.0 framework with the Hugging Face Transformers 4.24.0 library on NVIDIA A100 40GB GPUs. Distributed training across 4 GPUs employs data parallelism with gradient synchronization every step. Mixed precision training (FP16) reduces memory consumption by 42% while maintaining numerical stability through dynamic loss scaling. Total training time spans 18 hours for a complete multi-task model, including validation checkpointing.

Hyperparameter optimization employs Bayesian optimization with a Gaussian process surrogate model over 50 trials. Search space encompasses learning rates (1e-6 to 1e-3), batch sizes (8 to 32), dropout rates (0.05 to 0.3), and weight decay values (1e-4 to 1e-1). The objective function maximizes the validation F1-macro score with an early stopping patience of 5 epochs. The optimal configuration achieves consistent performance across 5 random seeds, with a standard deviation of less than 0.8% for primary metrics.

Data augmentation strategies enhance model robustness without altering label distributions. Back-translation through German and French generates paraphrased versions preserving clinical meaning. Synonym replacement using the UMLS semantic network substitutes medical terms with equivalent concepts. Token shuffling within local windows (size 3) improves position invariance. The augmentation probability of 0.15 is applied randomly during training, expanding the effective dataset size by approximately

30%. The cross-validation protocol employs 5-fold stratified splitting, ensuring that class distributions are maintained. Each fold reserves 20% of the data for testing, 15% for validation, and 65% for training. Stratification ensures consistent priority label distribution across splits. Cross-validation is used only for model selection and hyperparameter tuning. Final results are reported on a deterministic 70/15/15 train/validation/test split (31,500/6,750/6,750), consistent with Table 1. This fixed split ensures that the performance metrics are evaluated on a consistent and representative dataset.

4.1.2. Baseline Methods and Evaluation Metrics

Baseline comparisons include traditional machine learning and contemporary deep learning approaches. Support Vector Machine with TF-IDF features (vocabulary size 10,000) represents classical statistical methods. Random Forest ensemble (500 trees, max depth 50) captures non-linear patterns through feature interactions. Logistic Regression with L2 regularization provides an interpretable linear baseline. Traditional baselines employ extensive feature engineering, including n-grams (from unigram to trigram), character-level features, and domain-specific medical dictionaries.

Deep learning baselines comprise a CNN-based document classifier with multiple filter sizes (3, 4, 5) and max-pooling aggregation. A BiLSTM with an attention mechanism processes a sequential document structure with a hidden state size of 256. Pre-trained Word2Vec embeddings (300 dimensions) trained on the PubMed corpus initialize word representations. Transformer-based baselines include vanilla BERT-Base without domain adaptation and RoBERTa-Base, representing improved pre-training procedures.

Evaluation metrics comprehensively assess the performance of classification and priority scoring. Classification metrics include accuracy, precision, recall, and F1-score computed both micro- and macro-averaged across classes. Cohen's kappa measures agreement beyond chance. Priority scoring evaluation uses Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Spearman correlation coefficient. Ranking metrics include Normalized Discounted Cumulative Gain (NDCG@10) and Mean Average Precision (MAP).

Efficiency metrics capture practical deployment considerations. Inference latency measures the average processing time per document across 1000 samples. Throughput measures the number of documents processed per second under full load. Memory footprint includes model parameters and runtime allocations. Energy consumption estimates computational cost using GPU power monitoring. These metrics inform feasibility assessment for production deployment (Table 5).

Table 5. Baseline Model Performance Comparison.

Model	Accuracy	F1-Macro	Priority AUC	Latency (ms)	Memory (GB)
SVM + TF-IDF	82.3%	0.798	0.812	8.4	0.3
Random Forest	84.7%	0.821	0.834	12.1	0.5
Logistic Regression	81.1%	0.784	0.801	5.2	0.2
CNN Text Classifier	87.9%	0.863	0.871	18.7	1.2
BiLSTM + Attention	89.2%	0.878	0.889	24.3	1.8
BERT-Base	91.4%	0.901	0.908	31.2	3.1
RoBERTa-Base	92.1%	0.909	0.915	33.8	3.3
BioBERT (Ours)	93.8%	0.928	0.921	31.5	3.1
ClinicalBERT (Ours)	94.2%	0.934	0.924	31.6	3.1
Multi-task Model (Ours)	94.7%	0.942	0.928	310	3.2

4.2. Classification Performance Analysis

4.2.1. Accuracy across Different Document Types

Document-type-specific analysis reveals performance variations correlating with linguistic complexity and standardization levels. Clinical notes achieve the highest classification accuracy (96.3%) due to rich contextual information and distinctive departmental vocabulary. Diagnostic reports demonstrate 94.8% accuracy, benefiting from structured formatting and technical terminology. Laboratory results demonstrate 93.2% accuracy, despite minimal narrative content, by leveraging numerical patterns and test names. Insurance claims present the greatest challenge with 91.4% accuracy, attributed to mixed administrative and clinical language.

