Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.085268

ARTICLE

Knowledge Distillation for Biomedical Text Classification: A Systematic Comparative Analysis of Multiple Teacher–Student Architectures

Amine Gonca Toprak[1,*] **and Aytuğ Onan**[2]

[1]R&D Department, Türk Telekom, Ankara, Türkiye
[2]Computer Engineering Department, Izmir Institute of Technology, Izmir, Türkiye

*Corresponding Author: Amine Gonca Toprak. Email: aminegonca.toprak@turktelekom.com.tr

Received: 08 May 2026; Accepted: 23 July 2026; Published: 13 August 2026

ABSTRACT: Biomedical texts present significant challenges for natural language processing (NLP) due to their complex terminology, intricate contextual dependencies, and highly domain-specific semantics. This study investigates the effectiveness of knowledge distillation (KD) for biomedical text classification, aiming to develop lightweight, resource-efficient models that remain competitive with larger architectures. A balanced dataset of 25,000 PubMed records was constructed, equally distributed across five biomedical domains. Two teacher models (BERT and PubMedBERT) and five student models (DistilBERT, BioClinicalBERT, BioBERT, DistilBioBERT, and DistilRoBERTa) were evaluated across ten distinct KD configurations. Each student model was also directly fine-tuned to serve as a controlled baseline. Model performance was assessed using accuracy, precision, recall, and F1-score. The results show that KD can, under suitable teacher–student configurations, enable student models to surpass direct fine-tuning, while other configurations yield only marginal or comparable improvements. Among all configurations, BERT→DistilBERT achieved the highest performance, reaching 89% accuracy. Unexpectedly, the general-purpose BERT teacher outperformed the domain-specific PubMedBERT across multiple student models, suggesting that broader linguistic representations can transfer more effectively across diverse biomedical subdomains. Lower performance in certain KD settings, such as DistilBioBERT and DistilRoBERTa, was attributed to architectural mismatches and limited student capacity. These findings demonstrate that compact models can achieve strong biomedical classification performance through KD under compatible teacher–student pairings, while also highlighting that KD effectiveness varies substantially depending on the specific model combination.

KEYWORDS: Biomedical text classification; knowledge distillation; pretrained language models; natural language processing (NLP)

1 Introduction

Pretrained language models, particularly Transformer-based architectures, have transformed natural language processing (NLP) by enabling rich contextual and semantic representations across a wide range of tasks, including text classification, machine translation, information extraction, and dialogue systems [1–3]. By leveraging knowledge acquired during large-scale pretraining, LLMs can be adapted to downstream tasks through fine-tuning, reducing reliance on task-specific labeled data while maintaining strong generalization across diverse NLP applications [4,5].

The rapid digitalization of healthcare has led to an exponential growth in text-based biomedical resources, including clinical notes, electronic health records (EHRs), medical case reports, and large-scale bibliographic databases. These resources hold considerable potential for improving healthcare quality, supporting clinical decision-making, and accelerating biomedical research. However, since such content is typically available in unstructured natural language form, extracting actionable knowledge from it requires sophisticated NLP methods.

Among the core NLP tasks applied to biomedical text, classification occupies a particularly prominent role. It underpins a broad range of applications, including categorization of scientific literature by topic, disease classification, clinical document typing, and semantic indexing of biomedical records [6]. Biomedical text classification is especially critical in health informatics and clinical decision support, where timely and accurate information retrieval can directly influence research and patient outcomes. Nevertheless, biomedical text is characterized by high conceptual density, ambiguous terminology, rare domain-specific concepts, and context-sensitive language, all of which pose considerable challenges for general-purpose NLP models. Furthermore, the large parameter counts and substantial memory and compute requirements of LLMs make their direct deployment difficult in resource-constrained environments, underscoring the need for efficient, domain-adapted classification solutions [7].

Knowledge distillation (KD) has emerged as a promising approach to address these constraints. KD is a model compression technique in which a smaller student model is trained to approximate the output distributions of a larger, more capable teacher model, allowing the student to achieve competitive performance with significantly fewer parameters [8]. By learning from both ground-truth labels and the soft targets produced by the teacher, the student model can internalize richer representations than those obtained from supervised training alone [9]. Although KD has been widely adopted in general-domain NLP tasks and edge computing scenarios, its systematic evaluation in biomedical text classification remains limited, motivating further investigation.

This study investigates the effectiveness of KD for developing resource-efficient biomedical text classification models. A dataset comprising 25,000 PubMed records was constructed, and ten KD configurations involving two teacher models and five student models were evaluated under a unified experimental protocol. Each student model was also evaluated through direct fine-tuning to provide a controlled baseline. The analysis focuses on three main questions: how teacher type affects knowledge transfer, how student architecture influences KD performance, and to what extent teacher–student compatibility determines whether KD improves upon direct fine-tuning. Therefore, the primary contribution of this study lies not in proposing a new distillation objective, but in providing a systematic empirical comparison of KD behavior across multiple teacher–student combinations. The findings contribute to the understanding of efficient biomedical NLP under resource constraints, with implications for both research and deployment.

The remainder of this paper is organized as follows. Section 2 reviews related work on biomedical NLP and knowledge distillation. Section 3 describes the dataset construction, preprocessing, model architectures, and distillation strategy. Section 4 presents and discusses the experimental results. Section 5 concludes the paper with a summary of findings, and Section 6 outlines the limitations of the current study along with directions for future work.

2 Related Work

Biomedical text processing has attracted growing interest across both research and clinical communities, driven by the expanding volume of domain-specific textual resources such as health records, clinical documentation, and scientific literature.

Unlike general-purpose text, biomedical content is characterized by terminological density, semantic complexity, and a high degree of domain specificity, making standard NLP pipelines insufficient for reliable analysis. To address these characteristics, domain-adapted language models such as BioBERT [10], PubMedBERT [11], and ClinicalBERT [12] have been developed through pretraining on domain-specific corpora, enabling more accurate modeling of biomedical language compared to general-purpose counterparts.

Early approaches to biomedical text classification relied on statistical methods such as Support Vector Machines (SVM) and Naive Bayes, which were eventually superseded by deep learning architectures including Long Short-Term Memory (LSTM) networks, Convolutional Neural Networks (CNN), and Gated Recurrent Units (GRU) [13–15]. The subsequent adoption of Transformer-based models further advanced classification performance; however, these architectures impose substantial memory and computational demands that restrict their applicability in latency-sensitive or resource-constrained deployments [16–18].

Table 1 presents a comparative overview of the representative studies discussed in this section, highlighting their tasks, methodological approaches, key findings, and limitations. As shown in the table, classical deep learning methods are followed by transformer-based approaches, while more recent studies focus on knowledge distillation techniques to balance performance and computational efficiency.

Knowledge distillation (KD) has emerged as a widely adopted model compression strategy in NLP, with numerous studies demonstrating its potential in reducing the computational overhead of large language models while retaining competitive performance [19]. Architectures such as DistilBERT [20], TinyBERT [21], and MobileBERT [22] were specifically designed within this paradigm, achieving performance comparable to their teacher counterparts with considerably fewer parameters. Across the broader NLP literature, KD-based training has been explored for tasks including text classification, named entity recognition (NER), and question answering (Q&A) [23–25]. Recent KD studies have extended conventional logit-based supervision to intermediate-representation transfer, including attention-map, hidden-state, feature-level, and relational distillation strategies. Representative examples include TinyBERT [21], which combines attention-map distillation and hidden-state distillation to transfer both intermediate and output-level knowledge; MiniLM [26], which compresses transformer models through deep self-attention relation transfer; and Patient Knowledge Distillation (PKD) [27], which exploits multi-layer hidden-state supervision through progressive intermediate-layer learning. In addition, recent studies have explored representation matching, relational knowledge transfer, feature alignment, and multi-stage distillation strategies to preserve richer teacher representations during model compression. These approaches have demonstrated strong performance across a variety of NLP tasks by enabling the student model to imitate not only the teacher's output distributions but also its internal representational structure. However, such methods typically require architectural compatibility, explicit layer-mapping schemes, or auxiliary adaptation modules between teacher and student models. As a result, applying intermediate-representation distillation consistently across heterogeneous teacher–student architectures remains challenging, particularly in comparative studies involving multiple model families and compression settings. Beyond these general-domain developments, a growing body of work has examined KD in biomedical and clinical NLP settings, extending these distillation paradigms to domain-specific tasks and revealing the method's potential across diverse task types and model configurations. Gu et al. [28] leveraged KD for biomedical information extraction from unlabeled text, employing GPT-3.5 as a teacher to automatically annotate PubMed abstracts and subsequently training PubMedBERT

