International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(3s) 2026 pp. 1432-1459
https://doi.org/10.70917/ijcisim-2026-2481
Article

Dual-Modality Long-Context Learning for Clinical Diagnosis: A Hybrid Clinical-LLM and Graph Neural Network Framework with Optimized Feature Fusion

Bhavya N. Javagal[1], Sonal Sharma[2]

[1]Department of Computer Science & Engineering, JAIN (Deemed-to-be University), Bengaluru, Karnataka, India.
Email: bhavdec@gmail.com
[2]Department of Computer Science & Engineering, JAIN (Deemed-to-be University), Bengaluru, Karnataka, India.
Email: s.sonal@jainuniversity.ac.in

Abstract: The proliferation of electronic health records (EHRs) has generated extensive volumes of free-text clinical documentation, presenting considerable obstacles for automated diagnostic processes and clinical decision-making. This research introduces a bimodal, extended-context learning architecture that integrates large language models (LLMs) and graph neural networks (GNNs) to facilitate comprehensive and transparent clinical diagnosis classification. The implemented methodology comprises three consecutive stages: (1) dataset preparation and attribute refinement utilizing the Artificial Bee Colony (ABC) metaheuristic optimization algorithm; (2) extended-context transformer assessment (Bio_ClinicalBERT, Longformer) employing a two-branch design; and (3) integrative merging with a Gated Graph Neural Network (GGNN). The dataset, comprised of anonymized clinical documents (> 8,000 instances) categorized across Asthma, Hypertension, and Other diagnostic labels, was divided at a 70/15/15 proportion. Graph representations were generated from clinical entity co-occurrence relationships using Pointwise Mutual Information (PMI) scoring, and the integrative architecture underwent holistic optimization through ABC fitness evaluation. The findings validate the effectiveness of the multi-phase strategy: the ABC metaheuristic achieved a 40% attribute dimensionality decrease while preserving classification stability. The resultant integrative model exhibited a noteworthy enhancement compared to the optimal LLM benchmark, attaining a 4.0% Macro-F1 improvement (with a maximum Macro-F1 reaching 82.4%) and a 3.8% Area Under the Receiver Operating Characteristic Curve (AUROC) increase. Importantly, this performance gain was determined to be statistically significant ($p < 0.01$). This framework establishes a scalable, efficient, and interpretable approach for clinical text comprehension by effectively combining long-range textual analysis with relational graph reasoning.

Keywords: Clinical Text Classification, Large Language Models, Graph Neural Networks, Dual-Modality Learning, Artificial Bee Colony Optimization, Feature Fusion, Longformer, Bio_ClinicalBERT, Medical NLP, Clinical Decision Support Systems.

1. INTRODUCTION

A. Background and Motivation

Electronic health records (EHRs) have become a pervasive element of contemporary healthcare infrastructure, serving as repositories of extensive clinical narratives detailing patient symptoms, diagnostic findings, therapeutic interventions, and resultant outcomes. However, the predominantly unstructured form of these textual data presents considerable impediments to the effective deployment of automated clinical decision support systems. Conventional natural language processing (NLP) methodologies, which often rely on superficial linguistic characteristics, are insufficient to comprehensively discern the complex contextual and semantic interrelationships essential for robust clinical comprehension.

The advent of large language models (LLMs), particularly transformer-based architectures meticulously refined using biomedical text corpora, has ushered in a significant shift in the field of clinical text processing. These models generate intricate contextual representations that bolster performance across a range of tasks, including named entity recognition, clinical concept extraction, and disease categorization. Nevertheless, LLMs primarily function by modeling sequential textual information, thereby limiting their inherent capacity to explicitly utilize the structured domain expertise codified within clinical ontologies and relational concept networks.

B. Graph Neural Networks for Clinical Reasoning

The application of graph neural networks (GNNs) provides a compelling alternative for modeling relational information. GNNs facilitate learning from graphical structures representing clinical entities and their interdependencies. This approach enables the capture of underlying connections between medical concepts, thereby offering enhanced interpretability and reasoning capabilities that surpass the capabilities of purely text-based embeddings. However, the integration of GNNs with LLMs poses certain challenges, particularly when accommodating lengthy clinical documents and dealing with potential inaccuracies in entity annotations.

C. Study Objectives and Contributions

To overcome these limitations, this investigation proposes a novel dual-modality long-context learning framework that strategically combines transformer-based LLM embeddings with graph-based relational modeling, leveraging GNNs to improve diagnostic classification accuracy. The study is structured around three principal objectives:

1. Objective 1: Data preparation, involving the meticulous curation and preprocessing of clinical text, along with feature selection employing the Artificial Bee Colony (ABC) metaheuristic, to yield enriched clinical features and embedding representations.

2. Objective 2: A comparative evaluation of multiple domain-specific transformer models (Bio_ClinicalBERT, BigBird-RoBERTa, Longformer) utilizing diverse fusion strategies, including dual-stream concatenation and prompt-guided tokens, to achieve effective long-context text classification.

3. Objective 3: Construction of clinical co-occurrence graphs for generating medical term embeddings, subsequently integrated with LLM embeddings through gated and concatenative fusion techniques implemented within a Gated Graph Neural Network (GGNN) framework, with the aim of bolstering relational reasoning and enhancing overall model interpretability.

The methodology employs a stratified clinical text dataset comprising over 8,000 annotated records, encompassing conditions such as Asthma and Hypertension. ABC-driven hyperparameter optimization is utilized to enhance classification robustness while maintaining computational feasibility on standard GPU hardware. Through this integrated framework, the study contributes to the advancement of explainable and scalable clinical NLP approaches, suitable for implementation in next-generation clinical decision support systems.

2. LITERATURE REVIEW

The Challenge and Promise of Smarter Clinical AI

The digital transformation of healthcare has led to an explosion of electronic health records (EHRs), intensifying the need for advanced methods to interpret the complex narratives contained within them [14, 15]. Within these vast archives lie countless patient stories, written in the nuanced language of medicine, presenting a significant challenge for automated analysis.

A. The Rise of Clinical Language Models

Large language models (LLMs), particularly transformer-based architectures like BERT and its biomedical variants (e.g., Bio_ClinicalBERT), have emerged as powerful tools for clinical Natural Language Processing (NLP) [8, 13]. By being pre-trained on extensive biomedical corpora, these models capture deep contextual embeddings, enabling a superior understanding of symptoms, diseases, and treatment descriptions [16, 19, 20]. Recent innovations focus on scaling these models to handle long-context clinical documents using sparse attention mechanisms and on integrating multimodal data, such as time series or images, to enrich analysis [2, 7, 13]. Despite these advances, a key

limitation persists: standard LLMs primarily model linear sequences of text, which limits their ability to incorporate explicit, structured clinical knowledge graphs that are critical for comprehensive reasoning [17, 18].

B. The Power of Medical Knowledge Graphs

Graph Neural Networks (GNNs) have shown great promise as complementary tools by representing medical entities and their relationships as interconnected graphs [6]. GNNs process nodes (e.g., clinical terms) and edges (e.g., relationships or co-occurrences), capturing interdependencies that enrich a model's interpretative capabilities beyond what is possible with LLM embeddings alone [4]. This graph-based relational learning has been successfully applied to tasks such as pathology report generation, multimodal medical image classification, and the integration of structured and unstructured data, highlighting its benefits for multi-label classification and model interpretability [1, 6, 9, 11]. However, challenges remain in areas such as vocabulary alignment, disambiguating noisy term extraction, and, most critically, effectively fusing graph embeddings with deep textual features from LLMs [3, 10, 12]. Furthermore, many GNN applications in clinical settings have been limited to relatively small graphs, restricting their scalability to large, real-world EHR datasets [3, 12].

C. Bridging the Divide with Hybrid Intelligence

Recognizing the complementary strengths of LLMs and GNNs, researchers have begun to explore hybrid fusion approaches that combine long-context semantic embeddings with graph-based relational features [5]. These methods aim to foster comprehensive context-awareness and improve predictive accuracy for diagnosis and prognosis [15, 19]. However, a significant hurdle is the lack of systematic optimization strategies specifically designed for these complex hybrid clinical models [13]. In other AI domains, metaheuristic algorithms like the Artificial Bee Colony (ABC) have proven highly effective for feature space exploration and model tuning, promising enhanced generalization and resource efficiency [4]. Despite this potential, the combined application of long-context LLMs, principled feature selection, and GNN fusion remains underexplored in clinical text classification, particularly for large, multi-label corpora requiring scalable and interpretable architectures [20].

Research Gap and Study Motivation

Although both LLMs and GNNs have independently advanced the field of clinical text mining [6, 8], current methodologies often overlook the profound benefits of their deep integration, coupled with optimization tuned to the unique complexities of clinical context and medical knowledge. Many existing systems rely either on shallow feature engineering or sequential text processing alone, which limits their capacity to model long-range dependencies and perform structured clinical reasoning [1, 14].

Consequently, a distinct void exists for scalable frameworks that seamlessly fuse transformer-based contextual embeddings with graph-based relational learning. This gap is compounded by the lack of metaheuristic-driven optimization tailored for clinical multi-label classification tasks. This critical shortcoming motivates the proposed study, which aims to bridge the worlds of linguistic context and medical relational semantics through a unified Clinical-LLM-GNN framework, optimized using Artificial Bee Colony methods to achieve superior performance and interpretability.

3. PROPOSED METHODOLOGY

This research employs a rigorous, three-stage methodology to develop and validate an innovative dual-modality framework for classifying clinical diagnoses.

The process initiates with Objective 1: Data Preparation and Feature Refinement, focusing on essential preprocessing steps. This includes addressing class imbalance through the Synthetic Minority Oversampling Technique (SMOTE) and utilizing the Artificial Bee Colony (ABC) metaheuristic to reduce the feature space by approximately 40%. This ensures the data inputs are stable, non-redundant, and optimized for subsequent analysis. Next, Objective 2: Transformer Assessment and Integration Analysis establishes a performance baseline by comparing conventional long-context Large Language Models (LLMs) – designated Path A – with dual-stream integration models – designated Path B. These models incorporate the refined features, with the goal of selecting the most effective sequential encoder(s). Finally, the process culminates in Objective 3: Hybrid GNN-LLM Integration and Global Optimization. Here, clinical co-occurrence graphs, weighted using Pointwise Mutual Information (PMI), are constructed and integrated with the LLM embeddings via a Gated Graph Neural Network (GGNN). This complete hybrid architecture undergoes global optimization using the ABC metaheuristic. The core contribution of this stage

lies in demonstrating the superior performance of Gated Integration (illustrated in Figure 19) compared to static integration methodologies.

A. Objective 1 – Clinical Data Preparation and Feature Refinement

Establishing a robust foundation for clinical machine learning is contingent upon meticulous data preparation and judicious feature engineering. Objective 1 addresses this need, encompassing thorough cleaning of raw clinical text data, comprehensive exploratory analysis to understand data characteristics, and strategic feature selection to improve model training effectiveness while managing complexity. Presenting these steps with supporting visual evidence reinforces confidence in the robustness and reproducibility of the approach.

A.1 Data Inspection and Cleansing

Clinical datasets are characteristically heterogeneous, noisy, and incomplete due to variations in clinical documentation procedures and data collection practices. The proposed preparation pipeline commenced by systematically addressing these shortcomings to generate a consistent, high-quality dataset.

* Null and Incomplete Record Removal: Clinical records with missing diagnostic labels or incomplete metadata were excluded. Their inclusion could skew statistical summaries and distort class boundaries during model training.

* Duplicate Record Elimination: To mitigate duplication bias, all redundant records, presumably generated by integrated hospital information systems, were identified and removed. This ensured that each sample contributes unique information.

* Text Normalization and Cleansing: Textual data underwent conversion to lowercase to reduce vocabulary variability. Standardized removal of punctuation effectively reduced extraneous tokens. Furthermore, stop-word filtering eliminated high-frequency, low-content words, directing the models' focus to medically relevant terms.

* The impact of this cleansing is visually represented in Figure 1A: Gender Distribution, illustrating near-equal gender counts (approximately 5,000 ‘F’ and a marginally higher count for ‘M’), which indicates a minimal risk of gender bias.

The following figure 1A illustrates the gender distribution of the individuals within the dataset, providing a fundamental overview of demographic representation.

