electronics

MDPI

Article

MultiCaRe-CERGFormer: A Retrieval- and Concept-Guided Fusion Framework for Clinical Case Report Classification

Khalaf Alsalem

Department of Information Systems, College of Computer and Information Sciences, Jouf University, Sakaka 72388, Aljouf, Saudi Arabia; kosalem@ju.edu.sa

Abstract

The task of classifying clinical case reports by their specialty is challenging because of the heterogeneity of the text in the reports, overlapping terminology of specialties, class imbalance, and the use of proxy labels in the biomedical literature where the reports are located. We tested a MultiCaRe-CERGFormer multi-view classifier, which fuses word- and character-level TF–IDF features obtained via truncated SVD, precomputed biomedical document embeddings, and specialty-specific concept indicators and features learned from similar cases in the training partition. A gated fusion module is provided to fuse the encoded branches for multiclass prediction. There were 8564 case reports, assigned to 13 specialty categories created from the metadata. The models were evaluated with five-fold grouped cross-validation, where the reports in the same source article belong to the same fold. The proposed model achieved a mean accuracy of 0.8152 ± 0.0070, a macro-F1 of 0.6693 ± 0.0179, a weighted-F1 of 0.8147, a macro-AUC of 0.9706 ± 0.0049, and a macro-average precision of 0.7430 ± 0.0227. The mean macro-F1 (0.6705 ± 0.0264) of the word- and character-level TF–IDF SVM was slightly higher. Therefore, the approach proposed was competitive under the evaluated protocol but was not "leading" on all metrics. Due to the nature of the targets, which were based on metadata of the articles instead of clinician adjudication, these results provide a measure of the prediction of proxy labels and are not meant to indicate diagnostic validity, clinical utility, or external generalizability.

Keywords: clinical case-report classification; biomedical natural language processing; knowledge-guided multimodal fusion; medical concept knowledge graph; similar-case retrieval

Check for updates

Academic Editors: Athanasios Anastasiou, Antonios N. Gargalionis and George Dimitrakopoulos

Received: 17 September 2026
Revised: 24 October 2026
Accepted: 29 October 2026
Published: 3 November 2026

Copyright: © 2026 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Rare disease, unusual presentations, unusual complications, unexpected clinical responses, and diagnostic reasoning are all examples of clinical cases reported in the literature that are less likely to be captured in routine clinical datasets. They are useful resources for biomedical knowledge discovery, biomedical education, biomedical literature indexing, and biomedical clinically related evidence retrieval due to their narrative detail. However, it is difficult to perform a large-scale computational analysis of case reports due to differences in length, terminology, structure, specialty focus and reporting style. Specialty classification can be automated to provide a way to report, search, and analyze these for downstream purposes. There are a number of reasons why this is a challenging task. First, a report can include multiple specialties' symptoms, test results, differential diagnosis, interventions, and outcomes. Secondly, rare and unusual cases might not use the typical lexical patterns that are used in common clinical documents. Thirdly, specialty categories

might be unbalanced or be weakly labeled instead of expertly labeled. Lastly, reports from the same publication might have similar wording and context, so that related cases could be split into train and test sets and lead to optimistic reporting. The biomedical and clinical domain-specific transformer models (BioBERT, PubMedBERT, BlueBERT, and ClinicalBERT) have enhanced the representation learning for biomedical and clinical language tasks [1–5]. Other architectures, such as Longformer, have been developed to solve the short sequence length problem of standard BERT models [6]. Yet, tuning large transformers on large case-report collections is still very computationally intensive, and text-only predictions do not contain explicit concept-level signals or related-case evidence. MultiCaRe offers a massive archive of clinical cases from freely available PubMed Central publications [7–9]. Its clinical narratives and precomputed embeddings enable one to classify text, retrieve it semantically, and analyze it case by case. This combination of lexical, semantic, concept-oriented, and retrieval-derived representations is motivated by these characteristics. This integration could be useful for delivering information that is not represented in just one of the representations and may not be efficiently processed by a large transformer when treating each report as an end-to-end process.

The current clinical text classifiers mainly address either a sparse lexical representation or a dense representation of the document or the clinical context, such as contextual transformer representation. These methods can ignore the implicit medical concepts or training cases that are provided by the complementary indicators. Furthermore, simple concatenation assumes that all information sources are equally useful in all clinical narratives, whereas the utility of lexical, semantic, concept-based and retrieval-derived information can be significantly different across different clinical narratives. It encourages an adaptive multi-view approach that integrates complementary evidence, maintains computational efficiency and provides the ability to examine the contribution of each feature source. Although considerable progress has been made in biomedical text classification, the following are three limitations that are still relevant to large-scale clinical case-report indexing. The sparse lexical features, dense biomedical embeddings, and explicit medical concepts are commonly evaluated separately and not combined in a case-dependent architecture. Second, similar-case retrieval is usually considered a method for search or explanation, rather than a source of information. Third, early concatenation extends the same fusion operation to all the reports, though the value of lexical, semantic, conceptual, and retrieval-based evidence can be significantly different across cases. MultiCaRe-CERGFormer overcomes these drawbacks by incorporating branch-specific encoders, training-only similar-case retrieval, and a gating module for estimating the weights for the individual branches. The main contributions of this study are the following:

- A multi-view framework combining sparse lexical features, dense biomedical document embeddings, medical concept indicators, and similar-case retrieval representations.
- A retrieval branch to check all the candidates in the candidate index, which is built entirely from the training partition, used for leakage control.
- Projections of heterogeneous branch-specific inputs into a common latent space are provided by branch-specific encoders.
- A gating mechanism dependent on the case that can adaptively weight the encoded branches rather than applying fixed or uniform fusion.
- A comparative evaluation with classical classifiers, biomedical transformer models, frozen-embedding models, early-fusion models, and an ensemble selected according to the validation.

- Evaluation at the component level (ablation, calibration, class-specific performance, and analyses of errors) and the explicit discussion of the limitations brought in by proxy labels.

2. Related Works

2.1. Clinical Case Report Datasets and MultiCaRe

Clinical case reports have been a powerful tool for the treatment of rare diseases, unusual presentations, and medical education. They are presented, however, in the form of a story, which makes it hard to analyze using traditional statistical methods. This can be overcome by large-scale datasets that are gathered from biomedical literature to allow supervised, weakly supervised, and retrieval-based learning on several thousand clinical cases. The MultiCaRe dataset was compiled as a public dataset of clinical case reports from PubMed Central [7–9]. Previous descriptions of the dataset highlight its multimodal nature, such as clinical cases, images, labels, and captions, and more recent versions and dataset releases include clinical case text and embeddings [7–9]. Other advantages of MultiCaRe are its suitability for medical AI research and text classification for semantic similarity search. Although MultiCaRe is larger than many hand-curated datasets, it lacks clinical breadth and therefore can be used to develop scalable models that are able to generalize to the medical specialty that they were not trained on. There are, however, a number of methodological issues with the dataset. Clinical cases are not always easy to find labels for a specific downstream task and are heterogeneous, with varying length and writing style. Therefore, there are strong preprocessing, weak label issues, class imbalance, and interpretability issues that need to be taken into account for model design. These are the challenges that inspire the proposed retrieval- and knowledge-driven fusion approach.

2.2. Biomedical and Clinical Language Models

A novel architecture, which was developed, has revolutionized NLP, substituting the recurrent architecture with a self-attention mechanism [10]. BERT went one step further by conducting deep bidirectional pretraining and fine-tuning with tasks in the downstream [11]. This type of model has become the foundation of biomedical and clinical NLP. BioBERT is one of the first successful biomedical text adaptations of BERT. It has achieved excellent results in biomedical named entity recognition, biomedical relation extraction, and biomedical question answering with pretraining on large-scale biomedical corpora [1]. Later, PubMedBERT was demonstrated to be able to improve performance by training from scratch on biomedical literature instead of general-domain models [2]. To showcase the importance of biomedical and clinical-domain corpora for transfer learning, Blue BERT combined PubMed abstracts and MIMIC-III clinical notes. Moreover, Clinical BERT and Bio-Clinical BERT were able to capture patient-specific and medical-context information relevant to clinical prediction tasks [4,5] by training on the clinical notes. Transformer-based biomedical models are good but not necessarily appropriate for large-scale report mining. Clinical case reports tend to be longer than the typical BERT-based model. Longformer overcame the long-document problem by employing sparse attention mechanisms and linear scaling in the sequence length [6]. However, full fine-tuning of a transformer is still possible on a large case report dataset. This drives the development of lighter alternatives that integrate classical text representations, dense embeddings, retrieval features, and neural fusion.

2.3. Text Representation and Classical Baselines

It is essential to classify biomedical text, and classical machine learning techniques remain useful baselines. TF-IDF is a statistical method used to represent documents; it is

effective in finding discriminative words and phrases [12]. In the field of text classification, linear classifiers, such as logistic regression and support vector machines [13,14], are the most popular classifiers because they are efficient, robust, and interpretable. In biomedical NLP, however, in many cases, TF-IDF + linear SVM or logistic regression can be very effective as a baseline classifier because the number of documents is large, the number of labels is small, and the labels can be well separated by the words. Dimensionality reduction techniques can be used to reduce the number of features in TF-IDF, for example, by using latent semantic analysis (LSA) or truncated singular value decomposition (TSVD) [15]. Such representations minimize the space cost and could help with the downstream fusion of the local neurons. TF-IDF + SVD is used as the text representation branch of the proposed MultiCaRe-CERGFormer framework because it is an efficient and reproducible method for full-transformer representation, which is often very expensive.