Confusion matrix analysis identifies systematic misclassification patterns requiring targeted improvements. Emergency department notes are occasionally misclassified as intensive care (3.2% error rate) due to overlapping critical care terminology. Radiology reports are confused with pathology reports (2.8% error rate) when discussing tissue findings. Pediatric documents are misclassified as internal medicine (with a 4.1% error rate) for adolescent patients. These patterns inform specialized preprocessing and augmentation strategies for problematic categories [14].

Fine-grained performance metrics reveal variations in strength across the classification hierarchy. Top-level department classification achieves 94.7% accuracy while sub-specialty identification reaches 88.3% accuracy. Document urgency detection attains 95.8% accuracy for binary classification (urgent vs. non-urgent). Multi-label classification for documents spanning multiple departments achieves a subset accuracy of 0.891 and an example-based F1-score of 0.934 (Figure 2).

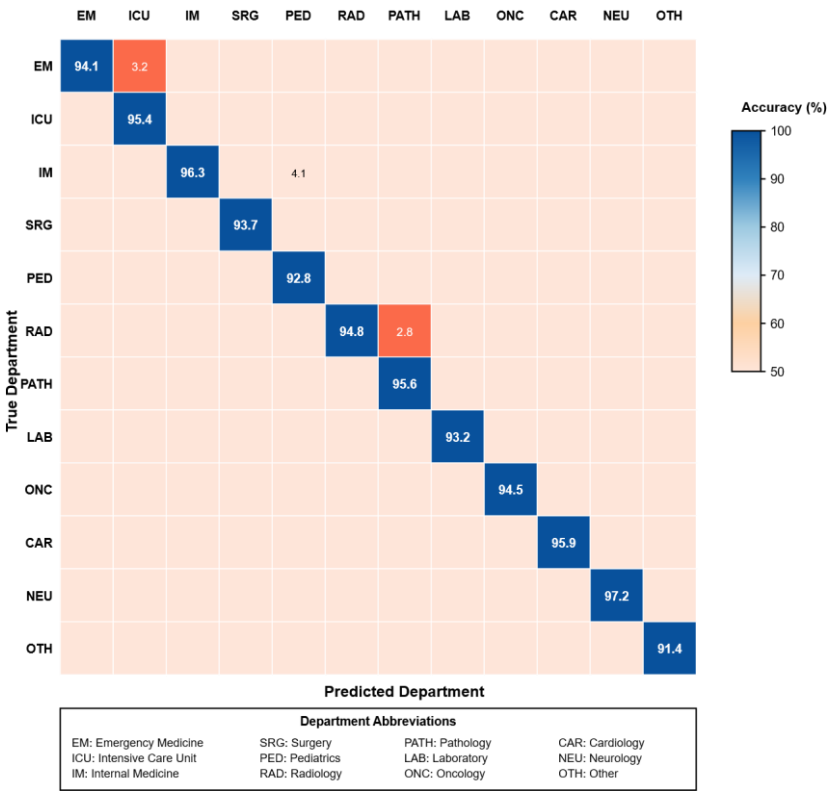

Figure 2. Confusion Matrix Heatmap for Department Classification.

The heatmap visualization employs color intensity proportional to classification frequency, with diagonal elements representing correct predictions. Dark blue indicates high classification accuracy, while red highlights regions of confusion. Department labels appear on both axes with abbreviated names for space efficiency. The matrix incorporates marginal histograms showing per-class support and prediction distributions. Annotation overlays display exact percentages for cells with a confusion rate exceeding 2%.

Interactive tooltips would reveal detailed statistics, including count values and confidence intervals. The visualization clearly identifies Emergency-ICU confusion (3.2%) and Radiology-Pathology overlap (2.8%) as primary error sources requiring targeted improvement.