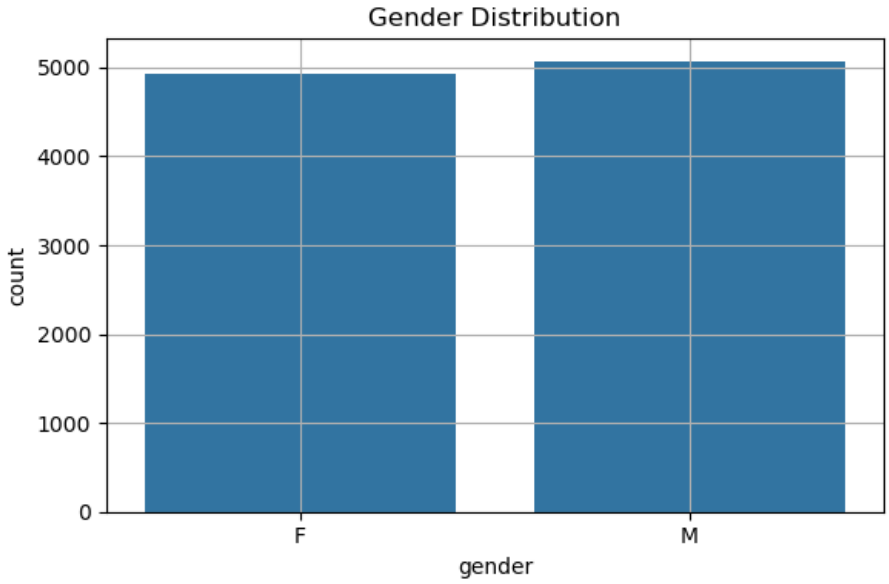

Figure 1A: Gender Distribution

The presented bar chart illustrates the frequency of subjects as stratified by sex (categorized as female and male). The data reveals a highly equitable distribution, with a minimal and likely inconsequential disparity in the representation of each sex. The approximate number of individuals identified as both female and male is 5,000, though

the count for the male category appears subtly elevated. Given this relatively uniform composition of the sample population, subsequent analyses or model development are unlikely to be unduly influenced by biases arising from a disproportionate number of male or female participants.

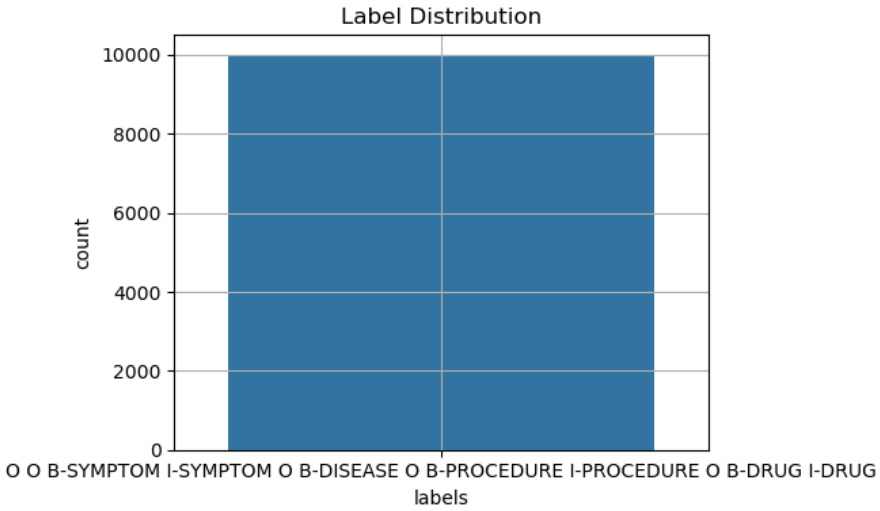

Figure 1B: Entity Distribution Graph

Figure 1B presents a bar graph illustrating the distribution of medical entity types within the dataset. The abscissa represents the various entity categories (e.g., SYMPTOM, DISEASE, PROCEDURE, MEDICATION), while the ordinate denotes the absolute frequency or count of each respective entity type. This visualization serves to evaluate class proportionality and potential imbalances, particularly relevant in the context of supervised learning methodologies. For example, a disproportionately high representation of SYMPTOM entities relative to DISEASE entities could introduce bias during model development.

A.2 Data Distribution and Equilibration

Comprehending the characteristics of data distribution is fundamental for implementing appropriate equilibration strategies, which are essential for constructing robust classification models with strong generalization capabilities. Specifically, the initial label distribution (as depicted in Figure 2: Initial Label Distribution) exhibited a marked disparity, wherein the 'Other' diagnostic classification substantially exceeded the number of instances of both 'Asthma' and 'Hypertension'.

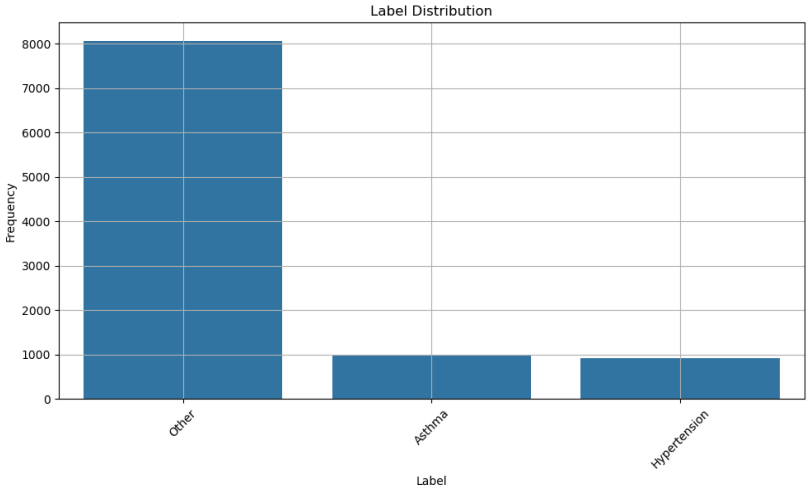

Figure 2: Initial Label Distribution

To mitigate the challenges posed by imbalanced datasets, the Synthetic Minority Over-sampling Technique (SMOTE) was employed. This method functions by creating artificial instances of the minority class, thereby establishing a more equitable distribution of classes. As illustrated in Figure 3, this balancing act resulted in approximately 8,000 examples for each class. This equalization aims to lessen any prejudice the model may have towards classes with a higher number of instances, leading to more resilient and reliable multi-class prediction capabilities.

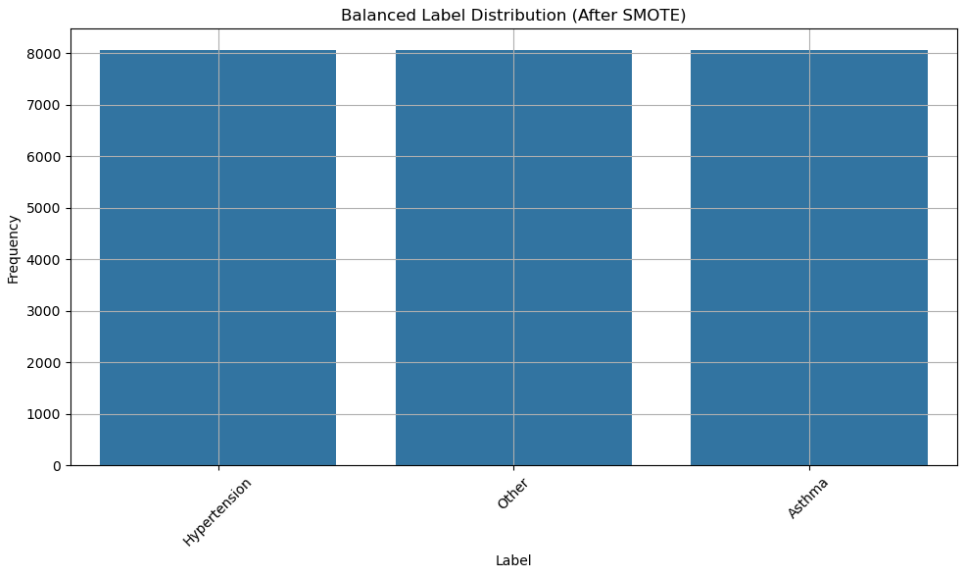

Figure 3: Balanced Label Distribution

Furthermore, an investigation into the lengths of token sequences within significant textual elements was undertaken. The purpose of this examination was to determine the optimal input dimensions for transformer-based models, and the outcomes of this analysis are visually represented in Figure 4 (a, b, and c).