and BioGPT student models for NER and relation extraction (RE). The resulting combined NER + RE architecture achieved higher F1-scores than both GPT-3.5 and GPT-4, while also providing improved interpretability. Zhang et al. [29] introduced a cross-domain KD framework in which the attention structures, latent representations, and output distributions of teacher models trained on multiple source domains are jointly transferred to a single student model, supported by auxiliary model training and domain adaptation mechanisms. Evaluated on multi-domain classification benchmarks including IMDB, Yelp, and Airline datasets, this approach yielded improvements in model size and inference speed, and achieved competitive results against strong baselines such as Multi-source Domain Adversarial Network in several experimental conditions. Sakai and Lam [30] proposed KDH-MLTC, a framework that integrates KD with Particle Swarm Optimization (PSO)-based hyperparameter tuning for multi-label healthcare text classification, using BERT as the teacher and DistilBERT as the student. Experiments conducted on three subsets of the Hallmarks of Cancer (HoC) dataset yielded F1-scores of up to 82.7%, with statistically validated performance gains confirmed through ablation studies, underscoring the framework's applicability in resource-limited healthcare environments. In the clinical domain, Rohanian et al. [31] developed lightweight transformer models for clinical NLP tasks through knowledge distillation, demonstrating that compact architectures can achieve competitive performance on clinical text processing benchmarks while substantially reducing computational requirements.

Table 1: Comparison of the proposed approach with representative biomedical NLP studies, summarizing the target tasks, model architectures, principal findings, and reported limitations, with particular emphasis on knowledge distillation and computational efficiency considerations.

Study	Task	Method/Model	Key Findings	Limitations
[13]	Short text classification	LSTM + GRU + CNN with attention	Improves accuracy using attention-based parallel RNNs; faster convergence	High computational cost; not suitable for long texts
[14]	Biomedical question classification	Bi-GRU + Temporal CNN	Captures contextual and temporal dependencies effectively; improves accuracy and F1-score	Limited semantic richness due to Word2Vec; increased model complexity
[15]	Biomedical text classification	LSTM + CNN + BioWordVec	Achieves very high performance using domain-specific embeddings	Potential overfitting; limited generalizability
[16]	Clinical text classification	CNN, RNN, LSTM, Transformer, BERT	Transformer achieves best performance; CNN provides efficient alternative	Sensitive to class imbalance; small dataset size
[17]	Clinical document classification	BERT vs. CNN, HiSAN	Simpler models outperform BERT in some cases; highlights inefficiency in long documents	BERT struggles with long text; high computational cost
[18]	Biomedical text classification	LSTM/RNN + BERT hybrid	Hybrid models improve sequential understanding and contextual representation	High model complexity; limited reproducibility details
[23]	Biomedical NER	BiLSTM-CRF/BioBERT + Knowledge Distillation	Knowledge distillation significantly improves recall; integrates multiple knowledge sources	Multi-stage pipeline complexity; noisy weak labels

(Continued)

Table 1 (continued)

Study	Task	Method/Model	Key Findings	Limitations
[24]	Biomedical QA	QA-aware model + Adversarial KD	Adversarial KD improves robustness and performance in low-resource settings	Computational overhead; sensitive to perturbations
[25]	Biomedical text summarization	Systematic literature review	Provides comprehensive analysis and BQA-integrated framework	No experimental validation; descriptive study
This study	Biomedical multi-class text classification	Knowledge Distillation (BERT, PubMedBERT → DistilBERT, BioBERT, BioClinicalBERT, DistilBioBERT, DistilRoBERTa)	Systematic comparison of 10 KD configurations; KD enables student models to match or surpass fine-tuning (up to 89% accuracy); general-domain BERT transfers knowledge more effectively than PubMedBERT	Performance depends on teacher–student compatibility; some KD configurations show degradation due to architectural mismatch

Kim and Joe [32] addressed the task of predicting SNOMED-CT codes from natural language clinical inputs, encoding domain knowledge through a BioBERT teacher model and training a lightweight LSTM student via KD, further supported by data augmentation. While the student model did not surpass the fine-tuned teacher (0.86 vs. 0.88 accuracy), it substantially outperformed the LSTM baseline trained without distillation (0.8695), demonstrating the practical value of KD in capturing structured medical terminology within compact architectures.

Despite the growing body of work on biomedical knowledge distillation, most existing studies focus on a single task or evaluate only a limited number of teacher–student configurations. Consequently, the interaction between different teacher and student architectures in biomedical multi-class text classification remains insufficiently understood.

To address this gap, the present study systematically evaluates multiple teacher–student configurations under a unified experimental protocol while directly comparing knowledge distillation with conventional fine-tuning.

3 Methods

In this study, two complementary model training approaches were adopted for the biomedical text classification task: direct fine-tuning and knowledge distillation. The overall methodological flow is illustrated in Fig. 1, encompassing dataset preparation, preprocessing, teacher and student model training, and the implementation of both training strategies. The resulting models were evaluated using multiple performance metrics, and the outcomes of the two approaches were compared systematically.

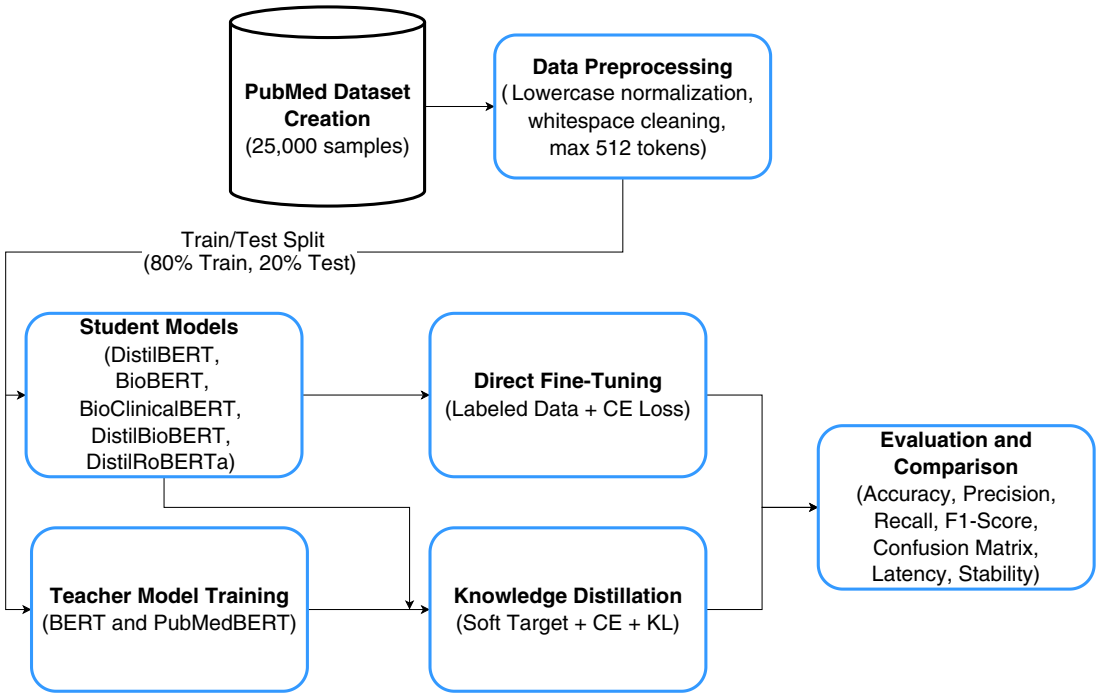

Figure 1: Overview of the experimental workflow, including dataset preparation, teacher model fine-tuning, student model training through both direct fine-tuning and knowledge distillation, and comparative evaluation of multiple teacher–student configurations using classification and efficiency metrics.

3.1 Dataset

The dataset used in this study was constructed from open-access biomedical article abstracts retrieved from the PubMed platform in Medline format via the BioPython library. Keywords representing five biomedical domains—cardiovascular disease, gynecology, mental health, neurological disorders, and breast cancer—were used as search queries, with 5000 records collected per keyword, yielding a multi-class text classification dataset of 25,000 examples in total. A selection of representative records is presented in Table 2. Each record contains the article title, abstract, and the query keyword used as the class label. The query keywords served exclusively for dataset retrieval and label assignment and were not provided to the models as explicit input features during training or evaluation. All models received only the article title and abstract as input. The query keywords were not explicitly removed from the retrieved abstracts, as the primary objective of the study was to evaluate KD behavior under a consistent and reproducible biomedical text classification setting rather than to construct a fully delexicalized benchmark dataset. This dataset structure was designed for supervised learning, with each record assigned to a single class label. Since the dataset was constructed using broad biomedical query keywords, a certain degree of semantic overlap between categories may naturally arise. Accordingly, the assigned labels should be interpreted as weakly supervised category indicators rather than clinically validated expert annotations. The dataset is fully balanced, with 5000 records per category, enabling controlled comparison across different teacher–student configurations without the influence of class imbalance.