2.4. Dense Embeddings and Similarity-Based Retrieval

Embeddings of dense sentences and documents can easily be used for semantic similarity, clustering, retrieval, and downstream classification. To create sentence embeddings that are efficiently compared by cosine similarity, Sentence-BERT used a Siamese network structure [16]. Since then, dense retrieval has been a main technique in modern information retrieval, which enables the search of information in large vector spaces. Retrieval-augmented models are models that combine parametric (model) knowledge with non-parametric (external) memory. The benefit of retrieving relevant documents for improving knowledge-intensive NLP tasks was shown in the study of Retrieval-Augmented Generation [17]. The basic idea here, although not a focus of this study, is that predictions can be enhanced by conditioning the model on relevant retrieved evidence; the classification of the present study is similar in this aspect. In addition, late interaction mechanisms were shown to be possible to integrate with contextual representations in order to obtain efficient retrieval [18]. To enable scalable embedding retrieval from large embedding collections, FAISS provides efficient vector similarity search and clustering [19]. Retrieval is also crucial for clinical case analysis because physicians are inclined to work in a comparative and associative fashion, that is, to relate the current case with previous ones that are comparable. Thus, similar case retrieval can benefit the classification both in terms of performance and interpretability. The model that is proposed in this work involves the retrieval branch generating a case-evidence representation and retrieved-label distribution from the top k similar cases that will be used with the other branches.

2.5. Knowledge-Guided Clinical Representation

There are a number of hierarchies in biomedical text: disease hierarchies, symptom hierarchies, procedure hierarchies, anatomical hierarchies, and others. Biomedical text relations exist between biomedical entities such as the hierarchy of diseases, the hierarchy of symptoms, the hierarchy of procedures, the anatomical hierarchy, etc. There exists a large number of biomedical knowledge resources successfully used for biomedical terminology normalization and linking biomedical text mentions to structured biomedical concepts like UMLS and MeSH [20,21] in the biomedical area. Biomedical named-entity recognition and concept-linking systems, like SciSpacy and MedCAT, can be utilized to "enrich" the clinical text data with interpretable concept-level indicators [22,23].

The representation of biomedical knowledge has also been carried out with graph-based learning. Structured relations between entities can be learned using graph convolutional networks and graph attention networks [24,25]. From a clinical NLP perspective, concept graphs can be used to help understand and strengthen models by linking symptoms, diseases, and treatments. However, there can be some challenges with the full

ontology and creating the graph: it can be time-consuming and may require a license agreement or resources.

The proposed system employs medical-concept representation that relies on the medical terms picked up in each case report based on specialties. This is a knowledge graph representation and is not meant to be interpreted as graph-based reasoning or as an input feature branch. The present study examines the utility of concept indicators in addition to lexical, biomedical-text, and retrieval features. The contribution of these terms needs to be taken with a grain of salt because they may be similar to the metadata employed to build the proxy labels, and they need to be evaluated with label-term-masked evaluation.

2.6. Multi-View and Multi-Branch Fusion

Fusion strategies are crucial when there is more than one representation for the same sample. Early fusion concatenates features before they are classified, and late fusion combines independent predictions. To achieve greater flexibility in learning the importance of individual modality/feature branches, several attention-based/fused methods have been proposed [26,27]. Fusion is widely used in medical AI systems to combine images and texts, structured variables and lab results, and clinical notes. Gated fusion is especially useful if there are varying reliabilities between the different branches across the samples. When the description of the case is long, the text branch may be more useful; when the description of the case is semantically similar, the embedding branch may be useful; and when the description of the case is similar but not included in the training set, the retrieval branch may be useful. This proposed model uses an adaptive-weight gated cross-branch fusion module to assign weights to text, embedding, concept and retrieval branches.

The studies reviewed show that biomedical transformer, dense retrieval, concept extraction and learned feature fusion are effective. However, these factors are generally examined individually or are used in other applications besides weakly supervised classification of specialty in case reports. There have been few works that explore the possibility of combining sparse lexical representations, dense biomedical embeddings, explicit medical concepts and similar case evidence with case-dependent gating in a computationally efficient model. To fill this void, MultiCaRe-CERGFormer takes a complementary perspective on the four representations of the same clinical story. Conventional early fusion is modified in such a way that each view is encoded independently prior to gated aggregation. Unlike post-hoc nearest-neighbor explanations, retrieved training cases directly shape the representation to make predictions. However, the idea and retrieval outputs are seen as model-inspection signals, rather than clinical explanations that are validated.

3. Proposed Approach

In this section, we introduce MultiCaRe-CERGFormer, a multi-view framework to predict the metadata-based specialty category of a clinical case report, as shown in Figure 1. Each case is represented with four modalities to the model: a sparse lexical vector, a dense biomedical document embedding, a medical concept vector, and a representation derived from other cases in the training partition that are similar to the present case. Independent encoders encode each of the four inputs into a common latent space. This is followed by a gating network, which estimates a normalized weight for each encoded branch, and the weighted representation is passed to a multiclass classification head.

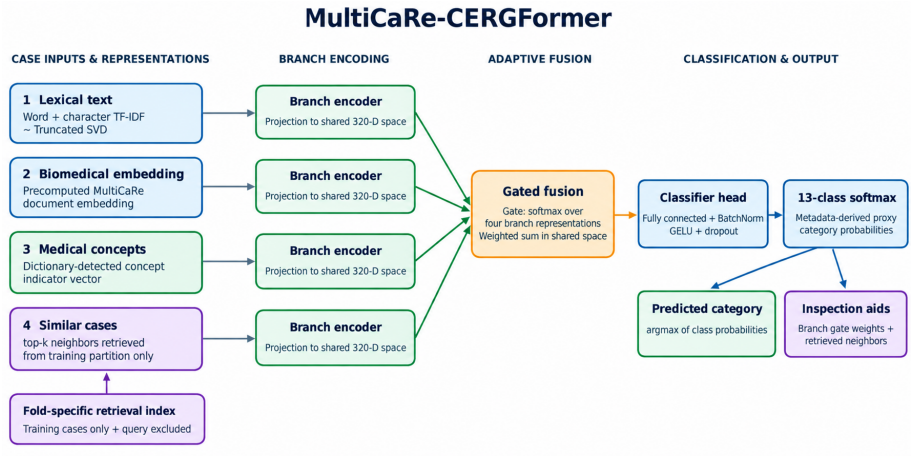


Figure 1. Architecture of the proposed MultiCaRe-CERGFormer framework.

The framework is designed for biomedical literature indexing, case report organization, retrieval in research, and so on. Retrieved neighbors, detected concepts, and branch weights can be used in model inspection and error analysis but should be considered as explanatory aids only, not as clinical facts or diagnostic suggestions.

3.1. Input Representation

For each clinical case (i), the model receives a clinical text report (Ti). When available, the model also uses the corresponding precomputed MultiCaRe embedding (Ei). The input representation consists of four branches:

3.1.1. Text Representation

The clinical case text is first cleaned and transformed into a numerical representation using TF-IDF followed by dimensionality reduction using Truncated Singular Value Decomposition.

$$X_i^{\text{text}} = \text{SVD}(\text{TFIDF}(T_i)) \tag{1}$$

This branch captures the discriminative textual patterns within clinical case reports.

3.1.2. Precomputed Embedding Representation

The MultiCaRe dataset provides clinical embeddings that encode semantic information from the clinical case reports. These embeddings are normalized and used as a second feature source:

$$X_i^{\text{emb}} = \text{Normalize}(E_i) \tag{2}$$

This branch helps preserve semantic similarity between clinical cases.

3.1.3. Medical Concept Representation

A medical concept vector is constructed by detecting terms associated with pre-defined clinical entities and semantic categories in each case report. Let $C = \{c_1, c_2, \ldots, c_m\}$ denote the concept vocabulary. The concept representation for case iis defined as

$$x_i^{\text{concept}} = [c_{i1}, c_{i2}, \ldots, c_{im}], \tag{3}$$

where $c_{ij} = 1$ when concept j is detected in case i, and $c_{ij} = 0$ otherwise. The representation provides explicit and inspectable domain-oriented features that complement the sparse and dense textual representations.

The present implementation should be described as a medical concept representation rather than a knowledge graph unless explicit concept nodes, typed relations, graph construction, and graph-based message passing are implemented.

3.1.4. Retrieval-Based Similar-Case Representation

For each query case, the retrieval module searches an index containing cases from the corresponding training partition. Cosine similarity is calculated between the normalized dense embedding of the query and the embeddings stored in the training index. The top kcases with the highest similarity scores are retained:

$$R_i = \mathrm{TopK}\Big(\cos(x_i^{\mathrm{emb}}, X_{\mathrm{train}}^{\mathrm{emb}})\Big). \tag{4}$$

Their embeddings are combined using similarity-normalized weights

$$x_i^{\mathrm{ret}} = \sum_{j \in R_i} \alpha_{ij} x_j^{\mathrm{emb}}, \qquad \alpha_{ij} = \frac{\exp(s_{ij}/\tau)}{\sum_{l \in R_i} \exp(s_{il}/\tau)}, \tag{5}$$

where s_{ij} is the cosine similarity between cases i and j, and τ is the temperature parameter.

For the query case, the query is not included in the training cases. The index used for the test cases was only built from the training partition, and no validation or test cases were included in the index. If several cases were linked to the same PubMed Central article, the data were split for those cases across the same data partition. As a result of this, there are no cases in source articles that appear in more than one of the training, validation, and test partitions.