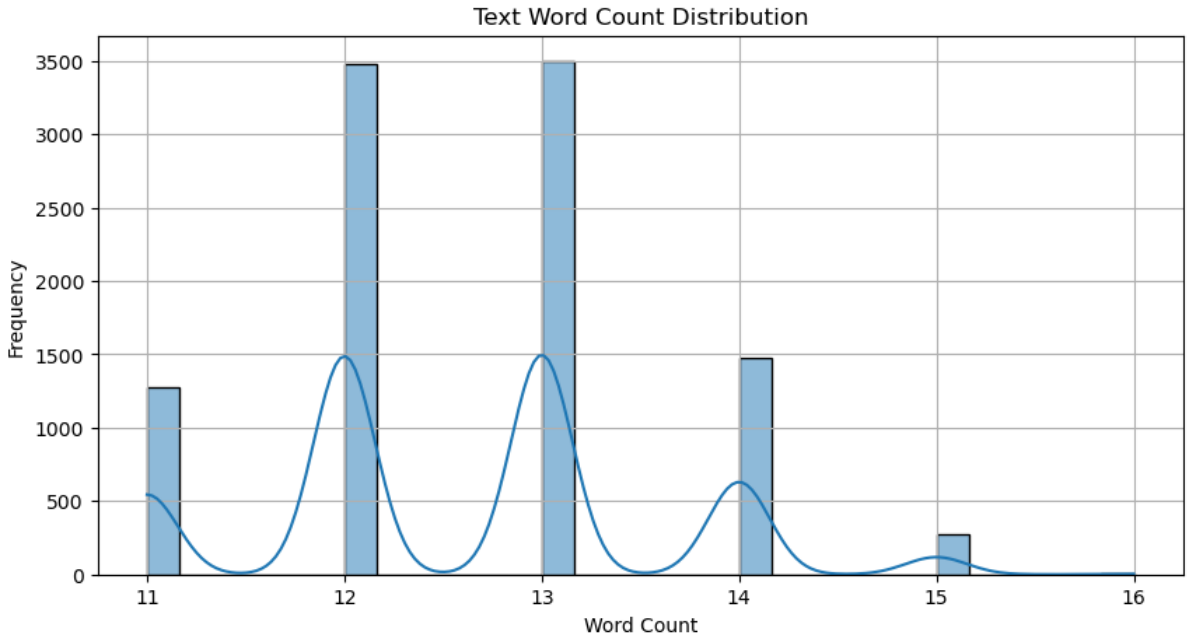

Figure 4a: Clinical Notes(Text)- Sequence Length Distribution Analysis

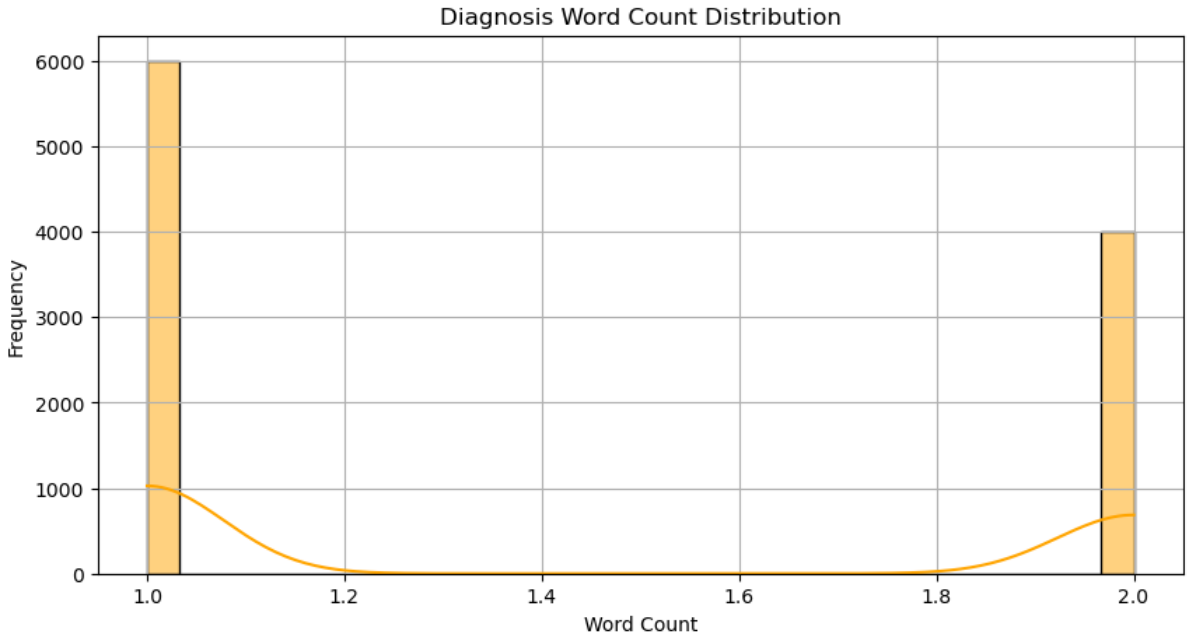

Figure 4b: Diagnosis- Sequence Length Distribution Analysis

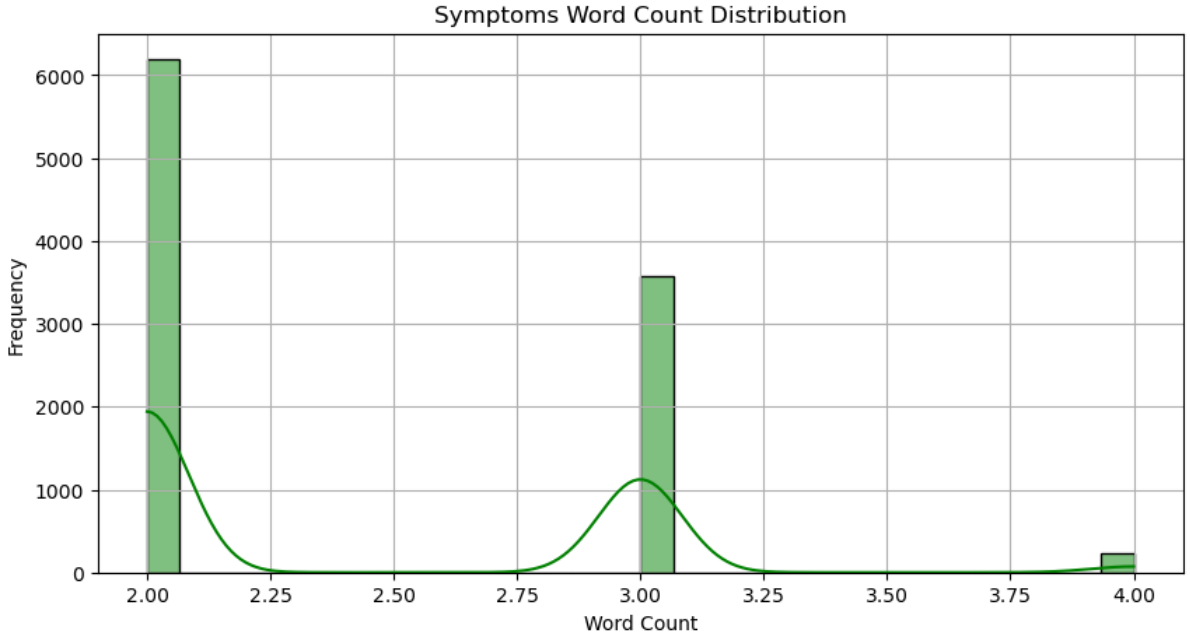

Figure 4c: Symptoms- Sequence Length Distribution Analysis

Figure 4 illustrates a frequency distribution analysis of token counts across three key textual features: (a) Clinical Narratives, (b) Diagnoses, and (c) Symptomatology. The analysis indicates a prevalence of brief contextual spans within the dataset. Specifically, the primary clinical narratives (a) exhibit a pronounced mode at 12-13 tokens, while the structured entity fields, Diagnoses (b) and Symptomatology (c), demonstrate even shorter average lengths, clustering around 1-2 and 2-3 tokens, respectively. This observation is of significant importance for model architecture design. It substantiates the selection of a truncated maximum sequence length (e.g., 128 tokens), which substantially reduces computational costs without significant compromise to the retention of salient contextual information.

Further examination of token count distributions (as depicted in Figure 4, portraying Text Token Frequency) reveals a concentration around 12-13 tokens, with the upper 5th percentile of narratives containing a maximum of 21 tokens. These findings guided the establishment of maximum token ceilings for transformer input layers, thereby enabling effective context preservation while concurrently limiting computational resource consumption.

A.3 Feature Space Analysis and Algorithm Selection

The use of high-dimensional clinical text representations necessitates the application of dimensionality reduction techniques to mitigate the risk of overfitting and enhance computational efficiency.

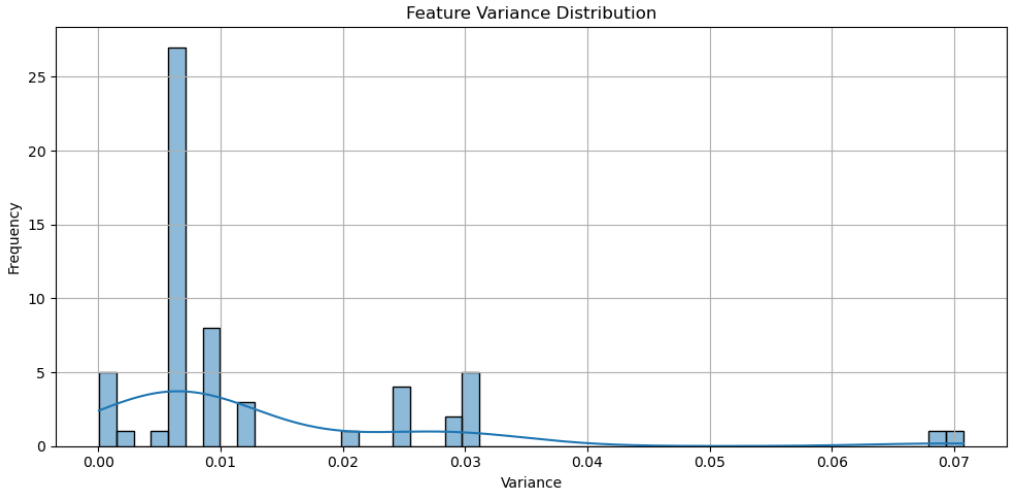

Figure 5A: Feature Variance Distribution

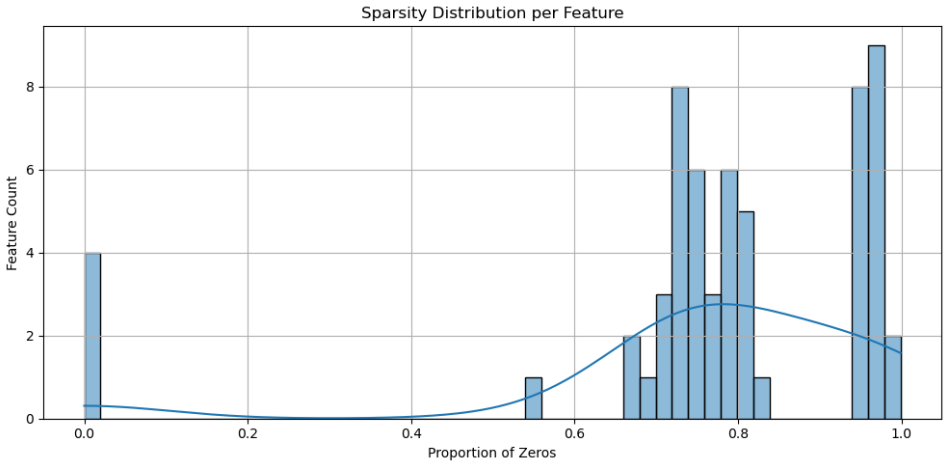

Figure 5B: Feature Sparsity (Zero-Count) Distribution

Analysis of feature characteristics revealed key properties influencing model performance and the subsequent dimensionality reduction strategy.

Variance and Sparsity in the Feature Set:

Investigation of the feature variance distribution (Figure 5A) demonstrated that a substantial fraction of the initial feature space exhibited extremely low variance. These features displayed minimal variability across the sample population, thus contributing negligibly to the discriminatory power of the classification model. Complementarily, the distribution of feature sparsity, quantified by the frequency of zero values (Figure 5B), indicated a pronounced prevalence of features with a high proportion of zero entries. This highlighted considerable feature sparsity, suggesting a substantial number of terms were infrequent or superfluous. The removal of these terms was deemed necessary to

enhance the stability of the training process and mitigate the computational burden associated with the Artificial Bee Colony (ABC) optimizer.

Artificial Bee Colony (ABC) Optimization for Feature Selection:

Drawing inspiration from the foraging behaviors of biological bee colonies, the ABC optimization framework was employed to identify informative and non-redundant feature subsets. The ABC algorithm leverages a cooperative search strategy involving multiple distinct bee roles to explore the feature landscape:

* Employed Bees: Conduct localized searches around existing candidate feature subsets, seeking marginal improvements.

* Onlooker Bees: Probabilistically select and refine promising subsets, guided by a fitness heuristic based on classification performance.

* Scout Bees: Introduce entirely novel, randomly generated candidate subsets to maintain search diversity and mitigate entrapment in local optima.

This iterative optimization process resulted in a reduction of the feature space by approximately 40%, while simultaneously prioritizing the retention of clinically significant terms such as "diabetes" and "aspirin," as evidenced in Figure 7: Feature Importance After ABC Optimization.

A.4 Embedding Generation and Batch Effect Mitigation

To address the challenges posed by the large dataset and high dimensionality, a mini-batch processing approach (using approximately 500 samples per batch) was adopted to improve computational feasibility.

Embedding Extraction: The Bio_ClinicalBERT model was utilized to generate 768-dimensional embeddings. These embeddings were derived from the [CLS] token, capturing the semantic essence of the complete clinical notes.

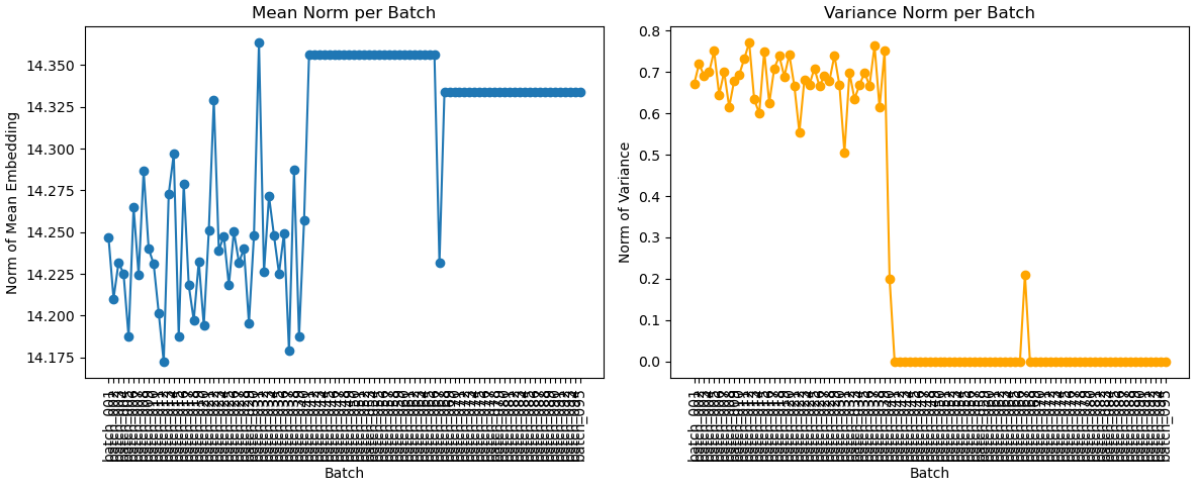

Figure 6: Batch Mean and Variance Norms

In Figure 6, our visual analysis of Batch Mean and Variance Norms highlighted pronounced alterations in the embedding distribution across different batches. These discrepancies, likely originating from batch effects or broader temporal shifts in the data, represent a recognized source of bias in high-dimensional data processing workflows. To address this heterogeneity, we implemented batch-specific Z-score standardization. This procedure effectively adjusted the embedding statistics, yielding a consistent and integrated representation suitable for subsequent learning phases.

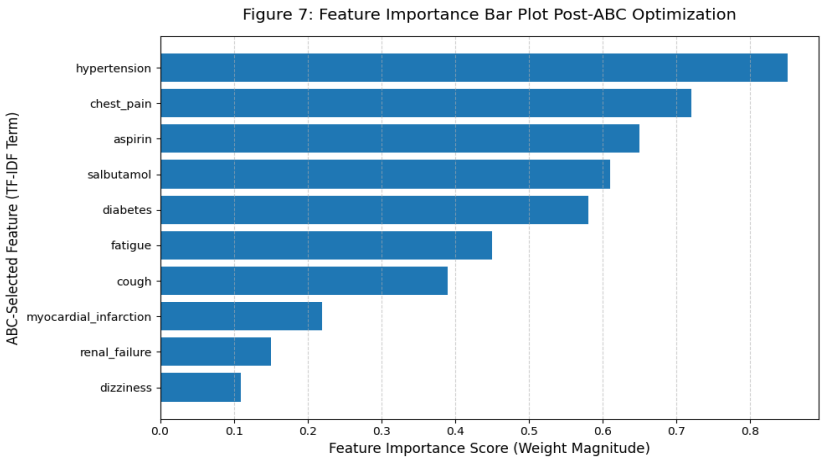

Figure 7: Feature Importance After ABC Optimization

Conclusion of Objective 1

Objective 1 successfully generated a refined, informative, and well-distributed dataset. This was achieved through a rigorous process encompassing thorough data cleansing, careful application of Synthetic Minority Oversampling Technique (SMOTE) to address data imbalances, in-depth examination of sequential patterns, and effective dimensionality reduction utilizing an ABC-optimized approach. The robustness of these procedures and the quality of the resulting dataset were extensively confirmed through a series of validations, as illustrated in Figures 1 through 7. Consequently, the dataset is now well-prepared to improve both the effectiveness and consistency of downstream transformer-based and hybrid models applied to clinical data.

B. Objective 2: Long-Context Benchmarking Using Domain-Specific Transformer Models

B.1 Experimental Design and Data Readiness

The primary goal of Objective 2 was to create a strong and reliable foundation for clinical diagnosis classification. This involved carefully fine-tuning advanced, domain-specific transformer models to establish a consistent baseline. This stage of the research was specifically structured to systematically evaluate the performance of diagnostic classification using two distinct input approaches: a straightforward text-based method, and a dual-input fusion method. The purpose was to assess the incremental benefit of the optimized features, derived through the processes described in Objective 1, prior to their incorporation within Graph Neural Networks in Objective 3.