The abstracts range from 28 to 510 words, with an average length of 193 words. Only two records with missing abstracts were removed (0.008%). The dataset was constructed to provide an open-access, multi-domain biomedical text classification resource spanning five biomedical categories. To further characterize the dataset, an inter-class TF-IDF cosine similarity analysis was conducted.

Table 2: Sample records from the dataset: each row corresponds to an article record and includes title, abstract, and keyword (class) information.

Keyword	Title	Abstract
Cardiovascular disease	Implications of Acanthosis Nigricans	Acanthosis nigricans is a dermatologic condition often overlooked in its role in cardiovascular implications...
Gynecology	ECEL1 Mutation in Arthrogryposis	Arthrogryposis involves joint contractures across various body parts due to ECEL1 mutation...
Mental health	Clean Energy and Elderly Mental Health	Clean energy transition has become a key strategy in combating global air pollution and improving mental health...
Neurological disorders	Brain Dynamics in Infarction	Evaluation of alterations in brain dynamics in patients suffering from cerebellar or brainstem infarctions...
Breast cancer	BRD3/4-Targeting Theranostic Probes	BET family proteins play critical roles in tumorigenesis and cancer progression, offering potential for imaging and therapy...

For the quantitative overlap analysis, TF-IDF vectors were extracted from all abstracts using unigram and bigram features after English stop-word removal. Class-level centroid vectors were then obtained by averaging the TF-IDF vectors of all records within each category, and pairwise cosine similarity was computed between these centroid vectors. The resulting inter-class similarity matrix is presented in Table 3. The highest inter-class similarity was observed between Cardiovascular Disease and Neurological Disorders (0.805), followed by Cardiovascular Disease and Gynecology (0.741). These results indicate that several biomedical categories share substantial terminology and contextual overlap, supporting the interpretation of the dataset as a challenging weakly supervised classification setting rather than a set of trivially separable topic labels.

Table 3: Inter-class TF-IDF cosine similarity matrix computed using class-level centroid vectors.

Class	BC	CVD	GYN	MH	ND
BC	1.000	0.565	0.661	0.460	0.560
CVD	0.565	1.000	0.741	0.635	0.805
GYN	0.661	0.741	1.000	0.670	0.670
MH	0.460	0.635	0.670	1.000	0.659
ND	0.560	0.805	0.670	0.659	1.000

Note: BC: Breast Cancer, CVD: Cardiovascular Disease, GYN: Gynecology, MH: Mental Health, ND: Neurological Disorders.

Overall, the analyses suggest moderate terminology overlap between several biomedical categories, particularly cardiovascular disease and neurological disorders, while generally supporting the consistency of document contents with their assigned labels. Nevertheless, semantic overlap, weak-label noise, and residual label ambiguity remain inherent limitations of keyword-based dataset construction. To further assess the

influence of keyword co-occurrence, a sensitivity analysis was performed by removing the five class-defining query phrases from all abstracts. The keyword-removed DistilBERT fine-tuning configuration achieved an accuracy of 0.766, compared to 0.792 in the original setting, suggesting that classification performance is influenced by broader semantic and contextual information in addition to keyword-based signals.

3.2 Data Preprocessing

Appropriate preprocessing of raw text is a fundamental step in NLP pipelines, directly influencing model performance and training stability. The remaining records were subsequently subjected to a standardized preprocessing procedure to ensure compatibility with model input requirements. Specifically, all text was converted to lowercase and redundant whitespace was normalized. Punctuation marks were retained to preserve potentially meaningful biomedical expressions, abbreviations, and domain-specific terminology. The processed dataset was then partitioned into training and test subsets using stratified sampling, with 80% of records allocated to training and 20% to testing, ensuring that class proportions were maintained across both splits.

3.3 Teacher and Student Models

This section presents technical details regarding the teacher and student model architectures used in the fine-tuning and knowledge distillation processes.

3.3.1 Teacher Models

BERT (Bidirectional Encoder Representations from Transformers) is a widely used Transformer-based language model pretrained on general-domain corpora using the Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) objectives [33]. In this study, BERT was employed as a teacher model to investigate the transferability of general linguistic representations to biomedical text classification.

PubMedBERT is a BERT-based language model pretrained on PubMed abstracts and PubMed Central (PMC) full-text articles, enabling it to learn domain-specific biomedical representations [34]. Owing to its biomedical pretraining, PubMedBERT was selected as the domain-specific teacher model for comparison with the general-domain BERT teacher.

3.3.2 Student Models

DistilBERT is a lightweight version of BERT obtained through knowledge distillation, providing substantially fewer parameters and faster inference while retaining competitive language understanding performance [20]. Owing to its compact architecture, it was selected as the primary student model in this study. **BioBERT** is a BERT-based language model pretrained on PubMed abstracts and PMC full-text articles, enabling improved representation of biomedical terminology [10]. It was included as a domain-specific baseline to evaluate the effect of biomedical pretraining on knowledge distillation. **BioClinicalBERT** extends BioBERT through additional pretraining on the MIMIC-III clinical notes corpus, making it particularly suitable for clinical text processing [35]. It was evaluated to investigate the behavior of clinically specialized language models under the KD framework. **DistilBioBERT** combines the lightweight DistilBERT architecture with biomedical domain adaptation through additional pretraining on biomedical corpora. It was selected to assess whether domain-specific pretraining benefits can be preserved in compressed biomedical models. **DistilRoBERTa** is a compressed RoBERTa-based language model designed for efficient inference while maintaining competitive performance across NLP tasks. It was included to examine the effect of an alternative Transformer architecture on knowledge distillation performance.

3.4 Knowledge Distillation Approach

Transformer-based pretrained language models achieve strong performance across a wide range of NLP tasks; however, their large parameter counts and substantial resource requirements impose computational costs that restrict their applicability in resource-constrained environments such as mobile devices, embedded systems, and edge computing platforms. Knowledge distillation (KD) addresses this limitation by enabling the transfer of knowledge from a large, high-capacity teacher model to a smaller, more computationally efficient student model [36].

Under the KD framework, the student model is trained not only on ground-truth class labels but also on the soft output distributions produced by the teacher model across all classes. This dual supervision may help the student model capture additional information regarding inter-class relationships beyond that available from hard labels alone, potentially improving performance in resource-constrained settings [37].

The applicability of KD has been demonstrated across a range of NLP tasks, including text classification, named entity recognition, and question answering. KD-derived models such as DistilBERT, TinyBERT, and MobileBERT reduce computational overhead while retaining performance levels close to those of their teacher counterparts [38]. Nevertheless, the majority of existing KD studies have been conducted on general-purpose datasets, and a systematic evaluation of KD in specialized domains such as biomedical text classification remains limited.

The present study addresses this gap by applying KD to biomedical text classification and conducting a systematic analysis across multiple teacher and student model combinations, evaluation metrics, and efficiency indicators, including model size and parameter count.

3.4.1 Knowledge Distillation Strategy

The KD strategy employed in this study involves transferring the soft output distributions of task-specific fine-tuned teacher models to smaller, more computationally efficient student models. The teacher models were initialized from pretrained language models and subsequently fine-tuned on the biomedical text classification task to produce task-specific representations. During distillation, the teacher model's logit outputs were smoothed using a temperature parameter and used as soft targets to guide student model training. The student model was optimized using a two-component loss function:

(i) Cross-Entropy (CE) loss, which ensures that the student model correctly predicts the task labels,
(ii) Kullback-Leibler (KL) divergence loss, which ensures that the student model imitates the output distributions of the teacher model.

$$L_{\text{total}} = \alpha \cdot L_{\text{CE}}(y, \hat{y}_s) + (1-\alpha) \cdot T^2 \cdot L_{\text{KL}}\left(\text{softmax}\left(\frac{\hat{y}_s}{T}\right), \text{softmax}\left(\frac{\hat{y}_t}{T}\right)\right) \quad (1)$$

In Eq. (1), $\hat{y}_s$ denotes the logit output of the student model and $\hat{y}_t$ denotes the logit output of the teacher model. T represents the temperature parameter ($T = 2.0$ in this study) and α represents the weighting coefficient applied to the two loss components ($\alpha = 0.5$ in this study). The temperature parameter produces smoother probability distributions, facilitating more informative knowledge transfer from teacher to student. All training procedures were carried out on the same fixed training/test split, ensuring that teacher and student models were evaluated under identical data conditions and enabling a systematic and fair comparison.