3.2. Branch-Specific Encoders

The proposed MultiCaRe-CERGFormer framework is based on four complementary branches: (i) the lexical branch that consists of word- and character-level TF–IDF features with truncated singular value decomposition, (ii) the biomedical embedding branch based on precomputed MultiCaRe document representations, (iii) the medical concept branch, which relies on the presence of specialty-associated terms, and (iv) the retrieval branch, which is built from similar cases in the corresponding training partition. Each branch is encoded separately with an encoder that transforms the input into a shared 320-dimensional latent representation vector before passing through a gated fusion and multiclass prediction steps (see Figure 1).

The medical concept branch relies on indicators, which are terms picked up in the case narratives from a pre-defined medical vocabulary that is associated with the domain. These are merely indicators that offer explicit terminology-based features but are not a knowledge graph or graph-based reasoning. However, caution should be used in interpreting the predictive contribution of some of these terms, as there may be some overlap with the metadata applied to give the specialty proxy labels. The retrieval branch also offers representations based on training-partition neighbors, but not on clinical evidence alone. These branches complement each other in a complementary manner for lexical, semantic, terminology, and similarity-based inputs for adaptive fusion.

Medical Concept Vocabulary Construction

The medical concept representation was created by utilizing a dictionary of medical words that are associated with each medical specialty. The concept branch produced binary indicators of the presence or absence of the dictionary terms in the clinical case report for each concept. The representation offers explicit terminology-based information without an ontology graph, without typed relations, or with graph-based message passing.

Overlap may be present between the concept branch terminology and the information used to generate the specialty proxy labels since the specialty proxy labels were created using information from the publication metadata such as MeSH terms, article titles, keywords, and journal information. This can be a potential source of circularity. There were concept-oriented baseline experiments conducted, but these were not systematically quantified, and there was no evaluation of model performance when all the terms associated with the label were masked. Thus, the contribution of the concept representation cannot be fully assessed based on the current results.

3.3. Gated Cross-Branch Fusion

The four branch representations are first projected to the same latent dimension. Their concatenation is defined as

$$h_i = [h_i^{\text{text}}; h_i^{\text{emb}}; h_i^{\text{concept}}; h_i^{\text{ret}}]. \tag{6}$$

The gating network generates one normalized weight for each branch

$$\alpha_i = \text{softmax}(W_g h_i + b_g), \tag{7}$$

where $\alpha_i = [\alpha_i^{\text{text}}, \alpha_i^{\text{emb}}, \alpha_i^{\text{concept}}, \alpha_i^{\text{ret}}]$, and the four weights sum to one. The fused representation is

$$z_i = \sum_{b \in \{\text{text,emb,concept,ret}\}} \alpha_i^b h_i^b. \tag{8}$$

This mechanism allows the relative contribution of the four encoded representations to vary across cases. All branch outputs must have the same dimensionality before the weighted sum is calculated.

3.4. Classification Head

The fused representation (Zi) is passed to a classifier head composed of fully connected layers, batch normalization, GELU activation, dropout, and a final softmax layer.

$$\hat{y}_i = \text{Softmax}(W_c Z_i + b_c) \tag{9}$$

where $\hat{y}_i$ is the predicted probability distribution over the clinical categories. The predicted class is:

$$y_i^{\text{pred}} = \arg\max(\hat{y}_i) \tag{10}$$

3.5. Training and Inference Algorithm

The proposed MultiCaRe-CERGFormer model, as shown in Algorithm 1, proposes a hybrid clinical case classification framework based on textual, semantic, concept-based, and retrieval-based information. The gated fusion mechanism adaptively combines these complementary branches, and the retrieval and concept modules enhance interpretability. The design allows the model to be used for clinical case mining in large-scale clinical data, clinical specialty prediction, and retrieval of clinical cases based on evidence.

3.6. Dataset and Proxy-Label Generation

The classification task was based on 13 pre-defined specialty categories, which were assigned using a weak-supervision approach based on metadata. Specialty scores were computed for each source article based on major Medical Subject Headings (MeSH), MeSH terms, author keywords, journal metadata, and title terms. An article was accepted if the highest (best) specialty score was at least 4.0, the difference between the highest (best)

and the second-highest (second-best) specialty scores was at least 1.5, and at least one supporting match occurred in either a major-MeSH field, a MeSH field, or a keyword field.

Algorithm 1 Training and inference procedure for MultiCaRe-CERGFormer

Input: Clinical case texts T = T1, T2, . . . , TN, Optional precomputed embeddings E = E1, E2, . . . , EN, Class labels Y = Y1, Y2, . . . , YN, Number of retrieved cases k
Output: Predicted clinical category ypred ,Class probability vector yprob , Retrieved similar cases, Branch-importance gate weights

1. Partition cases using the selected stratified or group-aware splitting procedure.
2. Fit the label-generation and preprocessing pipeline according to the stated protocol.
3. Fit TF-IDF, SVD, normalization, and retrieval-index components using training data only.
4. Exclude the query case from its training retrieval candidates.
5. Use validation data for hyperparameter selection and early stopping.
6. Evaluate the selected model once on the held-out test set.

To test the sensitivity of the labeling procedure to these criteria, several sets of minimum scores and separation margins were chosen, as detailed in Table 1. A minimum score of 4.0 and a margin of 1.0 were used, and 7117 articles were accepted, achieving a 96.16% agreement with the configured labeling assignments. With the margin increased to 1.5, the accepted articles decreased to 6844 (8.99% coverage) and 100% agreement with the configured labels. Narrower thresholds typically led to a decrease in coverage and did not increase this agreement measure.

So, the selected thresholds are an operational compromise between keeping articles eligible and discarding more "ambiguous" metadata-based assignments. They are not to be considered clinically optimized thresholds. Specifically, the 100% agreement reported here is consistent with aligning with the labeling rule established by the manufacturer and is not an independent indicator of the accuracy of the assigned specialty categories. Throughout this study, these categories are considered to be proxy labels rather than clinician-adjudicated diagnoses or specialty assignments that have been formally made based on the metadata.

Table 1. Metadata-label threshold sensitivity.

Minimum Score	Minimum Margin	Accepted Articles	Coverage	Agreement with Configured Labels
4.0	1.0	7117	9.35%	96.16%
4.0	1.5	6844	8.99%	100.00%
4.0	2.0	6841	8.99%	100.00%
4.0	3.0	6607	8.68%	100.00%
5.0	1.5	6153	8.08%	100.00%
6.0	1.5	3518	4.62%	100.00%

3.7. Implementation and Training Settings

Four active input branches, namely, lexical, biomedical embedding, medical concept, and retrieval, were proposed. Before gated fusion, each branch was independently projected to a common latent representation of 320 dimensions. In the fusion classifier, the hidden layer consisted of 512 neurons, GELU activation and normalization were used, dropout was used with a rate of 0.18, and a fusion dropout rate was used. The model was optimized with AdamW, with a maximum learning rate of 0.00015, a weight decay of 0.0005, a batch size of 48, and a maximum of 42 epochs. Gradient clipping, automatic mixed precision (AMP) (with CUDA support), a cosine OneCycle learning-rate schedule, and early stopping with a patience of 10 epochs were also used during training. Selection of checkpoints was

carried out only based on the validation set composite score (75/25 macro-F1/accuracy). After dense and lexical reranking, the retrieval branch was applied to the six most highly-ranked training cases. Experiments were performed with a random seed of 42 on an NVIDIA Tesla T4 GPU with 14.6 GiB VRAM, an Intel Xeon CPU at 2.00 GHz, and 31.3 GiB system RAM, using Ubuntu 22.04 LTS(Linux),Python 3.11.11, and PyTorch 2.5.1 with CUDA 12.4. The held-out test partition was only evaluated after model and checkpoint selection were made.

Five-fold article-grouped cross-validation was used to evaluate all classification models. To avoid direct source-article overlap between training and held-out data for case reports, all case reports from the same source article were placed on the same outer fold. The models were fitted using the training data (IDPs in the outer folds), and the checkpoint data were selected from the internal validation data (IDPs in the inner folds). The outer held-out fold was used solely for performance evaluation.

The training partition was used for data-dependent lexical preprocessing components, and the retrieval index was fitted. Query cases were also excluded from the retrieval index but were allowed to be retrieved from the training-only index. In creating representations of retrieval cases, cases were not included in self-matches. One prediction held out per case report was provided for the five folds. Data was summarized as fold-level means and standard deviations, and where suitable, pooled out-of-fold results were reported. The reported experiments were all repeated from the same random seed (42), so the variability at the fold level is not indicative of the variability of different random seeds.

4. Results

4.1. Dataset and Partition Characteristics

MultiCaRe includes clinical cases from open access PubMed Central publications with textual representations and metadata associated with them [7–9]. The original preprocessing data consisted of 93,816 rows of case-report data. After filtering out the clinical text for each required clinical report and the assignment of the proxy label and the grouping of the source articles in each clinical report, 8564 case reports remained for classification, while 85,252 rows were filtered out. A separate audit was conducted on metadata labels for 6844 accepted source articles, which equates to 8.99% coverage of the metadata records for the articles evaluated. These represent counts by different units (case report rows versus source articles) and should not be directly compared.

Article-grouped five-fold cross-validation was used to evaluate retained cases. There were no overlaps in the article-group identifier from the training partition to the validation partition and held-out test partition, and no overlaps in exact text across these partitions during the partition audits. Additionally, the retrieval audit confirmed that the neighbors retrieved were not from the source article of the query case. Additionally, an exploratory lexical-similarity analysis was conducted, but the results of this analysis rely on the representation and similarity measure employed. These observations are to support the procedure of separating articles in the set, but not to rule out all cases of semantic duplication or semantic similarity between articles in the publications. An independent external dataset was not assessed.

