# Domain-Adaptive Fine-Tuning of BioMedBERT for Medical Text Classification

**Ghulam Asrofi Buntoro**
Department of Electrical Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
7022201009@student.its.ac.id

**Oddy Virgantara Putra**
Department of Informatics, Universitas Darussalam Gontor, Ponorogo, Indonesia
oddy@unida.gontor.ac.id

**Mauridhi Hery Purnomo**
Department of Electrical Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia | Department of Computer Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
hery@ee.its.ac.id (corresponding author)

*Received: 28 June 2025 | Revised: 17 August 2025 and 9 September 2025 | Accepted: 11 September 2025*

 *DOI: https://doi.org/10.48084/etasr.13026*

## ABSTRACT

**Accurate classification of medical notes and texts is a critical task for improving biomedical information retrieval and decision support systems. In this study, we propose a hybrid deep learning model combining BioMedBERT with Cross-Attention and BiLSTM, aimed at enhancing the classification performance of disease-related abstracts across five categories. The proposed model was evaluated using a dataset comprising 14k annotated samples derived from scientific medical literature. The proposed architecture achieves a macro F1-score of 63.82, outperforming traditional methods such as sentence embedding models (SimCSE, SBERT), zero-shot entailment approaches, and BioBERT variants integrated with MLP classifiers. Findings show that while the model effectively distinguishes between categories such as neoplasms and cardiovascular diseases, challenges persist in classifying abstracts with overlapping semantics, particularly general pathological conditions. This research demonstrates the efficacy of combining domain-specific language models with sequence and attention mechanisms, proposing a viable method for scalable and interpretable biomedical text classification.**

***Keywords-BioMedBERT; domain-adaptive fine-tuning; machine learning; medical text classification; Natural Language Processing; text classification models***

## I. INTRODUCTION

The exponential growth of biomedical literature presents significant challenges for clinicians and researchers in efficiently accessing relevant information. Automated classification of medical abstracts into disease-specific categories has become essential for streamlining literature retrieval, supporting clinical decision-making, and improving biomedical knowledge management systems [1]. Traditional Machine Learning (ML), such as SVMs and Naive Bayes, often fall short in capturing the complex semantics and contextual subtleties of medical texts. While Deep Learning (DL) models like CNNs and LSTM have offered improvements, they still struggle to model long-range dependencies and detailed contextual relationships [2]. Models built on Transformer architecture, especially BERT and its domain-specific variants like BioBERT, ClinicalBERT, and PubMedBERT, have significantly advanced Natural Language Processing (NLP) in the biomedical domain. However, challenges persist in multi-class classification of abstracts, especially when disease categories exhibit overlapping semantic features [3]. Recent advancements in biomedical NLP have increasingly emphasized the importance of domain-adaptive fine-tuning for improving task-specific performance [4]. While general-purpose models such as SimCSE and SBERT provide efficient sentence embeddings, they often underperform in domain-specific, multi-class classification tasks due to their lack of targeted fine-tuning [5, 6]. To address this, models like BioBERT and BioALBERT have been developed using large biomedical corpora, demonstrating notable improvements in tasks such as Named Entity Recognition (NER) and medical text categorization [7]. Further studies have proposed strategies like pretraining from scratch on domain-specific corpora, outperforming generic model adaptation approaches [8]. ClinicalBERT has shown effectiveness when fine-tuned on clinical notes, reinforcing the value of task-specific tuning [9]. More sophisticated techniques such as multi-step transfer learning [10] have also been explored to incrementally adapt models to different biomedical tasks.

Additionally, the integration of external knowledge sources, such as knowledge graphs with BERT-based models, has been shown to enrich semantic understanding and further enhance classification performance [11]. These collective efforts underscore the trend toward leveraging domain-specific resources and architectures to improve the adaptability and efficacy of biomedical NLP systems.

The primary contributions of the proposed work are:

- Novel Hybrid Architecture: A new model, combining BioMedBERT, Cross-Attention, and BiLSTM, to capture both contextual and sequential features with improved interpretability is proposed.
- High Classification Accuracy: A macro F1-score of 63.82 is achieved, outperforming advanced baselines, such as SimCSE, SBERT, and zero-shot models.
- Robust in Complex Categories: Difficult medical categories (e.g., neoplasms, cardiovascular) are effectively classified, despite challenges with overlapping semantic classes.
- Customized Biomedical Corpus: Enhanced domain adaptation using a hybrid dataset of public biomedical literature and local Indonesian clinical texts.
- Validated Cross-Attention: An ablation study confirmed that Cross-Attention yields better and more stable performance than other attention mechanisms.
- Real-World Evaluation: The proposed model was tested on an imbalanced dataset of 14k+ clinical abstracts to prove its robustness.
- Scalability and Future Directions: Computational and generalization challenges are identified along with the proposed future work involving hierarchical labeling, data augmentation, and explainability.