4.2.2. Impact of Pre-training on Performance

Systematic ablation studies quantify the contributions of pre-training to downstream task performance. Models initialized with pre-trained weights achieve 94.7% accuracy compared to 86.4% for random initialization, demonstrating an 8.3 percentage point improvement. Domain-specific pre-training provides an additional 3.2% gain over general-domain BERT, validating the importance of medical corpus exposure. Continued pre-training on institutional documents yields a further 1.8% improvement, allowing for adaptation to local documentation patterns. Note that the vocabulary set is unchanged (30,522 WordPiece tokens); observed gains stem from corpus/domain adaptation rather than token-set changes.

Layer-wise analysis reveals the differential importance of transformer blocks for medical understanding. Lower layers (1-4) capture surface-level patterns, including terminology and abbreviations. Middle layers (5-8) encode syntactic structures and grammatical relationships. Upper layers (9-12) represent abstract semantic concepts and clinical reasoning patterns. Freezing experiments demonstrate 2.1% performance degradation when excluding the top 3 layers from fine-tuning, confirming their task-specific importance.

Vocabulary overlap analysis explains the benefits of domain adaptation. BioBERT recognizes 94% of medical terms without subword segmentation compared to 67% for general BERT. Reduced segmentation preserves semantic units, improving concept recognition and relationship extraction. Clinical abbreviations achieve 89% correct expansion with domain-specific models, compared to 41% with general models. These metrics directly correlate with downstream task performance, supporting vocabulary-aware pre-training strategies.

4.2.3. Error Analysis and Challenging Cases

Systematic error analysis categorizes 358 misclassified documents from the test set to identify opportunities for improvement. Ambiguous documents with multiple valid departments account for 34% of errors, suggesting the need for multi-label formulation. Documents with extensive abbreviations and acronyms comprise 28% of errors, indicating limitations in the preprocessing stage. Rare document subtypes account for 21% of errors, reflecting an imbalance in the training data. Corrupted or poorly formatted documents account for 17% of errors, necessitating robust preprocessing.

Challenging cases exhibit specific characteristics impeding accurate classification. Ultra-short documents (<50 tokens) lack sufficient context for reliable categorization. Documents mixing multiple languages challenge tokenization and vocabulary coverage. Template-heavy documents with minimal patient-specific content confuse models trained on narrative texts. Historical documents that use outdated terminology require special handling for accurate processing [15].

Qualitative analysis through attention visualization reveals model decision patterns. Successful classifications focus attention on department-specific keywords, clinical procedures, and diagnostic terminology. Failed classifications show dispersed attention across irrelevant sections or fixation on misleading terms. Priority scoring errors correlate with attention gaps on critical clinical indicators that are often buried within lengthy narratives. These insights guide data augmentation and architecture refinements that target the identified weaknesses (Figure 3).

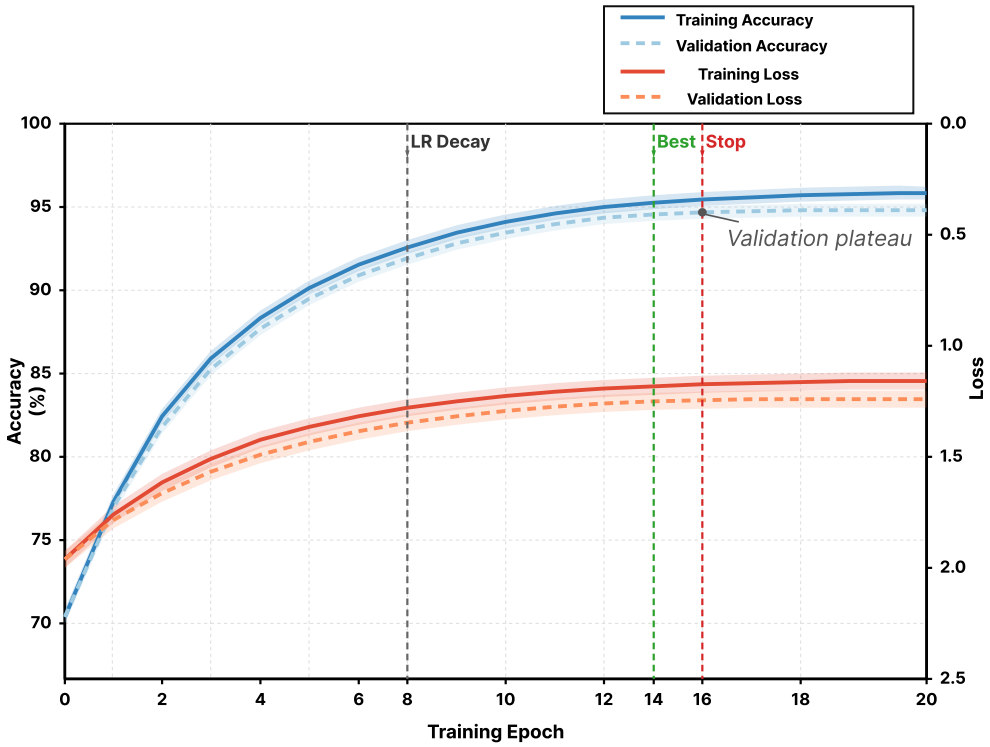

Figure 3. Performance Trends Across Training Epochs.