Figure 2 presents the distribution of the 8564 clinical case reports included in the pooled out-of-fold evaluation, revealing substantial imbalance across the 13 metadata-derived specialty categories. Oncology accounts for the largest proportion of cases ($n = 3841$), followed by infectious disease ($n = 1303$), whereas general medicine is represented by only 13 cases. This uneven distribution has important implications for model evaluation, as aggregate performance measures may be disproportionately influenced by categories with greater representation. Consequently, macro-averaged metrics and class-specific precision,

recall, and F1-scores are reported alongside overall accuracy to provide a more comprehensive assessment of classification performance. Results for sparsely represented categories require particular caution because relatively few prediction errors can substantially affect their estimated performance. Furthermore, these frequencies describe the composition of the retained metadata-derived proxy-label dataset rather than the underlying prevalence of medical specialties in clinical practice.

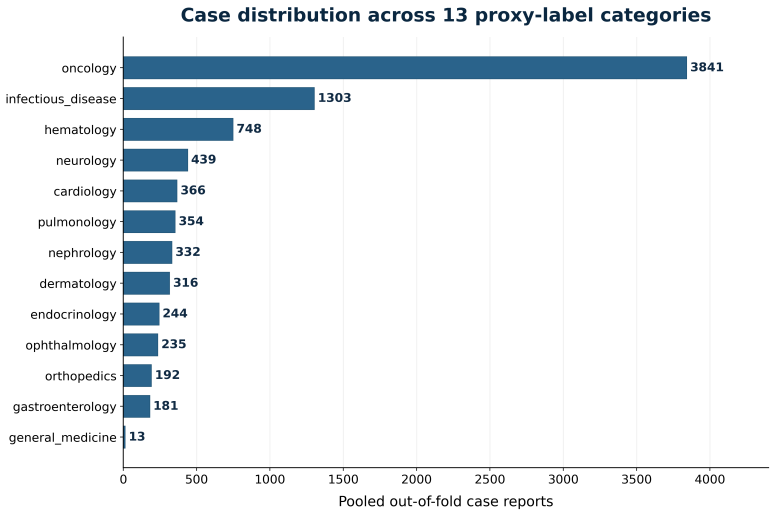


Figure 2. Distribution of pooled out-of-fold case reports across the 13 metadata-derived proxy-label categories.

4.2. Overall Classification Results

Table 2 summarizes the class-specific results obtained by aggregating the held-out predictions from the five article-grouped cross-validation folds. MultiCaRe-CERGFormer achieved an overall accuracy of 0.8152, a weighted-F1 score of 0.8146, and a macro-F1 score of 0.6727. Although the accuracy and weighted-F1 values indicate relatively strong aggregate classification performance, the lower macro-F1 reveals considerable variation in predictive effectiveness across the 13 specialty categories. This discrepancy is particularly relevant given the pronounced class imbalance, as weighted metrics assign greater influence to frequently occurring categories, whereas macro-F1 gives equal importance to each category regardless of its support.

At the individual-category level, oncology demonstrated the strongest performance, achieving an F1-score of 0.9187, precision of 0.9317, and recall of 0.9060. These results indicate that the model identified most oncology-labeled cases while maintaining a relatively low proportion of false-positive oncology predictions. Ophthalmology also achieved favorable performance, with an F1-score of 0.8148 and particularly high recall of 0.8894. Similarly, infectious disease yielded an F1-score of 0.7806, supported by recall of 0.8112 and precision of 0.7523. Together, these findings indicate that the proposed framework can effectively distinguish several well-represented or comparatively recognizable specialty categories within the evaluated dataset.

Nevertheless, performance was less consistent for several other categories. Gastroenterology recorded an F1-score of 0.4416, with precision of 0.5354 and recall of only 0.3757, indicating that a substantial proportion of gastroenterology-labeled cases were assigned to other categories. Orthopedics achieved a moderate F1-score of 0.6234, with closely matched precision (0.6218) and recall (0.6250). These results suggest that the model encounters greater difficulty distinguishing some specialty categories, potentially reflecting overlapping medical terminology, heterogeneous case descriptions, differences in category

representation, or uncertainty introduced by metadata-derived proxy labels. However, the relative contributions of these factors cannot be established from the class-level metrics alone.

Table 2. Class-specific classification performance based on pooled out-of-fold predictions across five article-grouped cross-validation folds.

Proxy-Label Category	Precision	Recall	F1-Score
Cardiology	0.6803	0.7732	0.7238
Dermatology	0.7072	0.7722	0.7383
Endocrinology	0.7576	0.7172	0.7368
Gastroenterology	0.5354	0.3757	0.4416
General medicine	0.0714	0.0769	0.0741
Hematology	0.7652	0.7059	0.7344
Infectious disease	0.7523	0.8112	0.7806
Nephrology	0.7174	0.7952	0.7543
Neurology	0.7338	0.7722	0.7525
Oncology	0.9317	0.9060	0.9187
Ophthalmology	0.7518	0.8894	0.8148
Orthopedics	0.6218	0.6250	0.6234
Pulmonology	0.7100	0.6017	0.6514
Overall Accuracy		0.8152	
Macro-F1		0.6727	
Weighted-F1		0.8146	

The most pronounced performance limitation was observed for general medicine, which achieved precision of 0.0714, recall of 0.0769, and an F1-score of 0.0741. Given that this category contained only 13 cases, its recall estimate is highly sensitive to individual prediction outcomes, and its low F1-score should not be interpreted as a stable estimate of performance on a broader population of general medicine reports. The very limited representation of this category also restricts the conclusions that can be drawn regarding the model's ability to recognize its associated textual characteristics.

Overall, the findings demonstrate that MultiCaRe-CERGFormer achieves comparatively strong aggregate performance for metadata-derived specialty classification but does not provide uniform predictive performance across all categories. The substantial difference between weighted-F1 and macro-F1 highlights the importance of evaluating minority categories independently rather than relying solely on overall accuracy. These observations also underscore the need for future investigations into class-aware learning strategies, improved representation of underrepresented specialties, and more rigorous assessment of proxy-label reliability. Importantly, the present results support the framework's potential utility for biomedical literature organization and specialty-oriented indexing within the evaluated dataset, but they should not be interpreted as evidence of clinically validated specialty identification or generalization to independent clinical populations.

As seen in Figure 3, correct predictions are clustered along the diagonal elements, especially in the oncology, infectious disease, and hematology categories (3480 of 3841, 1057 of 1303, and 528 of 748, respectively). In between, there were also some misclassifications (76 cases of hematology predicted as oncology, 73 cases of oncology predicted as hematology, and 93 cases of pulmonology predicted as infectious disease). In gastroenterology, 68 of 181 cases were correctly classified, with only 1 of 13 cases in general medicine correctly classified and 12 in infectious disease. The results should be considered in conjunction with the class supports, as the imbalance in the marking results in a less prominent visual representation of the categories; in the case of very small classes, the marking results are

less stable. The labels are derived from metadata; the matrix is a proxy for agreement with the labels and not necessarily for clinical diagnostic accuracy.

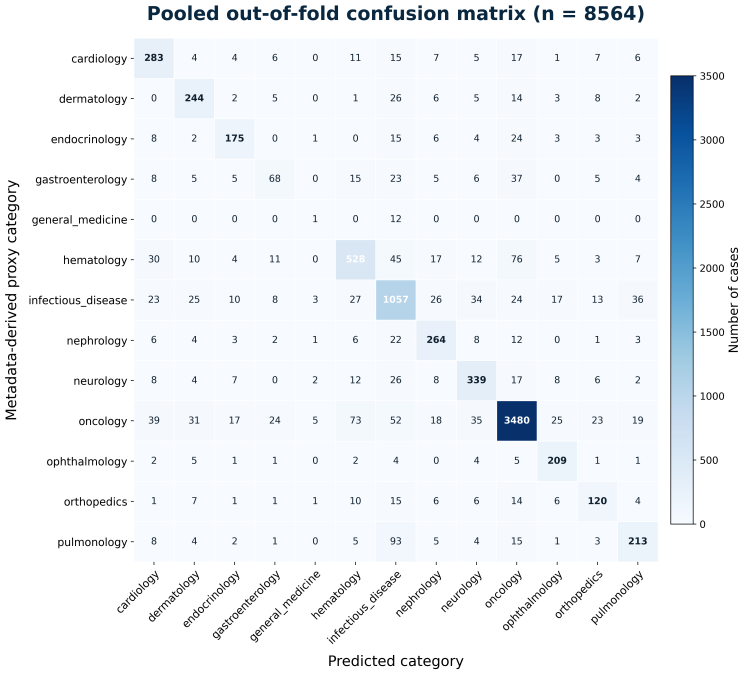


Figure 3. Pooled out-of-fold confusion matrix for the 13 metadata-derived proxy-label categories across 8564 case reports.

Figure 4 plots the training loss over time, which is seen to steadily decrease, and the validation loss over time, which is seen to start off at a low level and then increase to a higher level in the early epochs. The validation accuracy and macro-F1 also increase rapidly within a few epochs but then plateau and oscillate; macro-F1 decreases in the later epochs, while accuracy is relatively stable. The accuracy figure may mask changes in class-balanced performance, and using more training data may cause the gap between training and validation loss to widen. Hence, the prespecified validation criterion (macro-F1) should be used for checkpoint selection, not just the last epoch or accuracy.