## II. METHODOLOGY

The proposed framework leverages a multimodal architecture that integrates BioMedBERT, cross-attention, and BiLSTM to improve the classification of medical-related abstracts. The overall architecture of the proposed methodology is illustrated in Figure 1, which depicts the step-by-step process: It begins with data acquisition, collecting medical abstracts from publicly available datasets. After that, data preparation, text cleaning, tokenization, and normalization take place. The next steps involve BioMedBERT pre-training, input tokenization, embeddings generation, and contextual representation learning via Transformer Encoders with cross-attention. The next step is fine-tuning BioMedBERT, applying the model on the medical dataset, and incorporating cross-attention, BiLSTM for sequential dependency modeling, and a classification layer. The final stage is performance evaluation considering accuracy, precision, recall, and F1-score. The model's performance is evaluated against existing methods, including SimCSE+RoBERTa Large, SBERT(all-MiniLM-L6-v2), Lbl2Vec, Lbl2TransformerVec (SimCSE), Zero-shot Entailment (DistilBERT), (BART-large), (DeBERTa), BioMedBERT + Linear, BioMedBERT + MLP.

### A. Data Acquisition

The public dataset used in this study was obtained from [12]. The dataset contains 14k medical abstracts categorized into five types of diseases: neoplasms (NP), digestive system diseases (DSD), nervous system diseases (NSD), cardiovascular diseases (CD), and general pathological conditions (GPC). Each abstract is labelled with a "condition name," and the dataset was partitioned into separate training and testing subsets to facilitate model evaluation. To enhance domain adaptation, an additional biomedical corpus was constructed by combining public biomedical texts from the PubMed Central Open Access (PMC-OA) repository with anonymized clinical symptom descriptions from an Indonesian hospital. This combined dataset, after preprocessing and cleaning, resulted in approximately 50 million sentences and reflects both global and localized medical language variations.

### B. Preprocessing

Text preprocessing is essential in preparing data for domain-adaptive fine-tuning, particularly for models like BioMedBERT in medical text classification. Medical texts are often unstructured and noisy, requiring cleaning to remove irrelevant characters (e.g., symbols, punctuation) that could obscure meaningful patterns. Key steps include:

- Tokenization: Splitting text into tokens (words/subwords) using BioMedBERT's WordPiece tokenizer to preserve biomedical terminology and ensure compatibility with the model input [13].
- Normalization: Making text consistent by converting to lowercase and standardizing terms (e.g., replacing "HTN" with "hypertension") to reduce variation and ambiguity. These steps improve data quality, enabling the model to learn more effectively from the input [14].

### C. Pre-Training BioMedBERT

BioMedBERT represents a specialized variant of the BERT architecture, refined through fine-tuning on extensive biomedical text corpora to effectively manage the complex nature of medical language and terminology [15]. As a transformer-based model, it uses attention mechanisms to capture contextual relationships, making it suitable for applications such as medical text classification [16]. Its specialized training allows for more accurate interpretation of domain-specific terms (e.g., hypertension, myocardial infarction), improving performance in clinical NLP applications [17]. During the training phase, the model simultaneously minimizes the losses associated with masked language modeling and next sentence prediction tasks, employing cross-entropy loss functions to strengthen its ability to capture contextual relationships within the text [13].

$$\text{NSP loss} = \left(-\frac{1}{N}\sum_{i=1}^{N} y.\log(p) + (1-y)\log(1-p)\right) \quad (1)$$

$$\text{MaskedLM loss} = \left(-\sum_{i=1}^{N} q(y).\log\big(q(y)\big)\right) \quad (2)$$