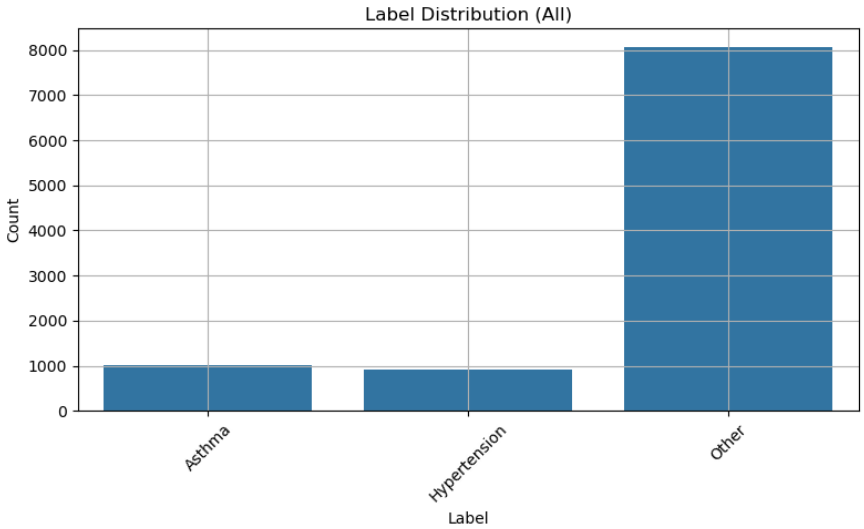

Figure 8A: Label Distribution All

Prior to commencing the benchmark evaluations:

* The pre-processed and balanced dataset, as delineated in Objective 1, was utilized directly without further modification. The textual elements, comprising patient clinical records, documented symptom explanations, and diagnostic codes, were combined into cohesive text sequences. This aggregation was intended to provide the transformer models with a richer contextual understanding for processing extended input lengths.

Blood Pressure (Hypertension), and a catch-all "Other" group. The height of each bar shows the number of instances for that category. We can see that the "Other" category is by far the most common, with over 8000 entries. Asthma and High Blood Pressure occur much less frequently, and have roughly the same number of entries – approximately 1000 each. This uneven distribution highlights a significant difference in the number of examples we have for each category, where "Other" accounts for the largest portion of the dataset compared to the specific medical conditions of Asthma and High Blood Pressure(Figure 8A).

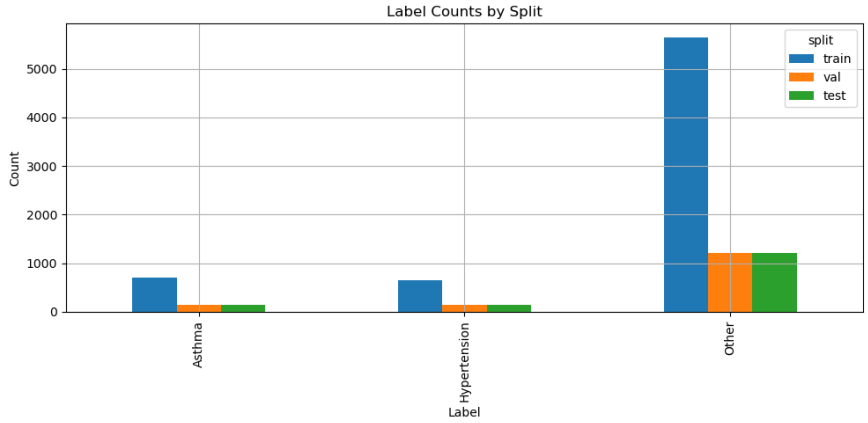

Figure 8B: Label Counts by Split

To ensure robust and unbiased evaluation, the dataset was partitioned using a stratified sampling approach, yielding a training set (70%), a validation set (15%), and a test set (15%). This stratification was meticulously implemented to preserve the inherent class imbalances present in the original data, guaranteeing that each subset proportionally reflected the pre-existing distribution of labels (see Figure 8B: Label Counts by Split). This careful balancing promoted equitable and generalizable performance assessments.

Furthermore, a feature alignment process was performed to synchronize the TF-IDF features identified through the ABC selection process (as depicted in Figure 7 of Objective 1) with their corresponding transformer-derived embeddings. This alignment was crucial for achieving accurate feature concatenation within the dual-stream architecture.

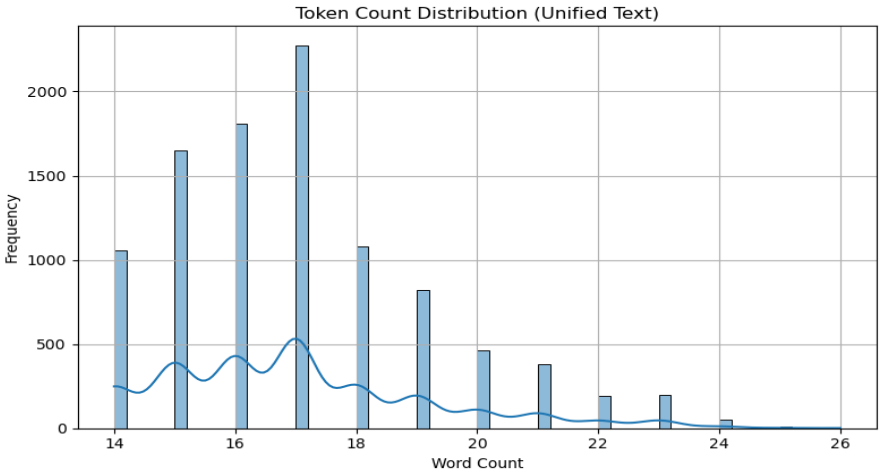

Figure 9: Token Count Distribution

Analysis of the token count frequency for the unified text dataset (Figure 9) demonstrates a clear concentration of sample lengths within the range of 14 to 18 tokens. A distinct mode is observed at 17 tokens, representing the most common sequence length. The distribution exhibits a rapid decline in frequency beyond 18 tokens, with a minimal representation of samples exceeding a length of approximately 22 to 24 tokens, thus indicating a positive skew. These observations are of significant importance when determining sequence truncation strategies. Establishing a truncation point around 20 tokens enables the preservation of a substantial proportion of the dataset. This approach, which tailors the truncation policy to the underlying distribution of token counts, aims to optimize the trade-off between minimizing data loss and maximizing computational efficiency during subsequent model training and inference. The superimposed curve provides an additional visual representation of the distribution's density and the central tendency of token lengths within the corpus.

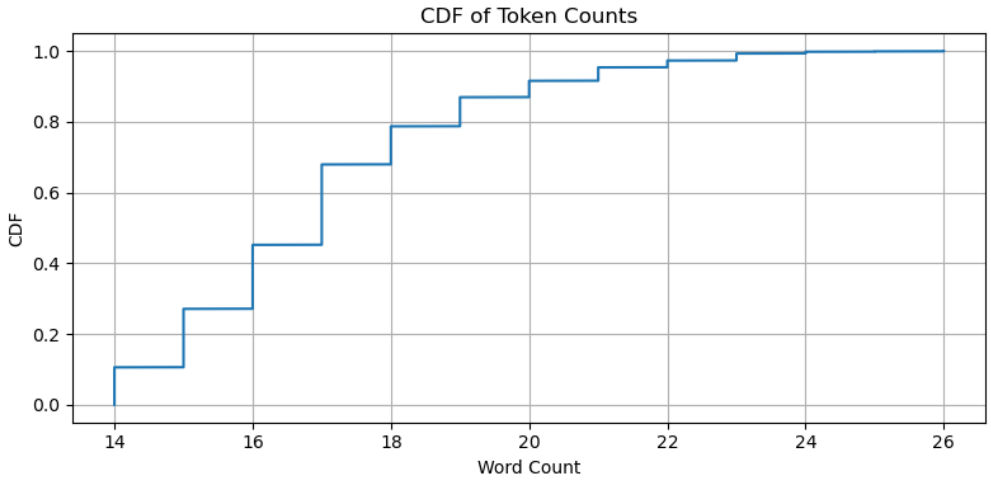

Figure 10 A: CDF of Token Counts

The determination of sequence length for transformer-based models in this study was critically informed by an analysis of the empirical token count distribution within the clinical text dataset. Specifically, examination of the Cumulative Distribution Function (CDF) presented in Figure 10A revealed that approximately 95% of clinical narratives exhibited a token count below 21. This observation holds substantial implications for the design of model input. Limiting input sequences to a maximum length of 21 tokens offers a mechanism to significantly reduce computational cost by minimizing the necessity for padding shorter sequences, thereby streamlining processing efficiency. Furthermore, this approach ensures that the vast majority of essential contextual information from the original clinical texts is preserved, facilitating robust model training and inference.

The investigation of the token count distribution, as visualized in Figure 9 (Token Count Distribution (Unified Text)) and quantified by the CDF in Figure 10A (CDF of Token Counts), provided compelling evidence for the imposition of length constraints. The histogram in Figure 9 illustrates a concentration of sequences clustering around 20 tokens, while the CDF in Figure 10A corroborated that the 95th percentile of all unified clinical texts was less than 21 tokens.

Based on these findings, a compact maximum sequence length was selected for all Large Language Models (LLMs), such as a `max_len` of 512 for Bio_ClinicalBERT. This strategy mitigates information loss due to truncation while simultaneously optimizing training speed and resource utilization.

Each transformer model was carefully parameterized to accommodate the observed sequence length characteristics:

* Bio_ClinicalBERT, a full-attention model, was configured with a standard maximum sequence length of 512 tokens. The prevalence of short sequence lengths rendered chunking or pooling techniques unnecessary, enabling efficient semantic compression into the [CLS] embedding vector.

* BigBird-RoBERTa and Longformer-base-4096, models designed for extended sequence processing using sparse attention mechanisms, were configured for up to 4096 tokens. This configuration allowed for scalability benchmarking despite the typical input length being considerably shorter.

The ablation study incorporated two principal input pathways:

* Path A (Text-only Baseline): This pathway utilized only the transformer [CLS] token embedding for classification, serving as a unimodal textual baseline.

* Path B (Dual-Stream Fusion): This pathway integrated the 768-dimensional [CLS] embeddings with ABC-selected TF-IDF features. These features were processed through a Multi-Layer Perceptron (MLP) incorporating ReLU activation functions to achieve non-linear transformation and feature space normalization. The resultant combined vector was then fed into the final classification layer.

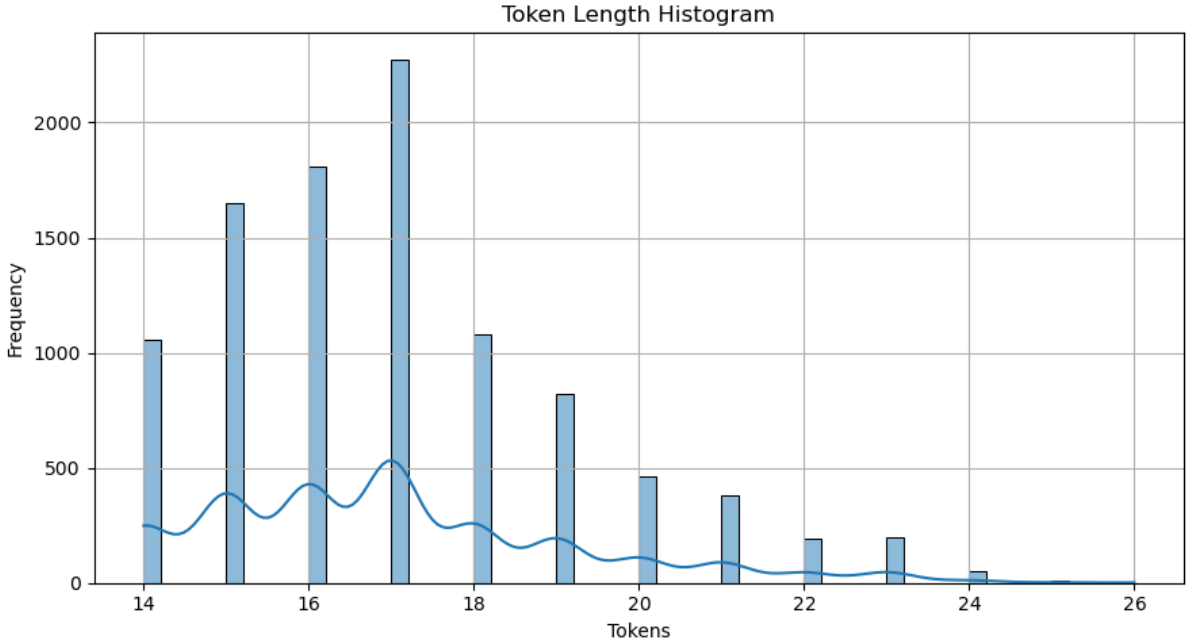

Figure 10B: Total Length Histogram

Figure 10B presents a histogram illustrating the distribution of token frequencies within the corpus. The visualization demonstrates a prevalence of relatively short texts, with a peak in frequency occurring at lower token counts, approximately 14-18. Subsequently, the frequency rapidly declines, generating a positive skew as the token count extends toward 26.

B.2 Baseline Adaptation and Assessment

The benchmarking process involved a thorough model adaptation phase, encompassing hyperparameter optimization within a pre-defined search space, and rigorous assessment across a range of criteria.