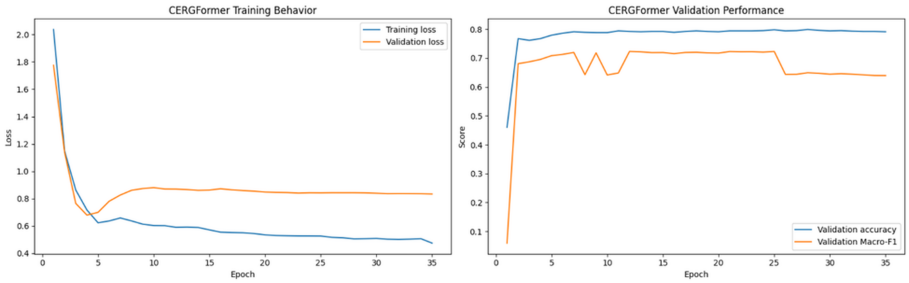


Figure 4. Training and validation behavior of MultiCaRe-CERGFormer across epochs.

The relationship between prediction coverage and classification accuracy, with cases kept by the model's uncertainty ranking, is shown in Figure 5. The accuracy improves as the coverage decreases, reaching about 0.99 for the most confidently ranked subset, and drops to about 0.82 when all evaluated cases are included. This trend shows that there are subsets that can be selected, which are more in agreement with the proxy labels derived from the metadata. However, the curve does not guarantee probability calibration

or uncertainty quantification or even clinical safety. It is also influenced by the imbalance of the specialty categories.

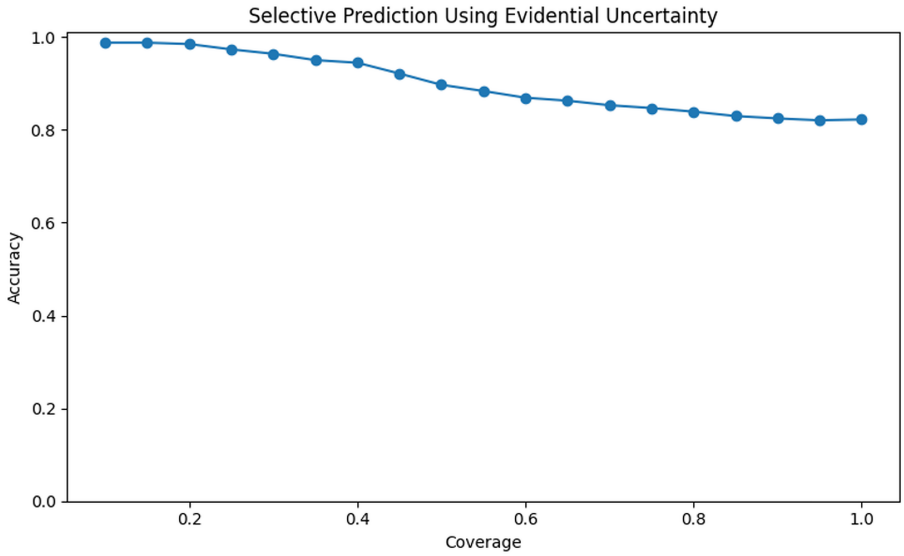


Figure 5. Selective prediction using evidential uncertainty.

Figure 6 summarizes the mean learned fusion weights and evidence-reliability scores for the four input branches of MultiCaRe-CERGFormer: lexical features, biomedical embeddings, medical concepts, and similar-case retrieval. The retrieval branch received the largest mean fusion weight (approximately 0.29) and the highest reliability score (0.95), followed by the medical concept branch (approximately 0.24 and 0.90, respectively). The lexical branch received a lower fusion weight (approximately 0.15) and reliability score (0.59), while the biomedical embedding branch had the smallest reported values (approximately 0.03 and 0.11, respectively). Within the evaluated dataset, these patterns indicate that the learned fusion mechanism placed greater emphasis on retrieval-derived and medical concept representations than on the lexical and biomedical embedding branches. The relatively high reliability scores assigned to the retrieval and concept branches reflect the model's internal evidence-scoring mechanism rather than independently verified evidence quality. Importantly, neither the fusion weights nor these reliability scores establish causal feature importance, clinical interpretability, or diagnostic validity. The contribution of the medical concept branch also warrants caution because its specialty-associated terminology may overlap with the publication metadata used to construct the proxy labels. Accordingly, these observations characterize the model's behavior on the evaluated dataset but do not independently demonstrate clinical validity or generalizability.

4.3. Baseline Comparison

Table 3 compares the proposed framework with classical lexical classifiers, alternative fusion strategies, component-restricted configurations, and biomedical embedding-based baselines under five-fold article-grouped cross-validation. The proposed MultiCaRe-CERGFormer ensemble achieved a mean accuracy of 0.8152 ± 0.0070, macro-F1 of 0.6693 ± 0.0179, macro-AUC of 0.9706 ± 0.0049, and macro-average precision (macro-AP) of 0.7430 ± 0.0227. It recorded the highest reported mean accuracy, weighted-F1 (0.8147), and macro-AUC among the evaluated configurations. Nevertheless, the word- and character-level TF-IDF SVM achieved a slightly higher macro-F1 (0.6705 ± 0.0264), while concatenation followed by MLP fusion yielded a marginally higher macro-AP (0.7433 ± 0.0344).

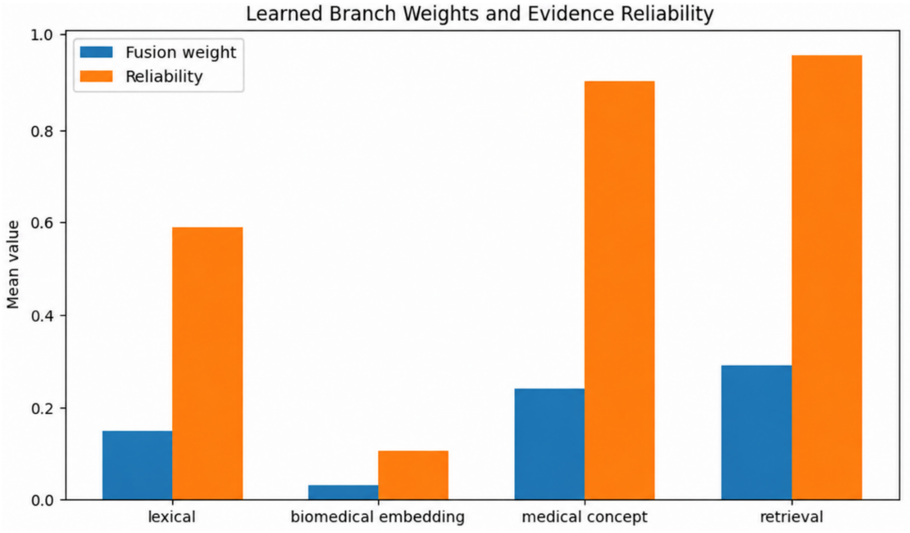


Figure 6. Mean learned branch weights and evidence-reliability scores.

Table 3. Five-fold article-grouped cross-validation comparison of MultiCaRe-CERGFormer and baseline configurations.

Model	Accuracy (Mean ± SD)	Macro-F1 (Mean± SD)	Weighted-F1 (Mean)	Macro-AUC (Mean ± SD)	Macro-AP (Mean ± SD)
MultiCaRe-CERGFormer, proposed ensemble	0.8152 ± 0.0070	0.6693 ± 0.0179	0.8147	0.9706 ± 0.0049	0.7430 ± 0.0227
Word + character TF-IDF SVM	0.8136 ± 0.0099	0.6705 ± 0.0264	0.8106	0.9639 ± 0.0066	0.7315 ± 0.0412
Validation-optimized ensemble	0.8137 ± 0.0061	0.6690 ± 0.0135	0.8131	0.9676 ± 0.0043	0.7361 ± 0.0334
Capacity-matched attentive softmax gate	0.8097 ± 0.0071	0.6650 ± 0.0179	0.8101	0.9656 ± 0.0035	0.7314 ± 0.0150
Concatenation + MLP fusion	0.8113 ± 0.0096	0.6605 ± 0.0143	0.8124	0.9678 ± 0.0045	0.7433 ± 0.0344
Equal-weight mean fusion	0.8105 ± 0.0116	0.6602 ± 0.0168	0.8109	0.9638 ± 0.0085	0.7331 ± 0.0386
MultiCaRe-CERGFormer neural core	0.8135 ± 0.0089	0.6600 ± 0.0177	0.8145	0.9653 ± 0.0042	0.7261 ± 0.0217
CERGFormer without retrieval	0.8099 ± 0.0079	0.6632 ± 0.0106	0.8105	0.9648 ± 0.0037	0.7298 ± 0.0057
CERGFormer without graph	0.8118 ± 0.0115	0.6616 ± 0.0172	0.8116	0.9656 ± 0.0043	0.7295 ± 0.0227
CERGFormer without counterfactual reliability training	0.8118 ± 0.0094	0.6556 ± 0.0166	0.8120	0.9630 ± 0.0033	0.7162 ± 0.0243
Hierarchical biomedical text only	0.7919 ± 0.0112	0.6496 ± 0.0161	0.7904	0.9610 ± 0.0072	0.6951 ± 0.0278
Learned hybrid retriever label posterior	0.7718 ± 0.0093	0.6133 ± 0.0091	0.7680	0.8539 ± 0.0120	0.6112 ± 0.0046
Frozen hierarchical BiomedBERT + logistic regression	0.6607 ± 0.0184	0.5362 ± 0.0198	0.6764	0.9328 ± 0.0095	0.5692 ± 0.0270

The relatively small differences among the leading models suggest that the proposed ensemble provides competitive, rather than uniformly superior, performance across the reported metrics. In particular, the strong results of TF-IDF SVM underscore the effectiveness of sparse lexical representations for the evaluated metadata-derived specialty classification task. The comparisons also suggest that the benefits of adaptive fusion depend on the performance criterion considered and should not be inferred solely from overall accuracy. Because these are descriptive differences in fold-aggregated results, they should not be interpreted as statistically significant improvements without appropriately designed paired comparisons. Taken together, the findings support the proposed ensemble as a viable approach to organizing case reports by metadata-derived specialty labels while highlighting the continued relevance of simpler, computationally efficient baselines.