$$\text{L} = \text{NSP loss} + \text{MaskedLM loss} \quad (3)$$

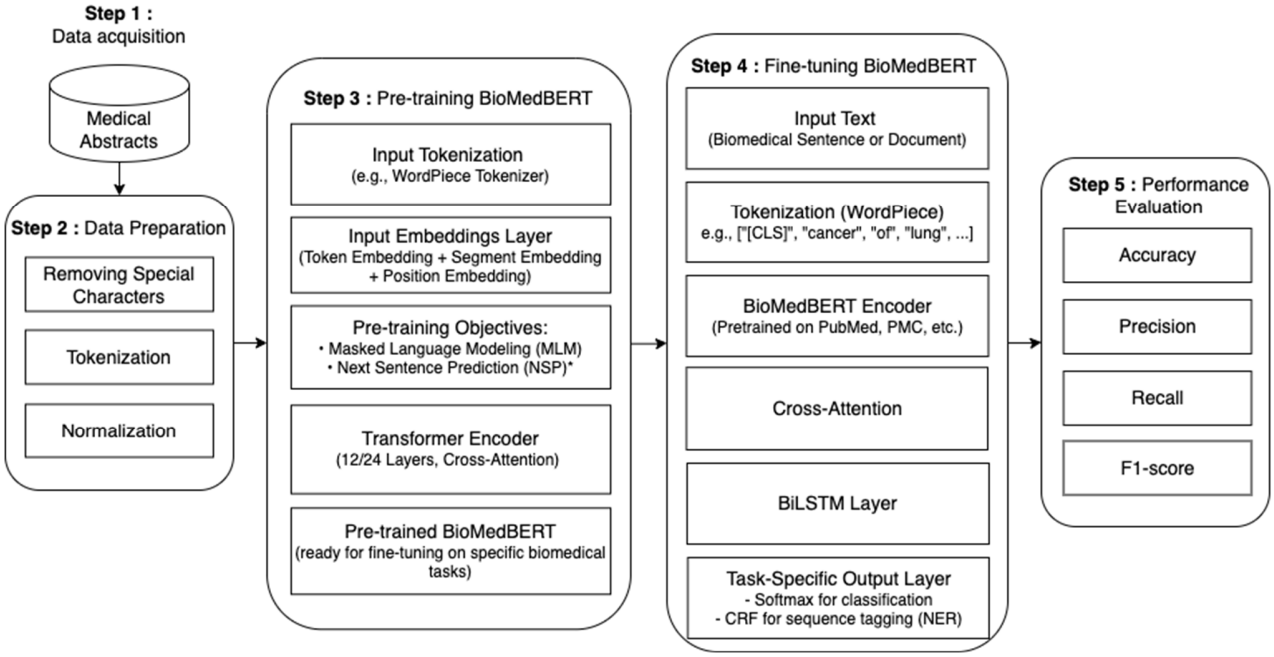


Fig. 1. The proposed framework uses a multimodal approach to improve text classification.

### *D. Attention Mechanism*

The attention mechanism enables neural networks to focus on the most relevant parts of an input sequence, addressing the limitations of models like RNNs and LSTM in capturing long-range dependencies [18]. It works by computing a weighted sum of input representations, where weights are determined by the similarity between queries and keys [19]. The commonly used form is the scaled dot-product attention [20]:

$$\text{Attention}(\text{Q}, \text{K}, \text{V}) = \text{softmax}\left(\frac{QK^T}{\sqrt{D}}\right)V \quad (4)$$

where Q (queries), K (keys), and V (values) are linear transformations of the input and D is the dimension of the key vectors (used for scaling).

Cross-attention, in particular, enhances model performance by aligning different input representations, improving contextual understanding in complex tasks [21]. Algorithm 1 showcases the followed steps:

```
Algorithm 1 Cross-Attention Fusion
Require: Context vectors c1, c2 ∈ R^(B×D)
Ensure: Fused output ∈ R^(B×D)
1: q ← Unsqueeze(c1)
2: k ← Unsqueeze(c2)
3: v ← Unsqueeze(c2)
4: score ← QK^T / √D
5: α ← softmax(score)
6: attended ← α · v
7: attended ← Squeeze(attended
8: fused ← LayerNorm(attended+c1)
9: return fused
```

### *E. Fine-Tuning BioMedBERT*

Fine-tuning serves as a transfer learning approach wherein a pre-trained model, such as BioMedBERT, is further trained on a smaller, task-specific labeled dataset. This additional training adjusts the model's internal parameters and classification layer, enhancing its capacity to recognize and interpret domain-specific linguistic patterns and specialized medical terminology. Success depends on the quality of the labeled data, which should represent the target classification categories (e.g., diagnosis, treatment) [22]. Fine-tuning enables the model to make accurate predictions by aligning its general language understanding with the specific needs of the classification task [23].

### *F. Performance Evaluation*

To ensure accuracy and generalizability, medical text categorization models must be evaluated on a separate validation set. Metrics like as accuracy, precision, recall, and especially F1-score provide a comprehensive view of performance [24]. The F1-score, defined as the harmonic mean of precision and recall, is especially well-suited for addressing class imbalance as it provides a balanced evaluation of both false positive and false negative rates [25].

All experiments were conducted with the following hyperparameters: Learning Rate = 1e-4, Batch size = 64, Input dimension = 768, Optimizer = Adam, Weight decay = 1e-4, Scheduler = Cosine Annealing, Epochs = 25, with the following machine specifications: GPU: RTX 4090, OS: Ubuntu 22.04 LTS, RAM: 32GB, Processor: 12th Gen Intel(R) Core(TM) i7-12700K.

## III. RESULTS AND DISCUSSION

The main objective of this study is to categorize medical abstracts into five groups that correspond to various patient

conditions. The analysis reveals that general pathological conditions are the most common, comprising 4.805 records or 33.28% of the data. Neoplasms follow at 21.9%, then cardiovascular diseases at 21.13%. Nervous system diseases make up 13.33%, and digestive system diseases account for the smallest share at 10.34%.

We conducted seven experimental scenarios combining CNN and LSTM with varying configurations as displayed in Table I. Scenarios 1–3 used only BiLSTM layers with 32, 64, and 128 units, respectively. Scenarios 4–7 integrated CNN layers with increasing filter sizes (64 to 512) followed by a BiLSTM with 128 units. These setups were designed to identify the optimal architecture for text classification prior to applying the Cross-Attention mechanism. The proposed hybrid classification model is expected to perform as well as existing models. The most effective setup of Scenario 1 uses a pure LSTM architecture with 32 units and no CNN layer. This model was tested, and the results are shown in Tables II and III. On average, the models across all scenarios achieved 59.13% accuracy. Scenario 1 performed the best with an accuracy of 63.61%, followed by Scenario 2 at 62%, and Scenario 3 at 60%. The Digestive System Diseases label had the weakest performance.