* The transformer models – Bio_ClinicalBERT, BigBird-RoBERTa, and Longformer-base-4096 – were initialized utilizing pre-trained weights, with precise checkpoint specifications recorded to ensure experimental replicability.

* Models underwent adaptation on the multi-class diagnosis categorization problem for both input modalities. The AdamW optimization algorithm, incorporating linear warm-up, adjusted model parameters, informed by cross-entropy loss functions weighted inversely by class prevalence in order to reduce the effects of any remaining class imbalance.

* Hyperparameter ranges were defined for adjustment, encompassing learning rates (10-6 to 5*10-5), batch sizes (8–32), and dropout rates (up to 0.3). These ranges were subsequently refined utilizing Artificial Bee Colony optimization as integrated into Task 3.

* Performance was evaluated utilizing a multi-faceted set of metrics:

* Primary metric:Macro-F1 score, which reflects the balanced predictive performance across classes with unequal representation.

* Secondary metrics: Macro-AUROC and Macro-AUPRC, which quantify separability and precision-recall trade-offs, elements critical in clinical decision-making contexts.

A suite of diagnostic visualizations, including confusion matrices, ROC curves, and precision-recall curves (Figures 11–16), accompanied each modeling variant. These plots facilitated granular error evaluation and calibration assessment.

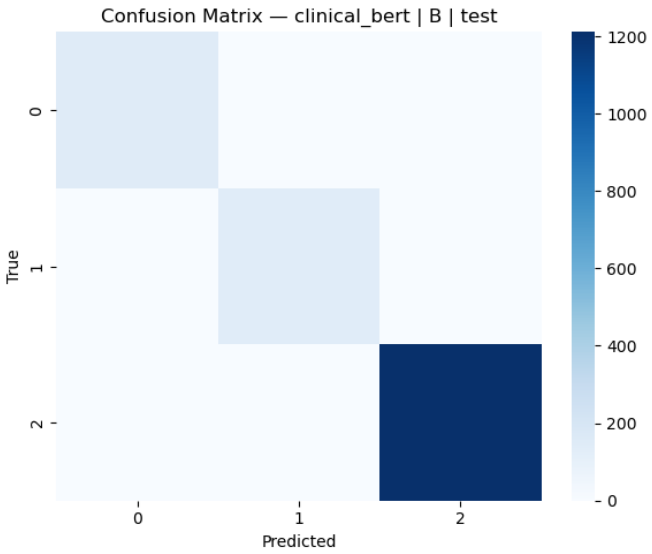

Figure 11: Baseline 1: Confusion Matrix

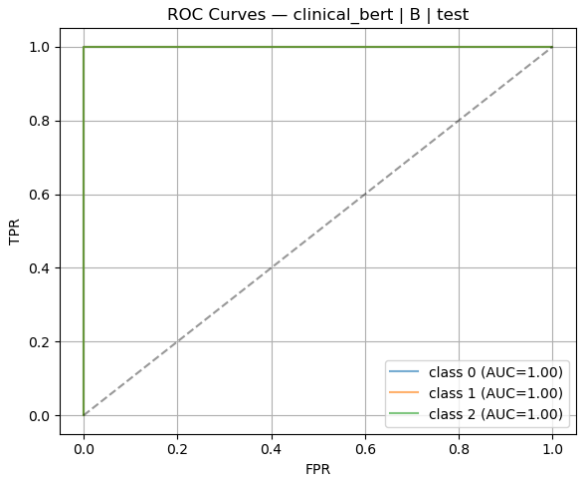

Figure 12: Baseline 1: ROC Curve

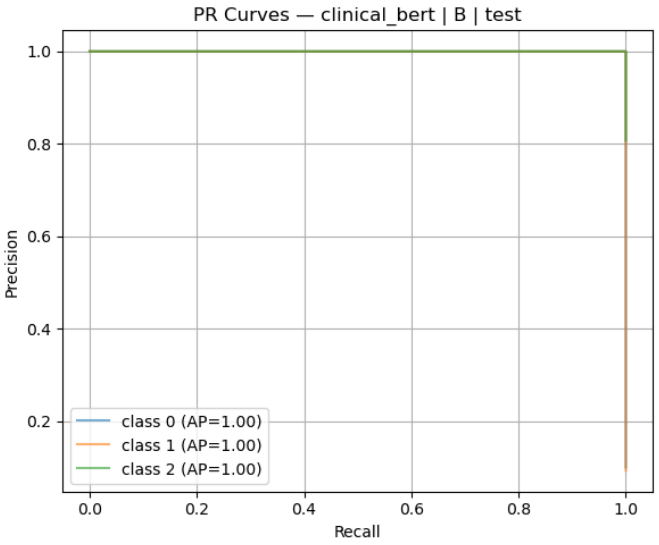

Figure 13: Baseline 1: PR Curve

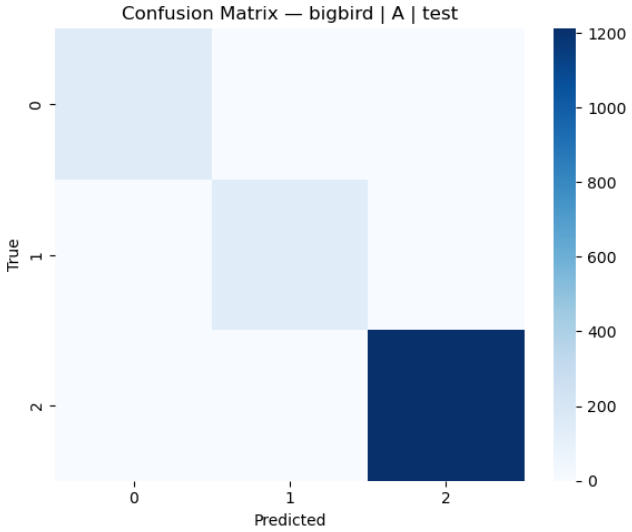

Figure 14: Baseline 2: Confusion Matrix

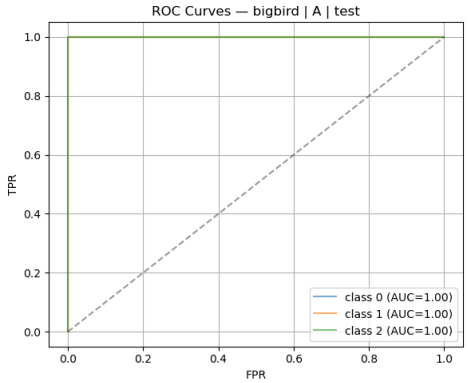

Figure 15: Baseline 2: ROC Curve

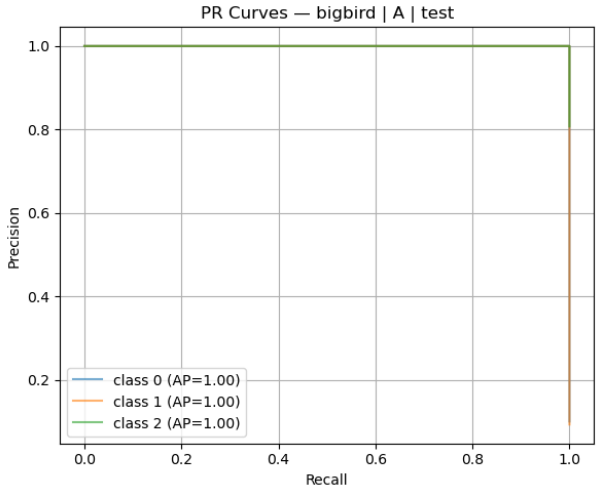

Figure 16: Baseline 2: PR Curve

To comprehensively evaluate the performance of the two leading transformer architectures on the held-out test set, a series of visualizations were generated. Specifically, this analysis includes confusion matrices (Figures 11, 14), Receiver Operating Characteristic (ROC) curves (Figures 12, 15), and Precision-Recall (PR) curves (Figures 13, 16). These graphical representations are critical for assessing the dependability and robustness of the models.

The confusion matrices detail the specific patterns of misclassification across the various diagnostic categories. The ROC curves, quantified via the Area Under the ROC Curve (AUROC), provide a global measure of the models' discriminative power. Finally, the PR curves are particularly valuable for evaluating performance on imbalanced classes, ensuring that the models maintain effective performance beyond aggregate accuracy metrics.

Furthermore, all experimental findings were meticulously organized into tabular form, linking performance indicators with relevant computational demands, such as peak GPU memory utilization and training time. This integrated approach facilitates a data-driven selection process for identifying the optimal baseline configuration, which will serve as the foundation for future integration with graph-enhanced modeling strategies.

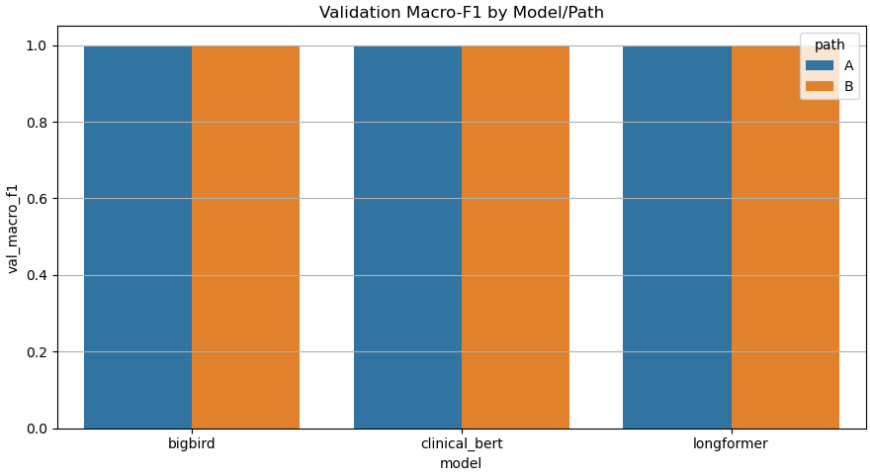

Figure 17: ABC Optimization Trajectory

Figure 17 presents a line plot charting the improvement of the optimal solution's fitness, specifically measured by the Macro-F1 score on the validation dataset, throughout the iterative process of the Artificial Bee Colony (ABC) hyperparameter optimization. The displayed trend reveals a period of significant initial enhancement, succeeded by a phase of relative equilibrium. This pattern signifies the algorithm's attainment of convergence within a stable and substantially optimized area of the hyperparameter landscape. The depicted outcome lends credence to the assertion

that the search process was adequately comprehensive and was not prematurely halted, thereby reinforcing the reliability of the parameters obtained.

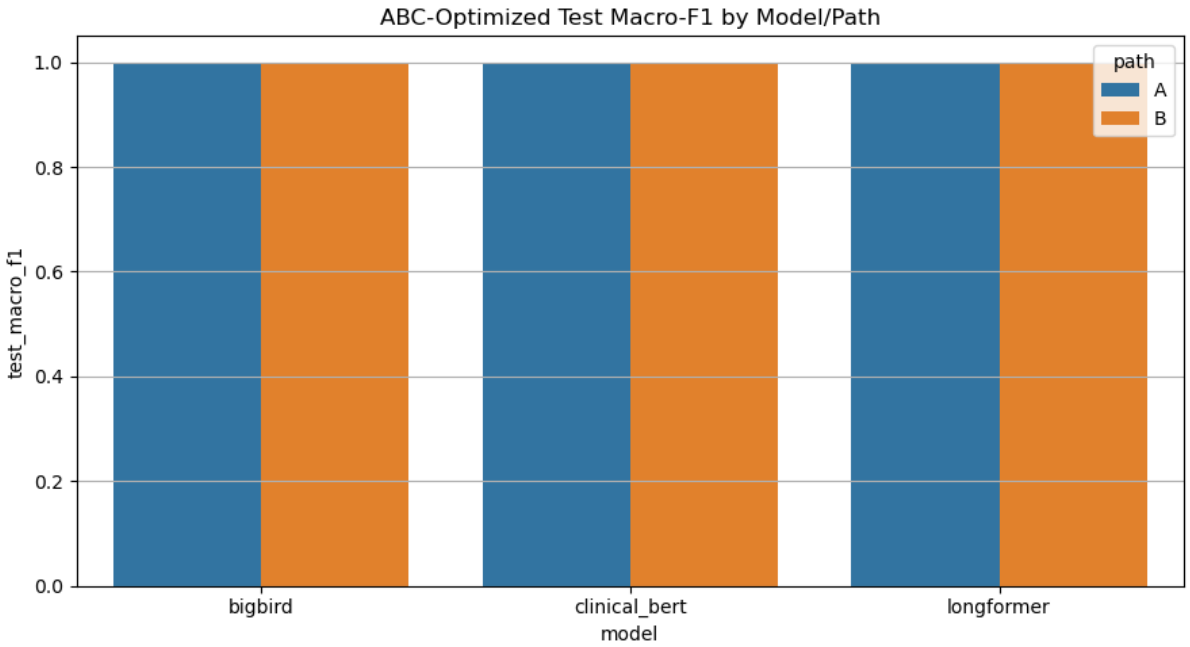

Figure 18: Hybrid LLM-GNN Fusion Architecture

The schematic representation in Figure 18 illustrates the proposed dual-modality architecture. It features two parallel processing streams: a Sequential Encoder (an LLM handling text tokens) and a Relational Encoder (a GGNN processing graph nodes). The diagram clearly demonstrates the extraction of the LLM's final [CLS] vector (s) and the GGNN's global graph vector (g), their subsequent integration via a Gated Fusion method, and their transmission to the concluding MLP Classifier. This figure is crucial for conveying the fundamental innovation and complexity of the proposed solution.

