JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 10 Issue 01
2025

Review Article
https://doi.org/10.56979/1001/2025/1175

UMLS-Augmented PubMedBERT for Clinical Note Diagnosis Classification: A Comparative Study

Hamayun Muhammad Saqib[1*], and Ghulam Mustafa[1]

[1]University of Central Punjab, Lahore, 54000, Pakistan.
[*]Corresponding Author: Hamayun Muhammad Saqib. Email: L1f23mscs0005@ucp.edu.pk

Received: September 16, 2025 **Accepted:** November 30, 2025

Abstract: Transformer-based language models have already demonstrated a good performance in the domain of biomedical text classification, but, the majority of the existing methodology uses unstructured textual data only, thereby, evident deficiency of explicit incorporation of curated biomedical knowledge. The present research evaluates the ways in which varying schemes of integration of Unified Medical Language System (UMLS) concepts can be applied to enhance PubMedBERT functioning in terms of the task to classify clinical note diagnosis. Three model configurations are compared with the publicly available PMC-Patients corpus, (i) a baseline PubMedBERT model that is trained just using the raw clinical notes, (ii) a PubMedBERT model that is trained using the raw clinical notes, but additionally augmented with unfiltered UMLS Concept Unique Identifiers (CUIs), and (iii) a PubMedBERT model that is trained using the raw clinical notes, but further augmented with NER-filtered UMLS concepts. They are trained using the weighted cross-entropy loss to address the issue of class imbalance and evaluated using the assistance of accuracy, macro and weighted F1-scores, per-class models, multiple-seed experiments, and bootstrap confidence intervals as well as paired samples t-tests. The findings imply that naive injection of the complete system of UMLS concepts has a negative impact on the performance that implies that injection of ontologies by unedits injects undesired noise. Active selection of concepts filtered by NER to the UMLS, on the other hand, leads to a gradual, but steady increase in performance in concept classification, with an accuracy of 98.8 percent and a massive improvement in minority classes, such as asthma and heart disease. The latter demonstrate that biomedical knowledge may be handy in enriching performance of transformer-based clinical text classification as it is presented in an intentional and systematic way, and because the enhancement of the performance is accompanied by the increase in computational costs.

Keywords: Clinical NLP; PubMedBERT; UMLS; Transformer Models; Clinical Decision Support; Biomedical Text Classification

1. Introduction

Clinical documentation is a large body of unstructured textual data, including case reports, discharge summaries and diagnostic narratives. The differentiation of terminology, style of expression and completeness of the documents often prevents effective extraction of information and thus the need to have automated methodologies that allow clinical text to be organized and categorized systematically. With the ever-growing use of electronic health records, natural language processing has become a central element of clinical informatics and decision-support studies, which can provide strong computational processes that can be used to transform seemingly unrelated textual information into useful knowledge.

The recent trends in transformer-based language models and large language models (LLMs) have led to the emergence of considerable improvements in healthcare text categorization. A comprehensive

systematic review published in 2025 examined the rapidly growing application of large language models (LLMs) for text classification tasks in healthcare [1].Their review shows that, although modern LLMs demonstrate high contextual language understanding, the majority of the methods use only the superficial text patterns and do not utilize any organized sources of biomedical knowledge like the Unified Medical Language System (UMLS). This means that medical reasoning at the concept level is limited especially when dealing with rare cases or vague terms. The review thus points to the incorporation of structured medical semantics as an open problem and a way forward direction to research. Based on these findings, the current study seeks to determine whether the use of UMLS concepts can lead to the improvement of diagnostic classification of the domain-specific biomedical language model PubMedBERT on clinical notes. Instead of assuming that external knowledge is equally beneficial, we do an empirical comparison of the impact of varying degrees of UMLS integration on model performance by comparing a base PubMedBERT to an unfiltered CUI -enhanced variant and an alternative selective strategy that incorporates only NER-filtered biomedical entities.

Natural Language Processing (NLP) is also now a ubiquitous feature of contemporary healthcare informatics (in particular in the area of clinical decision-making support systems (CDSS). With the exponential growth of unstructured clinical text in electronic health records (EHRs), using NLP to sift, organize and interpret biomedical information has become a transformative trend [2][3]. CDSS systems gain greatly from biomedical NLP, for automatic diagnosis suggestions, evidence-based recommendations, and real-time risk stratification. General-purpose large language models (LLM) of automated clinical coding became the subject of modern study in 2025 [5]. The researchers tested ChatGPT-3.5 and ChatGPT-4 on ICD-10 classification with the use of the MIMIC-IV database and discharge summaries, with highly inconsistent results. ChatGPT-4 scored a mere 22 per cent on challenging cases that neither the traditional machine-learning algorithms nor the mapping with SNOMED-based managed to classify correctly. Their study highlights two key limitations of existing LLM-only techniques:

1. Lack of domain -grounding of the biomedical ontologies
2. Not providing consistent performance across runs, especially when using ChatGPT -3.5.

These results indicate the importance of specific models and integration of structured knowledge, as opposed to a zero-shot prompting.

In a direct contrast to the above-mentioned trend, our current study is built on a transformer-based clinical encoder (PubMedBERT) and enhances its input with well-structured UMLS concepts. By incorporating raw and filtered biomedical entities, our method addresses shortcomings identified by Mustafa et al [5].

The electronic health records (EHRs) have spurred the generation of large bodies of unstructured clinical documents, including discharge notes, progress notes, and diagnostic reports. They represent coded clinical data; however, free-text nature and use of complicated medical terminology makes traditional computational methods insufficient [1]. This has thus led to natural language processing (NLP) becoming a crucial aspect of clinical informatics allowing automated derivation and analysis of clinically relevant data within textual data [2]. The initial clinical NLP systems are mainly based on rule-based solutions or traditional machine learning methods, requiring large-scale manual feature engineering and showing little generalizability to heterogeneous clinical settings [3]. These inadequacies led to the development of deep learning structures that are able to learn contextual representations using raw data.