TABLE I. EXPERIMENT SCENARIO

| Scenario | Characteristics of Hybrid Model |
|---|---|
| 1 | Pure LSTM (32 unit), no CNN layer |
| **2** | Pure LSTM (64 unit), no CNN layer |
| **3** | Pure LSTM (128 unit), no CNN layer |
| **4** | CNN (64 filters) + LSTM (128 units) |
| **5** | CNN (128 filters) + LSTM (128 units) |
| **6** | CNN (256 filters) + LSTM (128 units) |
| **7** | CNN (512 filters) + LSTM (128 units) |

TABLE II. CONFUSION MATRIX PERFORMANCE OF SCENARIO 1

| Scenario 1 | | Predicted | | | | |
|---|---|---|---|---|---|---|
| | | **NP** | **DSD** | **NSD** | **CD** | **GPC** |
| **Actual** | **NP** | **532** | 30 | 19 | 8 | 44 |
| | **DSD** | 30 | **211** | 3 | 3 | 52 |
| | **NSD** | 30 | 8 | **234** | 25 | 88 |
| | **CD** | 14 | 9 | 19 | **458** | 110 |
| | **GDC** | 150 | 127 | 104 | 172 | **408** |

TABLE III. TRAINING PERFORMANCE OF THE PROPOSED MODEL

| Scenario | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) |
|---|---|---|---|---|
| 1 | 63.61 | 62.38 | 66.42 | 63.82 |
| **2** | 62.08 | 61.09 | 63.88 | 62.22 |
| **3** | 60.77 | 59.46 | 60.92 | 59.99 |
| **4** | 59.66 | 59.02 | 59.99 | 59.28 |
| **5** | 56.96 | 56.88 | 56.48 | 56.57 |
| **6** | 56.93 | 58.35 | 56.93 | 56.31 |
| **7** | 53.95 | 53.24 | 55.59 | 54.25 |

Despite this, the hybrid model demonstrated a strong ability to understand each class as displayed in Table III. It achieved an average sensitivity of 66%, with the highest performance in cardiovascular diseases (75%), followed by neoplasms (70.4%), digestive system diseases (70%), nervous system diseases (60%), and general pathological conditions (42%). These results are supported by corresponding precision and F1 scores, which show similar patterns. This suggests that the Cross-Attention and BiLSTM model successfully leverages the strengths of both architectures. These findings prove the Cross-Attention and BiLSTM model's value in tackling complex text classification tasks like those involving medical abstracts.

The classification report shows strong model performance, particularly for major classes like NP and CD diseases, as reflected in high F1-scores and accurate predictions along the confusion matrix's performance of scenario 1 shown in the confusion matrix of Table II. However, the model struggles with the GPC class due to its broad and ambiguous nature, leading to frequent misclassifications. For example, abstracts describing non-specific inflammation contain keywords common to multiple classes. Future improvements could include label refinement or hierarchical classification to address this ambiguity.

To validate performance, the proposed model was compared with several baseline methods using the same Medical Abstracts dataset. Results across the experimental scenarios (S1–S7) confirm that the proposed model outperforms previous approaches in classification accuracy.

Figure 2 shows the validation loss curves for the experimental scenarios, where lower values indicate better generalization. The black curve (scenario_1_exp, Pure LSTM 32 units) performs best, with loss dropping from ~1.5 to below 0.85 and staying low. Scenario 7 (pink, CNN 512 + LSTM 128) shows severe overfitting after epoch 8, while scenario 6 and scenario 5 show milder overfitting from epoch 10, scenario 2 stays close to scenario 1, and scenarios 3 and 4 plateau near 0.9 after early gains. CNN-LSTM setups tend to overfit due to high parameter counts. In contrast, BiomedBERT + Cross-Attention + BiLSTM converges quickly and stabilizes near 0.88 with minor oscillations. Overall, simpler recurrent models maintain low loss, while high-capacity CNN-LSTM hybrids need stronger regularization.

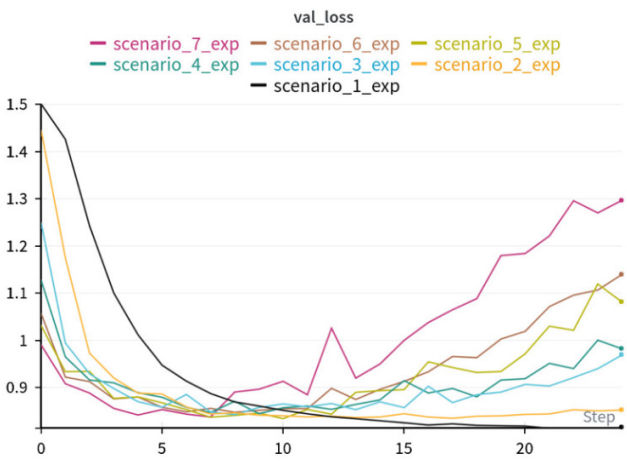


Fig. 2. Validation loss curves during model training.

