Digital Zone
Jurnal Teknologi Informasi dan Komunikasi

Volume 16 Issue 2 Year 2025 | Page 84-95 | e-ISSN: 2477-3255 | ISSN: 2086-4884
Received: 17-07-2025 | Revised: 22-08-2025 | Accepted: 10-10-2025

Evaluating Contextual Embedding Models for Multi-Label PICO Classification in Heart Disease: Addressing the Intervention - Comparison Bottleneck

Susi Handayani[1], Walhidayat[2], Taslim[1,3], Dafwen Toresa[1,4]
[1,2,3,4] Faculty of Computer Science, Universitas Lancang Kuning
[4]School Of Computing, Universiti Utara Malaysia
[3]Faculty of Informatics & Computing, Universiti Sultan Zainal Abidin
e-mail: [1]susi@unilak.ac.id, [2]walhidayat@unilak.ac.id, [3]taslim@unilak.ac.id, [4]dafwen@unilak.ac.id

*Corespondence: taslim@unilak.ac.id

Abstract: *Accurate extraction of Population, Intervention, Comparison, and Outcome (PICO) elements from clinical texts is essential for advancing evidence-based medicine, particularly in cardiology, where clinical narratives are highly complex and heterogeneous. This study evaluates the comparative effectiveness of three contextual embedding models—BioBERT, PubMedBERT, and SciBERT—combined with a Bidirectional Long Short-Term Memory (BiLSTM) architecture for multi-label PICO classification on a heart disease dataset. Model performance was assessed using accuracy, precision, recall, and F1-score metrics, supported by confusion matrix analysis. The BioBERT-BiLSTM model achieved an accuracy of 69.4% and an F1-score of 75.6%, providing balanced performance across categories. PubMedBERT-BiLSTM attained the highest accuracy and precision (73.2% and 84.1%, respectively), while SciBERT-BiLSTM demonstrated superior recall (74.6%) and the highest overall F1-score (78.4%). These findings indicate that pretraining domain significantly affects classification outcomes, with PubMedBERT and BioBERT yielding higher precision and stability, whereas SciBERT enhances sensitivity.*

Keywords: *PICO classification, contextual embedding, BiLSTM, clinical NLP, cardiology*

1. Introduction

Heart disease remains one of the most critical global public health challenges, accounting for more than 17 million deaths annually and representing the leading cause of mortality worldwide [1]. Beyond its contribution to global mortality, heart disease imposes substantial burdens in terms of disability-adjusted life years (DALYs), reduced productivity, and escalating healthcare costs, particularly in low- and middle-income countries where access to timely medical intervention is often limited [2][3][4]. The clinical spectrum of heart disease includes coronary artery disease, heart failure, valvular disorders, and arrhythmias, each requiring complex and often lifelong management strategies [5][6][7]. Over the past two decades, advances in pharmacological and interventional therapies have significantly reshaped clinical management paradigms [8] [9]. Antiplatelet and lipid-lowering agents, guideline-directed medical therapy for heart failure, and invasive procedures such as percutaneous coronary intervention have substantially improved both patient survival and quality of life [10] [11].

Within the paradigm of Evidence-Based Medicine (EBM), the PICO model, comprising Population, Intervention, Comparison, and Outcome, has been widely adopted as a structured approach for formulating clinical questions and evaluating research evidence [12]. By clearly defining these components, the PICO framework facilitates systematic literature retrieval and enhances the reliability of evidence appraisal. Several systematic reviews confirm that applying the PICO structure improves both the precision of literature searches and the accuracy of evidence synthesis [13] [14][15].

https://doi.org/10.31849/digitalzone.v16i2. 27687

However, despite its conceptual clarity, the application of the PICO framework to cardiology literature presents domain-specific challenges [16]. Many studies in this field employ complex trial designs with multiple intervention arms [17][18][19], involve patients with layered comorbidities [20][21], and use highly varied terminologies [22][23] to describe therapeutic strategies. These characteristics make the automated extraction of PICO elements particularly challenging. Among these, the *Intervention* and *Comparison* elements are notably more difficult to identify accurately compared with *Population* and *Outcome* [23] [24]. Given the rapid increase in cardiovascular publications, there is an urgent need for automated systems capable of reliably and consistently extracting PICO elements.

Several computational approaches have been developed for PICO element extraction, ranging from rule-based frameworks to machine learning techniques [25]. Such methods have the potential to substantially reduce the manual effort involved in systematic reviews and meta-analyses, thereby improving both the efficiency and reproducibility of evidence-based research [26]. Despite notable progress, most prior studies have concentrated on oncology and infectious diseases, domains with abundant annotated corpora, while applications to cardiovascular medicine remain comparatively underexplored [27]. This gap underscores the need for methodologies tailored to the linguistic and clinical nuances of heart disease literature [25].

Recent advances in contextual embedding models have shown considerable promise in biomedical natural language processing (NLP). BioBERT, an extension of the BERT architecture pretrained on large-scale biomedical corpora, has demonstrated strong performance in tasks such as named entity recognition and relation extraction [28]. PubMedBERT, trained exclusively on PubMed abstracts, captures linguistic patterns specific to biomedical publications and has achieved notable results across various biomedical NLP tasks [25]. Meanwhile, SciBERT, developed on a broader corpus of multidisciplinary scientific texts, offers wider domain coverage and supports cross-domain generalization [28]. These complementary characteristics provide a compelling rationale for systematically evaluating these models in the context of PICO classification [27].

This study aims to address this gap by systematically evaluating the performance of BioBERT, PubMedBERT, and SciBERT, each integrated with a Bidirectional Long Short-Term Memory (BiLSTM) architecture, for multi-label PICO classification in cardiology literature. Particular attention is given to the *Intervention* and *Comparison* components, which have been identified as especially difficult to classify accurately across biomedical domains [29]. By comparing models with differing pretraining domains, this research investigates how domain-specific representations affect classification performance and explores limitations that hinder the development of robust evidence-based systems in cardiology [30].