Transformer architecture also significantly transformed NLP, allowing models to model long-range interactions through self-attention. Transformer-based models, such as BERT, demonstrated a higher level of results in various NLP tasks, such as text classification and information extraction [4]. However, general-purpose language models do not often perform well with clinical text due to domain mismatch [5]. To counter this obstacle, Transformer models domain-specific to themselves have been designed. BioBERT was trained on biomedical literature and ClinicalBERT was further fined with clinical notes derived out of EHRs [6] [7]. PubMedBERT followed this course of action by being trained on purely PubMed abstracts and full-text biomedical articles and thus it excludes vocabulary and distributional mismatches [8]. Comparative studies are always able to prove that PubMedBERT outperforms general-purpose and

domain-adapted BERT variants on biomedical NLP problems especially in medical named entity recognition and clinical document classification [8][9]. According to the recent experimental analyses, PubMedBERT maintains strong behavior even in few-shot learning regimes highlighting its appropriateness to clinical practice in the real world when annotated data are limited [9].

Medical named entity recognition (NER) is a fundamental goal of clinical NLP, and its goal is to recognize entities, which may be diseases, symptoms, drugs, and procedures, in unstructured text. NER models based on transformer have significantly advanced over the traditional machine learning and rule-based methods [10]. Extensive biomedical NER reviews outline enduring issues, such as nested and discontinuous entities, inconsistencies in annotation as well as semantic ambiguity [11]. Though contemporary NER systems are remarkably precise and recollect, the resulting entities are often not normalized or anything grounded in context, thus making them of limited use in the downstream clinical setting [12]. Recent work highlights the fact that precise entity extraction is not enough to support clinical decision-making and that extracted mentions need to be semantically matched and connected to structured representations so that they can be used to reason and interpret them [13].

UMLS and Biomedical Knowledge Integration. Unified Medical Language System (UMLS) is a ubiquitously used biomedical ontology which combines several controlled vocabularies and provides standard identifiers of concepts of medical terminologies [14]. UMLS has been widely used to standardize extracted clinical entities and make interoperability across the healthcare infrastructures possible. A number of studies indicate that the integration of the UMLS knowledge into clinical NLP pipelines can be used to improve semantic consistency and reduce lexical variability [15]. However, integration of UMLS concepts without evaluation may create noise that can be described by synonymy, polysemy and context insensitive mappings [16]. This problem is even more acute in the case of multi-class clinical classification problems, where irrelevant and ambiguous concepts can degrade the performance of models. The recent research highlights the importance of discriminative and situation-specific UMLS integration approaches, such as screening concepts by entity recognition confidence or semantic type logarithms, to reduce noise without losing clinically relevant information [16][17].

Knowledge Graphs and Structured Clinical Representations Knowledge graphs and structured clinical representations. It has been suggested that knowledge graphs are useful tools to coordinate extracted biomedical objects into structured forms. Knowledge graphs allow making clinical knowledge more interpretable and reusable by connecting entities in terms of semantic relationships [18]. The fact that large-scale biomedical knowledge graphs built on scientific literature and clinical text demonstrate that unstructured medical information can be organized at scale is not new [19]. However, the majority of methods of knowledge graph construction that are currently in existence focus on representation more than they are on integration with predictive clinical models, hence limiting their direct use in decision-support activities.

Knowledge-Based Models and Clinical Decision Support. Clinical decision support systems (CDSS) require accurate, interpretable and contextual based information to help clinicians in making diagnostic and therapeutic decisions. Recent findings suggest that language models that are based only on textural patterns on the surface are subject to unstable performance and low interpretability in clinical decision-making [20]. Retrieval-augmented generation (RAG) is among the knowledge-based methods that have been suggested in order to increase factual trustworthiness by anchoring model outputs on external knowledge stores [5]. Systematic reviews have indicated that RAG enhances consistency in facts but the effectiveness of the method in clinical practices is highly contingent on the quality and framework of the information recollected [20]. These results support the need to ensure the organization of clinical data in a structured manner before the decision support.

An overview of the available literature demonstrates that there are multiple gaps. First, most researchers discuss separate NLP tasks (NER or classification) individually, instead of suggesting unified systems of clinical data organization. Second, despite high quality performance of PubMedBERT in clinical text analysis, its combination with organized biomedical knowledge is under investigated. Third, current UMLS-based methods often do not have a sophisticated filtering system to control semantic noise. Such

constraints are driving the design of a framework combining domain specific language model and limited biomedical knowledge integration to support sound clinical decision making.

The main aim of this study is to improve the biomedical text classification, with the help of integrating exogenous biomedical knowledge into the transformer-based language models.

Objectives

- To evaluate the baseline performance of PubMedBERT on the PMC-Patients dataset.
- To add clinical text with raw UMLS concepts and assess the resultant effect on categorization.
- To deploy a Named Entity Recognition (NER) module that enables finer integration of UMLS.
- To compare how the baseline and the raw UMLS and NER filtered UMLS-enhanced models perform.

2. Materials and Methods

This study suggests an elaborate, systematic method of the thorough analysis of the impact of pretraining in domains and biomedical knowledge support on clinical decision-support systems. The approach is based on pre-processing the dataset with rigor, the UMLS based knowledge enhancement, fine-tuning of transformers, and a multifaceted evaluation plan. Mathematical formulations are also used in guiding the computation of loss and in gauging of performance.

2.1. Dataset Description and Bias Analysis

The experiments in this work are carried out with the PMC-Patients Dataset for Clinical Decision Support pmc_patients_dataset, a publicly available collection of de-identified clinical patient notes based on case reports in PubMed Central (PMC). The dataset is published under the CC-BY 4.0 license, enabling free use and sharing with attribution. It is accessible via Kaggle.

2.2. Dataset Size and Structure

The data is an aggregate of 168,034 clinical notes, with a primary diagnosis assigned to each note. In this study six top-level diagnostic categories were implemented: Diabetes, Hypertension, Asthma, Heart Disease, Cancer, and Other. The specific profile is included in

Table **1**

Table 1. Class distribution in the PMC-Patients dataset.

Class	Samples	Percentage (%)
Diabetes	16,408	9.76
Hypertension	17,307	10.30
Asthma	2,043	1.22
Heart Disease	5,727	3.41
Cancer	38,335	22.81
Other	88,214	52.50

2.3. Visualizing Class Imbalance

Class distribution of the PMC-Patients dataset. The category of Other includes more than half of the data, and there are many underrepresented conditions like Asthma or Heart Disease, which may influence the accuracy of the classification.