Table V shows that the proposed BioMedBERT + Cross-Attention + BiLSTM model achieved the highest performance with an F1 score of 63.82%, outperforming other configurations such as BioMedBERT + Linear (58.2%) and BioMedBERT + MLP (60.7%). This model also outperforms the results of [12], who also used a dataset of medical abstracts using a zero-shot entailment approach, which achieved a result of 57.28%. This study confirms the superiority of integrating domain-adaptive pre-training with the Cross-Attention mechanism for medical abstract classification. Furthermore, the domain-specific model clearly outperformed SimCSE, SBERT, and the zero-shot approach (F1: 34–46%) . Even advanced zero-shot models such as DeBERTa and BART-large (F1: 56–57%) were unable to outperform the proposed approach .

TABLE IV. COMPARATIVE PERFORMANCE METRICS

| Label | Precision (%) | Recall (%) | F1-Score (%) |
|---|---|---|---|
| **NP** | 70 | 84 | 76 |
| **FDF** | 54 | 70 | 61 |
| **NSD** | 61 | 60 | 61 |
| **CD** | 68 | 75 | 71 |
| **GPC** | 58 | 42 | 49 |

TABLE V. PERFORMANCE COMPARISON

| Model | F1-Score (%) |
|---|---|
| SimCSE+RoBERTa Large | 34.94 |
| SBERT(all-MiniLM-L6-v2) | 46.53 |
| Lbl2Vec | 43.03 |
| Lbl2TransformerVec (SimCSE) | 39.6 |
| Zero-shot Entailment (DistilBERT) | 27.74 |
| Zero-shot Entailment (BART-large) | 56.86 |
| Zero-shot Entailment (DeBERTa) | 57.28 |
| BioMedBERT + Linear | 58.2 |
| BioMedBERT + MLP | 60.7 |
| **BioMedBERT + Cross-Attention + BiLSTM** | **63.82** |

### *A. Ablation Study*

To evaluate the role of the attention mechanism, an ablation study was conducted by replacing the Cross-Attention module with alternatives GMU, Single Attention, and Stacked Attention while keeping the base model (BioMedBERT + CNN + LSTM) unchanged. Each variant was tested using 5-fold cross-validation, with performance measured by accuracy and F1-score. The ablation study results shown in Table VI, demonstrate the superior effectiveness of Cross-Attention.

The boxplots of accuracy, as shown in Figure 3, and F1 score, as in Figure 4, across five folds consistently highlight Cross-Attention as the best-performing model, achieving the highest median values (~64.2% accuracy and ~64.0% F1) with low variability. This indicates that Cross-Attention not only performs well in terms of classification accuracy but also maintains a strong balance between precision and recall, making it the most reliable and stable choice among the evaluated models. In comparison, GMU and Attention Layer exhibit lower performance and higher inconsistency, while Stacked Attention offers competitive results, but with slightly more fluctuation across folds.

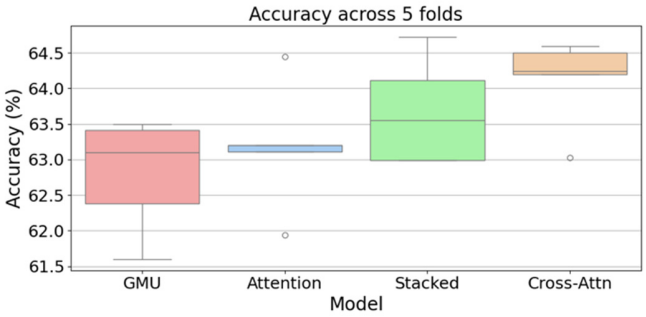


Fig. 3. Accuracy boxplot for four models (GMU, Attention, Stacked Attention, and Cross-Attention).

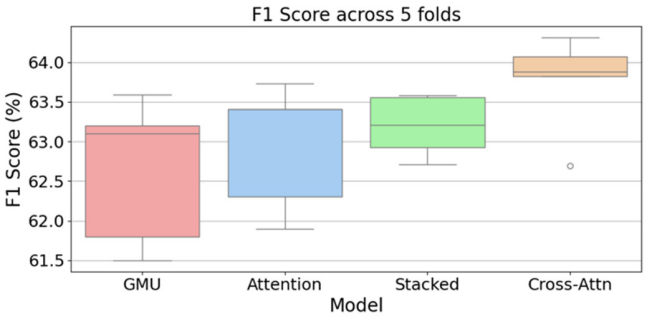


Fig. 4. F1 score boxplot across 5 folds for four models (GMU, Attention, Stacked Attention, and Cross-Attention).