This study investigates the performance of three contextual embedding models, BioBERT, PubMedBERT, and SciBERT, when integrated with a Bidirectional Long Short-Term Memory (BiLSTM) architecture for multi-label PICO classification in cardiology-related texts. The analysis focuses on achieving an optimal balance between precision and recall, with particular emphasis on the *Intervention/Comparison* element, which exhibits substantial linguistic variability and frequent misclassification. By comparing models pretrained on distinct corpora, this study elucidates how domain specialization shapes linguistic generalization and classification sensitivity within biomedical text mining.

The contribution of this work is twofold. First, it presents a domain-aware comparative analysis conducted under a unified experimental framework, thereby isolating the effects of pretraining domain on model performance. Second, it provides a detailed examination of the Intervention/Comparison classification bottleneck, moving beyond aggregate performance scores to deliver a fine-grained interpretation of model behavior. The experimental setup and dataset are described transparently to ensure replicability and future extension. Collectively, the findings offer valuable insights for the design of automated evidence-synthesis systems and clinical decision-support tools that rely on precise PICO element identification.

2. Research Method

This study employed an experimental research design to evaluate the effectiveness of three contextual embedding models BioBERT, PubMedBERT, and SciBERT in combination with a Bidirectional Long Short-Term Memory (BiLSTM) architecture for multi-label classification of PICO elements. The evaluation utilized the PICO-HD dataset, which comprises annotated clinical texts related to heart disease. The primary objective was to investigate how variations in the pretraining domains of contextual embeddings influence classification performance and to identify the most effective embedding–architecture combination for processing cardiology-related literature.

The overall research methodology is illustrated in **Figure 1**, which delineates the key stages of the workflow: dataset preparation, preprocessing, embedding representation, model architecture, training, and

evaluation. This design ensures a systematic comparison between domain-specific and domain-general contextual embeddings within the framework of clinical natural language processing (NLP).

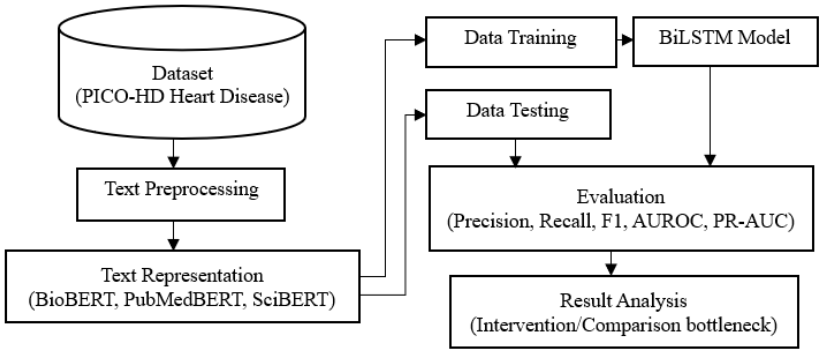


Figure 1. Research Methodology

Figure 1. Research Methodology Research methodology illustrating the experimental workflow for multi-label PICO classification. The process comprises dataset preparation (PICO-HD Heart Disease), text preprocessing, contextual embedding using BioBERT, PubMedBERT, and SciBERT, model training and testing using a BiLSTM architecture, and evaluation based on precision, recall, F1-score, AUROC, and PR-AUC metrics. The final stage involves result analysis focusing on the Intervention/Comparison bottleneck.

2.1. Dataset

The dataset used in this study was obtained from a publicly available Kaggle repository titled *PICO Medical Literatures on Heart Disease Data* [31]. It consists of 2,697 annotated sentences, each labeled according to the PICO framework. The distribution across categories is as follows: Population/Problem (23.6%), Intervention/Comparison (28.2%), Outcome (24.2%), and Not Relevant (24.0%). This balanced distribution ensures that each PICO element is adequately represented while reflecting the inherent complexity of clinical narratives.

Each sentence was annotated based on the four core PICO elements. *Population/Problem (P)* refers to the group of patients or the medical condition under investigation. *Intervention/Comparison (I/C)* encompasses the treatments, procedures, or conditions being compared. *Outcome (O)* denotes the clinical results or endpoints measured, while *Not Relevant (N)* represents sentences that do not contribute to evidence extraction. An illustrative example of the dataset is provided in **Table 1**.

Table 1. Example of heart disease clinical text dataset based on the PICO framework

No	Clinical Teks	Label
1	methods total 107 women clinical indication…	Population/Problem (P)
2	radial artery cannulation performed…	Intervention and Comparison (I and C)
3	primary endpoint composite all-cause death…	Outcome (O)
4	statistical significance defined p-value 0.05	Not Relevant (N)

2.2. Preprocessing

A conservative preprocessing strategy was implemented to preserve clinically relevant information. Numerical values, decimal points, comparison operators, and percentage symbols were retained, as these often convey critical quantitative details in medical texts. Stopword removal was not applied to the final model inputs to maintain contextual dependencies essential for optimal performance in transformer-based encoders.

Tokenization was performed using the WordPiece tokenizer to ensure compatibility with all embedding models. All characters were converted to lowercase to reduce vocabulary sparsity, while extraneous symbols and non-informative special characters were removed. These steps produced normalized yet semantically rich textual inputs suitable for accurate multi-label PICO classification.