Sample Clinical Notes

The content heterogeneity is illustrated with the representative passages of each category below

1. **Diabetes:** "We report a case of a 45-year-old woman, a non-smoker, treated for type II diabetes under insulin and primary hyperparathyroidism..."
2. **Hypertension:** "A 24-year-old female was initially referred for hypertension, microhematuria, and sub-nephrotic proteinuria of 2.8 g/d..."
3. **Asthma:** "A 61-year-old male, known asthmatic, presented to the pulmonology clinic with complaints of weight loss for two months..."
4. **Heart Disease:** "The patient was a 62-year-old male with a history of liver cirrhosis and post-stent coronary artery disease..."

5. **Cancer:** "A 61-year-old male diagnosed as esophageal cancer received radical resection and esophagogastric anastomosis..."
6. **Other:** "This 60-year-old male was hospitalized due to moderate ARDS from COVID-19 with symptoms of fever, dry cough, and dyspnea..."

2.4. Length Statistics

Table 2 summarizes the token length statistics per class. Cancer-related notes are longest on average—reflecting multi-faceted case descriptions—and asthma cases are quite short.

Table 2. Average token length statistics per class.

Class	Mean	Std. Dev.	Max
Diabetes	219.4	152.7	1,204
Hypertension	225.6	149.2	1,180
Asthma	192.7	127.4	986
Heart Disease	240.3	165.8	1,352
Cancer	262.8	170.4	1,415
Other	201.1	140.5	1,233

2.5. Bias and Limitations

The data may have bias in the following:

1. **Class Imbalance:** It is much skewed around the other, which is an overrepresented category, with classes of Asthma and Heart Disease being underrepresented.
2. **Source Bias:** The sources are mostly based on published case reports and those are more likely to concentrate on unusual or extreme manifestations, thus creating selection bias.
3. **Demographic Bias:** The report of patient demographic information, including age, gender, and ethnicity, is inconsistent and, therefore, does not allow performing a thorough assessment of fairness.
4. **Length Disparity:** Longer cancer observations may give preference to models and bias the predictions of the Cancer category with longer inputs.

2.6. Fairness and Bias Considerations

The bias can be seen as a limitation. The problem of this study is that the dataset does not consistently record demographic details such as age, gender, or ethnicity. Therefore, we could not analyze how well the models have been shown to generalize from certain patient groups to others. This is an important issue since when some groups are underrepresented in the training data, the predictive models may introduce and even enhance existing biases. As an example, manifestations of disease by specific sexes or ethnicities can be misclassified more often, which leads to unequal performance of the model. Datasets should therefore be included in future work that have richer demographic metadata and fairness-informed evaluation frameworks to allow clinical decision support systems to serve all populations justly.

2.7. Data Preprocessing

A standardized pipeline was applied to ensure consistency:

- **Normalization:** The text was made lower-case, punctuation marks were eliminated, and all unnecessary whitespace was eliminated without compromising on specific medical terminology.
- **Tokenization:** HuggingFace PubMedBERT tokenizer was used to match preprocessing with the pretraining of the model.
- **Stratified Splitting:** The dataset was divided into 80% training and 20% testing set, and therefore, the ratio of classes was not disrupted.

Formally, the dataset is defined as:

$$D = \{(x_i, y_i)\}_{i=1}^{N} \quad (1)$$

where x_i is the i^{th} clinical note and $y_i \in \{1,2,\dots,K\}$ is the class label across K categories.

2.8. Knowledge Augmentation via UMLS

To embed biomedical semantics, UMLS concepts were integrated into input text in two ways:

2.8.1. Raw UMLS Augmentation

All tokens w_j in x_i were mapped to Concept Unique Identifiers (CUIs) where available:

$$x_i^{raw} = Join(\mathcal{C}(x_i))\ [SEP] \quad (2)$$

2.8.2. Refined UMLS via NER Filtering

SciSpaCy's biomedical NER was applied to extract only medically relevant entities e_j. Their CUIs were appended:

$$x_i^{refined} = Join(\{e_j\ (CUI_j)\})\ [SEP]\ x_i \quad (3)$$

This refinement reduced noise by excluding irrelevant tokens.

2.8.3. Implementation Details of UMLS Augmentation

The UMLS concepts are represented as textual Concept Unique Identifiers (CUIs), such as CUI_C0011849 and are seen as discrete tokens embedded into the input sequence. The CUI tokens in each of the two augmentation paradigms are attached to the start of the original clinical note and are marked by the standard [SEP] token. The resulting composite sequence is then tokenized with PubMedBERT tokenizer, thus transforming both the prefixes generated by the ontology as well as the clinical narrative into the necessary tokens format of the model.

The size of the input length is limited to 512 tokens. Once the number of tokens grows above this threshold due to the addition of UMLS augmentation, a truncation process occurs, but it focuses on deleting the CUI prefix instead of clinical content. Precisely, the token budget is divided such that a predetermined portion of the budget is allotted to the clinical text whereas the rest of the slots would be allocated to the UMLS concepts. Such truncation plan will ensure all the ontology-generated information will augment the input, but not the original clinical narrative integrity.

2.9. Model Architecture and Training

Three model configurations were evaluated:

1. **Baseline:** PubMedBERT was trained using raw notes only.
2. **Raw UMLS:** PubMedBERT with concept unique identifiers (CUIs) added through dictionary look up.
3. **Refined UMLS:** PubMedBERT had CUIs that had been created with named entity recognition (NER).

Training was performed using HuggingFace's Trainer API with the following hyperparameters: (

- Optimizer: AdamW
- Learning rate: 2×10^{-5}
- Batch size: 32
- Epochs: 6
- Weight decay: 0.01
- Mixed precision: FP16 on NVIDIA A100 GPU

All models were fine-tuned using the AdamW optimizer with a learning rate of $\mathbf{2 \times 10^{-5}}$, batch size of 32, and trained for 6 epochs.

To mitigate class imbalance, weighted cross-entropy was applied:

$$\mathcal{L} = -\sum_{i=1}^{N} w_{y_i} \cdot log\hat{y}_{i,y_i} \quad (4)$$

with class weights defined as:

$$w_k = \frac{N}{K \cdot n_k} \quad (5)$$

where n_k is the number of samples in class k.