3.5 Comparative Fine-Tuning Experiments

Fine-tuning is a transfer learning approach in which the weights of a pretrained language model are updated on a task-specific dataset, allowing the model to adapt its general linguistic representations to the target task. This strategy is widely applied in supervised NLP tasks such as text classification, named entity recognition, and relation extraction [39].

To provide a meaningful reference point for evaluating the student models trained via KD, direct fine-tuning was also applied to each student model independently. In this experimental condition, no knowledge was transferred from any teacher model; instead, each student model was trained directly on the labeled dataset for the target classification task. This allowed the effects of KD and fine-tuning to be compared under identical dataset and architectural conditions.

Fine-tuning was applied to all five student models (DistilBERT, BioBERT, BioClinicalBERT, DistilBioBERT, and DistilRoBERTa) using the same 80/20 data split, the AdamW optimizer, and a consistent set of hyperparameters: 5 training epochs, a learning rate of 5e−5, and a batch size of 16. This configuration ensured a fair and reproducible comparison between the two training strategies.

In the fine-tuning scenario, all model parameters were updated without any teacher model involvement. The resulting performance metrics were directly compared against the corresponding KD-trained versions at the class and aggregate levels, enabling a systematic assessment of both model compression efficiency and the contribution of knowledge transfer to task performance.

3.6 Experimental Setup

All experiments were conducted in a Google Colab Pro environment using an NVIDIA Tesla T4 GPU. Model training and evaluation were performed using the HuggingFace Transformers library (v4.39) with the PyTorch (v2.2) framework under Python 3.10. Data processing and result visualization were supported by auxiliary libraries including pandas, scikit-learn, seaborn, matplotlib, and tqdm.

All teacher and student models were trained and evaluated on the same 80/20 data split. Hyperparameters were set in accordance with configurations commonly reported in the literature: 5 training epochs, a batch size of 16, a learning rate of 5e−5, a maximum token length of 512, and the AdamW optimizer. In the KD experiments, the distillation temperature was set to 2.0 and the loss weighting coefficient (α) was set to 0.5.

A random seed of 42 was used in all primary experiments to reduce stochastic variation and ensure controlled and reproducible comparisons. Subsequently, an auxiliary stability analysis was conducted using five different random seeds (42, 123, 999, 2024, and 3407) applied to the best-performing configuration to assess the robustness of the observed KD behavior. Teacher models were fine-tuned on the target task using the BERT and PubMedBERT architectures; student models were evaluated under both distillation and direct fine-tuning conditions. The output logits and final classification results of all models were systematically recorded and analyzed using standard classification report outputs.

This experimental setup was designed to support a fair and systematic comparison of knowledge distillation and direct fine-tuning across all evaluated model configurations.

3.7 Performance Metrics

In this study, the classification performance and computational efficiency of the teacher and student models were assessed using multiple metrics.

The following standard classification metrics were computed for each model:

- **Accuracy:** Defined as the proportion of correctly classified instances across all classes relative to the total number of instances:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative counts, respectively.

- **Precision:** Measures the proportion of instances predicted as belonging to a given class that are correctly classified:

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall:** Measures the proportion of actual instances of a given class that are correctly identified:

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1-Score:** The harmonic mean of precision and recall, providing a balanced assessment of classification performance across classes:

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Precision, recall, and F1-score are reported using class-wise and aggregate averaging strategies, as specified in the corresponding tables, while accuracy is reported as the overall proportion of correctly classified instances. These metrics collectively support a multidimensional comparison of knowledge distillation and direct fine-tuning strategies across all evaluated model configurations.

4 Experimental Results

This section presents the outcomes of the knowledge distillation and direct fine-tuning experiments conducted for biomedical multi-class text classification. Knowledge transfer was performed from two large-scale teacher models, BERT and PubMedBERT, to five student models: DistilBERT, BioBERT, BioClinicalBERT, DistilBioBERT, and DistilRoBERTa. Each student model was additionally trained via direct fine-tuning to serve as a controlled baseline, enabling a systematic comparison of the two training strategies.

All experiments were conducted using the preprocessing, data partitioning, training configuration, and evaluation protocol described in Sections 3.2–3.7. This section reports the standalone performance of the teacher models and compares the student models trained through direct fine-tuning and knowledge distillation under the same experimental conditions.

The following subsections present comparative performance results for both directly fine-tuned student models and student models trained via knowledge distillation. Overall classification accuracy together with class-wise precision, recall, and F1-score are reported to assess the effectiveness of knowledge transfer under different teacher–student configurations.

4.1 Teacher Model Performance

To support a more informed interpretation of the downstream knowledge distillation experiments, the standalone classification performance of each teacher model was evaluated on the test set. The results are summarized in Table 4.

Table 4: Standalone performance results of the teacher models on the test set, reported in terms of accuracy, precision, recall, and F1-score prior to knowledge distillation.

Teacher Model	Accuracy	Macro-Precision	Macro-Recall	Macro-F1
BERT	0.79	0.79	0.79	0.79
PubMedBERT	0.80	0.80	0.80	0.80

As shown in Table 4, PubMedBERT achieved marginally higher standalone classification performance than BERT under the current experimental conditions. However, this advantage was not consistently reflected in the downstream KD experiments, where the BERT→DistilBERT configuration yielded the highest student-level performance among all evaluated settings. These observations indicate that standalone teacher accuracy alone may not fully account for downstream distillation behavior, and that additional factors such as teacher–student architectural compatibility and representation alignment are likely to contribute to downstream KD effectiveness.

4.2 Fine-Tuning

This subsection reports the classification performance of each student model trained through direct fine-tuning (FT) on labeled data. The evaluated models include BioBERT, BioClinicalBERT, DistilBERT, DistilBioBERT, and DistilRoBERTa, each trained on the five-class biomedical classification task and assessed using accuracy, precision, recall, and F1-score.

Table 5 presents the per-class precision, recall, and F1-score values for each model, along with overall accuracy. While aggregate accuracy values were broadly comparable across models, ranging from 0.78 to 0.81, notable differences emerged at the class level. BioClinicalBERT achieved the most consistent performance across the evaluated categories, yielding more uniform results than the other models.

Across all models, the *Breast Cancer* and *Mental Health* classes yielded the highest F1-scores, with DistilRoBERTa and DistilBioBERT performing comparably to BioBERT in these categories. The *Gynecology* class yielded relatively lower F1-scores across all models, indicating greater classification difficulty, potentially because of semantic overlap and ambiguous terminology shared with related biomedical categories. This pattern may also reflect the presence of semantic overlap and ambiguous terminology across related clinical domains, which can contribute to increased inter-class confusion during classification. This class-level variability should be considered when evaluating biomedical text classification systems for clinical or decision-support applications. BioClinicalBERT achieved the highest overall accuracy (81%) in the fine-tuning scenario. The distilled models, DistilBERT, DistilBioBERT, and DistilRoBERTa, attained overall accuracy values of approximately 80%, demonstrating competitive performance relative to their larger counterparts despite having considerably fewer parameters. These results indicate that direct fine-tuning on labeled data alone can yield competitive classification performance across the evaluated models, and the fine-tuning results serve as a controlled baseline for assessing the contribution of knowledge distillation in subsequent experiments.

4.3 Knowledge Distillation

This subsection presents the classification results obtained when each student model was trained via knowledge distillation, using BERT and PubMedBERT as teacher models in separate experimental conditions.

Table 5: Class-wise performance of the fine-tuned student models on the test set, reported in terms of precision, recall, and F1-score for each biomedical category, together with overall model accuracy.