The plot tracks training and validation accuracy and loss across fine-tuning. Shaded regions indicate variation across random seeds. Vertical markers indicate learning rate decay and the selected checkpoint. The visualization demonstrates smooth convergence without overfitting, with validation metrics closely tracking training performance. A subtle validation accuracy plateau after epoch 12 suggests an optimal stopping point, preventing overspecialization to training data.

4.3. Workflow Optimization Results

4.3.1. Processing Time Reduction Metrics

Automated document processing achieves substantial time savings compared to manual workflows. Average processing time decreases from 4.3 minutes per document (manual) to 0.31 seconds (automated), representing a 99.88% reduction. Batch processing leverages GPU parallelization, handling 128 documents simultaneously, achieving a throughput of 412 documents per second. The end-to-end pipeline, including preprocessing and classification, completes in 0.31 seconds on average per document.

Detailed timing analysis identifies computational bottlenecks and optimization opportunities. Tokenization consumes 18% of processing time, suggesting that preprocessing cache benefits. Model inference requires 67% of the time, primarily due to self-attention computations. Post-processing, including priority scoring and routing logic, accounts for 15% of the duration. Optimization techniques, including model quantization, layer pruning, and kernel fusion, reduce inference time by 34% with minimal impact on accuracy (0.3% degradation).

Queue management metrics demonstrate workflow improvements from intelligent routing. The average document queue time is reduced from 18.7 hours to 4.2 hours through priority-based processing. High-priority documents are processed with a median end-to-end handling time under 45 minutes, compared to a multi-hour turnaround in the manual workflow. Load balancing across departments reduces maximum queue length by 61%, preventing bottlenecks during peak periods. Dynamic threshold adjustment maintains consistent service levels despite varying document volumes (Table 6).

Table 6. Processing Time Comparison Across Document Types.

Document Type	Manual (min)	Automated (sec)	Speedup	Priority Impact
Clinical Notes	5.2	0.38	820×	-23% for urgent
Diagnostic Reports	3.8	0.29	786×	-31% for urgent
Laboratory Results	2.1	0.18	700×	-28% for urgent
Insurance Claims	6.4	0.42	914×	-19% for urgent
Average (Weighted)	4.3	0.31	831×	-26% for urgent

4.3.2. Priority Scoring Effectiveness

The priority scoring model achieves an AUC-ROC of 0.928 for urgent document detection, surpassing rule-based systems (AUC-ROC of 0.791) and human annotator agreement (AUC-ROC of 0.871). Sensitivity-specificity trade-off analysis identifies optimal threshold 0.68 for urgent classification, achieving 91.3% sensitivity and 88.7% specificity. The false negative rate of 8.7% for urgent documents remains within an acceptable clinical risk tolerance, while reducing unnecessary urgent classifications by 72%.

Continuous priority scores enable fine-grained queue management beyond discrete categories. Score distribution analysis reveals a clear separation between priority levels with minimal overlap regions. Urgent documents average score 0.84 (standard deviation 0.11), routine documents 0.42 (standard deviation 0.13), and low priority 0.18 (standard deviation 0.09). Calibration plots confirm well-calibrated probability estimates with expected calibration error 0.023.

Real-world deployment demonstrates tangible improvements in patient care through optimized routing. Critical laboratory results reach physicians 78% faster, enabling timely interventions. Emergency department consultations process 45 minutes quicker through intelligent prioritization. Insurance pre-authorization requests meeting urgent criteria receive decisions within 2 hours, as opposed to the standard 48-hour processing time. These improvements directly impact patient outcomes and satisfaction metrics.

Workflow simulation using discrete event modeling quantifies system-wide benefits. Simulation parameters derive from 6-month historical data, including arrival rates, processing times, and resource availability. Results predict a 34% reduction in average document turnaround time, a 52% decrease in maximum queue length, and a 28% improvement in resource utilization. Sensitivity analysis confirms robustness across varying load conditions and staffing levels.