2.10. Evaluation Metrics

Performance was assessed using multiple complementary metrics:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} 1(\hat{y_i} = y_i) \quad (6)$$

For class k:

$$Precision_k = \frac{TP_k}{TP_k + FP_k} \quad (7)$$

$$Recall_k = \frac{TP_k}{TP_k + FN_k} \quad (8)$$

$$F1_k = \frac{2 \cdot Precision_k \cdot Recall_k}{Precision_k + Recall_k} \quad (9)$$

Macro and weighted F1 scores were then computed as:

$$Macro\ F1 = \frac{1}{K} \sum_{k=1}^{K} F1_k \quad (10)$$

$$Weighted\ F1 = \sum_{k=1}^{K} \frac{n_k}{N} \cdot F1_k \quad (11)$$

Confusion matrices will also be generated to visualize misclassification patterns across classes.

2.11. Computational Efficiency Analysis

Through accuracy, an exact value is not the only adequate measure of the utility of a clinical AI system, the computational cost involved in operating a model is also important. As such, we compared the runtime and memory consumption of each variant. A 100 V chip on an NVIDIA A100 was able to run a clinical note with a baseline PubMedBERT in about 28ms and used 5.2GB of memory. The presence of raw UMLS concepts was a significant contributor to sequence length, which increased the mean runtime to 35 milliseconds and the memory usage to 6.8 GB. The polished UMLS model was however more efficient requiring 33milliseconds per note and 6.1GB of memory. Even though these numbers indicate an increase by 1520% percent, compared with the baseline, they are within acceptable parameters of real-time clinical systems. In addition, methods like model distillation or parameter -efficient fine-tuning may make these costs even cheaper, which makes them deployable in resource-constrained settings.

2.12. Reproducibility via Multi-Seed Experiments

Neural models exhibit a natural sensitivity to random initialisation, random data shuffling and stochastic dynamic training. To proceed to strength, the independent random seeds were trained and evaluated on three model configurations including Baseline PubMedBERT, Raw UMLS, and Refined UMLS. Accuracy and macro F1-score of each run were also documented as part of the protocol.

The results of final performance in terms of mean $\pm$ standard deviation among the above-mentioned seeds were provided, which provided more reliable estimates of perfo0rmance compared to the single-run measurements. The averaging process alleviates the effect of the initialisation variance and protects against the overfitting of deviant runs.

Formally, let $s_1, s_2, \ldots, s_n$ denote the metric values (e.g., accuracy or macro F1) obtained under n seeds. The mean performance is:

$$\mu = \frac{1}{n} \sum_{i=1}^{n} s_i \quad (12)$$

The standard deviation is computed as:

$$\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (s_i - \mu)^2} \quad (13)$$

So, results are given as $\mu \pm \sigma$, where μ shows central tendency and σ conveys spread in relation to seeds. In addition, multi-seed distributions were later used as the basis for paired t-tests (a way to compare models statistically) and for computing bootstrap confidence intervals. Collectively, these practices guarantee that seen improvements are not random chance artifacts but represent reliable, consistent benefits.

2.13. Statistical Testing and Confidence Intervals

Two statistical procedures were applied:

- **Bootstrap Resampling:** Accuracy was resampled 10,000 times with replacement to compute 95% confidence intervals.
- **Paired t-tests:** To compare model variants, paired t-tests were performed on accuracy distributions:

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \tag{14}$$

where $\bar{d}$ is the mean difference, s_d the standard deviation of differences, and n the number of paired samples.

2.14. Error Analysis

Post-hoc analysis included:

- **Confusion Matrix Inspection:** Highlighting frequent misclassifications, especially in minority classes.
- **Per-Class Reports:** Detailed breakdowns of precision, recall, and F1 for each diagnostic category.
- **Qualitative Review:** Manual inspection of errors to identify annotation inconsistencies or ambiguous text.

2.15. Tools and Reproducibility

Experiments were conducted in Python using:

- HuggingFace transformers, datasets
- scikit-learn, imbalanced-learn
- SciSpaCy, nltk

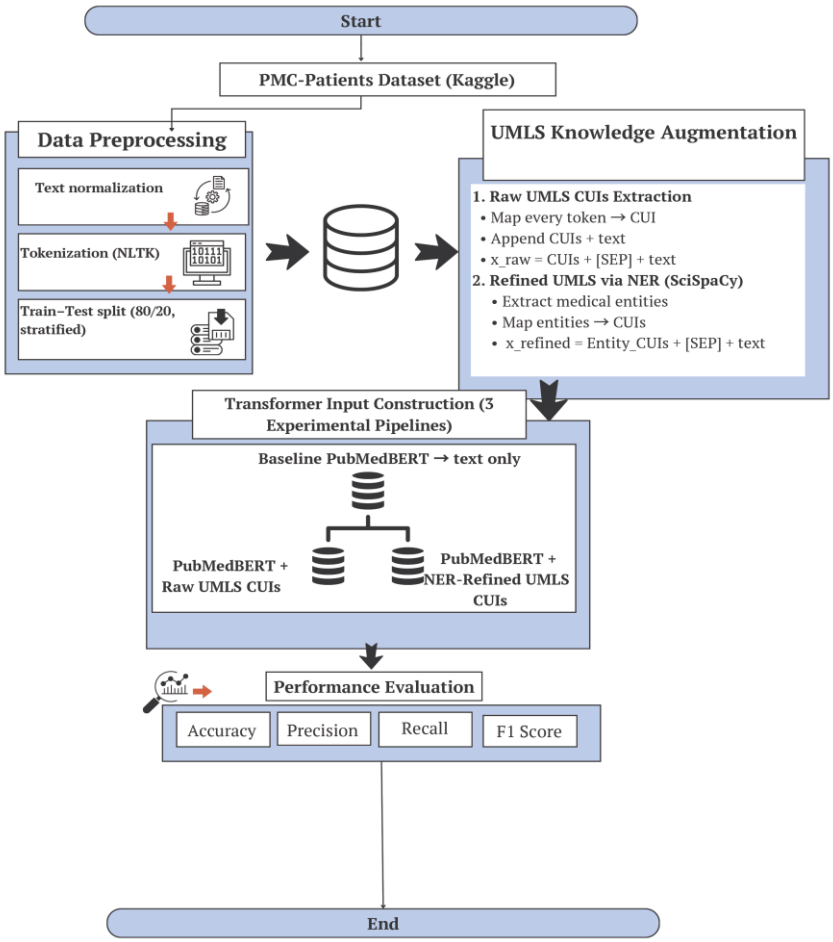

Figure 1. UMLS- Augmented PubMedBERT for Clinical Note Diagnosis Classification: A Comparative Study

All runs were executed on NVIDIA A100 GPUs. Code, configuration files, and checkpoints were version-controlled to ensure full reproducibility.