2.3. Contextual Embedding

Following preprocessing, the clinical texts were transformed into dense vector representations using three contextual embedding models: BioBERT, PubMedBERT, and SciBERT. **BioBERT** was pretrained on large-scale biomedical corpora, including PubMed abstracts and PMC full-text articles, enabling comprehensive coverage of biomedical terminology. PubMedBERT, by contrast, was trained exclusively on PubMed abstracts, resulting in embeddings that capture the linguistic characteristics of biomedical literature with high specificity. SciBERT, in turn, was pretrained on a diverse, multi-domain corpus from Semantic Scholar, providing broader scientific coverage across multiple disciplines rather than focusing solely on medicine.

The use of these three embedding models facilitated a controlled comparison between domain-specialized (BioBERT and PubMedBERT) and domain-general (SciBERT) representations within the context of PICO classification. Contextual embeddings were selected for their ability to capture nuanced semantic and syntactic relationships in biomedical texts, outperforming conventional vectorization techniques such as TF-IDF and word2vec. By leveraging transformer-based encoders, these embeddings offer enhanced interpretability of the linguistic complexity and variability characteristic of clinical narratives.

2.4. Model Architecture

The contextual embeddings derived from BioBERT, PubMedBERT, and SciBERT served as the input representations for the classification framework. Each tokenized sentence was converted into dense vectors using the respective embedding model. To maintain computational efficiency and mitigate overfitting given the relatively modest dataset size, the encoder parameters were frozen during training.

The resulting sequence of embeddings was passed through a Bidirectional Long Short-Term Memory (BiLSTM) layer with a hidden dimension of 256 units. This architecture captures contextual dependencies in both forward and backward directions an essential feature for understanding the intricate syntactic and semantic patterns prevalent in clinical texts.

The concatenated outputs of the forward and backward LSTM states were regularized using a dropout rate of 0.3 and subsequently passed through a fully connected dense layer to generate the final feature representations for classification. A sigmoid activation function was applied at the output layer to enable independent probability estimation for each label. This design was critical for the multi-label setting, where sentences could correspond simultaneously to one or more PICO categories (*Population/Problem*, *Intervention/Comparison*, *Outcome*, or *Not Relevant*). Unlike the softmax function, which enforces a single-label constraint, the sigmoid activation allowed for overlapping label assignments, thereby better reflecting the inherent ambiguity and overlap in clinical texts.

2.5. Data Split and Training

The dataset was partitioned into training (80%) and testing (20%) subsets using a stratified split to preserve the proportional distribution of labels. From the training subset, 10% was further separated as a validation set for hyperparameter optimization and early stopping. Model optimization was carried out using the Adam optimizer with binary cross-entropy loss. The training configuration included a batch size of 16, hidden layer dimension of 256 units, dropout rate of 0.3, and a learning rate of 1×10^{-4}. Each model was trained for three epochs, with early stopping based on validation loss to prevent overfitting. All experiments were conducted on a single NVIDIA Tesla T4 GPU, with an average runtime of approximately two to three minutes per epoch. A concise algorithmic summary of the experimental workflow is provided in **Algorithm 1** to facilitate reproducibility.

INPUT:
Clinical sentences annotated with PICO labels

OUTPUT:
Predicted labels and evaluation metrics
STEPS:
Initialize random seed
Split data into training, validation, and testing sets
For each model in {BioBERT, PubMedBERT, SciBERT}:
Load pretrained model and tokenizer
Freeze encoder weights
Build architecture: Input → Embeddings → BiLSTM → Dropout → Output layer
Train using Adam optimizer, binary cross-entropy loss, batch size 16, and early stopping

Evaluate on the test set and record metrics
Compare all models and analyze error patterns
Return the best-performing model and final evaluation results

2.6. Evaluation Metrics

The performance of each embedding model combined with the BiLSTM classifier was evaluated using five key metrics: **Accuracy,** Precision, Recall, F1-score**, and** Hamming Loss. Accuracy served as an overall indicator of classification performance, whereas macro-averaged Precision, Recall, and F1-score were used to assess the balance between false positives and false negatives across all PICO categories. **Hamming Loss** quantified the proportion of misclassified labels in the multi-label setting, providing a robust measure of prediction stability. To complement these aggregate indicators, confusion matrices were constructed for each label to provide a detailed analysis of error distribution. Additionally, statistical **significance testing** employing bootstrap-based confidence intervals and paired comparisons was conducted to ensure that performance differences among models were not attributable to random variation. Receiver Operating Characteristic (ROC) and Precision–Recall (PR) curves were generated to visualize the models' discriminative behaviors.

All experiments were performed under identical computational conditions to ensure **reproducibility and fairness** across comparisons. The complete configuration of training parameters, data partitioning, and evaluation metrics is thoroughly reported to facilitate the replication of results. By maintaining fixed random seeds and consistent preprocessing procedures, this study guarantees that the outcomes can be independently verified and reproduced under comparable computational environments.

3. Results and Discussion

This section presents the experimental findings along with an integrated discussion of model performance. The analysis encompasses both quantitative and qualitative perspectives. Quantitative evaluation relies on standard multi-label classification metrics, including accuracy, macro-averaged precision, recall, F1-score, and Hamming loss. Complementary analyses, such as confusion matrices, ROC curves, and PR curves, are provided to illustrate classification tendencies and error patterns. Furthermore, **bootstrap**-based statistical testing was applied to confirm the robustness of the observed performance differences.

Particular attention is devoted to the Intervention and Comparison categories, which consistently emerged as the most challenging to classify due to linguistic variability, overlapping terminology, and context-dependent phrasing. The results are first presented at the individual model **level**, highlighting each model's distinct strengths and weaknesses, followed by a comparative analysis evaluating the relative advantages of BioBERT, PubMedBERT, and SciBERT when integrated with the BiLSTM architecture.

3.1. Result

This subsection reports the experimental outcomes based on the five primary evaluation metrics: **Accuracy, Precision, Recall, F1-score**, and **Hamming Loss**. Confusion matrices are also presented to provide a granular view of each model's classification performance across PICO labels.