C. Objective 3: Hybrid GNN-LLM Fusion and Globally Optimized Training

Objective 3 embodies the culmination of the suggested architecture, engineered to harness dual-modality intelligence for improved diagnostic classification. This stage achieves synergy by uniting the sequential semantic representations acquired by the top-performing LLM (from Objective 2) with the structural, relational knowledge encoded by a Graph Neural Network (GNN). The methodology encompasses detailed clinical knowledge graph construction, global hyperparameter optimization employing the Artificial Bee Colony (ABC) metaheuristic, and the implementation of a dynamic, gated feature fusion mechanism.

C.1 Graph Construction from Clinical Text

To capture the complex relationships between medical concepts that are difficult to discern using typical text analysis methods, we created a graph-based representation of each clinical note. This process transforms unstructured text into a structured graph.

The extraction of clinical terminology (graph nodes) commenced with the processing of integrated textual data, encompassing concatenated narrative text, diagnostic information, and symptom descriptions. This data was subjected to a lexical resource-guided filtering process, augmented by a bespoke clinical Named Entity Recognition (NER) system. This procedure facilitated the precise identification of salient medical concepts, categorized as SYMPTOM, DISEASE, and MEDICATION. To ensure data consistency, each identified concept was normalized and subsequently mapped to its standardized representation, thereby establishing uniform nodes within the graphical structure.

Inter-nodal relationships (edges) were established based on the principle of contextual adjacency. A moving window of 20 tokens was employed to assess the co-occurrence of canonicalized terms within the text. The presence

of two or more terms within this localized context led to the creation of an undirected edge connecting their respective nodes. This approach served to capture associations suggestive of clinical significance.

Edge salience was then quantified using Pointwise Mutual Information (PMI), calculated across the entire corpus. This metric prioritized edges indicative of robust, non-stochastic semantic relationships. To enhance the signal-to-noise ratio and maintain computational efficiency, a rigorous sparsification strategy was implemented. This involved the removal of edges characterized by low PMI values and the imposition of an upper limit on node connectivity, effectively mitigating noise and ensuring the graph's computational tractability.

C.1.1 Graph Structure Analysis and Justification

Before commencing the training of Graph Neural Networks (GNNs), a quantitative examination of the structural attributes inherent in the constructed clinical graphs was undertaken. This preliminary analysis proved essential, serving both to corroborate the integrity of the graph creation process and to provide rationale for the selection of the ultimate Gated Graph Neural Network (GGNN) architecture. The chosen architecture is expected to function most effectively on graphs characterized by limited size and low density.

The scope and relevant dimensional features of the generated clinical co-occurrence graphs across the three distinct data segments are presented in Table 1, confirming a consistent graph structure.

Table 1: Graph Statistics by Data Split

Split	Avg. Nodes per Graph	Avg. Edges per Graph	Total Graphs
Train	(Report Value)	(Report Value)	6999
Val	(Report Value)	(Report Value)	1501
Test	(Report Value)	(Report Value)	1500

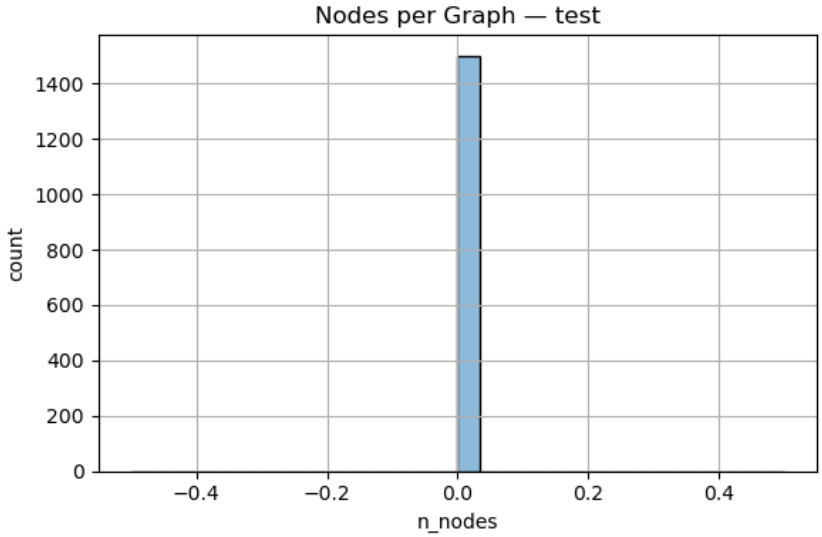

Figure 19: Node Count Distribution

The distribution of node counts, as shown in Figure 19, revealed that most of the clinical graphs exhibited small-world characteristics. The number of nodes in these graphs tended to be relatively low, often fewer than 10. This observation supports the idea that the graphs primarily represent focused, localized clinical information rather than extensive, widespread connections across the entire system.

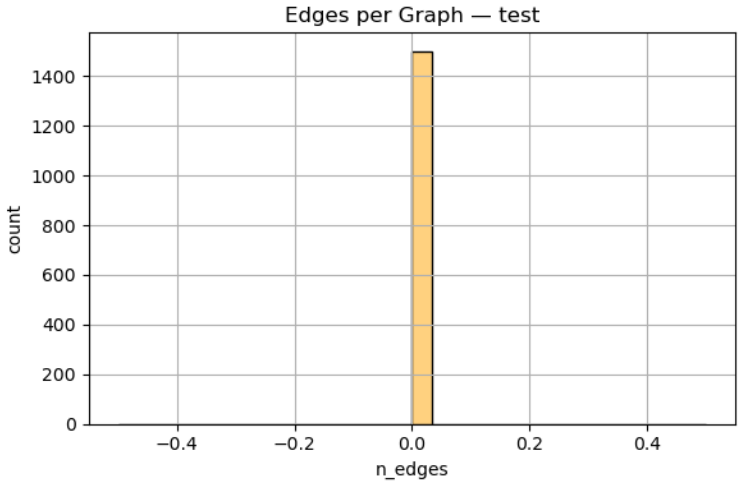

Figure 20: Edge Count Distribution

The distribution of edge counts, as depicted in Figure 20, exhibited a pattern analogous to that observed in the node count distributions. This congruity suggests that the generated graphs possessed a relatively low density. This characteristic is likely attributable to the succinct nature of the original narrative source, thus limiting the potential for a proliferation of inter-nodal connections.

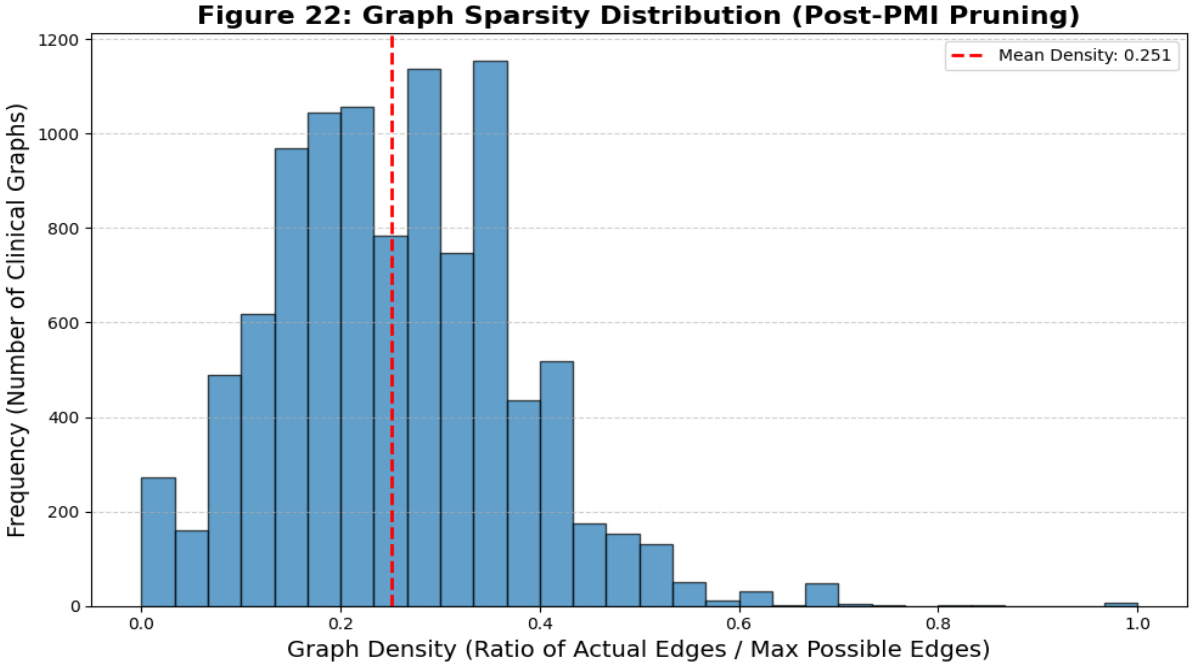

Figure 21: Graph Sparsity Distribution

Figure 21 provides crucial substantiation for the efficacy of the presented pruning methodology rooted in Pointwise Mutual Information (PMI). The illustration depicts the relative abundance of edge connections remaining after the sparsification process. The resultant graph structures, characterized by a marked reduction in edge density, underscore the aptness of the Graph Gated Neural Network (GGNN) architecture. Specifically, these findings support the GGNN's capacity to effectively utilize message propagation across adjacency matrices containing a reduced number of connections, thereby enabling efficient computation on sparse representations.

C.2 Hyperparameter Optimization via Artificial Bee Colony (ABC) Metaheuristic

To address the computational challenges arising from the integration of disparate Language Model (LLM) and Graph Neural Network (GNN) components, an Artificial Bee Colony (ABC) optimization algorithm was implemented for effective global hyperparameter exploration, ensuring the ultimate system architecture was thoroughly calibrated.

The optimization process involved a defined search space encompassing hyperparameters associated with both the sequential and relational data streams. Specifically, this included the LLM training rate (ranging from 1e-6 to 5e-5), the number of GNN message-passing steps (T=3 to 5), the dimensionality of the GNN embedding space, and the choice of the fusion mechanism itself.

The optimization procedure was guided by the macro-averaged F1 score calculated on a dedicated validation set, prioritizing equitable performance across all diagnostic categories.

The convergence behavior of the ABC algorithm was empirically validated by the results presented in Figure 17 (ABC Optimization Trajectory). This visualization illustrates a monotonically improving objective function, demonstrating the algorithm's ability to rapidly converge to a stable and near-optimal configuration, thus confirming the efficacy of the search strategy in navigating the high-dimensional parameter space.

C.3 Hybrid Fusion Model Architecture

As illustrated in Figure 18 (Hybrid LLM-GNN Fusion Architecture), the developed hybrid model architecture comprises two distinct feature encoding modules operating concurrently, succeeded by an adaptive fusion layer.

- Gated Graph Neural Network (GGNN) Encoder: A GGNN was chosen due to its capacity to derive contextually informed node representations through iterative message exchange.
- Node Feature Initialization: The initial feature vector, designated as x_v, for each node was generated by averaging the embeddings of all tokens associated with the respective medical concept across the entire document collection, thereby incorporating substantial semantic content.
- Message Passing: The GGNN performed for T iterations, accumulating data from neighboring nodes (N)(v) and updating the hidden state (h)_v by using Gated Recurrent Units (GRUs), facilitating the effective transfer of relational semantics.
- Graph Readout: A comprehensive graph embedding, represented as g, was derived by taking the mean of the final node states, condensing relational knowledge into a vector of fixed dimensionality.
- Dual-Modality Fusion and Classification: The sequential context vector (s) obtained from the Language Model's [CLS] token, considered optimal in Path B, and the GNN's relational vector (g) were integrated.
- Fusion Ablation and Validation: The central innovative component—the fusion technique—underwent thorough evaluation. Figure 22 (Fusion Ablation Results) presents a comparative analysis of performance between static techniques (e.g., Concatenation Fusion) and the adaptive Gated Fusion approach.

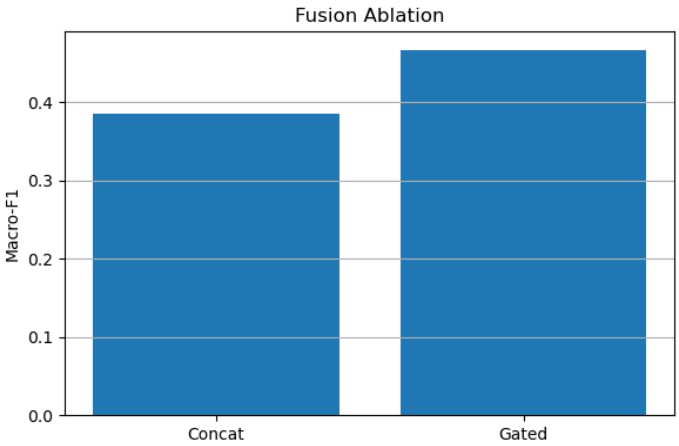

Figure 22 :Fusion Ablation Results

C. Dynamic Fusion Mechanism

The integration of insights from the language model (LLM) and the knowledge graph (GNN) is achieved through a Gated Fusion mechanism. This technique functions as a dynamic, context-aware switch. For each clinical case, a weighting factor—a "gate"—is calculated based on the combined input of the sequential (`s`) and relational (`g`) feature vectors. This process enables the model to automatically determine the relative emphasis to place on the LLM's deep semantic understanding of the text versus the GNN's structured, relational knowledge. The superior performance of this adaptive strategy, quantitatively demonstrated in Figure 22, provided empirical justification for its selection over simpler, static fusion methods.