The Cross-Attention mechanism outperformed all alternative attention variants, achieving the highest average accuracy and F1-score with the lowest variance. This confirms its ability to better align abstract features with class representations, capturing subtle semantic nuances. While GMU and Stacked Attention showed moderate results, they lacked the precision and adaptability of Cross-Attention, reinforcing its value for domain-specific tasks like medical abstract classification [26].

TABLE VI. ABLATION STUDY RESULT

| No. | Model (BioMedBERT+CNN+LSTM) | Metric | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Mean | Std |
|---|---|---|---|---|---|---|---|---|---|
| **1.** | **Gated Multimodal Unit (GMU)** | Acc | 63.1 | 63.41 | 63.5 | 62.36 | 61.6 | 62.8 | ± 0.8013 |
| | | F1 | 63.2 | 63.59 | 63.1 | 61.5 | 61.8 | 62.64 | ± 0.9264 |
| **2.** | **Attention Layer** | Acc | 63.11 | 64.45 | 63.2 | 63.2 | 61.94 | 63.18 | ± 0.8883 |
| | | F1 | 63.41 | 63.73 | 62.3 | 63.41 | 61.9 | 62.95 | ± 0.7995 |
| **3.** | **Stacked Attention** | Acc | 63.55 | **64.72** | 64.11 | 62.99 | 62.99 | 63.67 | ± 0.7475 |
| | | F1 | 63.56 | 63.56 | 63.21 | 62.93 | **62.93** | 63.20 | ± 0.3831 |
| **4.** | **Cross-Attention** | Acc | **64.20** | 64.50 | **64.59** | **63.03** | **63.03** | **64.11** | ± **0.6273** |
| | | F1 | **63.88** | **64.31** | **64.07** | **63.82** | 62.69 | **63.75** | ± **0.6255** |

### *B. Sample of Misclassification and Feature Ambiguity*

Figure 5 illustrates the relationship between prediction confidence and accuracy on the test set. Accuracy remains below 60% for confidence scores under 0.60, after which it increases steadily, surpassing 80% for confidence above 0.85. The choice of 0.60 as the low-confidence threshold was empirically derived by identifying the inflection point in this curve, where accuracy transitions from consistently low to steadily improving, rather than being chosen arbitrarily. This approach aligns with selective classification principles, where predictions below a calibrated threshold are flagged as uncertain to reduce predictive risk. In [27], it was shown that abstaining from low-confidence predictions is particularly beneficial in top-ambiguity cases where the top two class probabilities are similar. Authors in [28] demonstrated that softmax probability, despite known calibration limitations, remains an effective and computationally efficient method.

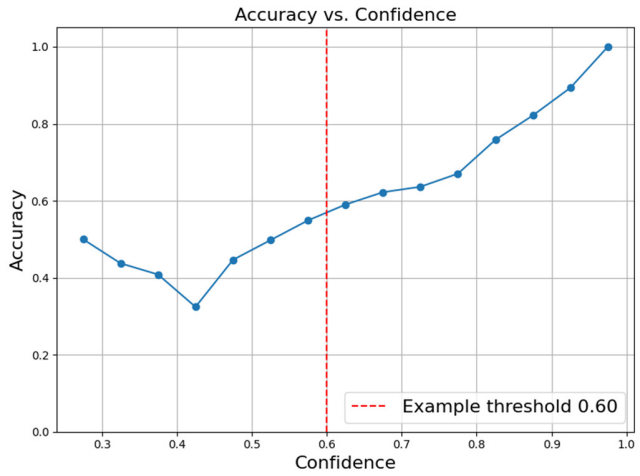


Fig. 5. Prediction Confidence distribution.

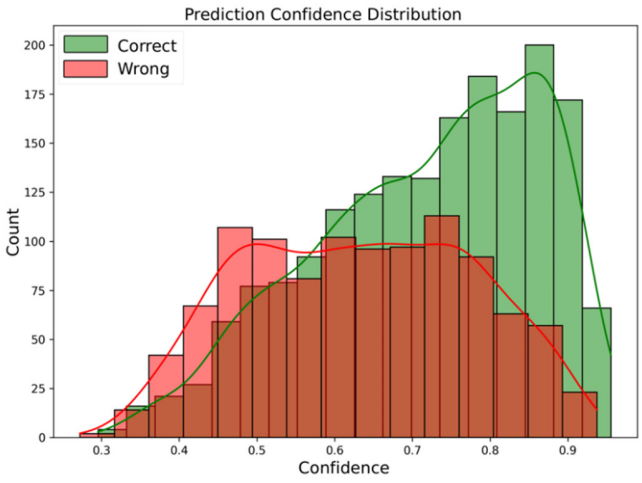


Fig. 6. Accuracy vs confidence plot.