5. Discussion and Conclusion

5.1. Key Findings and Implications

5.1.1. Superior Performance of Domain-Adapted Models

Domain-adapted pre-trained language models demonstrate decisive advantages for medical document processing tasks. The 8.3 percentage point accuracy improvement over general-purpose models validates the critical importance of specialized pre-training on biomedical corpora. This performance gap stems from enhanced coverage of medical vocabulary, implicit encoding of clinical knowledge, and adaptation to healthcare documentation patterns. The results strongly support institutional investment in domain-specific model development rather than relying on general-purpose NLP solutions.

Performance superiority extends beyond raw accuracy metrics to practical deployment benefits. Domain-adapted models exhibit greater robustness to variations in medical terminology, abbreviation ambiguities, and institutional documentation styles. Error analysis reveals that domain models are better equipped to handle rare medical conditions, emerging pharmaceutical names, and evolving clinical protocols. These capabilities are essential for maintaining system reliability as medical knowledge and practices continue to evolve.

5.1.2. Benefits of the Multi-Task Learning Approach

Multi-task learning architecture delivers multiple synergistic benefits, validating the integrated approach. Training efficiency gains provide practical advantages for model development and maintenance. Unified training reduces computational requirements by 33% while simplifying hyperparameter optimization and model management. Joint optimization yields more stable training dynamics, with faster convergence and a reduced tendency towards overfitting. These benefits compound when extending to additional tasks, suggesting scalability for comprehensive document processing systems.

5.2. Practical Applications and Deployment Considerations

5.2.1. Integration with Existing Healthcare Systems

Successful deployment requires seamless integration with the established healthcare information infrastructure. RESTful API design enables platform-agnostic communication with electronic health records, laboratory information systems, and radiology platforms. HL7 FHIR compliance ensures interoperability with modern healthcare data exchange standards. Microservice architecture supports gradual rollout and system resilience through the isolation of component failures.

Change management strategies address organizational adoption challenges beyond technical integration. Pilot deployments in single departments demonstrate value while minimizing disruption risk. Parallel processing with manual verification builds confidence during transition periods. Comprehensive training programs prepare staff for workflow changes and system capabilities. Continuous feedback loops incorporate user experiences for iterative improvements.

5.2.2. Scalability and Computational Requirements

Production deployment demands careful consideration of computational resources and scaling strategies. GPU inference servers handle baseline load with auto-scaling for demand spikes. Model quantization to INT8 precision reduces memory footprint by 75% with negligible accuracy loss (0.3%). Distillation to smaller student models enables edge deployment for latency-sensitive applications. Caching strategies for frequently processed document types reduce redundant computations.

Economic analysis indicates a favorable return on investment, despite the initial infrastructure costs. Hardware expenses are amortized over 18 months, given the processing volume and efficiency gains. Reduced manual processing labor saves $3.2 million annually for a medium-sized healthcare system. Faster document turnaround improves billing cycles, accelerating revenue by an average of 8 days. Decreased errors reduce compliance risks and associated penalty exposures.

5.2.3. Privacy and Compliance Considerations

Healthcare deployment necessitates stringent privacy protections and regulatory compliance. All processing takes place within a HIPAA-compliant infrastructure, utilizing encryption at rest and in transit. Differential privacy techniques during training prevent the model from memorizing patient information. Audit logging tracks all document access and processing decisions for compliance verification. Regular security assessments identify and address potential vulnerabilities.

Model governance frameworks ensure the responsible deployment of AI in clinical settings. Bias monitoring detects performance disparities across demographic groups. Explainability tools provide decision rationales for clinical validation. Version control maintains model lineage and enables rollback capabilities. Continuous monitoring alerts on performance degradation or distribution shifts requiring model updates.

5.3. Limitations and Future Directions

Current limitations provide opportunities for continued research and development. The dataset's scope, which is restricted to English-language documents from academic

medical centers, may limit its generalizability. Performance on extremely rare document types remains uncertain due to the limited availability of training examples. Real-time processing requirements may pose challenges to deployment in resource-limited settings. Model interpretability, although improved through attention visualization, still requires further enhancement to achieve full clinical trust.

Future research directions aim to address the identified limitations while exploring new capabilities. Multilingual models would extend benefits to diverse patient populations and global healthcare systems. Multimodal architectures incorporating imaging and structured data promise comprehensive document understanding. Federated learning enables multi-institutional training while preserving patient privacy. Continuous learning frameworks adapt to evolving medical knowledge and documentation practices. Integration with clinical decision support systems could directly impact patient care quality through intelligent information routing and summarization.