Class	Metric	BioBERT	BioClinicalBERT	DistilBERT	DistilBioBERT	DistilRoBERTa
Cardiovascular Disease	Precision	0.77	0.81	0.73	0.73	0.77
	Recall	0.79	0.73	0.79	0.83	0.79
	F1-Score	0.78	0.77	0.76	0.78	0.78
Gynecology	Precision	0.62	0.74	0.74	0.72	0.76
	Recall	0.82	0.78	0.77	0.81	0.74
	F1-Score	0.71	0.76	0.75	0.77	0.75
Mental Health	Precision	0.85	0.84	0.89	0.84	0.85
	Recall	0.76	0.83	0.77	0.81	0.82
	F1-Score	0.80	0.83	0.82	0.83	0.83
Neurological Disorders	Precision	0.80	0.75	0.75	0.82	0.74
	Recall	0.65	0.80	0.77	0.65	0.78
	F1-Score	0.71	0.77	0.76	0.72	0.76
Breast Cancer	Precision	0.92	0.90	0.91	0.90	0.90
	Recall	0.87	0.90	0.89	0.89	0.90
	F1-Score	0.89	0.90	0.90	0.89	0.90
Overall	Accuracy (Weighted)	0.78	0.81	0.80	0.80	0.80

4.3.1 Teacher Model: BERT

This subsubsection evaluates the performance of student models trained via KD using BERT as the teacher. The five student models (BioBERT, BioClinicalBERT, DistilBERT, DistilBioBERT, and DistilRoBERTa) were each trained on the five-class biomedical classification task using the teacher's soft output distributions, and assessed using accuracy, precision, recall, and F1-score. Table 6 presents per-class and overall performance results for each student model.

KD produced the largest performance improvement for DistilBERT, which achieved F1-scores of 0.93 for both the Breast Cancer and Mental Health classes. The shared BERT-based encoder design may have facilitated knowledge transfer by reducing teacher–student representation mismatch. Despite its lower parameter count, DistilBERT effectively acquired task-relevant representations under KD supervision. However, the current experimental design does not determine whether these representations primarily capture domain semantics or topic-associated patterns.

Considering aggregate performance, DistilBERT achieved the highest overall accuracy (89%), followed by BioClinicalBERT, DistilRoBERTa, and DistilBioBERT. BioBERT recorded the lowest accuracy among the student models (79%). This result may reflect limited alignment between BioBERT's domain-specific pretrained representations and the supervisory signals produced by the general-domain BERT teacher. The moderate performance of DistilRoBERTa may reflect architectural and pretraining differences between the BERT teacher and the RoBERTa-based student, which could limit the effectiveness of knowledge transfer.

Table 6: Class-wise performance of student models trained via knowledge distillation using BERT as the teacher model. Precision, recall, and F1-score are reported for each biomedical category, together with overall classification accuracy on the test set.

Class	Metric	BioBERT	BioClinicalBERT	DistilBERT	DistilBioBERT	DistilRoBERTa
Cardiovascular Disease	Precision	0.87	0.83	0.93	0.81	0.84
	Recall	0.64	0.67	0.81	0.69	0.68
	F1-Score	0.74	0.74	0.87	0.74	0.75
Gynecology	Precision	0.69	0.70	0.82	0.68	0.70
	Recall	0.83	0.81	0.90	0.82	0.81
	F1-Score	0.75	0.75	0.86	0.75	0.75
Mental Health	Precision	0.88	0.88	0.93	0.86	0.88
	Recall	0.81	0.79	0.87	0.79	0.79
	F1-Score	0.85	0.83	0.90	0.83	0.83
Neurological Disorders	Precision	0.72	0.72	0.84	0.73	0.70
	Recall	0.84	0.84	0.92	0.78	0.86
	F1-Score	0.78	0.78	0.88	0.76	0.77
Breast Cancer	Precision	0.91	0.91	0.93	0.91	0.94
	Recall	0.88	0.88	0.94	0.87	0.85
	F1-Score	0.89	0.90	0.93	0.89	0.89
Overall	Accuracy (Weighted)	0.80	0.80	0.89	0.79	0.80

An examination of per-class results indicates that all models achieved strong F1-scores for the *Breast Cancer* class, consistent with the relatively homogeneous terminology and distinctive contextual cues associated with this domain. Conversely, lower F1-scores across models in the *Gynecology* class reflect greater conceptual diversity and increased classification difficulty within that category. One possible explanation for this disparity is that the Breast Cancer category contains relatively focused terminology, whereas the Gynecology category encompasses a broader range of clinical topics. This broader thematic scope may increase intra-class variability and vocabulary overlap with other biomedical categories, potentially making classification more challenging.

Overall, the student models trained through BERT-guided KD achieved competitive accuracy and F1-scores, including those based on compact architectures. These findings suggest that KD can serve not only as a model compression technique but also as an informative framework for examining knowledge transfer dynamics in compact models under architecturally compatible teacher–student configurations. However, the results also indicate that such performance gains are not uniformly observed across all model combinations.

Fig. 2 presents t-SNE visualizations of the [CLS] token representations extracted from the BERT teacher, DistilBERT FT, and BERT→DistilBERT KD models on the test set. Across all three configurations, the *Breast Cancer* class forms the most compact and well-separated cluster, consistent with its consistently high F1-scores reported in Table 6. The *Gynecology* and *Neurological Disorders* classes exhibit the greatest overlap across all models, providing visual confirmation of the classification difficulty discussed above. The BERT→DistilBERT KD configuration maintains cluster compactness comparable to the BERT teacher

model, suggesting that the distillation process successfully preserved the teacher's representational structure despite a 40% reduction in parameter count.

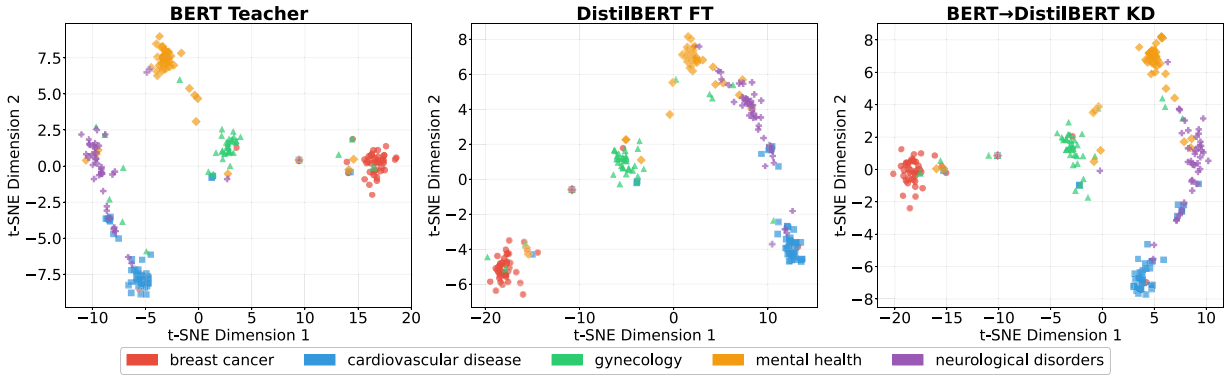

Figure 2: t-SNE visualization of the [CLS] token representations extracted from the BERT teacher, DistilBERT FT, and BERT→DistilBERT KD models on the test set. Each point corresponds to a document, and colors indicate the associated biomedical category labels.

To complement the qualitative observations obtained from the t-SNE visualization, we further quantified the similarity between the learned representations using Linear Centered Kernel Alignment (Linear CKA). As shown in Table 7, the BERT→DistilBERT KD model achieved a higher representation similarity to the teacher model (Linear CKA = 0.9641) than the directly fine-tuned DistilBERT model (Linear CKA = 0.9444). This result suggests that knowledge distillation enables the student model to preserve the internal feature representations learned by the teacher more effectively than conventional fine-tuning. Overall, the quantitative CKA analysis is consistent with the qualitative t-SNE visualization and further supports the conclusion that knowledge distillation preserves the teacher's internal representation structure more effectively than direct fine-tuning.

Table 7: Representation similarity analysis between the teacher and student models using Linear Centered Kernel Alignment (Linear CKA) and mean paired cosine similarity computed on 1000 randomly sampled test instances.

Comparison	Samples	Linear CKA	Mean Cosine Similarity
Teacher vs. FT	1000	0.9444	0.2484
Teacher vs. KD	1000	0.9641	0.2313
FT vs. KD	1000	0.9623	0.2614

4.3.2 Teacher Model: PubMedBERT

This subsubsection presents the classification performance of student models trained via KD using PubMedBERT as the teacher. The same five student models—BioBERT, BioClinicalBERT, DistilBERT, DistilBioBERT, and DistilRoBERTa—were evaluated on the five-class biomedical classification task using accuracy, precision, recall, and F1-score.

Table 8 presents the per-class and overall performance results. Overall accuracy across the student models ranged from 0.79 to 0.81, with BioBERT and DistilBioBERT attaining the highest values (0.81), while the remaining models delivered competitive but slightly lower performance.

Table 8: Class-wise performance of student models trained via knowledge distillation using PubMedBERT as the teacher model. Precision, recall, and F1-score are reported for each biomedical category, together with overall classification accuracy on the test set.