Figure 7 compares the mean accuracy and macro-F1 scores of the evaluated classification models. The proposed MultiCaRe-CERGFormer ensemble achieved the highest mean accuracy (0.8152), whereas its macro-F1 score (0.6693) was comparable to that of the word- and character-level TF–IDF SVM (0.6705). The validation-optimized ensemble and attentive softmax-gated fusion also produced closely comparable results. In contrast, the biomedical-text-only and learned-retrieval configurations showed lower performance, particularly in macro-F1, while the frozen BiomedBERT–logistic-regression and label-term-only models exhibited weaker results on both measures. The predictive performance of the label-term-only model suggests that terminology associated with the metadata-derived specialty labels

contains useful classification signals. This observation reinforces the need to consider potential overlap between input terminology and proxy-label construction when interpreting the results. Overall, the figure indicates that the proposed ensemble is competitive across the reported metrics, although it does not consistently outperform all alternatives, and the small differences should not be interpreted as evidence of statistical superiority.

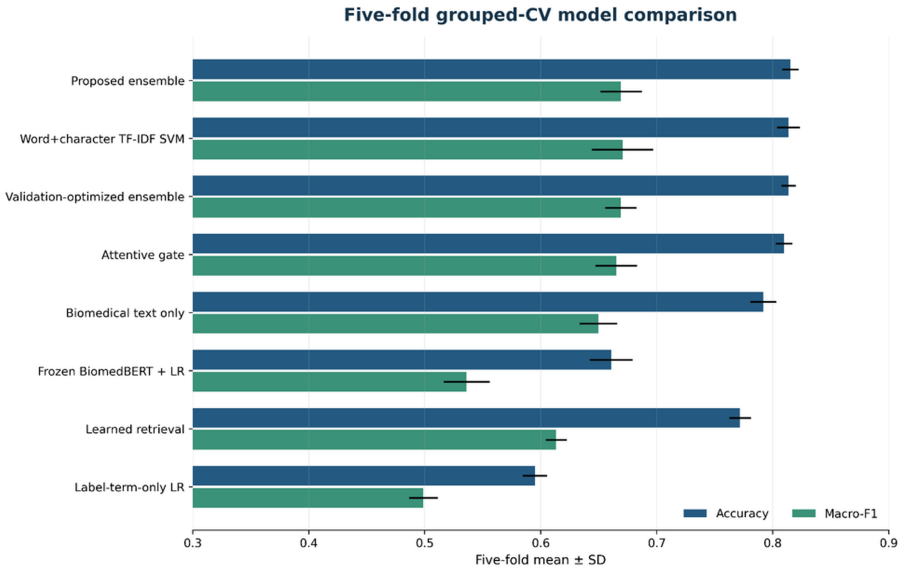


Figure 7. Five-fold grouped cross-validation comparison of accuracy and macro-F1.

4.4. Ablation Study

Table 4 evaluates the contributions of individual components and alternative fusion mechanisms within the MultiCaRe-CERGFormer framework. The proposed ensemble achieved the highest mean accuracy (0.8152 ± 0.0070) and macro-F1 (0.6693 ± 0.0179) among the configurations included in the ablation study. Removing counterfactual reliability training produced the largest macro-F1 reduction among the three component-removal experiments, decreasing the score to 0.6556 ± 0.0166, with a reported difference of 0.0137 ± 0.0089. Excluding the medical concept branch reduced macro-F1 to 0.6616 ± 0.0172, while removing the retrieval branch yielded 0.6632 ± 0.0106. These results are consistent with complementary contributions from the model components, although the observed differences alone do not establish the statistical significance of those contributions.

Table 4. Component-ablation and fusion-strategy comparison for MultiCaRe-CERGFormer using five-fold article-grouped cross-validation.

Variant	Accuracy (Mean ± SD)	Macro-F1 (Mean ± SD)	Δ Macro-F1 (Mean ± SD)
Proposed MultiCaRe-CERGFormer ensemble	0.8152 ± 0.0070	0.6693 ± 0.0179	—
Without counterfactual reliability training	0.8118 ± 0.0094	0.6556 ± 0.0166	0.0137 ± 0.0089
Without medical concept branch	0.8118 ± 0.0115	0.6616 ± 0.0172	0.0077 ± 0.0155
Without retrieval branch	0.8099 ± 0.0079	0.6632 ± 0.0106	0.0062 ± 0.0223
Attentive softmax-gated fusion	0.8097 ± 0.0071	0.6650 ± 0.0179	0.0043 ± 0.0242
Concatenation + MLP fusion	0.8113 ± 0.0096	0.6605 ± 0.0143	0.0088 ± 0.0196
Equal-weight mean fusion	0.8105 ± 0.0116	0.6602 ± 0.0168	0.0091 ± 0.0097
Biomedical embedding branch only	0.7919 ± 0.0112	0.6496 ± 0.0161	0.0197 ± 0.0203
Medical concept branch only	0.5992 ± 0.0146	0.3868 ± 0.0189	0.2825 ± 0.0185

Note: Δ Macro-F1 denotes the reported difference between the proposed ensemble and each comparison variant. Values are presented as the mean ± standard deviation across the five article-grouped cross-validation folds. The reported standard deviations of the differences are preserved as provided.

Among the alternative fusion configurations, attentive softmax-gated fusion achieved the highest macro-F1 (0.6650 ± 0.0179), followed by concatenation with an MLP (0.6605 ± 0.0143) and equal-weight mean fusion (0.6602 ± 0.0168). The biomedical embedding-only configuration obtained a macro-F1 of 0.6496 ± 0.0161, whereas the medical concept-only configuration performed substantially less well (0.3868 ± 0.0189). The latter result indicates that terminology-based concept indicators alone were insufficient to reproduce the performance of the combined model, despite their potential predictive overlap with the metadata-derived labels. Taken together, the ablation results support the value of combining heterogeneous representations, while the modest differences among several leading variants call for cautious interpretation. Paired statistical analyses would be needed to determine whether these differences are distinguishable from cross-validation variability.

Figure 8 shows the five-fold mean macro-F1 compared between the proposed ensemble, component ablations, and fusion controls (Figure 8) with error bars of ±1 standard deviation. The attentive softmax gate is close (around 0.665), and the proposed ensemble has the highest mean score (around 0.669). The mean values for retrieval training, graph training, and counterfactual training are about 0.663, 0.662, and 0.656, respectively, without these types of reliability training. The biomedical-text-only version is around 0.650, and the graph-only concept representation is significantly lower at about 0.387.

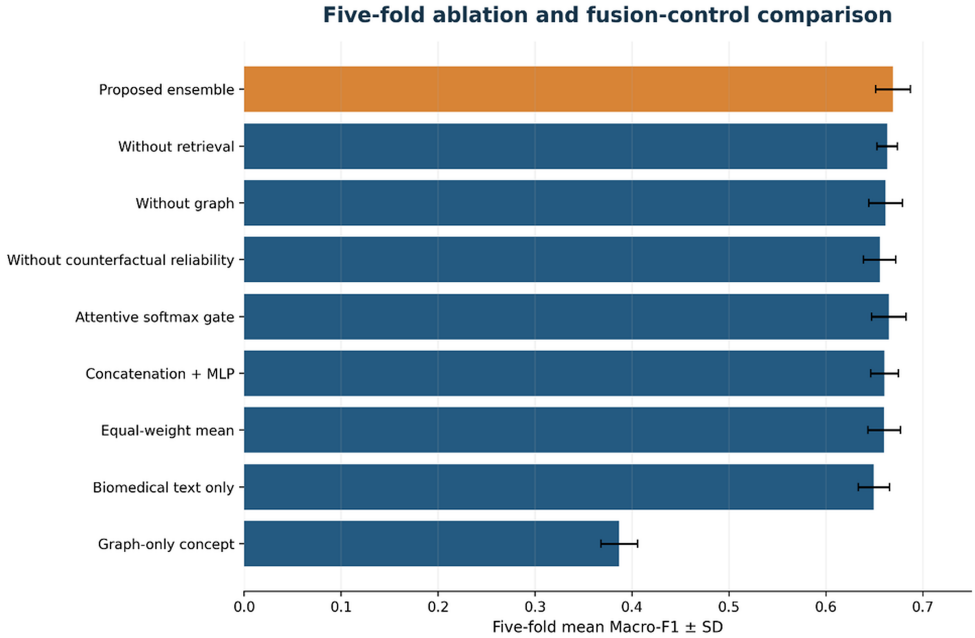


Figure 8. Macro-F1 comparison across the full and ablated models.