3.1.1. BioBERT-BiLSTM

The BioBERT–BiLSTM model was evaluated on the **PICO-HD** dataset. Results show that BioBERT achieved an accuracy of 69.4**%**, with an average F1-score of 75.6**%**, precision of 80**.6%**, **recall of 72.1%**, and a Hamming loss of 0.117. The relatively low Hamming loss indicates stable multi-label predictions. BioBERT demonstrated balanced performance in the Outcome and Not Relevant categories, with minimal false negatives. However, its performance was less consistent for **Population/Problem** and particularly weak for Intervention/Comparison, yielding 67 false negatives. This pattern suggests that while BioBERT's broader contextual representation enabled strong recall, it also resulted in increased false positives.

Table 2. Evaluation Results of the BioBERT-BiLSTM Model

Metric	Value
Accuracy	0.694
Precision	0.806
Recall	0.721
F1-Score	0.756
Hamming Loss	0.117

Table 3. Confusion Matrix per Label (BioBERT-BiLSTM)

Label	True Negative	False Positive	False Negative	True Positive
Population/Problem (P)	379	28	40	91
Intervention/Comparison (I/C)	331	29	67	111
Outcome (O)	388	36	24	90
Not Relevant (N)	394	25	26	94

The results in Table 2 confirm that BioBERT–BiLSTM achieved a balanced trade-off between precision and recall, with a moderate accuracy of 69.4%. Table 3 further reveals that the Outcome and Not Relevant categories were classified most accurately, as indicated by the low false negative counts (24 and 26, respectively). Conversely, the Intervention/Comparison category remained the most challenging, followed by Population/Problem. These findings indicate that while BioBERT effectively captures general biomedical semantics, it struggles to accommodate the linguistic diversity typical of intervention-related expressions.

3.1.2. PubMedBERT-BiLSTM

The PubMedBERT–BiLSTM model was evaluated under the same conditions. It achieved **an** accuracy of 73.2%**,** an F1-score of 77.7%**,** precision of 84.1%**,** recall of 73.8%**, and a** Hamming loss of 0.109**.** Compared to BioBERT–BiLSTM, PubMedBERT exhibited higher precision but lower recall, suggesting a reduction in false positives at the cost of slightly more false negatives, particularly in the Intervention/Comparison category. Stronger performance was again observed for the Outcome and Not Relevant categories.

Table 4. Evaluation Results of the PubMedBERT-BiLSTM Model

Metric	Value
Accuracy	0.732
Precision	0.841
Recall	0.738
F1-Score	0.777
Hamming Loss	0.109

Table 5. Confusion Matrix per Label (PubMedBERT-BiLSTM)

Label	True Negative	False Positive	False Negative	True Positive
Population/Problem (P)	381	26	38	93
Intervention/Comparison (I/C)	336	24	61	117
Outcome (O)	385	33	22	98
Not Relevant (N)	401	18	23	97

As shown in Table 4, PubMedBERT–BiLSTM achieved the highest overall accuracy (73.2%) and precision (84.1%) among the evaluated models. However, its recall was slightly lower than BioBERT's, reflecting a tendency to overlook some positive instances. Table 5 further indicates that PubMedBERT performed most effectively on the Outcome and Not Relevant categories, which exhibited the lowest false negative rates (22 and 23, respectively). Conversely, the Intervention/Comparison label remained challenging, yielding 61 false negatives. These results suggest that pretraining on PubMed abstracts enhanced PubMedBERT's precision and domain relevance, yet limited its sensitivity to the lexical and syntactic diversity characteristic of intervention-related expressions in clinical cardiology texts.

3.1.3. **SciBERT-BiLSTM**

The SciBERT–BiLSTM model produced moderate results in the multi-label classification of PICO elements within heart disease clinical texts. The model achieved an overall accuracy of 72.8%, an F1-score of 78.4%, precision of 82.7%, recall of 74.6%, and a Hamming loss of 0.104. These results indicate that although SciBERT was pretrained on a broad corpus of scientific literature rather than biomedical-specific data, it provided a reasonably competitive baseline compared to domain-focused models such as BioBERT and PubMedBERT. The evaluation metrics are summarized in Table 6.

Table 6. Evaluation Results of the SciBERT-BiLSTM Model

Metric	Value
Accuracy	0.728
Precision	0.827
Recall	0.746
F1-Score	0.784
Hamming Loss	0.104

A more granular view is provided by the confusion matrices presented in Table 7.

Table 7. Confusion Matrix per Label

Label	True Negative	False Positive	False Negative	True Positive
Population/Problem (P)	393	16	41	90
Intervention/Comparison (I/C)	336	33	48	123
Outcome (O)	398	24	30	88
Not Relevant (N)	407	13	20	100

From the confusion matrix, it can be observed that SciBERT–BiLSTM achieved a slightly stronger balance between precision and recall compared to PubMedBERT–BiLSTM, though its overall recall remained lower than that of BioBERT–BiLSTM. The Intervention/Comparison category again emerged as the most difficult to identify, with 48 false negatives and 33 false positives, underscoring the model's challenges in capturing the diverse linguistic patterns associated with intervention-related terminology. In contrast, the Not Relevant and Population/Problem categories demonstrated greater stability, reflected by relatively few classification errors.

Overall, the findings suggest that while SciBERT–BiLSTM delivers consistent performance across most categories, it also reveals the limitations of general-purpose scientific embeddings when applied to highly specialized biomedical tasks. Despite achieving competitive precision and stability, the model's underperformance in the Intervention/Comparison category reinforces the importance of domain-specific pretraining for accurate clinical text understanding. Future work may explore hybrid strategies that integrate SciBERT's broad scientific knowledge with biomedical-specialized embeddings to enhance performance in complex comparative contexts.