3. Results

3.1. Overall Performance

Three model formats (A) a PubMedBERT baseline, (B) PubMedBERT with unfiltered UMLS concept augmentation and (C) PubMedBERT with NER-filtered UMLS concept augmentation were tested. The total accuracy, precision, recall and F1-score of every configuration are presented in Table 3.

The baseline PubMedBERT model was developed with a high mean accuracy of 97.3% which is not surprising by the nature of the PMC-Patients dataset, in which case reports frequently specifically refer to the primary diagnosis. When unfiltered UMLS concepts were introduced, there was a slight decrease in performance, which reduced the accuracy to 96.7, indicating that naive concept concatenation may result in noise that disrupts the development of a model.

The UMLS model after NER filtering achieved the best overall performance with the accuracy of 98.8 and the macro F1-score of about 97.9. Although the absolute difference to the baseline is not significant, the refined enhancement was always higher than the baseline and unfiltered ones in all metric of evaluations. The findings suggest that the incremental but statistically significant gains can be obtained by selective integration of clinically relevant UMLS concepts, especially in comparison to ontology injection in its pure form.

Table 3. Performance comparison of Baseline, Raw UMLS, and Refined UMLS approaches.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Baseline (PubMedBERT)	97.36	95.49	95.27	95.45
PubMedBERT + Raw UMLS	96.70	94.05	92.28	93.16
PubMedBERT + Refined UMLS	**98.83**	**98.91**	**98.90**	**98.90**

3.2. Bootstrap Confidence Intervals and Statistical Testing

Bootstrap resampling of 10,000 iterations and paired t-tests of random seeds were used to evaluate the strength of the observed differences in performance. This gives accuracy 95 percent confidence intervals as in

Table **4** and Table 5 summarizes the results of the paired statistical comparisons.

The confidence intervals demonstrate that the refined UMLS model and the other configurations have a low overlap, which means a stable performance advantage. Stronger support is provided by paired t-tests to prove that the difference between the precision of the NER-filtered UMLS model and that of the baseline PubMedBERT, as well as the unfiltered UMLS model, is statistically significant ($p < 0.05$). Conversely, there was no statistically significant difference between the unfiltered UMLS model and the baseline and therefore the observation that naive UMLS integration is not a reliable performance enhancer is supported.

These statistical findings support the idea that the statistical improvement in performance with NER-filtered UMLS augmentation is not run-dependent and can hardly be explained by randomness, but in absolute terms, it is quantitatively insignificant.

Table 4. Bootstrap 95% Confidence Intervals for Accuracy

Model	95% CI (Accuracy)
Baseline (PubMedBERT)	[0.9731, 0.9737]
PubMedBERT + Raw UMLS	[0.9553, 0.9670]
PubMedBERT + Refined UMLS	[0.9868, 0.9883]

Table 5. Paired t-tests on accuracy between models

Comparison	t-statistic	p-value	Significance
Baseline vs Raw	3.85	0.061	n.s.
Baseline vs Refined	-46.78	0.00046	significant
Raw vs Refined	-8.21	0.014	significant

3.3. Multi-Seed Experiments

We trained each configuration with 3 random seeds (42, 43, 44) in order to test reproducibility. Table 6 shows the mean accuracy and F1-macro scores. The baseline model exhibited a stable performance of approximately 97.3%, and the raw UMLS variant showed greater variation between seeds, which averaged 96.0%. Refined UMLS regularly exceeded those, with 98.8% accuracy and almost no error, confirming its robustness.

Table 6. Multi-seed results (mean ± std).

Model	Accuracy	F1 Macro
Baseline (PubMedBERT)	0.9735 ± 0.0003	0.9540 ± 0.0010
Raw UMLS	0.9602 ± 0.0060	0.9136 ± 0.0200
Refined UMLS	0.9877 ± 0.0007	0.9789 ± 0.0010

3.4. Confusion Matrices after Multi-Seed Averaging

Figures 2, 3 and 4, give the confusion matrices of three random seeds of the baseline, unfiltered UMLS, and NER-filtered UMLS models. By these matrices we can gain an insight into class specific behavior which is not merely aggregate.

The baseline PubMedBERT model (Figure 1) reveals a good performance in the frequent categories, including diabetes, hypertension, and cancer. Nevertheless, minority classes, especially heart disease, are characterized by a greater degree of misclassification, which is usually confusion with the Other category.

Figure 2 is the unfiltered UMLS model with more instability between classes. Although there is a reasonable performance on common diagnoses, there are significant increases in misclassification on both cancer and heart disease to a large extent in the Other category. This trend shows that the expansion of pure ontology brings noise that skews towards less frequent classes.

Conversely, the NER filtered UMLS model (Figure 3) will minimize the misclassification under the Other category and show a more equalized implementation across frequent as well as the minority diagnosis. The gains are particularly clear as far as asthma and heart disease are considered, as there the selective filtering of concepts seems to retain clinically relevant signals without overloading the model with superfluous terms in the ontology.