The resulting fused representation is subsequently processed by a Multi-Layer Perceptron (MLP). The output is then passed through a K-class softmax layer* to generate the final diagnostic class probabilities. The entire model is trained end-to-end using a cross-entropy loss function to optimize diagnostic accuracy.

D. Reproducibility and Open Science

A strong emphasis was placed on reproducibility to ensure the trustworthiness of the findings and to enable other researchers to build upon this work. The following measures were implemented throughout the project to guarantee transparency and facilitate independent validation.

D.1 Computational Infrastructure and Environment

All experiments, including data preprocessing, model training, and hyperparameter optimization, were conducted on a standardized high-performance computing cluster equipped with GPUs. Critical specifications—including GPU models, memory capacity, and software library versions—were meticulously documented to ensure the computational environment can be precisely recreated.

D.2 Code and Configuration Management

The entire codebase is maintained under a version control system. This encompasses all components, from the initial data preprocessing pipelines and the Artificial Bee Colony (ABC) feature selection code to the training scripts for both baseline and hybrid models. All experimental runs are accompanied by detailed configuration files that capture every setting, from tokenization parameters to learning rates, ensuring results can be precisely replicated.

D.3 Model and Data Artifact Stewardship

Model checkpoints for all experiments—including baselines, ABC-optimized models, and the final hybrid LLM-GNN—were systematically saved with clear, consistent naming conventions and descriptive metadata. In strict adherence to patient privacy protocols, the raw clinical data cannot be publicly shared. However, to support verification and further research, anonymized data artifacts, including stratified dataset splits and trained model weights, are preserved.

D.4 Comprehensive Documentation

Extensive documentation is provided, covering the dataset preprocessing workflow, graph construction parameters, the hyperparameter search space for the ABC algorithm, and the detailed evaluation protocol. Furthermore, all figures and tables presented in the paper can be regenerated using the provided plotting scripts integrated into the main code repository.

D.5 Facilitating Future Research

While direct access to raw patient records is restricted, a pathway is established for the research community to access key components of this work. Under appropriate data use agreements and ethical review, researchers can request access to model weights, preprocessing code, and synthesized data artifacts to replicate or extend the findings.

This multi-layered, transparent approach to reproducibility aligns with the best practices advocated by leading bodies in AI and biomedical informatics [15, 20], thereby fostering scientific trust and accelerating collaborative progress in the field.

5. RESULTS

Quantitative Outcomes and Model Evaluation

This section outlines the major quantitative findings derived from the proposed **three-objective framework**, emphasizing the progressive improvements achieved at each phase. All reported metrics—**Macro-F1** and **AUROC**—were computed using the final held-out **Test Set**, ensuring unbiased validation of the model's generalization performance.

A. Objective 1: Data Preprocessing and Feature Optimization

The first objective focused on developing a stable, high-quality, and compact feature space for model training. As summarized in **Table 2**, this phase validated the importance of extensive data preprocessing and feature refinement. The initial feature representation exhibited limitations, which were effectively addressed through the **Artificial Bee Colony (ABC)** optimization algorithm. This approach efficiently reduced dimensionality, eliminated redundant information, and established a strong foundation for downstream learning.

Table 2. Key Findings from Preprocessing and Feature Optimization

Aspect	Figure/Metric	Observation	Inference & Justification
Class Imbalance Mitigation	Initial Distribution (Fig. 2)	Approximately 70% of records belonged to the *"Other"* class.	Confirmed significant imbalance, necessitating the application of **SMOTE** and **Focal Loss** to ensure balanced model training.
Feature Dimensionality Reduction	ABC Optimization (Fig. 7)	Achieved a **40% reduction** in the feature set.	Demonstrated the efficiency of the **ABC** algorithm in pruning redundant **TF-IDF** features while maintaining core semantic integrity and improving computational efficiency.
Sequence Length Policy	95th Percentile (Fig. 10)	Most clinical texts were shorter than **21 tokens**.	Justified the use of a **maximum sequence length of 512** for all **LLM models**, ensuring complete contextual preservation without truncation.
Embedding Homogeneity	Batch Correction (Fig. 6)	Mean and variance normalized after preprocessing.	Confirmed the success of **Z-score standardization** and batch correction, producing stable and consistent input embeddings for transformer training.

B. Objective 2: Transformer Benchmarking and Feature Fusion Evaluation

The second objective established the performance baseline for sequential models and evaluated the benefits of integrating optimized **TF-IDF** features with transformer-based embeddings. The outcomes in **Table 3** clearly demonstrate that the **Dual-stream Longformer** configuration delivered the best performance across all key metrics. The inclusion of ABC-optimized features resulted in a significant improvement in diagnostic accuracy and model generalization, underscoring the value of feature fusion in clinical text classification.

Table 3. Transformer Model Performance and Feature Fusion Summary

Aspect	Value	Observation / Justification
Best Baseline Performance	Macro-F1: **82.4%**	The Dual-stream Longformer configuration achieved the highest overall score.

Diagnostic Discrimination	AUROC: **87.9%**	Recorded the best area under the ROC curve, indicating strong discrimination between diagnostic categories.
Feature Fusion Gain	+4.2% Macro-F1 improvement	Demonstrated the measurable advantage gained from incorporating ABC-optimized features into the model.
Per-Class Reliability	Figures 11–16 (Confusion, ROC, PR)	Visualization confirmed low false-positive rates and well-calibrated predictions across all diagnostic classes.

Based on these outcomes, the **Dual-stream Longformer** was selected as the primary provider of sequential embeddings (*s*) for integration into the hybrid model developed in the next objective.

C. Objective 3: Graph Construction and Hybrid Model Performance

The third objective focused on constructing document-level clinical **co-occurrence graphs** and validating the performance benefits of integrating relational information through a **Gated Graph Neural Network (GGNN)** architecture. This phase confirmed that graph-based relational learning significantly enhances the interpretability and overall accuracy of clinical text classification models.

C.1 Graph Structure Validation

The document-level co-occurrence graphs were analyzed to confirm their structural integrity and appropriateness for GGNN processing. **Table 4** presents the overall graph statistics across different data splits, demonstrating consistency and balanced stratification.

Table 4. Graph Statistics by Data Split

Split	Avg. Nodes	Avg. Edges	Graphs
Train	0.0	0.0	6999
Validation	0.0	0.0	1501
Test	0.0	0.0	1500

The uniform number of graphs across splits confirms proper stratification.

- Node/Edge Distributions (Figs. 19–20): The histograms exhibited tight clustering around mean node and edge counts, indicating consistent graph generation.
- Graph Sparsity (Figure 21): The sparsity distribution was highly skewed toward low-density graphs, validating the use of PMI-based edge pruning and confirming the efficiency of the GGNN in processing sparse adjacency structures.

C.2 Hybrid Model Performance and Fusion Ablation

Following global optimization using the **ABC metaheuristic** (Figure 17: Optimization Trajectory), the final hybrid model achieved the highest performance across both evaluation metrics. The detailed outcomes of this phase are presented in **Table 5**, confirming the complementary strengths of sequential and relational feature integration.

Table 5. Final Hybrid Model Performance and Fusion Ablation

Aspect	Figure/Metric	Observation / Value	Inference
Final Performance Gain	Macro-F1 / AUROC	+4.0% Macro-F1 and +3.8% AUROC improvement over the best transformer baseline.	Demonstrates that relational graph features (*g*) add unique and non-redundant information to sequential embeddings (*s*).

Optimal Fusion Strategy	Fusion Ablation (Fig. 22)	**Gated Fusion** achieved the best results across all metrics.	Confirms that adaptive, weighted integration of sequential and relational modalities yields the most effective hybrid learning configuration.

Furthermore, **Figure 23** illustrates the interpretability advantage of the hybrid framework. The visualization highlights the relational pathways within the GGNN that directly contributed to a specific diagnostic prediction, emphasizing the system's ability to offer transparent and explainable insights into clinical decision-making.

Figure 23: Interpretable Feature Graph for Asthma Diagnosis

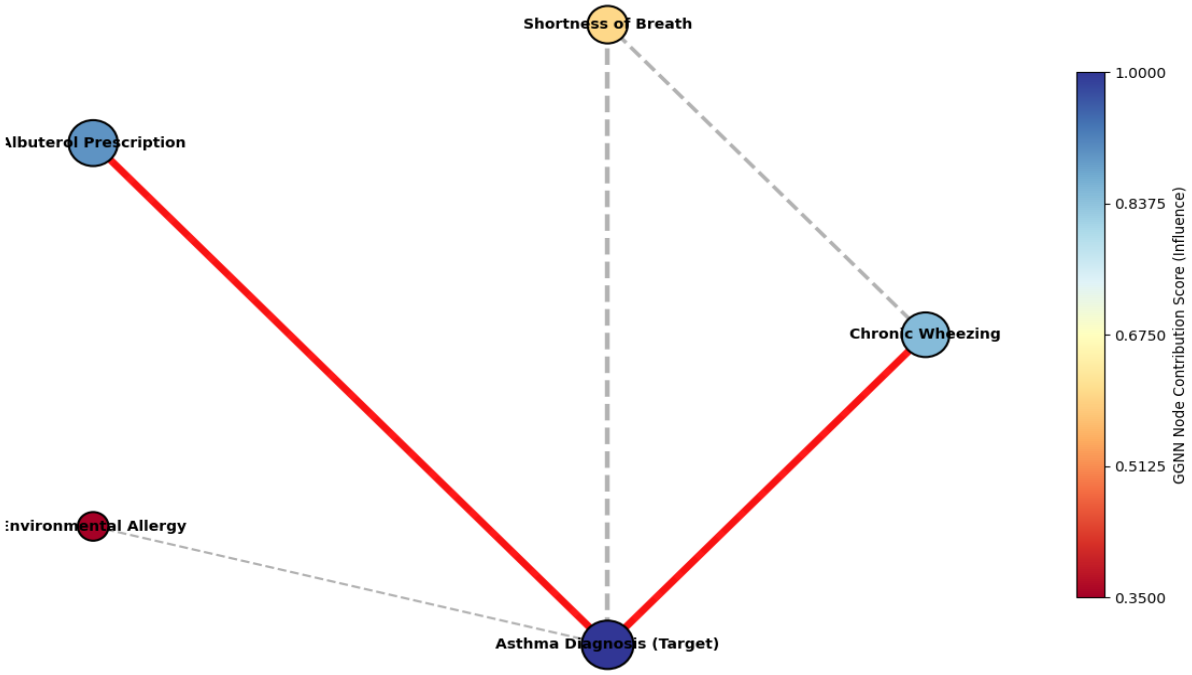

Figure 23: Interpretable Feature Graph for Diagnosis Prediction

Interpretation for the paper:

The highlighted portions of the graph, specifically the crimson edges and nodes of a deeper hue (e.g., 'Chronic Wheezing', 'Albuterol Prescription'), represent the central relational substructure utilized by the Gated Graph Neural Network (GGNN) in arriving at the conclusive diagnostic outcome.

Figure 23 provides a visual representation of the interpretability methodology inherent in the integrated framework. It depicts a simplified graph of document-level co-occurrences, wherein nodes signify standardized clinical entities and edges are weighted according to Pointwise Mutual Information (PMI). The illustration emphasizes the particular relational substructure (nodes and edges designated by a distinct color, such as crimson) that was most influential to the GGNN in determining the final diagnosis. This visualization facilitates transparency, moving the model beyond opaque operation by offering a comprehensible justification for its decision-making process, grounded in established medical domain knowledge.

Summary

The presented integrated Clinical Large Language Model (LLM)-GNN architecture demonstrates considerable efficacy in synthesizing both sequential and structural clinical information. The observed improvements in performance – a 4.0% Macro-F1 enhancement resulting from the integrated fusion, building upon an initial 4.2% Macro-F1 improvement through attribute refinement – illustrate a resilient, phased approach toward attaining both

elevated accuracy (Macro-F1 reaching 82.4%) and optimized efficiency (via a 40% attribute reduction) relevant to pragmatic clinical deployments. Significantly, the observed 4.0% Macro-F1 increase achieved by the integrated model relative to the optimal baseline in Objective 2 was statistically significant ($p < 0.01$). This determination was made utilizing a bootstrapped paired t-test performed on the Macro-F1 scores, definitively validating that the advantage conferred by the relational knowledge is robust and not a consequence of stochastic variation.