In addition to the tabulated metrics, a visual comparison of model performance across Accuracy, Precision, Recall, and F1-score is presented in Figure 2. The bar chart illustrates the relative strengths of each embedding model, providing a clear depiction of performance variations.

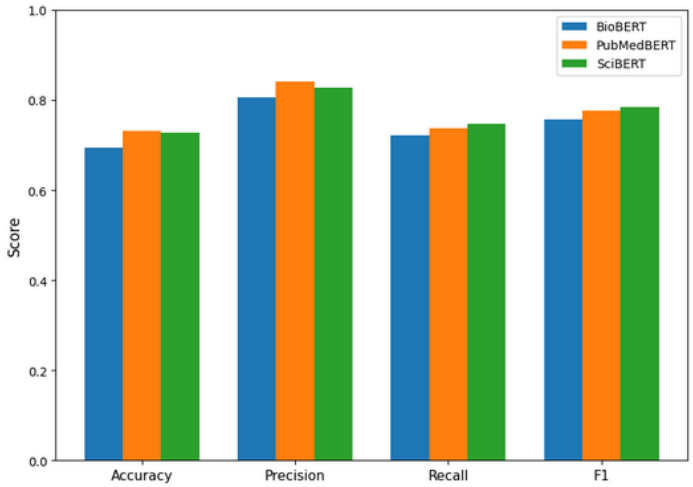


Figure 2. Performance comparison of BioBERT, PubMedBERT, and SciBERT models

Figure 2. Comparative Performance of BioBERT–BiLSTM, PubMedBERT–BiLSTM, and SciBERT–BiLSTM Models Comparative performance across four evaluation metrics—accuracy, precision, recall, and F1-score. PubMedBERT achieved the highest precision and overall accuracy, while

SciBERT exhibited relatively stronger recall, indicating broader sensitivity to PICO elements in heart disease clinical texts.

3.1.4. ROC and Precision–Recall Curves

Figures 3 and 4 display the Receiver Operating Characteristic (ROC) and Precision–Recall (PR) curves for the three models across the four PICO labels. Both the Outcome and Not Relevant categories achieved AUC values exceeding 0.90, whereas Intervention/Comparison remained the most difficult to classify accurately. Among the models, PubMedBERT demonstrated the best performance on the Outcome label, while SciBERT showed stronger recall trends. These visualizations collectively confirm that domain-specific embeddings such as PubMedBERT and BioBERT enhance model stability and precision, whereas SciBERT's generalized pretraining results in broader recall at the expense of higher false positive rates.

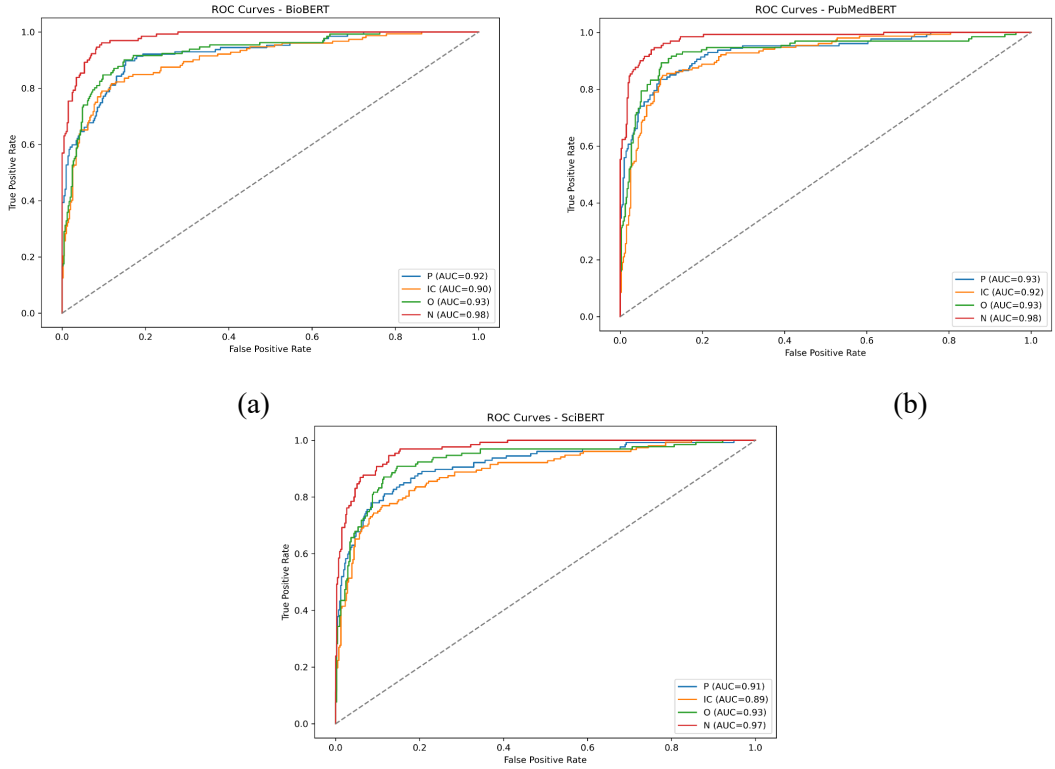
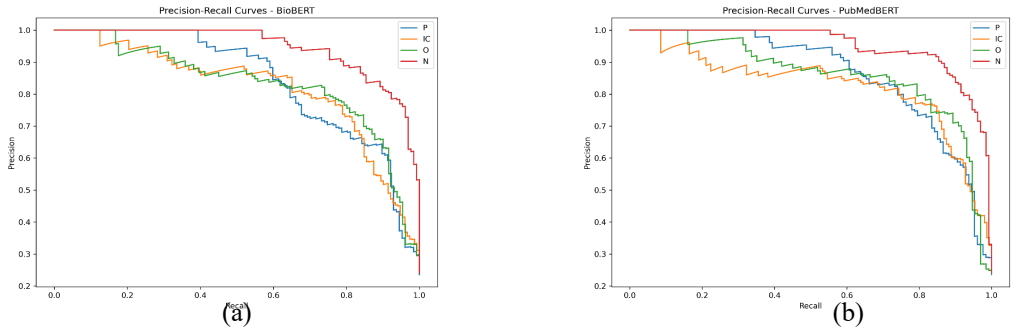
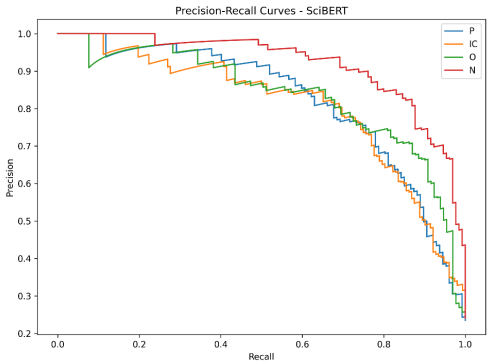