Class	Metric	BioBERT	BioClinicalBERT	DistilBERT	DistilBioBERT	DistilRoBERTa
Cardiovascular Disease	Precision	0.82	0.75	0.81	0.83	0.83
	Recall	0.72	0.74	0.68	0.67	0.67
	F1-Score	0.77	0.74	0.74	0.74	0.74
Gynecology	Precision	0.75	0.74	0.72	0.68	0.87
	Recall	0.76	0.77	0.76	0.83	0.68
	F1-Score	0.76	0.75	0.74	0.75	0.76
Mental Health	Precision	0.88	0.84	0.86	0.88	0.72
	Recall	0.80	0.83	0.79	0.78	0.87
	F1-Score	0.84	0.83	0.82	0.82	0.79
Neurological Disorders	Precision	0.72	0.76	0.69	0.72	0.70
	Recall	0.85	0.75	0.82	0.83	0.85
	F1-Score	0.78	0.75	0.75	0.77	0.77
Breast Cancer	Precision	0.89	0.90	0.90	0.93	0.91
	Recall	0.90	0.89	0.88	0.85	0.90
	F1-Score	0.90	0.89	0.89	0.89	0.90
Overall	Accuracy (Weighted)	0.81	0.80	0.79	0.79	0.79

As in the BERT teacher experiments, the *Breast Cancer* and *Mental Health* classes yielded the highest F1-scores across all models. BioBERT and DistilBioBERT were particularly strong in these categories, with F1-scores in the range of 0.89–0.90. The *Gynecology* and *Cardiovascular Disease* classes generally produced lower F1-scores, a pattern consistent with the greater terminological diversity and contextual complexity associated with these domains.

On aggregate, most student models trained under PubMedBERT guidance reached accuracy levels comparable to the fine-tuning baseline. However, PubMedBERT's domain-specific pretraining did not consistently translate into improved student performance in this multi-domain classification setting. Nevertheless, lower-parameter models such as DistilRoBERTa maintained competitive performance in certain classes, even under less compatible teacher–student configurations. The absence of consistent accuracy gains for models such as DistilBioBERT and DistilRoBERTa may be associated with architectural mismatches between teacher and student, limited student model capacity, and potential information loss during the distillation process.

A cross-strategy comparison of overall accuracy values is presented in Table 9. The BERT→DistilBERT configuration achieved the highest performance across all evaluated settings, reaching 89% accuracy. This outcome suggests that, under the evaluated experimental conditions, the general-purpose BERT teacher provided more transferable supervisory signals than PubMedBERT across the biomedical categories considered in this study. Conversely, the PubMedBERT teacher produced no consistent accuracy improvements over direct fine-tuning across the evaluated student models, indicating that domain specificity alone does not ensure effective knowledge transfer, particularly in multi-domain classification tasks with varied data

distributions. Taken together, these findings indicate that the effectiveness of knowledge distillation is strongly conditioned on teacher–student architectural compatibility and representation alignment, rather than being a universally applicable training advantage. While the BERT→DistilBERT configuration demonstrated substantial performance gains, several other KD configurations yielded only marginal improvements or outcomes comparable to direct fine-tuning. The results should therefore be interpreted as reflecting the variability of KD behavior across different teacher–student combinations and task contexts, rather than as evidence of KD as a universally superior training strategy.

Table 9: Comparison of classification accuracy obtained through direct fine-tuning and knowledge distillation with BERT and PubMedBERT teachers across all evaluated student models.

Model	FT	BERT-KD	PubMed-KD
BioBERT	0.78	0.80	0.81
BioClinicalBERT	0.81	0.80	0.80
DistilBERT	0.80	0.89	0.79
DistilBioBERT	0.80	0.79	0.79
DistilRoBERTa	0.80	0.80	0.79

4.4 Stability Analysis Across Multiple Random Seeds

To assess the sensitivity of the observed KD behavior to random initialization and data partitioning, an additional stability analysis was conducted using five different random seeds (42, 123, 999, 2024, and 3407). Given computational cost considerations, this auxiliary analysis was restricted to the best-performing configuration, BERT→DistilBERT KD, and its corresponding direct fine-tuning baseline.

The mean performance values obtained across repeated runs are summarized in Table 10. The BERT→DistilBERT KD configuration yielded consistent performance across seed conditions, with mean accuracy and macro-F1 values of 0.8870 (±0.0085) and 0.8911 (±0.0084), respectively. The corresponding fine-tuning baseline achieved mean accuracy and macro-F1 values of 0.7990 (±0.0094) and 0.7998 (±0.0093), respectively.

Table 10: Results of the multi-seed stability analysis for the BERT→DistilBERT knowledge distillation configuration and the corresponding fine-tuning baseline, reported as mean ± standard deviation across five independent runs.

Method	Accuracy	Macro-Precision	Macro-F1	Weighted-F1
FT DistilBERT	0.7990 ± 0.0094	0.8027 ± 0.0091	0.7998 ± 0.0093	0.7998 ± 0.0093
BERT→DistilBERT KD	0.8870 ± 0.0085	0.8920 ± 0.0083	0.8911 ± 0.0084	0.8911 ± 0.0084

As shown in Table 10, consistent performance trends were maintained across all seed conditions, indicating that the observed performance differences between the KD configuration and the fine-tuning baseline were not attributable to a single random initialization. Furthermore, a paired t-test applied to the repeated-run accuracy results indicated that the performance difference between the BERT→DistilBERT KD configuration and the fine-tuning baseline was statistically significant ($p < 0.001$).

4.5 Temperature and Alpha Sensitivity Analysis

To further investigate the sensitivity of the KD framework to hyperparameter choices, additional experiments were conducted by varying the temperature and alpha values in the BERT→DistilBERT KD configuration. The results are summarized in Table 11.

Table 11: Sensitivity analysis results for different temperature (T) and loss-weighting coefficient (α) settings in the BERT→DistilBERT knowledge distillation configuration, reported using classification performance metrics on the test set.

Temperature	Alpha	Accuracy	Macro-F1
1.0	0.5	0.88	0.88
2.0	0.5	0.89	0.89
4.0	0.5	0.89	0.89
2.0	0.3	0.89	0.89
2.0	0.7	0.88	0.88

The results indicate that performance remained broadly stable across the evaluated temperature values. Variation in the alpha parameter produced similarly limited effects, with a slight performance decrease observed at $\alpha = 0.7$. These observations suggest that the BERT→DistilBERT KD configuration is relatively robust to the specific temperature and alpha settings examined in this analysis.

To further investigate the robustness of the best-performing BERT→DistilBERT knowledge distillation configuration, additional ablation experiments were conducted by varying the number of training epochs and batch size while keeping the distillation temperature ($T = 2.0$) and the loss-weighting coefficient ($\alpha = 0.5$) fixed. The corresponding results are presented in Table 12. Among the evaluated epoch settings, the highest performance was achieved after three training epochs, yielding an accuracy of 0.8924 and a macro-F1 score of 0.8963. Extending the training to five and seven epochs resulted in only a slight reduction in performance, suggesting that the student model effectively acquires the teacher's knowledge during the early stages of training and that prolonged optimization provides limited additional benefit for this KD configuration. Similarly, the batch-size analysis demonstrated highly consistent performance across the evaluated settings. Batch sizes of 8 and 16 produced nearly identical results, while increasing the batch size to 32 resulted in only a marginal decrease in performance. Overall, these findings indicate that the proposed knowledge distillation framework is relatively insensitive to reasonable training hyperparameter choices, supporting the robustness of the experimental conclusions.

Table 12: Additional hyperparameter ablation results for the best-performing BERT→DistilBERT knowledge distillation configuration. The distillation temperature and loss-weighting coefficient were fixed at $T = 2.0$ and $\alpha = 0.5$, while the number of training epochs and batch size were varied.

Setting	Value	Accuracy	Macro-Precision	Macro-F1
Epoch	3	0.8924	0.8970	0.8963
Epoch	5	0.8870	0.8916	0.8911
Epoch	7	0.8838	0.8878	0.8877
Batch Size	8	0.8872	0.8928	0.8911
Batch Size	16	0.8872	0.8920	0.8913
Batch Size	32	0.8856	0.8910	0.8898

Parameter count, GPU inference latency, CPU inference latency, and peak GPU memory usage were additionally measured for the BERT teacher model, the fine-tuned DistilBERT baseline, and the BERT→DistilBERT KD configuration. The results are presented in Table 13.