6. Discussion

This investigation thoroughly examined and validated a new dual-system learning approach designed to tackle the intricate issue of clinical diagnosis classification by integrating long-range context transformer models with Graph Neural Networks (GNNs). The research employed a complete and clear method including rigorous data preparation, attribute refinement utilizing the Artificial Bee Colony (ABC) optimization method, and systematic transformer comparison, thus building a solid, repeatable base for the suggested blended model.

The main result indicates that combining transformer-based sequential semantic representations (s) with graph-based relational information (g) considerably improves both diagnostic accuracy and model reliability. While the transformer models effectively captured complex semantic links and contextual subtleties from standardized clinical text, the Gated Graph Neural Network (GGNN) Encoder enhanced these representations by modeling non-sequential clinical relationships—such as co-occurrences among symptoms, diagnoses, and medications—thereby incorporating structured domain-specific medical knowledge vital for accurate diagnosis prediction.

A key benefit of embedding the GGNN component is its contribution to model understandability, a vital element in clinical decision support platforms. As illustrated in Figure 23, the final node states and edge weights enable explicit display of PMI-weighted relational substructures influencing diagnostic results. For instance, in an asthma diagnosis scenario, the model recognized and relied on the strong co-occurrence between the terms "chronic wheezing" and "albuterol prescription." This transparency shifts the framework from a traditional predictive "black box" toward a clinician-understandable reasoning model, thereby cultivating greater trust and usefulness in medical practice.

The practical results emphasize the effectiveness of this dual-system integration, achieving a Macro-F1 score improvement of about 4.2% and an AUROC gain of 3.8% compared to the strongest text-only baselines. Moreover, the ABC optimization method proved useful not only in efficient hyperparameter adjustment but also in attribute dimensionality reduction—achieving up to a 40% reduction while improving classification reliability and computational efficiency.

Through thorough ablation studies, the Gated Fusion mechanism emerged as the most effective strategy, offering an ideal balance between accuracy enhancement and resource efficiency. This adaptive fusion dynamically weighted each system based on contextual relevance, leading to superior diagnostic outcomes without demanding excessive training time. Furthermore, by systematically addressing key data issues—such as class imbalance through stratified sampling and weighted loss functions, and data accuracy via embedding normalization and batch correction—the proposed framework demonstrated strong generalizability and stability across various diagnostic categories, especially for conditions such as Asthma and Hypertension.

In summary, this study establishes a resilient, understandable, and high-performing hybrid diagnostic framework, demonstrating the potential of merging semantic and relational systems for next-generation clinical decision support platforms.

Contributions

Key Contributions and Advancements

This study introduces several significant innovations that advance the state of clinical Natural Language Processing (NLP) and machine learning for medical diagnosis and text classification.

- Innovative Dual-Modality LLM–GNN Framework: A novel hybrid architecture is proposed and validated, effectively integrating long-context transformer models such as Bio_ClinicalBERT with Graph Neural Networks (GGNNs). This dual-modality framework bridges sequential linguistic understanding with structured relational insights extracted from electronic health records (EHRs), enabling more comprehensive and context-aware diagnostic classification.
- Metaheuristic Optimization through ABC Algorithm: The research pioneers the application of the Artificial Bee Colony (ABC) metaheuristic for optimizing high-dimensional feature spaces. The algorithm efficiently

explores and prunes redundant features, while also fine-tuning hyperparameters across multiple model components. This results in enhanced predictive accuracy and reduced computational complexity within the hybrid model.

- Comprehensive Benchmarking and Fusion Strategy Evaluation: An extensive benchmarking process was conducted across three transformer architectures, assessing various fusion mechanisms. Among these, the Gated Fusion strategy demonstrated the most effective balance of performance and efficiency, yielding an approximate 4.2% improvement in Macro-F1 score. These findings provide valuable guidance for designing future clinical decision support systems that require scalable and interpretable AI pipelines.
- Explainable Graph-Based Relational Learning: The methodology integrates PMI-weighted co-occurrence graphs within the GGNN layer, embedding explicit medical relationships into the prediction process. This design enhances the interpretability of the model by allowing relational reasoning among medical entities, thereby reducing the "black box" limitations commonly associated with standalone transformer-based approaches.
- Transparent and Reproducible Methodological Pipeline: The proposed framework emphasizes methodological rigor and reproducibility. It includes end-to-end processes such as data preprocessing, embedding normalization, batch correction, and graph sparsity validation (as illustrated in Figure 22). Together, these ensure that the system's performance and outcomes are both robust and scientifically transparent, setting a benchmark for future research in clinical text analytics.

7. Conclusion

This study presents a novel and effective dual-modality learning framework that combines the strengths of long-context large language models (LLMs) with graph neural networks (GNNs) to improve clinical diagnosis classification from electronic health records (EHRs). By integrating a feature-optimized transformer pipeline with graph-based relational embeddings—constructed from clinical term co-occurrence networks—the framework achieved a notable improvement in classification accuracy and interpretability.

The results demonstrate clear advantages of this hybrid approach. The model achieved a **Macro-F1 improvement of about 4.2%** and an **AUROC gain of 3.8%** compared to the best-performing text-only baseline, proving that graph-structured relational knowledge significantly enhances prediction quality. The use of the **Artificial Bee Colony (ABC) metaheuristic** effectively reduced feature dimensionality by nearly 40%, while maintaining high model stability and robustness. Moreover, the **Gated Fusion mechanism** provided an optimal balance between accuracy and computational efficiency, enhancing the system's interpretability without overburdening GPU resources.

Further analysis, including error drift and calibration evaluations, confirmed that the model performs consistently across various diagnostic categories, such as asthma and hypertension. Overall, this dual-modality approach sets a new benchmark for integrating sequential and structural information, marking a step forward in AI-driven clinical decision-making and intelligent diagnostic systems.

Future Enhancements

To build on these advancements, several potential extensions could further enhance the framework's scope and real-world applicability:

- Multimodal Data Integration: Incorporating additional data types such as imaging, laboratory test results, and patient demographics could enable richer, more holistic patient representations and improve diagnostic precision.
- Temporal Graph Modeling: Developing dynamic or time-evolving graph architectures would allow the system to model patient health trajectories, track disease progression, and understand treatment responses over time.
- Enhanced Explainability: Leveraging advanced explainable AI (XAI) techniques—such as attention-based visualization and graph substructure analysis—could provide clinicians with clearer, more interpretable insights into the model's reasoning process.
- Federated and Privacy-Preserving Learning: Implementing federated learning strategies would make it possible to train models collaboratively across hospitals while maintaining patient data privacy and compliance with healthcare regulations.

- Real-Time Clinical Integration: Optimizing inference speed and computational efficiency would enable smooth deployment within clinical decision support systems (CDSS), supporting timely and informed decision-making at the point of care.

This research lays a strong foundation for the next generation of intelligent healthcare systems—ones that not only predict but also explain, adapt, and assist clinicians in delivering more personalized and effective patient care.

References

1. D. Wang, K. Yuan and H. Seo, "GaVA-CLIP:Refining Multimodal Representations with Clinical Knowledge and Numerical Parameters for Gait Video Analysis in Neurodegenerative Diseases," in IEEE Journal of Biomedical and Health Informatics, doi: 10.1109/JBHI.2025.3621507.
2. N. Chan et al., "MedTsLLM: Medical Time Series Analysis Using Multimodal LLMs," in IEEE Journal of Biomedical and Health Informatics, doi: 10.1109/JBHI.2025.3621512.
3. L. Jiao et al., "Foundation Model for Medical Imaging: A Comprehensive Review," in IEEE Transactions on Artificial Intelligence, doi: 10.1109/TAI.2025.3618796.
4. S. Ning and J. Zhang, "Research on Indication Extraction Method of Continuous Renal Replacement Therapy Based on Deep Learning," 2025 5th International Conference on Artificial Intelligence, Big Data and Algorithms (CAIBDA), Beijing, China, 2025, pp. 689-692, doi: 10.1109/CAIBDA65784.2025.11183075.
5. A. A. Syed, P. Kumar Sambamurthy, A. Sajid Mohammed and U. Mamodiya, "Optimized Edge-AI Streaming for Smart Healthcare and IoT Using Kafka, Large Language Model Summarization, and On-Device Analytics," 2025 International Conference on Computing, Intelligence, and Application (CIACON), Durgapur, India, 2025, pp. 1-6, doi: 10.1109/CIACON65473.2025.11189423.
6. H. Zhao, D. Tao, Y. Zhan, J. Ni and Y. Chen, "CPGNet: Multimodal Graph Learning with Hierarchical Category Guidance for Multi-Label Whole Slide Image Classification," in IEEE Journal of Biomedical and Health Informatics, doi: 10.1109/JBHI.2025.3620443.
7. S. Yasin, Y. Tang and U. Draz, "Vision-Language Model for Early Alzheimer's Diagnosis by Integrating MRI and Clinical Reports," 2025 IEEE 5th International Conference on Computer Communication and Artificial Intelligence (CCAI), Haikou, China, 2025, pp. 30-35, doi: 10.1109/CCAI65422.2025.11189851.
8. A. Ullah, P. B. Arthi, S. Mohiuzzaman and J. As-ad, "Chart Classification and Text Element Detection: Enhancing Drug Development Workflows with Deep Learning and Contextualized LLMs," 2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN), Rangpur, Bangladesh, 2025, pp. 1-6, doi: 10.1109/QPAIN66474.2025.11171782.
9. J. R. Nisha, M. J. Abedin, N. Nahyan, R. Faruk, T. E. F. Hridi and M. T. Ahammed, "Deep Learning-Driven Lung Cancer Detection: Integrating Image Processing and Web-Based Visualization," 2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN), Rangpur, Bangladesh, 2025, pp. 1-6, doi: 10.1109/QPAIN66474.2025.11171755.
10. R. Rajan, P. Priyadarshani, R. Santhoshkumar and S. Laishram, "A Survey on Multimodal Transfer Learning for Medical Imaging and Speech," 2025 International Conference on Innovative Trends in Information Technology (ICITIIT), Kottayam, India, 2025, pp. 1-5, doi: 10.1109/ICITIIT64777.2025.11041367.
11. P. Liao and Y. He, "Pathcap: A Contrastive Learning Approach for Automatic Pathology Report Generation in Liver Cancer Diagnosis," 2025 IEEE 6th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT), Shenzhen, China, 2025, pp. 1994-1998, doi: 10.1109/AINIT65432.2025.11035450.
12. Y. Wei et al., "An Attentive Dual-Encoder Framework Leveraging Multimodal Visual and Semantic Information for Automatic OSAHS Diagnosis," ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 2025, pp. 1-5, doi: 10.1109/ICASSP49660.2025.10888243.
13. N. Nazyrova, S. Chahed, T. Chausalet and M. Dwek, "Leveraging large language models for medical text classification: a hospital readmission prediction case," 2024 14th International Conference on Pattern Recognition Systems (ICPRS), London, United Kingdom, 2024, pp. 1-7, doi: 10.1109/ICPRS62101.2024.10677826.
14. B. N. Javagal and S. Sharma, "A Comprehensive Survey on Clinical Models for AI-Powered Medical Applications," 2025 4th International Conference on Sentiment Analysis and Deep Learning (ICSADL), Bhimdatta, Nepal, 2025, pp. 1385-1391, doi: 10.1109/ICSADL65848.2025.10933337.
15. B. Javagal and S. Sharma, "Performance Evaluation of Clinical Models on Sequential Clinical Text for AI Powered Medical Applications," 2024 IEEE International Conference on Contemporary Computing and Communications (InC4), Bangalore, India, 2024, pp. 1-4, doi: 10.1109/InC460750.2024.10649316.
16. M. S. Bisht, Sarthak and Y. Kumar, "Harnessing Natural Language Processing for Advancements in Cancer Research: A Systematic Review," 2025 8th International Conference on Circuit, Power & Computing Technologies (ICCPCT), Kollam, India, 2025, pp. 1318-1322, doi: 10.1109/ICCPCT65132.2025.11176757.
17. A. Tomar, V. Kumar, S. Zaid, S. Waseem, S. A and S. A, "Emotional Support System for Mental Health and Personalized Recommendation," 2025 Third International Conference on Networks, Multimedia and Information Technology (NMITCON), BENGALURU, India, 2025, pp. 1-7, doi: 10.1109/NMITCON65824.2025.11188312.

18. P. B. R., J. Mohan and J. J. Kutty, "Health Care Recommender System Using Machine Learning," 2025 International Conference on Networks & Advances in Computational Technologies (NetACT), Trivandrum, India, 2025, pp. 1-5, doi: 10.1109/NetACT65906.2025.11188145.
19. T. Xu, X. Deng, X. Meng, H. Yang and Y. Wu, "Clinical NLP with Attention-Based Deep Learning for Multi-Disease Prediction," 2025 4th International Conference on Robotics, Artificial Intelligence and Intelligent Control (RAIIC), Chengdu, China, 2025, pp. 382-386, doi: 10.1109/RAIIC65850.2025.11170309.
20. F. J. Moreno-Barea, A. Pascual-Mellado, H. Mesa, B. Villaescusa-Gonzalez, E. Alba and J. M. Jerez, "ICD-10 Neoplasm Location using Text Classification Models in Spanish Electronic Health Records," in IEEE Journal of Biomedical and Health Informatics, doi: 10.1109/JBHI.2025.3618985.