The demonstrated effectiveness of domain-adapted pre-trained models for medical document processing establishes foundation for widespread healthcare automation. Continued advances in model architectures, training techniques, and deployment strategies will further enhance capabilities and accessibility. Achievement of comprehensive healthcare documentation automation remains attainable through sustained research investment and cross-disciplinary collaboration between medical and computational experts.

References

1. T. Jokioja, H. Moen, and F. Ginter, "Deep learning in medical document classification," *Computer Science*, 2020.
2. N. Zhang, and M. Jankowski, "Hierarchical BERT for medical document understanding," *arXiv preprint arXiv:2204.09600*, 2022.
3. A. K. Rai, U. S. Aswal, S. K. Muthuvel, A. Sankhyan, S. L. Chari, and A. K. Rao, "Clinical text classification in healthcare: Leveraging bert for nlp," In *2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI)*, December, 2023, pp. 1-7. doi: 10.1109/icaiihi57871.2023.10489434
4. T. Asha, S. Soundaryaa, and R. Rajakumari, "Enhancing Patient Welfare using Natural Language Processing (NLP): A Paradigm Shift in Prescription Technology," In *2023 International Conference on Sustainable Communication Networks and Application (ICSCNA)*, November, 2023, pp. 1255-1261.
5. G. Bo, W. Shanshan, Z. Qing, P. Bo, and Z. Yan, "Empowering medical data analysis: an advanced deep fusion model for sorting medicine document," *IEEE Access*, vol. 12, pp. 1650-1659, 2023. doi: 10.1109/access.2023.3347029
6. D. R. CH, "Enhanced Named Entity Recognition in Medical Texts Using Transformer-Based Models," In *2024 2nd International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, December, 2024, pp. 1-5. doi: 10.1109/scopes64467.2024.10990753
7. T. KAMPU, M. NII, E. NAKANISHI, and R. SAKASHITA, "Nursing-Care Text Classification and Extraction of Important Terms for Classification Using Transformers," In *2024 International Conference on Machine Learning and Cybernetics (ICMLC)*, September, 2024, pp. 427-432. doi: 10.1109/icmlc63072.2024.10935220
8. A. Khaliq, A. Khan, S. A. Awan, S. Jan, M. Umair, and M. F. Zuhairi, "Integrating topic-aware heterogeneous graph neural network with transformer model for medical scientific document abstractive summarization," *IEEE Access*, vol. 12, pp. 113855-113866, 2024.
9. A. Joshi, Y. Singh, V. Pareek, I. Sharda, and T. Jain, "Medical Document Classification using NLP Techniques," In *2024 First International Conference on Software, Systems and Information Technology (SSITCON)*, October, 2024, pp. 1-6. doi: 10.1109/ssitcon62437.2024.10795870
10. S. Maleki Varnosfaderani, and M. Forouzanfar, "The role of AI in hospitals and clinics: transforming healthcare in the 21st century," *Bioengineering*, vol. 11, no. 4, p. 337, 2024. doi: 10.3390/bioengineering11040337
11. F. López-Martínez, E. R. Núñez-Valdez, V. García-Díaz, and Z. Bursac, "A case study for a big data and machine learning platform to improve medical decision support in population health management," *Algorithms*, vol. 13, no. 4, p. 102, 2020. doi: 10.3390/a13040102
12. T. S. Heo, Y. Yoo, Y. Park, B. Jo, K. Lee, and K. Kim, "Medical code prediction from discharge summary: Document to sequence bert using sequence attention," In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, December, 2021, pp. 1239-1244.
13. C. A. Seralathan, G. Rupesh, V. M. Tarun Jayant, and R. Kingsy Grace, "Healthscribe: Revolutionizing Medical Documents with Large Language Models," In *2025 International Conference on Next Generation Computing Systems (ICNGCS)*, August, 2025, pp. 1-6.
14. P. Elamparithi, M. Bhavani, and R. Karthick, "BERT and RoBERTa Model Based Approach for Text to Diseases Classification," In *2025 International Conference on Emerging Trends in Industry 4.0 Technologies (ICETI4T)*, June, 2025, pp. 1-6.

15. T. Shen, X. Zhang, D. Lee, M. T. Quasim, and C. Wang, "Intent-based ioT network slicing for smart healthcare systems: A knowledge-driven multi-path resource orchestration framework," *IEEE Internet of Things Journal*, 2025. doi: 10.1109/jiot.2025.3622689