Figure 6 presents the confidence distribution for correct (green) and incorrect (red) predictions. Correct predictions cluster at higher confidence levels (0.80–0.90), whereas incorrect predictions are more prevalent in the mid-confidence range, particularly between 0.45–0.55 and 0.65–0.75. The overlap between correct and incorrect predictions from 0.55 to 0.75 marks a zone of uncertainty, supporting the use of the 0.60 threshold. A small but notable fraction of high-confidence errors (>0.85) points to potential label quality issues or systematic feature overlap; for example, some abstracts labeled as cardiovascular diseases contain terminology ("ischemic," "vascular") that strongly overlaps with general pathological conditions.

Table VII lists the five most frequent misclassification patterns after excluding correct classifications. The model most often confused general pathological conditions with cardiovascular diseases (7.55% of all errors), followed by general pathological conditions with neoplasms (4.81%), and general pathological conditions with nervous system diseases (4.33%). These frequent confusions align with the notion of feature ambiguity, where overlapping clinical terminology and shared symptom descriptions make categories inherently difficult to distinguish. For instance, abstracts labeled as *general* pathological conditions often contain cardiovascular-related terms such as "ischemic" or "vascular," contributing to cross-class similarity.

TABLE VII. MOST FREQUENT MISCLASSIFICATIONS

| True Label | Predicted Label | Count | Errors (%) |
|---|---|---|---|
| GPC | CD | 218 | 7.55 |
| GPC | NP | 139 | 4.81 |
| GPC | NSD | 125 | 4.33 |
| NSD | GPC | 63 | 2.18 |
| NP | GPC | 58 | 2.01 |

In summary, this analysis reveals that many misclassifications stem from semantically overlapping categories and that a significant proportion of errors occur in low-confidence or small-margin cases. Targeted data cleaning, expanding the representation of ambiguous class pairs, and further model calibration could improve both predictive accuracy and interpretability.

## IV. CONCLUSION

This study proposed a medical text classification model that enhances the BioMedBERT architecture with Cross-Attention and BiLSTM layers to improve contextual and sequential representation learning. Evaluated on a multi-class dataset of over 14k clinical abstracts, the model outperformed several state-of-the-art baselines, achieving an F1-score of 63.82. The results highlight the model's effectiveness in distinguishing disease-specific categories, especially in complex domains like neoplasms and cardiovascular diseases. Despite these improvements, challenges remain, particularly in classifying abstract categories with semantic overlap and imbalanced data, such as general pathological conditions. Moreover, the use of complex architectures introduces higher computational demands and risks of overfitting. The model also inherits the interpretability limitations common to deep learning approaches. Future work will focus on improving generalizability and

transparency through hierarchical labeling, domain-specific data augmentation, and explainability techniques to enhance the model's applicability across broader clinical and biomedical text classification tasks.

## REFERENCES


[1] H. Lu, L. Ehwerhemuepha, and C. Rakovski, "A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance," *BMC Medical Research Methodology*, vol. 22, no. 1, Jul. 2022, Art. no. 181, https://doi.org/10.1186/s12874-022-01665-y.

[2] D. Kurniasari, Warsono, M. Usman, F. R. Lumbanraja, and Wamiliana, "LSTM-CNN Hybrid Model Performance Improvement with BioWordVec for Biomedical Report Big Data Classification," *Science and Technology Indonesia*, vol. 9, pp. 273–283, Apr. 2024, https://doi.org/10.26554/sti.2024.9.2.273-283.

[3] J. Lee *et al.*, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020, https://doi.org/10.1093/bioinformatics/btz682.

[4] P. Su and K. Vijay-Shanker, "Investigation of improving the pre-training and fine-tuning of BERT model for biomedical relation extraction," *BMC Bioinformatics*, vol. 23, Dec. 2022, Art. no. 120, https://doi.org/10.1186/s12859-022-04642-w.

[5] O. M. Alyasiri and Y.-N. Cheah, "Multi-Class Text Classification using Machine Learning Techniques," *Engineering, Technology & Applied Science Research*, vol. 15, no. 3, pp. 22598–22604, June 2025, https://doi.org/10.48084/etasr.9994.

[6] D. Zheng, R. Han, F. Yu, and Y. Li, "Biomedical named entity recognition based on multi-cross attention feature fusion," *PLOS ONE*, vol. 19, no. 5, 2024, Art. no. e0304329, https://doi.org/10.1371/journal.pone.0304329.

[7] U. Naseem, M. Khushi, V. Reddy, S. Rajendran, I. Razzak, and J. Kim, "BioALBERT: A Simple and Effective Pre-trained Language Model for Biomedical Named Entity Recognition," in *2021 International Joint Conference on Neural Networks (IJCNN)*, Shenzhen, China, Jul. 2021, https://doi.org/10.1109/IJCNN52387.2021.9533884.

[8] H. Jin, C. Yao, W. Zhang, and H. Chen, "Strategic Medical Text Classification with Improved Blending Ensemble Learning," in *Artificial Intelligence and Robotics*, Singapore, 2025, pp. 296–305, https://doi.org/10.1007/978-981-96-2914-5_27.

