A Hybrid AI-Approach for High-Precision Binary and Multilabel Classification: Patient Preference Studies in Focus

Cesar A. Seminario Yrigoyen
Politecnico di Torino
Turin, Italy
s317031@studenti.polito.it

Francesco G. Gillio*
Politecnico di Torino
Turin, Italy
s305909@studenti.polito.it

Abstract

Medical research hinges on precision and efficiency, yet the overwhelming volume of biomedical literature challenges researchers to extract actionable insights, particularly in specialized fields like patient preference studies (PPS). Current retrieval systems, such as PubMed, often fail to provide precise, relevant results, hampering research workflows. This paper introduces a novel hybrid framework combining advanced AI-driven techniques and traditional machine learning models to streamline the identification and categorization of PPS. The first stage employs a high-precision binary classification model using PubMed-BERT and BioMed-BERT embeddings, coupled with k-NN and SVM classifiers, achieving a 95.9% true positive rate and 91.8% precision. This ensures high inclusivity while filtering irrelevant papers. The second stage extends the framework to multilabel classification, categorizing studies by clinical areas and intervention types. Here, Binary Relevance with Support Vector Classifiers achieves a remarkable Hamming loss of 0.022 and an F1-score of 0.81. Additionally, BERTopic facilitates the discovery of new categories, significantly reducing manual labeling effort. This framework not only enhances the precision and scalability of biomedical literature retrieval but also demonstrates the potential for integrating domain-specific embeddings and semi-supervised approaches. Future work will focus on refining these systems for broader applications in medical research, fostering faster and more reliable decision-making.

Keywords

Text Classification, BERT, LLM, Machine Learning, Patient Preference Studies, Text Embeddings, Binary Classification, Multilabel Classification, Imbalance, BERTopic, Topic Modeling

ACM Reference Format:
Cesar A. Seminario Yrigoyen and Francesco G. Gillio. 2024. A Hybrid AI-Approach for High-Precision Binary and Multilabel Classification: Patient Preference Studies in Focus. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX).* ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Medicine extends beyond the conventional boundaries of empirical sciences, merging the study of natural phenomena with a commitment to social well-being. While medical research seeks to uncover biological processes, its ultimate mission often lies in improving human health and quality of life. This dual purpose underscores the critical importance of precision and efficiency throughout the research process. Any deviation—whether in data collection, analysis, or validation—may lead to profound implications, potentially endangering lives. Unlike purely laboratory-based sciences, medicine uniquely incorporates social dimensions, relying heavily on patient perspectives and community feedback. This marriage of scientific rigor and social responsibility creates a complex landscape in which artificial intelligence (AI) can play a transformative role.

In this study, we propose an AI-driven approach to address a key challenge in medical research: the inefficiencies in semantic search and information retrieval from biomedical digital libraries. Existing systems, such as PubMed, often inundate users with an overwhelming volume of results. For example, a query like "Studies on Patient Preferences 2023" may generate thousands of documents, only a small fraction of which are truly relevant. These systems struggle to distinguish between papers that substantively explore patient preferences and those that merely mention the topic, creating significant barriers to efficient and precise data collection.

To tackle this issue, we evaluate a combination of contemporary and traditional natural language processing (NLP) techniques, aiming to refine PubMed search results for patient preference studies. Specifically, we investigate whether modern large language models (LLMs) can transcend the limitations of existing systems. However, practical constraints guide our study toward a more targeted approach. Fine-tuning LLMs, while promising, demands extensive datasets to transition from pattern memorization to comprehension. Our dataset, provided by the University of Turin, contains only 1,215 documents, of which approximately 200 focus on patient preference studies. While data augmentation offers a way to expand the dataset, it demands significant manual effort, domain expertise, and time to maintain clinical relevance. Given the resource-intensive nature of manual data augmentation—in terms of effort, domain expertise, and time—we explore a hybrid approach that balances innovation and practicality.

Our proposed solution combines the semantic strength of advanced models like BERT [5] with the classification capabilities of traditional methods such as Support Vector Machines (SVM) and k-Nearest Neighbors (k-NN). These traditional models have demonstrated robustness in clinical and biomedical data classification [3] [2], making them suitable candidates for integration into a unified

[*] Both authors contributed equally to this research.

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

system. While this hybrid approach initially targets binary classification to filter relevant studies on patient preferences, it reveals a broader challenge intrinsic to biomedical research: the need for multilabel classification. Biomedical datasets often reflect the multifaceted nature of medical studies, where documents span multiple categories simultaneously. For example, a single study might examine diabetes management, patient preferences, and intervention outcomes, necessitating multilabel categorization to capture its full scope.