Table 13: Comparison of parameter count, inference latency, CPU inference latency, and memory consumption for the evaluated teacher and student models. Latency values represent the average inference time per sample measured under the experimental hardware configuration described in Section 4.

Model	Params (M)	Mean (ms)	Median (ms)	CPU (ms)	Mem. (MB)
BERT Teacher	109.49	8.05	7.91	167.88	949.0
DistilBERT FT	66.96	4.77	4.22	87.39	955.7
BERT→DistilBERT KD	66.96	4.24	4.16	90.27	951.7

As shown in Table 13, both DistilBERT-based configurations contained fewer parameters and achieved lower GPU and CPU inference latency than the BERT teacher. The KD-based DistilBERT model maintained latency comparable to the directly fine-tuned baseline while achieving higher classification performance. Peak GPU memory consumption was similar across all configurations, ranging from 949.0 to 955.7 MB. CPU inference latency was additionally evaluated to assess deployment under resource-constrained conditions. Both DistilBERT-based models substantially reduced CPU latency relative to the BERT teacher, and the KD model introduced no meaningful inference overhead compared with the directly fine-tuned baseline. Direct GPU energy measurements could not be obtained because the experiments were conducted in the managed Google Colab Pro environment, which does not provide reliable access to GPU power consumption statistics. Therefore, GPU latency, CPU latency, parameter count, and memory consumption are reported as practical deployment-oriented efficiency indicators.

Fig. 3 presents the confusion matrices of the directly fine-tuned DistilBERT model and the BERT→DistilBERT KD configuration evaluated on the same test set. The KD-based configuration exhibited stronger diagonal concentration across several biomedical categories, suggesting reduced class confusion relative to the fine-tuning baseline. Nevertheless, residual inter-class confusions remained observable, particularly among semantically related biomedical categories.

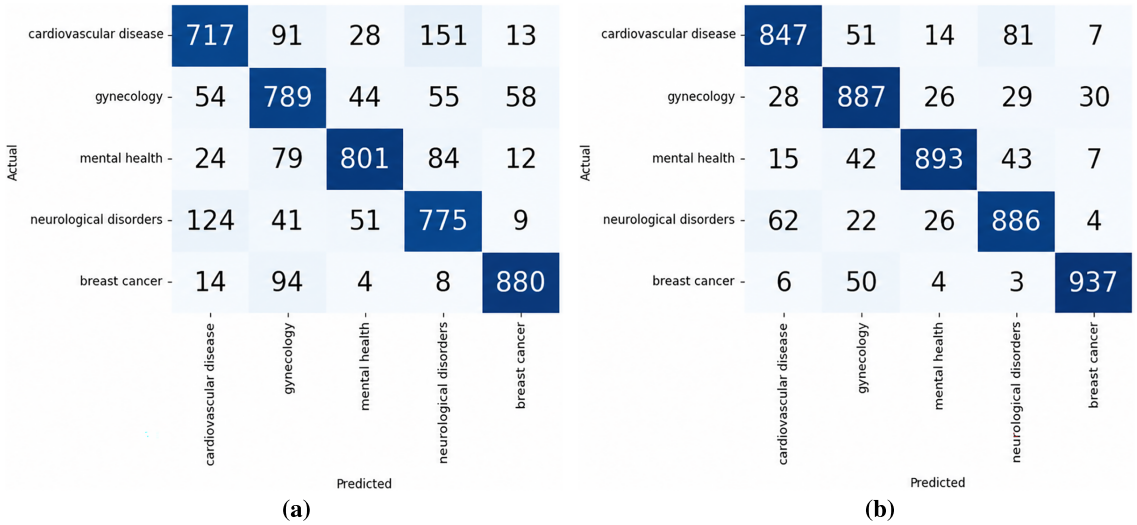

Figure 3: Confusion matrices of (**a**) the directly fine-tuned DistilBERT model and (**b**) the BERT→DistilBERT KD configuration evaluated on the same test set.

5 Conclusion

This study presents a comprehensive evaluation of multiple Transformer-based language models for biomedical multi-class text classification under two training paradigms: direct fine-tuning, in which models are trained solely on labeled data, and knowledge distillation (KD), in which compact student models are trained under the guidance of larger teacher models. This dual experimental framework supports a systematic assessment of both standalone model performance and the extent to which knowledge can be effectively transferred from high-capacity to lightweight architectures.

Under direct fine-tuning, the full-sized models (BioBERT and BioClinicalBERT) and the distilled models (DistilBERT, DistilBioBERT, and DistilRoBERTa) achieved competitive classification performance. BioClinicalBERT yielded the highest overall accuracy, while the distilled models demonstrated strong F1-scores in the *Breast Cancer* and *Mental Health* classes. It should be noted that BioBERT and BioClinicalBERT were evaluated as full-sized domain-specific baseline models rather than lightweight compressed architectures, whereas the distilled models were considered within the scope of computationally efficient configurations. These results indicate that models with reduced parameter counts can still achieve competitive classification performance, making them promising candidates for resource-constrained deployment settings.

BERT and PubMedBERT were employed as teacher models in the KD experiments. Although PubMedBERT achieved marginally higher standalone classification performance, this advantage did not consistently translate into improved student performance. In contrast, the general-purpose BERT teacher produced higher or comparable results across several configurations, with BERT → DistilBERT yielding the best overall performance. These findings indicate that teacher accuracy alone does not determine distillation effectiveness; teacher–student architectural compatibility, representation alignment, student capacity, and task characteristics also influence knowledge transfer.

Overall, KD effectively transferred task-specific knowledge to smaller models when the teacher and student were appropriately matched. The multi-seed analysis confirmed that the performance gains of BERT→DistilBERT were consistent across different random initializations. However, because KD performance varied across model combinations, careful teacher–student selection remains necessary, particularly in heterogeneous classification settings.

6 Limitations and Future Work

While this study provides a controlled experimental framework for comparing KD strategies, several limitations should be acknowledged. The dataset comprises five balanced classes, which facilitates model comparisons but does not fully reflect the class imbalance and semantic heterogeneity of real-world biomedical NLP settings. Class labels are derived from PubMed query keywords, which may introduce weak-label noise and inter-category semantic overlap. Furthermore, the primary experiments rely on a fixed train/test split, although supplementary multi-seed analyses were conducted to assess the stability of the results.

Future work may explore KD using more realistic imbalanced biomedical datasets, multi-teacher distillation frameworks, and complementary compression techniques, including pruning, quantization, and TinyBERT- or MobileBERT-style compression strategies. The present study was conducted on a single controlled, weakly supervised PubMed classification setting rather than a fully expert-annotated biomedical benchmark. Evaluation on established public biomedical multi-class classification benchmarks remains an important direction for future work to further validate the generalizability of the observed KD behavior across diverse biomedical corpora and annotation schemes. Directly comparable multi-class benchmarks

were not available within the scope of this study. The Hallmarks of Cancer (HoC) dataset uses a multi-label annotation scheme and is therefore not directly compatible with the present single-label framework. Similarly, the i2b2 datasets primarily address named entity recognition and relation extraction rather than document-level classification.

Furthermore, the present study employs vanilla logit-based KD, which was deliberately selected to enable architecture-agnostic comparison across all ten teacher–student configurations. Future work may investigate intermediate-representation techniques, including attention and hidden-state distillation, for architecturally compatible teacher–student pairs. Such methods may provide additional gains when layer-level alignment is feasible. While this study incorporates both t-SNE visualization and quantitative representation similarity analysis using Linear Centered Kernel Alignment (Linear CKA), future work may further investigate intermediate representation transfer through layer-wise analyses, attention alignment, and additional feature-space similarity measures across a broader range of teacher–student architectures.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Amine Gonca Toprak and Aytuğ Onan; methodology, Amine Gonca Toprak and Aytuğ Onan; software, Amine Gonca Toprak; validation, Amine Gonca Toprak and Aytuğ Onan; formal analysis, Amine Gonca Toprak and Aytuğ Onan; investigation, Amine Gonca Toprak; resources, Amine Gonca Toprak; data curation, Amine Gonca Toprak; writing–original draft preparation, Amine Gonca Toprak; writing–review and editing, Aytuğ Onan; visualization, Amine Gonca Toprak; supervision, Aytuğ Onan. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: Data available on request from the authors.

Ethics Approval: Not applicable.

Conflicts of Interest: Given his role as Editorial Board Member of this journal, Aytuğ Onan had no involvement in the peer review of this article and had no access to information regarding its peer review. Full responsibility for the editorial process for this article was delegated to another journal editor. The authors declare no other conflicts of interest.