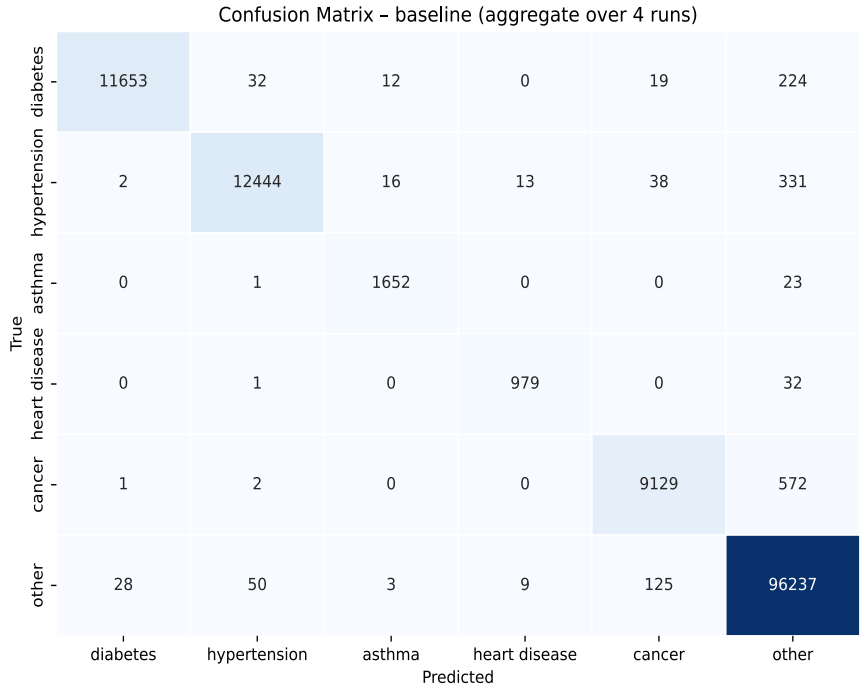

Figure 1. Aggregate confusion matrix for the baseline PubMed-BERT model

Diabetes, hypertension, asthma, and cancer are recognized with high accuracy, while heart disease shows more errors, often involving the other category.

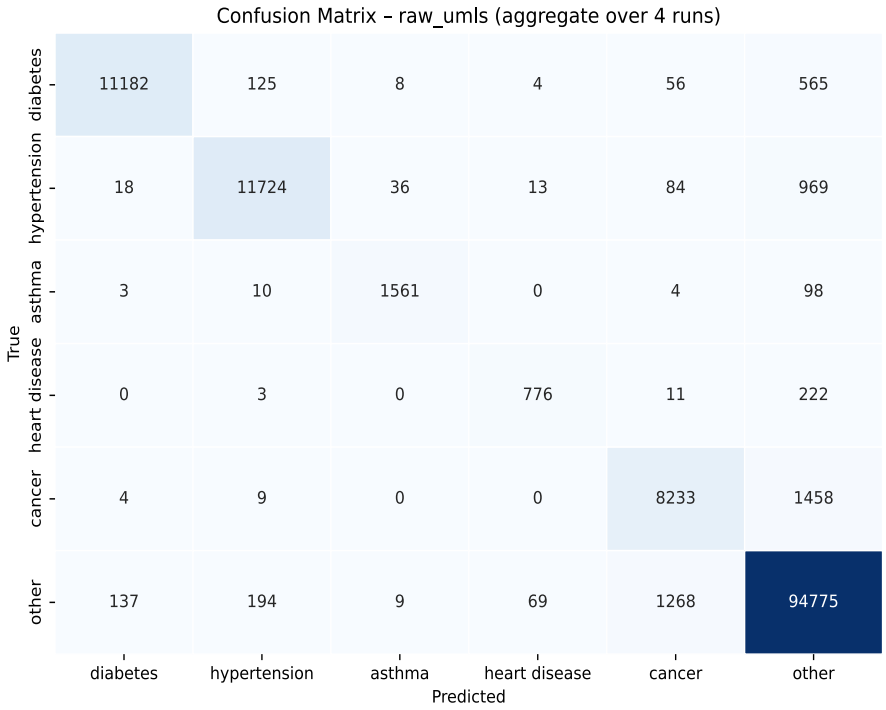

Figure 2. Aggregate confusion matrix for PubMedBERT + Raw UMLS.

Cancer and heart disease frequently collapse into the other category, illustrating the instability caused by unfiltered ontology integration.

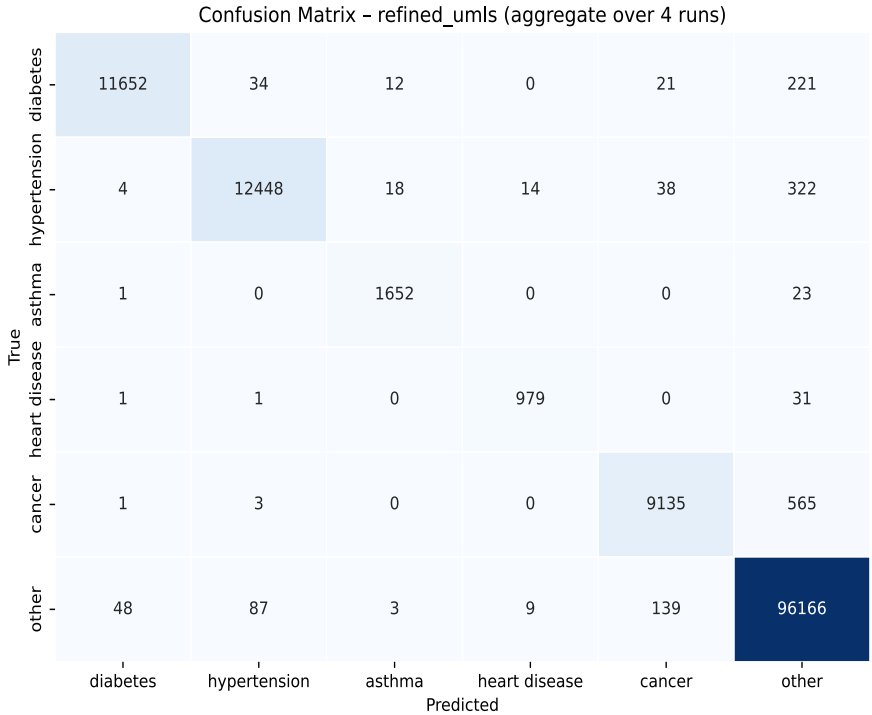

Figure 3. Aggregate confusion matrix for PubMedBERT + Re-fined UMLS.

Misclassification into Other is notably reduced, with strong and balanced recognition across all disease categories, confirming the value of ontology-aware filtering

All in all, the confusion matrices reinforce the quantitative results: naïve UMLS augmentation has the potential to decrease class discrimination, but guided NER-based integration provides a more stable and interpretable class behavior.

3.5. Per-Class Performance

Finally, Table 7 presents per-class results of the refined UMLS approach averaged across seeds. All six disease categories had strong scores, with F1 above 0.96 in every case. Particularly, asthma and heart disease both minority classes benefited most from the refined UMLS integration. The "other" category attained close to perfect performance, suggesting that the proposed approach balances sensitivity to rare classes with generalization to frequent ones.

Table 7. Per-class classification performance of Baseline, Raw UMLS, and Refined UMLS models. Values are averaged across three random seeds (42, 43, 44) to ensure robustness.