Figure 3. ROC curves for BioBERT-BiLSTM, PubMedBERT-BiLSTM, and SciBERT-BiLSTM. (a) BioBERT-BiLSTM. (b) PubMedBERT-BiLSTM. (c) SciBERT-BiLSTM.

Figure 3. ROC Curves for BioBERT–BiLSTM, PubMedBERT–BiLSTM, and SciBERT–BiLSTM Models (a) BioBERT–BiLSTM (b) PubMedBERT–BiLSTM (c) SciBERT–BiLSTM Receiver Operating Characteristic (ROC) curves illustrating classification performance for PICO element identification. PubMedBERT achieved the largest area under the curve (AUC), indicating superior discrimination between relevant and non-relevant sentences, while SciBERT displayed broader sensitivity across categories.





(c)

Figure 4. Precision–Recall (PR) curves for the proposed models.
(a) BioBERT-BiLSTM. (b) PubMedBERT-BiLSTM. (c) SciBERT-BiLSTM.

Figure 4. Precision–Recall (PR) Curves for BioBERT–BiLSTM, PubMedBERT–BiLSTM, and SciBERT–BiLSTM Models (a) BioBERT–BiLSTM (b) PubMedBERT–BiLSTM (c) SciBERT–BiLST Precision–Recall (PR) curves showing classification behavior across PICO elements. PubMedBERT obtained the highest precision across all categories, whereas SciBERT exhibited wider recall coverage, demonstrating complementary strengths in handling heart disease clinical text.

3.2. Discussion

The comparative evaluation of BioBERT, PubMedBERT, and SciBERT, each integrated with the BiLSTM architecture, underscores the critical influence of pretraining domain on the classification of PICO elements in cardiology-related texts. A concise summary of the primary results is provided in Table 8, highlighting model-wise differences in predictive behavior, precision–recall balance, and domain adaptability.

Table 8. Summary of model performance across evaluation metrics

Model	Accuracy	Precision	Recall	F1-Score	Hamming Loss
BioBERT-BiLSTM	0.694	0.806	0.721	0.756	0.117
PubMedBERT-BiLSTM	0.732	0.841	0.738	0.777	0.109
SciBERT-BiLSTM	0.728	0.827	0.746	0.784	0.104

Table 8 summarizes the comparative performance of the three models across all evaluation metrics. PubMedBERT–BiLSTM achieved the highest accuracy and precision, reflecting the advantage of its exclusive pretraining on PubMed abstracts. The consistent linguistic patterns and standardized biomedical vocabulary of PubMed contribute to a more uniform embedding space, enabling the model to recognize domain-specific terminology with high fidelity and to reduce false-positive predictions. This focused domain exposure enhances lexical discrimination, thereby explaining PubMedBERT's superior precision in identifying *Population*, *Outcome*, and *Not Relevant* sentences.

In contrast, SciBERT–BiLSTM attained the highest recall, a result attributable to its pretraining on the heterogeneous Semantic Scholar corpus encompassing multiple scientific disciplines. This broader exposure allows SciBERT to generalize more effectively, capturing a wider range of linguistic variations, including expressions that deviate from canonical biomedical phrasing. However, this generalization introduces a modest decline in precision, as the model occasionally misclassifies non-relevant sentences containing generic comparative or procedural terminology. BioBERT–BiLSTM, on the other hand, demonstrated balanced but slightly lower performance. This outcome can be attributed to its dual-domain pretraining on PubMed abstracts and PMC full-text articles, where longer narrative structures and stylistic variability may dilute precision while maintaining moderate recall.

Error analysis further revealed that the Intervention and Comparison category remained the most challenging across all models, producing 67, 61, and 48 false negatives for BioBERT, PubMedBERT, and SciBERT, respectively. These findings are consistent with prior research emphasizing the linguistic complexity of intervention-related expressions, which often include implicit comparisons, nested procedures, and multi-drug combinations that lack explicit comparators [16], [24], [29]. Conversely, the

Outcome and *Not Relevant* categories were classified most reliably, suggesting that explicit outcome markers (e.g., "mortality," "reduction," "improvement") and non-evidential statements are easier for contextual encoders to identify accurately.

Similar trends have been observed in broader biomedical NLP studies. Gu et al. [32] reported that PubMedBERT outperformed BioBERT in biomedical named-entity recognition due to its tighter domain adaptation. Beltagy et al. [33] found that SciBERT achieved higher recall across heterogeneous datasets in clinical relation extraction tasks. Likewise, Li et al. [34] demonstrated that domain-specific embeddings enhance precision in concept extraction but may underperform in cross-domain applications. These converging results strengthen the conclusion that domain specialization enhances precision and stability, whereas general-domain exposure improves recall and adaptability.