[9] K. Huang, J. Altosaar, and R. Ranganath, "ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission." arXiv, Nov. 29, 2020, https://doi.org/10.48550/arXiv.1904.05342.

[10] T. Manaka, T. V. Zyl, D. Kar, and A. Wade, "Multi-step Transfer Learning in Natural Language Processing for the Health Domain," *Neural Processing Letters*, vol. 56, Jun. 2024, Art. no. 177, https://doi.org/10.1007/s11063-024-11526-y.

[11] Q. Lu, A. Wen, T. Nguyen, and H. Liu, "Enhancing Clinical Relevance of Pretrained Language Models Through Integration of External Knowledge: Case Study on Cardiovascular Diagnosis From Electronic Health Records," *JMIR AI*, vol. 3, no. 1, Aug. 2024, Art. no. e56932, https://doi.org/10.2196/56932.

[12] T. Schopf, D. Braun, and F. Matthes, "Evaluating Unsupervised Text Classification: Zero-shot and Similarity-based Approaches," in *Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval*, New York, NY, USA, Mar. 2023, pp. 6–15, https://doi.org/10.1145/3582768.3582795.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv, Oct. 2018, https://doi.org/10.48550/arXiv.1810.04805.

[14] C. P. Chai, "Comparison of text preprocessing methods," *Natural Language Engineering*, vol. 29, no. 3, pp. 509–553, 2023, https://doi.org/10.1017/S1351324922000213.

[15] S. Chakraborty, E. Bisong, S. Bhatt, T. Wagner, R. Elliott, and F. Mosconi, "BioMedBERT: A Pre-trained Biomedical Language Model for QA and IR," in *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online), Sep. 2020, pp. 669–679, https://doi.org/10.18653/v1/2020.coling-main.59.

[16] S. Madan, M. Lentzen, J. Brandt, D. Rueckert, M. Hofmann-Apitius, and H. Fröhlich, "Transformer models in biomedicine," *BMC Medical Informatics and Decision* Making, vol. 24, no. 1, July 2024, Art. no. 214, https://doi.org/10.1186/s12911-024-02600-5.

[17] L. Fang, Q. Chen, C.-H. Wei, Z. Lu, and K. Wang, "Bioformer: an efficient transformer language model for biomedical text mining," arXiv, 2023, https://doi.org/10.48550/arXiv.2302.01588.

[18] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, 2021, https://doi.org/10.1016/j.neucom.2021.03.091.

[19] L. Feng, F. Tung, H. Hajimirsadeghi, M. O. Ahmed, Y. Bengio, and G. Mori, "Attention as an RNN," arXiv, 2024, https://doi.org/10.48550/arXiv.2405.13956.

[20] A. Vaswani *et al.*, "Attention Is All You Need." arXiv, Aug. 02, 2023, https://doi.org/10.48550/arXiv.1706.03762.

[21] D. Li, L. Zhang, J. Huang, N. Xiong, L. Zhang, and J. Wan, "Enhancing zero-shot relation extraction with a dual contrastive learning framework and a cross-attention module," *Complex & Intelligent Systems*, vol. 11, no. 1, Jan. 2025, Art. no. 42, https://doi.org/10.1007/s40747-024-01642-6.

[22] R. Tinn *et al*., "Fine-tuning large neural language models for biomedical natural language processing," *Patterns*, vol. 4, no. 4, Apr. 2023, Art. no. 100729, https://doi.org/10.1016/j.patter.2023.100729.

[23] B. Nguyen and S. Ji, "Fine-Tuning Pretrained Language Models With Label Attention for Biomedical Text Classification," arXiv, https://doi.org/10.48550/arXiv.2108.11809.

[24] M. Lepore, E. Plenzich, R. Tufano, R. Cerulli, and R. Maccioni, "Improving patient's medical history classification using a feature construction approach based on situation awareness and granular computing," *Neural Computing and Applications*, vol. 36, no. 35, pp. 22461–22484, Dec. 2024, https://doi.org/10.1007/s00521-024-10413-w.

[25] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, "Handling imbalanced medical datasets: review of a decade of research," *Artificial Intelligence Review*, vol. 57, no. 10, Oct. 2024, Art. no. 273, https://doi.org/10.1007/s10462-024-10884-2.

[26] M. Gheini, X. Ren, and J. May, "Cross-Attention is All You Need: Adapting Pretrained Transformers for Machine Translation." arXiv, Sep. 14, 2021, https://doi.org/10.48550/arXiv.2104.08771.

[27] Q. Ding, Y. Cao, and P. Luo, "Top-Ambiguity Samples Matter: Understanding Why Deep Ensemble Works in Selective Classification," in *37th Conference on Neural Information Processing Systems (NeurIPS 2023)*, 2023.

[28] G. Xia and C.-S. Bouganis, "Augmenting the Softmax with Additional Confidence Scores for Improved Selective Classification with Out-of-Distribution Data," *International Journal of Computer Vision*, vol. 132, no. 9, pp. 3714–3752, Sept. 2024, https://doi.org/10.1007/s11263-024-02029-3.