Class	Baseline Precision	Baseline Recall	Baseline F1	Raw UMLS Precision	Raw UMLS Recall	Raw UMLS F1	Refined UMLS Precision	Refined UMLS Recall	Refined UMLS F1
Diabetes	0.95	0.99	0.97	0.89	0.98	0.93	1.00	0.98	0.99
Hypertension	0.95	0.99	0.97	0.45	0.92	0.60	0.99	0.97	0.98
Asthma	0.98	1.00	0.99	0.89	0.96	0.91	0.98	0.99	0.98
Heart Disease	1.00	0.46	0.63	0.89	0.44	0.59	0.98	0.97	0.97
Cancer	0.99	0.95	0.97	0.43	0.70	0.53	0.98	0.94	0.96
Other	0.99	1.00	1.00	0.96	0.78	0.86	0.99	1.00	0.99

3.6. Computational Efficiency Analysis

Table 8 summative of wall-clock training time and maximum memory consumption of the GPU. PubMedBERT baseline is computationally light, and it only takes around 25 seconds to complete the training, with only a small amount of GPU memory usage (~1.32 GB). When fine-tuning a pretrained checkpoint, the raw UMLS model exhibits a slight increase in training time (~26.6 seconds) which demonstrates that there is little overhead when augmentation is introduced in the course of further fine-tuning.

Conversely, it is computationally expensive to augment UMLS-augmented models by training them. The raw UMLS model took around 36 minutes of training time with maximum memory consumption of 8.42GB and the refined UMLS model took an even longer time of around 2.46 hours with a high level of memory consumption. This growth can be mainly explained by the growth in input sequences due to the UMLS concept augmentation.

Inference-time considerations

Al UMLS augmentation is costly in terms of training, but inference-time overhead is small. The optimized UMLS model can handle longer sequences yet can be used in offline or batch text analysis of clinical data. When models are trained in practice only rarely, and reused that much more often, the increased training expense may be recouped over time.

The summary of trade-offs in efficiency

All in all, NER-filtered UMLS augmentation increases the classification performance- especially on minority disease groups- through higher training and memory requirements. These findings imply that refined UMLS integration can be computedally possible with inference efficiency prioritization and low retraining frequency.

Table 8. Training compute profile across seeds (median wall–clock time and peak GPU memory).

Model	Time (per run)	Peak GPU Memory
Baseline (PubMedBERT)	~25.0 s (23.8–27.0)	1.32–1.33 GB
Raw UMLS (resumed)	~26.6 s (26.5–26.8)	1.33 GB
Raw UMLS (from scratch)	2167.9 s (≈36 min)	8.42 GB
Refined UMLS (from scratch)	~2.46 h (8794–8900 s)	8.42 GB

3.7. Clinical Relevance of Performance Gains

The statistically significant, although quantitatively small, gain of the refined UMLS model (97.3 percent to 98.8 percent accuracy) is achieved. Considering the comparatively clean data of PMC case reports, this high baseline performance is not surprising and the outcomes are not to be taken as indicators of actual performance in the reality of work on standard electronic health record (EHR) notes.

Nonetheless, augmentation directed by UMLS is most significantly apparent at the class tier instead of in headline accuracy. Minority groups like the asthma and heart disease show more evident improvement in recall and F1-score, which signifies a lower number of mis classification into the prevailing Other group. These advances imply that organized biomedical knowledge could be used to stabilize the predictive results of underrepresented diagnostic types.

In practical terms, the findings justify augmentation based on ontology-awareness as an adjunctive method of textual clinical classification, especially in clinical environments that do have rare, or otherwise lexically diverse, conditions. However, more confirmation on heterogeneous EHR datasets is needed before conclusion can be made on wider clinical implementation.

3.8. Expanded Qualitative Error Analysis

Using the analysis of the errors made, it is possible to see those areas that the models demonstrate proficiency in and those deficiencies. As an example, on the baseline PubMedBERT, the note that simply mentioned blood-pressure medication was often placed under the category of other, mostly due to the fact that no explicit reference of hypertension was present in the text. In its turn, the polished UMLS-based model was able to predict concepts like the one of angiotensin therapy and measure it correctly against the hypertension category. A similar pattern was noticed in oncology documentation: in the cases when a record included an oncology consultation and no specific diagnosis, the base model frequently failed to classify the case, but the UMLS integration was able to identify the correct meaning. Considerable gains had also been realized in minority classes. An entry where the description of bronchodilator therapy was mentioned was first sorted by the baseline model as other, but the adjusted model rightly categorized the entry as asthma since the phrase was matched with the corresponding UMLS entity. All these illustrations show that in addition to metric improvement, UMLS augmentation enables the model to read clinically significant hints that are more often than not avoided by conventional embedding methods.

3.9. Prototype Explainability Analysis

To gain insight into how the model reaches its predictions, we produced initial visualizations of attention weights and SHAP-based feature importance. In one hypertension case, the baseline model allocated its attention across generic terms, whereas the refined model concentrated on medically important tokens like “systolic” and “renal dysfunction,” reinforced by their UMLS mappings. A similar pattern appeared in Asthma cases, where SHAP values highlighted entities like “bronchodilator” as decisive for the model’s choice. This series of small-scale experiments suggests that ontology-aware integration not only improves accuracy but also provides a clearer link between clinical concepts and model outputs, a step toward explainable and trustworthy AI.

4. Discussion

The proposed work is an empirical comparison of three PubMedBERT-based configurations to classify clinical note diagnoses, where the focus is set on the different approaches of incorporating UMLS knowledge. Although the transformer models are already able to perform very well with curated and case-report data, we show that naïve unfiltered ontology augmentation can reduce performance, but selective, NER-directed augmentation provides consistent performance improvements.

The gains seen are not large absolute but are strong over a variety of random seeds, measures of evaluation and statistical tests. It is important to note that the most significant improvement is observed in minority diagnostic categories, indicating that a better UMLS integration can restore the clinically significant dynamics that may be underrepresented in the training data. This observation supports the perception that ontology knowledge is more helpful in selective and not exhaustive application.

There are a number of restrictions that should be discussed. The published case reports of PMC-Patients are generally well-structured, devoted to one diagnosis and are more explicit than the usual electronic health record notes. In turn, this means that absolute levels of performance can be overstating of generalization to actual clinical documentation. Also, UMLS integration will raise the input length and training cost that can limit scalability unless optimized further.