The lower the minimum top category score, the higher the acceptance rate, as can be seen in Figure 9. The acceptance rate drops rapidly between 5 and 6 top category scores. Increasing the separation between the score in the top category and the score in the second-highest category also decreases coverage, as more ambiguous articles are filtered out at each of the score thresholds. This illustrates the compromise between keeping more articles and using more restrictive criteria in assigning proxy labels. Its currently configured threshold of 4.0 and minimum margin of 1.5 resulted in the preservation of 6844 articles (8.99% coverage); this 100% agreement reported is consistent with its configured labeling rule and not independent clinical validation.

As shown in Figure 10, most variants converge towards good performance within a few epochs and then settle, and the validation macro-F1 is in the range of 0.65–0.71. The concept representation in the form of a graph is still significantly lower, around 0.37–0.39. The "MultiCaRe-CERGFormer | Neural Core" curve briefly hits approximately

0.72, followed by a drop to approximately 0.64 at the end of training, and several other variants hover around 0.68–0.71. The pattern is useful for choosing checkpoints based on a prespecified criterion for validation, instead of automatically picking the final epoch.

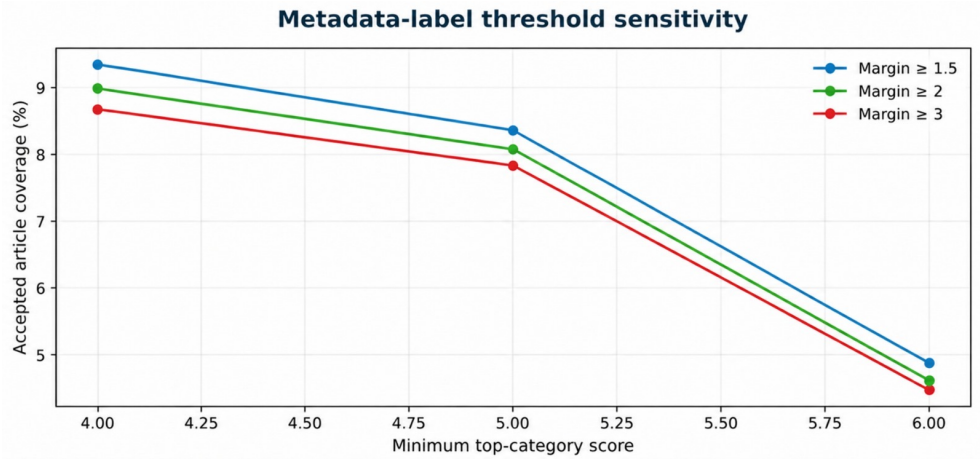


Figure 9. Label-threshold sensitivity.

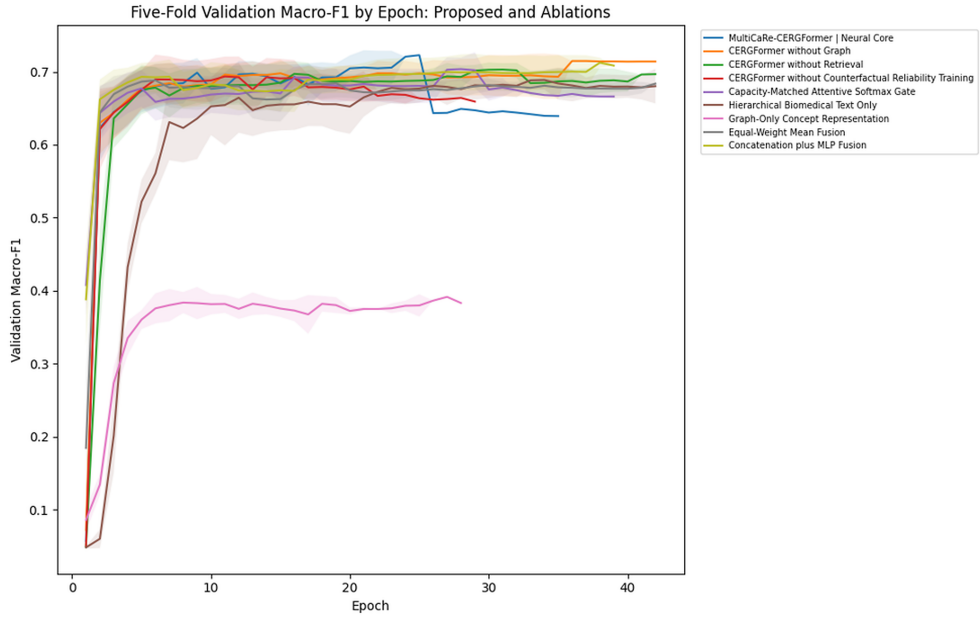


Figure 10. Five-fold validation macro-F1 across training epochs for the proposed model and ablation variants.

Figure 11 shows that most variants present a steep increase in validation accuracy in the initial epochs and then gradually level off between around 0.79 and 0.83. One of the highest mean-fusion curves, which is equal in weight, is shown, while the curve of the graph-only concept representation is still low, at approximately 0.58. The hierarchical biomedical-text-only variant has a less sudden improvement and gentler leveling off compared to most fusion variants. We can see from these trajectories that a lot of models have stable accuracy from early on, but we also note that accuracy is not necessarily a good indicator of performance on imbalanced classes; it needs to be taken in conjunction with the macro-F1 and class-level results.

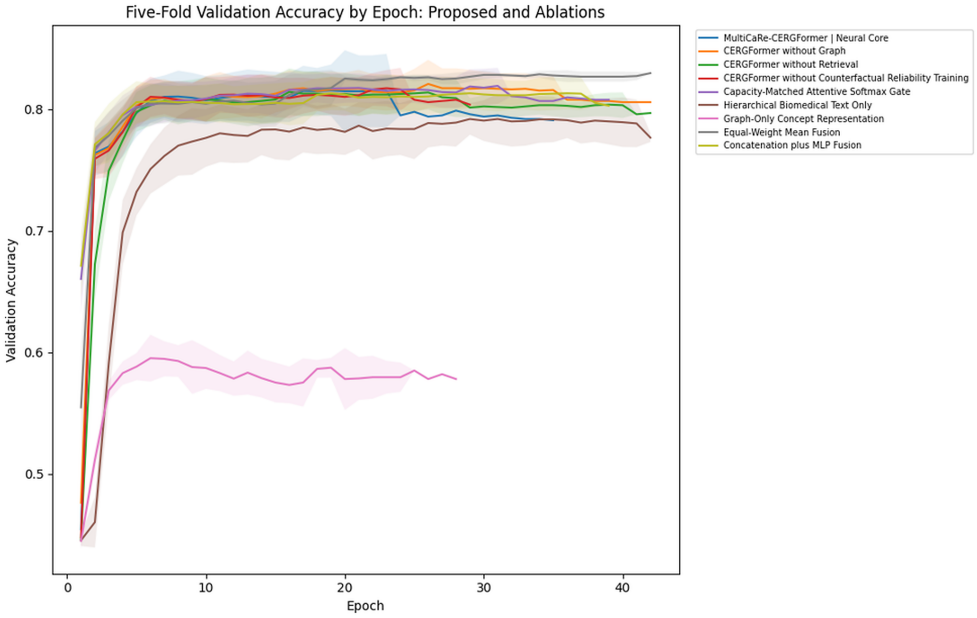


Figure 11. Five-fold validation accuracy across training epochs for the proposed model and ablation variants.

5. Discussion

The five-fold article-grouped cross-validation results indicate that MultiCaRe-CERGFormer is competitive for classifying clinical case reports into 13 metadata-derived specialty proxy categories. The proposed ensemble achieved a mean accuracy of 0.8152 ± 0.0070 and a macro-F1 score of 0.6693 ± 0.0179, alongside a macro-AUC of 0.9706 ± 0.0049. Although the proposed ensemble obtained the highest reported mean accuracy and macro-AUC among the compared configurations, the word- and character-level TF–IDF support vector machine (SVM) achieved a slightly higher macro-F1 score (0.6705 ± 0.0264). Concatenation with multilayer perceptron (MLP) fusion also yielded a marginally higher macro-average precision (0.7433 ± 0.0344) than the proposed ensemble (0.7430 ± 0.0227). These small numerical differences should not be interpreted as statistically significant improvements in the absence of paired inferential analyses. Overall, the findings support the feasibility of complementary multi-view representations but do not establish consistent superiority over strong lexical or conventional fusion baselines.

The discrepancy between macro-F1 and weighted-F1 highlights the importance of category imbalance when interpreting aggregate performance. The pooled out-of-fold evaluation comprised 8564 cases, including 3841 oncology cases and 1303 infectious disease cases, whereas general medicine was represented by only 13 cases. Class-specific F1-scores ranged from 0.9187 for oncology to 0.4416 for gastroenterology and 0.0741 for general medicine. The general-medicine estimate is particularly unstable because individual prediction outcomes can substantially affect metrics calculated from such a small sample. Accordingly, overall accuracy and weighted-F1 should be considered together with macro-F1 and class-specific results. These measures describe prediction of the retained metadata-derived proxy labels; they do not quantify the prevalence of specialties in clinical practice or establish clinically validated specialty recognition.

The strong performance of the TF–IDF SVM underscores the value of sparse lexical features for this benchmark. The proposed ensemble achieved modestly higher mean accuracy and macro-AUC, whereas the SVM exhibited a marginal advantage in macro-F1. This metric-dependent ranking suggests that combining lexical, semantic, concept-based, and retrieval-derived representations can be useful, but also that a computationally simpler

lexical classifier remains a demanding comparator. Similarly, the biomedical-text-only and learned-retrieval baselines yielded lower macro-F1 scores than the proposed ensemble. Such comparisons are consistent with the potential value of complementary evidence sources, although they do not establish that every branch or fusion mechanism makes a statistically significant contribution.