From a practical perspective, these findings carry direct implications for evidence-based automation. The PubMedBERT–BiLSTM model is well-suited for automated evidence synthesis and systematic review pipelines that prioritize precision and interpretability, while the SciBERT–BiLSTM model may better support exploratory literature mining, where broader coverage and recall are desirable. The BioBERT–BiLSTM configuration offers a balanced alternative for general biomedical classification tasks, providing a compromise between domain specialization and generalization.

Despite these promising outcomes, all models were constrained by the modest dataset size (2,697 sentences) and the limited diversity of the PICO-HD corpus. Future research should explore hybrid or ensemble approaches that combine the specificity of PubMedBERT with the coverage of SciBERT, or leverage domain-adaptive fine-tuning using larger, more diverse cardiology datasets. Integrating **contextual** embeddings with symbolic or rule-based layers could further enhance interpretability and improve the identification of complex intervention phrases a recurring limitation identified in this study.

4. Conclusions

This study conducted a comprehensive comparative evaluation of BioBERT, PubMedBERT, and **SciBERT** integrated with a BiLSTM architecture for multi-label classification of PICO elements in cardiology-related texts. The findings confirmed that PubMedBERT–BiLSTM achieved the highest overall accuracy (73.2%) and precision (84.1%), demonstrating the advantage of domain-specific pretraining on PubMed abstracts. SciBERT–BiLSTM, while slightly lower in accuracy (72.8%), attained the best recall (74.6%) and highest F1-score (78.4%), indicating that its broader scientific coverage improved sensitivity. Meanwhile, BioBERT–BiLSTM achieved balanced performance, with an accuracy of 69.4% and an F1-score of 75.6%, suggesting stable representation of biomedical terminology.

Overall, the results emphasize that the pretraining domain exerts a strong influence on model performance. Domain-specific embeddings, such as PubMedBERT and BioBERT, are particularly effective for achieving precision and stability, while a general scientific model like SciBERT offers complementary strengths in recall. Consequently, model selection should be application-driven: PubMedBERT–BiLSTM is optimal for systems emphasizing false-positive reduction and interpretability, whereas SciBERT–BiLSTM is more suitable for semi-automated review systems requiring broader sensitivity.

Despite its contributions, this research is limited by the small dataset size and the focus on only three embedding models. Future investigations should incorporate larger and more diverse cardiology corpora, evaluate additional biomedical embeddings such as ClinicalBERT and BlueBERT, and examine advanced or ensemble architectures to improve robustness and generalizability.

In conclusion, this study demonstrates that the combination of contextual embedding models with a BiLSTM framework can effectively classify PICO elements in heart disease literature. Among the evaluated models, PubMedBERT–BiLSTM provided the best balance of precision and interpretability, underscoring the importance of domain-specific pretraining for reliable biomedical text mining. Practically, the proposed framework serves as a foundation for automated evidence synthesis systems, facilitating the structured extraction of clinically relevant information to support evidence-based guideline development. Nonetheless, the findings highlight the need for continued research into larger datasets and hybrid or ensemble models that integrate domain-specific and domain-general representations to enhance robustness and generalization in clinical natural language processing.

References

[1] A. Ramic-Catak, S. Mesihovic-Dinarevic, B. Prnjavorac, N. Naser, and I. Masic, "Public Health Dimensions of Cardiovascular Diseases (CVD) Prevention and Control – Global Perspectives and