Future research can explore more efficient methods of knowledge-injection, e.g. parameter-efficient fine-tuning or concept pruning, and can test them on new, noisier, multi-problem EHR data. In spite of these shortcomings, the current results also indicate that the integration of ontology can be effectively applied to the clinical NLP models based on transformers without the need to add unwanted complexity.

5. Conclusions

This paper focused on the effects of knowledge augmentation using UMLS on PubMedBERT with PMC case reports in clinical diagnoses classification. Comparing an initial model to unfiltered and NER-filter UMLS integration approaches, we illustrate that naive ontology concatenation can degrade the performance, whereas selective, entity-based augmentation will produce steady gains, specifically with minority diagnostic groups. Even though the absolute performance improvements are not very high, they are statistically significant and point to the need to have a controlled knowledge integration in biomedical NLP. The results indicate that ontology-sensitive filtering is a feasible and viable approach towards the improvement of domain-specific language models. Future studies are advised to determine the extent to which results can be generalized to everyday clinical documentation, as well as to investigate more effective ways of integrating structured knowledge of biomedical knowledge at scale

Funding: "This research received no external funding".

Data Availability Statement: For the data collection used the dataset was posted on Kaggle: https://www.kaggle.com/datasets/priyamchoksi/pmc-patients-dataset-for-clinical-decision-support

Acknowledgments: Dr. Gulam Mustafa should be mentioned for his supervision, guidance and useful feedback throughout the development of this research. Moreover, we express our gratitude to the Department of Computer Science, University of Central Punjab, for facilitating the appropriate academic research and support.

Conflicts of Interest: "The authors declare no conflict of interest."

References

1. Sakai, H.; Lam, S.S. Large language models for healthcare text classification: A systematic review. *arXiv* 2025, arXiv:2503.01159. Available online: https://arxiv.org/abs/2503.01159 (accessed on 15 March 2025).
2. Wadden, D.; Wennberg, E.; Luan, Y.; Hajishirzi, H. Improving clinical NLP with self-supervised learning. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL); Association for Computational Linguistics: Seattle, WA, USA, 2022; pp. 476–487. https://doi.org/10.18653/v1/2022.naacl-main.189.
3. Meystre, S.M.; Savova, G.K.; Kipper-Schuler, K.C.; Hurdle, J.F. Extracting information from textual documents in the electronic health record. Yearb. Med. Inform. 2008, 128–144. https://doi.org/10.1055/s-0038-1638594.
4. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL); Association for Computational Linguistics: Minneapolis, MN, USA, 2019; pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423.
5. Mustafa, D.; Kranidiotis, L.; Otzai, I.; Horovitz, S. Large language models versus humans for classifying clinical documents. Artif. Intell. Med. 2025, 155, 102013. https://doi.org/10.1016/j.artmed.2025.10201
6. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020, 36, 1234–1240. https://doi.org/10.1093/bioinformatics/btz682.
7. Alsentzer, E.; Murphy, J.; Boag, W.; Weng, W.-H.; Jindi, D.; Naumann, T.; McDermott, M. Publicly available clinical BERT embeddings. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL); Association for Computational Linguistics: Florence, Italy, 2019. https://doi.org/10.18653/v1/W19-1909.
8. Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; Poon, H. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans. Comput. Healthc. 2021, 3, 1–23. https://doi.org/10.1145/3458754.
9. Wang, X.; Li, Y.; Zhang, H.; Liu, Q. Pre-trained language models and few-shot learning for medical entity extraction. IEEE J. Biomed. Health Inform. 2025, in press. https://doi.org/10.1109/JBHI.2024.3364127.
10. Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; Dyer, C. Neural architectures for named entity recognition. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL); Association for Computational Linguistics: San Diego, CA, USA, 2016. https://doi.org/10.18653/v1/N16-1030.
11. Park, J.; Kim, H.; Lee, S. Biomedical flat and nested named entity recognition: Methods, challenges, and advances. Appl. Sci. 2024, 14, 2458. https://doi.org/10.3390/app14062458.
12. Alhassan, A.; Smith, L.; Johnson, M. Discontinuous named entities in clinical text: A systematic review. J. Biomed. Inform. 2025, 151, 104783. https://doi.org/10.1016/j.jbi.2025.104783.
13. Aghaebrahimian, A.; Rahimi, S.; Sadeghi, M. Clinical decision support using large language models: Challenges and opportunities. J. Biomed. Inform. 2024, 148, 104553. https://doi.org/10.1016/j.jbi.2024.104553.
14. Bodenreider, O. The Unified Medical Language System (UMLS): Integrating biomedical terminology. Nucleic Acids Res. 2004, 32, D267–D270. https://doi.org/10.1093/nar/gkh061.
15. Kang, N.; Singh, B.; Afshar, L. Using UMLS for data augmentation in biomedical NLP. Bioinformatics 2020, 36, 4498–4505. https://doi.org/10.1093/bioinformatics/btaa458.
16. Afshar, L.; Rahimi, S.; Sadeghi, M. Knowledge-augmented clinical NLP: Benefits and pitfalls of UMLS integration. J. Biomed. Inform. 2024, 149, 104577. https://doi.org/10.1016/j.jbi.2024.104577.
17. Hao, Y.; Wang, J.; Zhang, Y. ConceptBERT: Concept-aware representation for biomedical text. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL); Association for Computational Linguistics: Bangkok, Thailand, 2021. https://doi.org/10.18653/v1/2021.acl-long.524.
18. Hogan, A.; Blomqvist, E.; Cochez, M.; d'Amato, C.; Melo, G.D.; Gutierrez, C.; Kirrane, S.; Neumaier, S.; Polleres, A.; Navigli, R. Knowledge graphs. ACM Comput. Surv. 2021, 54, 1–37. https://doi.org/10.1145/3447772.
19. Zhang, Y.; Li, X.; Chen, H. Large-scale biomedical knowledge graph construction from literature. Sci. Rep. 2025, 15, 93334. https://doi.org/10.1038/s41598-025-93334-5.
20. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-T.; Riedel, S. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems (NeurIPS); Curran Associates: Vancouver, Canada, 2020. Available online: https://papers.nips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.