The ablation experiments provide additional descriptive evidence about the framework's components. Relative to the proposed ensemble, removing counterfactual reliability training was associated with a macro-F1 reduction of 0.0137, removing the medical concept branch with a reduction of 0.0077, and removing the retrieval branch with a reduction of approximately 0.0062. Alternative fusion strategies exhibited reported macro-F1 differences of 0.0043–0.0091 relative to the proposed ensemble. The medical-concept-only configuration achieved a substantially lower macro-F1 of 0.3868 ± 0.0189, suggesting that concept indicators alone are insufficient to reproduce the performance of the combined framework. Nevertheless, the modest differences among the strongest configurations, and their variation across folds, require cautious interpretation. Paired statistical testing would be needed to determine whether the observed differences are distinguishable from sampling variability. Moreover, the concept-only result does not demonstrate that concept-based features are independent of the proxy-label construction process.

A central methodological limitation concerns the use of specialty proxy labels generated from publication metadata. Labels were derived from information such as article titles, author keywords, journal metadata, and Medical Subject Headings (MeSH), whereas the model's medical concept branch identified specialty-associated terminology within case narratives. Although labels and model inputs originated from different representations, they could share terminology and topical signals. Consequently, part of the observed predictive performance may reflect the model's ability to recover regularities associated with the metadata-labeling rules rather than independently validated clinical specialty distinctions. The comparatively weak concept-only baseline does not eliminate this potential circularity. In particular, a systematic masking experiment targeting label-associated terminology and independent expert adjudication of specialty categories were not performed. The present findings should therefore be understood as results for an operational metadata-derived classification task, not as evidence of clinical diagnostic validity.

The retrieval and fusion components also require careful interpretation. Similar-case neighbors, detected medical concepts, and learned branch weights can support model inspection and error analysis by showing which representations or retrieved cases accompany particular predictions. However, they are not causal explanations, independently verified medical evidence, or diagnostic recommendations. Likewise, high internal evidence-reliability scores indicate the behavior of the model's scoring procedure, not externally validated clinical reliability. The usefulness of these outputs for expert review has not been established.

Several additional limitations constrain the conclusions. First, article-level metadata-derived labels may contain assignment errors and introduce terminology-related circularity. Second, evaluation was conducted on a selected subset of cases from a single source, without an independent external dataset. Third, pronounced category imbalance, particularly the 13-case general-medicine category, limits the stability of minority-class estimates. Fourth, article-grouped splitting prevented direct overlap of source-article groups across training and held-out partitions, but it cannot exclude semantic similarities between different publications or demonstrate robustness under distribution shift. Finally, the reported experiments used a single random seed; multi-seed variability, paired statistical significance analyses, and independent clinical validation were not established.

Future work should therefore prioritize independently adjudicated specialty labels, controlled masking of terminology associated with proxy-label generation, and evaluation on external collections with different sources and category distributions. Additional investigations should assess sensitivity to retrieval choices, quantify uncertainty in paired model comparisons, and test whether retrieved neighbors and fusion weights provide useful, reproducible information to domain experts. These analyses would help determine whether the additional complexity of MultiCaRe-CERGFormer provides meaningful benefits relative to well-tuned lexical and simpler fusion methods. Until such evidence is available, the framework is best viewed as a research approach to metadata-oriented biomedical case-report indexing and organization rather than a clinically validated classification or diagnostic system.

6. Conclusions

This study investigated MultiCaRe-CERGFormer, a multi-view framework integrating sparse lexical representations, precomputed biomedical document embeddings, specialty-associated medical concept indicators, and training-partition similar-case retrieval through adaptive gated fusion. Its objective was to examine whether complementary representations could support classification and organization of clinical case reports according to metadata-derived specialty proxy categories.

Under five-fold article-grouped cross-validation, the proposed ensemble achieved competitive results, with a mean accuracy of 0.8152 ± 0.0070 and a macro-F1 of 0.6693 ± 0.0179. However, a strong word- and character-level TF–IDF SVM produced a comparable macro-F1 score, and the proposed model did not lead on every evaluation metric. These findings demonstrate the feasibility of the multi-view design within the evaluated benchmark while emphasizing the need to assess any additional complexity against efficient lexical alternatives.

Interpretation is limited by weakly supervised metadata-derived labels, possible overlap between label-generation terminology and model features, substantial class imbalance, single-source evaluation, and the absence of external or clinician-adjudicated validation. Accordingly, the results support research into specialty-oriented biomedical literature indexing but do not establish clinical diagnostic utility or deployment readiness. Future studies should incorporate expert-validated labels, systematic label-term masking, external-dataset evaluation, multi-seed and paired uncertainty analyses, and rigorous assessments of the retrieval and fusion components' reliability and explanatory usefulness.

Funding: This work was funded by the Deanship of Graduate Studies and Scientific Research at Jouf University under grant no. DGSSR-2025-02-01030.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset is publicly available on Kaggle at https://www.kaggle.com/datasets/mauronievasoffidani/multicare/data (accessed on 1 June 2026).

Acknowledgments: The author extend his appreciation to the Deanship of Graduate Studies and Scientific Research at Jouf University for funding this research work.

Conflicts of Interest: The author declare no conflicts of interest.

References

1. Beltagy, I.; Peters, M.E.; Cohan, A. Longformer: The Long-Document Transformer. *arXiv* **2020**, arXiv:2004.05150.
2. Nievas Offidani, M.; Roffet, F.; González Galtier, M.C.; Massiris, M.; Delrieux, C. An Open-Source Clinical Case Dataset for Medical Image Classification and Multimodal AI Applications. *Data* **2025**, *10*, 123. https://doi.org/10.3390/data10080123.

3. Nievas Offidani, M.A.; Delrieux, C.A. Dataset of Clinical Cases, Images, Image Labels and Captions from Open-Access Case Reports from PubMed Central (1990–2023). *Data Brief.* **2024**, *52*, 110008. https://doi.org/10.1016/j.dib.2023.110008.
4. Nievas Offidani, M. *MultiCaRe: An Open-Source Clinical Case Dataset for Medical Image Classification and Multimodal AI Applications*; Zenodo: Geneva, Switzerland, 2025. https://doi.org/10.5281/zenodo.10079369.
5. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
6. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
7. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining. *Bioinformatics* **2020**, *36*, 1234–1240.
8. Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; Poon, H. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans. Comput. Healthc.* **2021**, *3*, 2.
9. Peng, Y.; Yan, S.; Lu, Z. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. In Proceedings of the BioNLP 2019 Workshop, Florence, Italy, 1 August 2019.
10. Alsentzer, E.; Murphy, J.R.; Boag, W.; Weng, W.-H.; Jin, D.; Naumann, T.; McDermott, M. Publicly Available Clinical BERT Embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop, Minneapolis, MN, USA, 7 June 2019.
11. Huang, K.; Altosaar, J.; Ranganath, R. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. *arXiv* **2019**, arXiv:1904.05342.
12. Salton, G.; Buckley, C. Term-Weighting Approaches in Automatic Text Retrieval. *Inf. Process. Manag.* **1988**, *24*, 513–523.
13. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297.
14. Hosmer, D.W.; Lemeshow, S.; Sturdivant, R.X. *Applied Logistic Regression*, 3rd ed.; Wiley: Hoboken, NJ, USA, 2013.
15. Deerwester, S.; Dumais, S.T.; Furnas, G.W.; Landauer, T.K.; Harshman, R. Indexing by Latent Semantic Analysis. *J. Am. Soc. Inf. Sci.* **1990**, *41*, 391–407.
16. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. In Proceedings of the EMNLP-IJCNLP, Hong Kong, China, 3–7 November 2019; pp. 3982–3992.
17. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-T.; Rocktäschel, T.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33.
18. Khattab, O.; Zaharia, M. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, 25–30 July 2020.
19. Douze, M.; Guzhva, A.; Deng, C.; Johnson, J.; Szilvasy, G.; Mazaré, P.-E.; Lomeli, M.; Hosseini, L.; Jégou, H. The Faiss Library. *arXiv* **2024**, arXiv:2401.08281.
20. Bodenreider, O. The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. *Nucleic Acids Res.* **2004**, *32*, D267–D270.
21. Lipscomb, C.E. Medical Subject Headings (MeSH). *Bull. Med. Libr. Assoc.* **2000**, *88*, 265–266.
22. Neumann, M.; King, D.; Beltagy, I.; Ammar, W. SciSpacy: Fast and Robust Models for Biomedical Natural Language Processing. In Proceedings of the BioNLP 2019 Workshop, Florence, Italy, 1 August 2019.
23. Kraljevic, Z.; Searle, T.; Shek, A.; Roguski, L.; Noor, K.; Bean, D.; Mascio, A.; Zhu, L.; Folarin, A.A.; Roberts, A.; et al. Multi-Domain Clinical Natural Language Processing with MedCAT: The Medical Concept Annotation Toolkit. *Artif. Intell. Med.* **2021**, *117*, 102083.
24. Kipf, T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
25. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph Attention Networks. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
26. Lu, J.; Yang, J.; Batra, D.; Parikh, D. Hierarchical Question-Image Co-Attention for Visual Question Answering. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2016; Volume 29.
27. Arevalo, J.; Solorio, T.; Montes-y-Gómez, M.; González, F. Gated Multimodal Units for Information Fusion. In Proceedings of the International Conference on Learning Representations Workshop, Toulon, France, 24–26 April 2017.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.