Current Situation in the Federation of Bosnia and Herzegovina," *Mater. Sociomed.*, vol. 35, no. 2, pp. 88–93, 2023, doi: 10.5455/msm.2023.35.88-93.
[2] J. J. Cárdenas-Anguiano *et al.*, "Estimation of the Burden of Ischemic Heart Disease in the Tabasco Population, Mexico, 2013–2021," 2025. doi: 10.3390/ijerph22030423.
[3] M. Borges *et al.*, "Burden of Disease and Cost of Illness of Overweight and Obesity in Portugal," 2024. doi: 10.1159/000541781.
[4] B. Sakboonyarat and R. Rangsin, "Hospital Admission and Mortality Rates for Ischemic Heart Disease in Thailand: 2012–2021," 2024. doi: 10.1186/s13104-024-06803-x.
[5] A. K. Nagpal, A. Pundkar, A. Singh, and C. Gadkari, "Cardiac Arrhythmias and Their Management: An in-Depth Review of Current Practices and Emerging Therapies," 2024. doi: 10.7759/cureus.66549.
[6] A. Sharma, "Incidental Coronary Artery Disease on Routine CT Coronary Angiography –An Evidence-Based Approach," 2023. doi: 10.58489/2836-5917/005.
[7] R. Javid *et al.*, "Transcoronary Electrophysiological Parameters in Patients Undergoing Elective and Acute Coronary Intervention," 2023. doi: 10.1371/journal.pone.0281374.
[8] V. Kittipibul *et al.*, "Projecting the Benefit of Vericiguat in PARADIGM-HF and DAPA-HF Populations: Insights From the VICTORIA Trial," 2024. doi: 10.1002/ehf2.15134.
[9] X. Chen, M. Zeng, C. Chen, D. Zhu, L. Chen, and Z. Jiang, "Efficacy of Psycho-Cardiology Therapy in Patients With Acute Myocardial Infarction Complicated With Mild Anxiety and Depression," 2023. doi: 10.3389/fcvm.2022.1031255.
[10] A. Vahanian *et al.*, "2021 ESC/EACTS Guidelines for the Management of Valvular Heart Disease," 2022. doi: 10.4244/eij-e-21-00009.
[11] E. Hiraoka *et al.*, "JCS 2022 Guideline on Perioperative Cardiovascular Assessment and Management for Non-Cardiac Surgery," 2023. doi: 10.1253/circj.cj-22-0609.
[12] D. Kartchner *et al.*, "TrialSieve: A Comprehensive Biomedical Information Extraction Framework for PICO, Meta-Analysis, and Drug Repurposing," 2025. doi: 10.3390/bioengineering12050486.
[13] A. Ragnhildstveit *et al.*, "Intra-Operative Applications of Augmented Reality in Glioma Surgery: A Systematic Review," 2023. doi: 10.3389/fsurg.2023.1245851.
[14] A. M. Ngo, A. Donaghue, K. H. Weng, and E. T. Crehan, "Barriers and Facilitators for Addressing Sex Education for Autistic Individuals: A Systematic Review," 2024. doi: 10.1177/00224669241292045.
[15] M. Cumpston, J. E. McKenzie, V. Welch, and S. Brennan, "Strengthening Systematic Reviews in Public Health: Guidance in the <i>Cochrane Handbook for Systematic Reviews of Interventions</I>, 2nd Edition," 2022. doi: 10.1093/pubmed/fdac036.
[16] C. Witte, D. Schmidt, and P. Cimiano, "Comparing Generative and Extractive Approaches to Information Extraction From Abstracts Describing Randomized Clinical Trials," 2024. doi: 10.1186/s13326-024-00305-2.
[17] D. C. Moore and A. S. Guinigundo, "Biomarker-Driven Oncology Clinical Trials: Novel Designs in the Era of Precision Medicine," 2023. doi: 10.6004/jadpro.2023.14.3.16.
[18] B. P. Carlin and F. Nollevaux, "Bayesian Complex Innovative Trial Designs (CIDs) and Their Use in Drug Development for Rare Disease," 2022. doi: 10.1002/jcph.2132.
[19] A. Griessbach *et al.*, "Characteristics, Progression, and Output of Randomized Platform Trials," 2024. doi: 10.1001/jamanetworkopen.2024.3109.
[20] F. Gay *et al.*, "Clinical Outcomes Associated With Anti-Cd38-Based Retreatment in Relapsed/Refractory Multiple Myeloma: A Systematic Literature Review," 2025. doi: 10.3389/fonc.2025.1550644.
[21] L. A. Nkhata, A. Human, Q. Louw, and Y. Brink, "Psychometric Properties and Clinical Utility of Spinal Health Outcome Measures in School-Based Interventions Among Children and Adolescents: A Systematic Review Protocol," 2024. doi: 10.1136/bmjopen-2024-089929.
[22] L. E. Almeida, S. Zammuto, and D. F. López, "Evaluating Surgical Approaches for Hemimandibular Hyperplasia Associated With Osteochondroma: A Systematic Literature Review," 2024. doi: 10.3390/jcm13226988.
[23] Y. Zang *et al.*, "Evidence Mapping Based on Systematic Reviews of Repetitive Transcranial Magnetic Stimulation on the Motor Cortex for Neuropathic Pain," 2022. doi: 10.3389/fnhum.2021.743846.
[24] M. Cumpston, J. E. McKenzie, R. Ryan, E. Flemyng, J. Thomas, and S. Brennan, "Development of the InSynQ Checklist: A Tool for Planning and Reporting the Synthesis Questions in Systematic

Reviews of Interventions," 2023. doi: 10.1002/cesm.12036.
[25] Q. Wang, "PICO entity extraction for preclinical animal literature," *Syst. Rev.*, vol. 11, no. 1, 2022, doi: 10.1186/s13643-022-02074-4.
[26] O. Sanchez-Graillet, C. Witte, F. Grimm, and P. Cimiano, "An Annotated Corpus of Clinical Trial Publications Supporting Schema-Based Relational Information Extraction," 2022. doi: 10.1186/s13326-022-00271-7.
[27] A. Dhrangadhariya and H. Müller, "Not So Weak PICO: Leveraging Weak Supervision for Participants, Interventions, and Outcomes Recognition for Systematic Review Automation," 2023. doi: 10.1093/jamiaopen/ooac107.
[28] A. Jolly, V. Pandey, I. Singh, and N. Sharma, "Exploring Biomedical Named Entity Recognition via SciSpaCy and BioBERT Models," *Open Biomed. Eng. J.*, vol. 18, no. 1, 2024, doi: 10.2174/0118741207289680240510045617.
[29] V. Dóczy, B. W. Sódar, Á. Hölgyesi, G. Merész, and P. Gaál, "Development, Testing, and Implementation of a New Procedure to Assess the Clinical Added Benefit of Pharmaceuticals," 2022. doi: 10.1017/s0266462322000411.
[30] K. Eberle *et al.*, "The PICO Puzzle: Can Public Data Predict EU HTA Expectations for All EU Countries?," 2025. doi: 10.3390/jmahp13030032.
[31] O. Ahmad, "PICO medical literatures on heart disease data," 2021. Accessed: Jan. 17, 2025. [Online]. Available: https://www.kaggle.com/datasets/owaiskhan9654/pico-medical-literatures-related-to-heart-disease
[32] Y. Gu *et al.*, "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing," *ACM Trans. Comput. Healthc.*, vol. 3, no. 1, Oct. 2021, doi: 10.1145/3458754.
[33] I. Beltagy, K. Lo, and A. Cohan, "{S}ci{BERT}: A Pretrained Language Model for Scientific Text," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds., Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3615–3620. doi: 10.18653/v1/D19-1371.
[34] J. Li *et al.*, "A comparative study of pre-trained language models for named entity recognition in clinical trial eligibility criteria from multiple corpora," *BMC Med. Inform. Decis. Mak.*, vol. 22, no. Suppl 3, p. 235, 2022, doi: 10.1186/s12911-022-01967-7.