References

1. Qin L, Chen Q, Feng X, Wu Y, Zhang Y, Li Y, et al. Large language models meet NLP: A survey. arXiv:2405.12819. 2024.
2. Varan M, Yatkınoğlu A, Toprak AG, Soygazi F, Mocan B. Fenomen-hedef kitle eşleştirmesinin otomatikleştirilmesi: sosyal medya gönderilerinin Siniflandirilmasi ile reklama Yönelik hedef kitle analizi. J Intell Syst Theory Appl. 2024;7(2):159–73. doi:10.38016/jista.1509968.
3. Çepni S, Toprak AG, Yatkınoğlu A, Mercan Ö.B, Ozan Ş. Performance evaluation of a pretrained BERT model for automatic text classification. J Artif Intell Data Sci. 2023;3(1):27–35.
4. Yin W, Ye Q, Liu P, Ren X, Schütze H. LLM-driven instruction following: progresses and concerns. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts; 2023 Dec 6–10; Singapore. p. 19–25. doi:10.18653/v1/2023.emnlp-tutorial.4.
5. Patil R, Gudivada V. A review of current trends, techniques, and challenges in large language models (LLMs). Appl Sci. 2024;14(5):2074. doi:10.3390/app14052074.
6. Sakai H, Lam SS. Large language models for healthcare text classification: a systematic review. arXiv:2503.01159. 2025.
7. Jahan I, Laskar MTR, Peng C, Huang JX. A comprehensive evaluation of large language models on benchmark biomedical text processing tasks. Comput Biol Med. 2024;171(4):108189. doi:10.1016/j.compbiomed.2024.108189.

8. Tian Y, Pei S, Zhang X, Zhang C, Chawla NV. Knowledge distillation on graphs: a survey. ACM Comput Surv. 2025;57(8):1–16. doi:10.1145/3711121.
9. Chen P, Liu S, Zhao H, Jia J. Distilling knowledge via knowledge review. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2026 Jun 3–7; Denver, CO, USA. p. 5008–17.
10. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2019;36(4):1234–40. doi:10.1093/bioinformatics/btz682.
11. Han Q, Tian S, Zhang J. A PubMedBERT-based classifier with data augmentation strategy for detecting medication mentions in tweets. arXiv 2112.02998. 2021.
12. Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. arXiv:1904.05342. 2020.
13. Yu S, Liu D, Zhu W, Zhang Y, Zhao S. Attention-based LSTM, GRU and CNN for short text classification. J Intell Fuzzy Syst. 2020;39(1):333–40. doi:10.3233/JIFS-191171.
14. Gupta T, Kumar E. Fusion of Bi-GRU and temporal CNN for biomedical question classification. Int J Comput Appl. 2023;45(6):460–70. doi:10.1080/1206212X.2023.2235458.
15. Kurniasari D, Warsono U, Lumbanraja M, Wamiliana FR. LSTM-CNN hybrid model performance improvement with BioWordVec for biomedical report big data classification. Sci Technol Indones. 2024;9(2):273–83. doi:10.26554/sti.2024.9.2.273-283.
16. Lu H, Ehwerhemuepha L, Rakovski C. A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance. BMC Med Res Methodol. 2022;22(1):181. doi:10.1186/s12874-022-01665-y.
17. Gao S, Alawad M, Young MT, Gounley J, Schaefferkoetter N, Yoon HJ, et al. Limitations of transformers on clinical text classification. IEEE J Biomed Health Inform. 2021;25(9):3596–607. doi:10.1109/JBHI.2021.3062322.
18. Mandal BK, Majumder P, Tewari BP. Role of BERT model for sequential text classification in biomedical abstracts. In: Real-world applications and implementations of IoT. Singapore: Springer Nature; 2025. p. 67–82. doi:10.1007/978-981-97-8627-5_5.
19. Mansourian AM, Ahmadi R, Ghafouri M, Babaei AM, Golezani EB, Ghamchi ZY, et al. A comprehensive survey on knowledge distillation. arXiv:2503.12067, 2025. doi:10.48550/arXiv.2503.12067.
20. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108. 2020.
21. Jiao X, Yin Y, Shang L, Jiang X, Chen X, Li L, et al. TinyBERT: distilling BERT for natural language understanding. In: Findings of the Association for Computational Linguistics: EMNLP 2020. Stroudsburg, PA, USA: Association for Computational Linguistics; 2020. p. 4163–74. doi:10.18653/v1/2020.findings-emnlp.372.
22. Sun Z, Yu H, Song X, Liu R, Yang Y, Zhou D. MobileBERT: a compact task-agnostic BERT for resource-limited devices. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020 Jul 5–10; Online. p. 2158–70. doi:10.18653/v1/2020.acl-main.195.
23. Zhou H, Liu Z, Lang C, Xu Y, Lin Y, Hou J. Improving the recall of biomedical named entity recognition with label re-correction and knowledge distillation. BMC Bioinform. 2021;22(1):295. doi:10.1186/s12859-021-04200-w.
24. Bai J, Yin C, Zhang J, Wang Y, Dong Y, Rong W, et al. Adversarial knowledge distillation based biomedical factoid question answering. IEEE/ACM Trans Comput Biol Bioinform. 2023;20(1):106–18. doi:10.1109/TCBB.2022.3161032.
25. Xie Q, Luo Z, Wang B, Ananiadou S. A survey for biomedical text summarization: from pre-trained to large language models. arXiv:2304.08763. 2023. doi:10.48550/arXiv.2304.08763.
26. Wang W, Wei F, Dong L, Bao H, Yang N, MiniLM ZM. Deep self-attention distillation for task-agnostic compression of pre-trained transformers. arXiv:2002.10957. 2020. doi:10.48550/arXiv.2002.10957.
27. Sun S, Cheng Y, Gan Z, Liu J. Patient knowledge distillation for BERT model compression. 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); 2019 Nov 3–7; Hong Kong, China. p. 4323–32. doi:10.18653/v1/D19-1441.
28. Gu Y, Zhang S, Usuyama N, Woldesenbet Y, Wong C, Sanapathi P, et al. Distilling large language models for biomedical knowledge extraction: a case study on adverse drug events. arXiv:2307.06439. 2023.

29. Zhang S, Jiang L, Tan J. Cross-domain knowledge distillation for text classification. Neurocomputing. 2022;509:11–20. doi:10.1016/j.neucom.2022.08.061.
30. Sakai H, Lam SS. KDH-MLTC: knowledge distillation for healthcare multi-label text classification. arXiv:2505.07162. 2025.
31. Rohanian O, Nouriborji M, Jauncey H, Kouchaki S, Nooralahzadeh F, Clifton L, et al. Lightweight transformers for clinical natural language processing. Nat Lang Eng. 2024;30(5):887–914. doi:10.1017/S1351324923000542.
32. Kim H, Joe I. Effective SNOMED-CT concept classification from natural language using knowledge distillation. In: Data science and algorithms in systems. Cham, Switzerland: Springer International Publishing; 2023. p. 54–64.
33. Devlin J, Chang MW, 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: Human Language Technologies, Volume 1; 2019 Jun 2–7; Minneapolis, MN, USA. p. 4171–86. doi:10.18653/v1/N19-1423.
34. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthc. 2021;3(1):1–23. doi:10.1145/3458754.
35. Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, et al. Publicly available clinical BERT embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop; 2019 Jun 6–7; Minneapolis, MN, USA. p. 72–8. doi:10.18653/v1/W19-1909.
36. Yuan Y, Shi J, Zhang Z, Chen K, Zhang J, Stoico V, et al. The impact of knowledge distillation on the energy consumption and runtime efficiency of NLP models. In: Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering—Software Engineering for AI; 2024 Apr 14–15; Lisbon, Portugal. p. 129–33. doi:10.1145/3644815.3644966.
37. Gou J, Yu B, Maybank SJ, Tao D. Knowledge distillation: a survey. Int J Comput Vis. 2021;129(6):1789–819. doi:10.1007/s11263-021-01453-z.
38. Vakili YZ, Fallah A, Sajedi H. Distilled BERT model in natural language processing. In: Proceedings of the 2024 14th International Conference on Computer and Knowledge Engineering (ICCKE); 2024 Nov 19–20; Mashhad, Iran. p. 243–50. doi:10.1109/ICCKE65377.2024.10874673.
39. Tinn R, Cheng H, Gu Y, Usuyama N, Liu X, Naumann T, et al. Fine-tuning large neural language models for biomedical natural language processing. Patterns. 2023;4(4):100729. doi:10.1016/j.patter.2023.100729.