Transitioning from binary to multilabel classification aligns with the overarching goal of improving biomedical data retrieval. Binary classification models, while straightforward and reliable, often oversimplify the complexity of medical research. Multilabel classification, by contrast, allows for richer representations of biomedical data but introduces challenges in dataset construction, annotation, and algorithmic design. Recent advancements in multilabel classification algorithms and evaluation metrics have made these approaches more feasible [12] [10] [7], paving the way for their application in complex biomedical tasks.

In this study, we bridge the gap between binary and multilabel classification by exploring two multilabel datasets derived from the same corpus of research papers, categorized under Clinical Areas and Interventions. To analyze these datasets, we evaluate two primary approaches:

(1) **Problem Transformation Approaches**: Decomposing the multilabel task into multiple binary classification problems, enabling the use of traditional classifiers on each subset.
(2) **Algorithm Adaptation Approaches**: Extending traditional classifiers to directly handle multilabel data while preserving inter-label relationships.

Among the models tested, the Binary Relevance (BR) approach [6], paired with a Support Vector Classifier (SVC), demonstrated the best performance, achieving a Hamming loss of 0.02 and a micro-averaged F1-score of 0.82. Similarly, the RAkEL (Random k-labelsets) approach, combined with a Gaussian Naïve Bayes classifier, showed comparable efficacy, underscoring its potential for multilabel tasks. To complement classification, we applied topic modeling using the BERTopic model [8], uncovering novel categories such as skin and urinal diseases under Clinical Areas. This highlights the utility of topic modeling for refining multilabel datasets and identifying emerging trends in biomedical research.

By tackling inefficiencies in biomedical information retrieval, this study directly addresses a critical gap in ensuring that medical research supports timely and accurate decision-making, thereby reducing potential risks to patient care. Our hybrid approach, integrating advanced AI models with traditional classifiers, seeks to refine semantic search systems, minimizing the likelihood of overlooking or misinterpreting key studies. In doing so, we aim to enhance the precision and reliability of research workflows, ensuring that the knowledge guiding medical practice is both accessible and actionable, ultimately contributing to the protection and improvement of human health.

2 Related Work

The convergence of large language models (LLMs) and traditional machine learning (ML) techniques represents a pivotal domain in clinical prediction and biomedical natural language processing (NLP). Recent research has delved into the comparative strengths and limitations of these approaches, shedding light on their distinct roles and trade-offs. For example, Chen et al. (2024) introduced **ClinicalBench**, a comprehensive benchmarking framework to evaluate LLMs against traditional ML models. Their results highlight that while LLMs excel in general-purpose tasks, they often fall short in achieving the domain-specific precision required for clinical applications [3]. Complementing this, Brown et al. (2024) emphasize that traditional ML methods, which rely heavily on feature engineering and task-specific optimization, frequently outperform LLMs in specialized healthcare tasks [2].

Further advancing this field, Gu et al. (2020) developed **PubMedBERT**, a pre-trained language model tailored specifically to biomedical literature. By leveraging domain-specific corpora, PubMedBERT demonstrates the potential of task-specific pre-training to significantly enhance performance in specialized NLP applications [9]. Collectively, these studies underscore the critical importance of aligning model selection and evaluation strategies with the specific demands of the task, particularly in the nuanced realm of biomedical research.

While much of this work has focused on binary classification, multilabel classification presents analogous challenges. In these scenarios, assigning multiple labels to a single instance introduces added complexity, particularly when working with imbalanced datasets where some categories are severely underrepresented. Effectively addressing these issues requires not only robust classification techniques but also evaluation strategies sensitive to dataset imbalances.

In this context, our study aims to bridge these gaps by developing and evaluating multilabel classification models explicitly designed to handle imbalanced datasets. Multilabel classification often amplifies the difficulties associated with imbalance, demanding specialized methods that go beyond traditional frameworks. We built on the foundational contributions of Tarekegn et al. [12], Haixiang et al. [10], and Tsoumakas et al. [7], whose works explore problem transformation methods, algorithm adaptation, and ensemble approaches for multilabel classification. These studies also introduce critical metrics and evaluation criteria tailored to imbalanced data scenarios.

Our methodology leveraged insights from these foundational works to guide the selection and testing of models. Among the strategies explored, we incorporated advancements such as the **Value-Weighted k-Nearest Neighbors (VW-kNN)** model [14], specifically designed to address label imbalance. The selection process was informed by a detailed analysis of dataset characteristics, focusing on multilabel properties and the degree of imbalance across categories. To evaluate model performance, we adopted a diverse array of metrics designed to capture the nuanced requirements of biomedical applications. This comprehensive evaluation provides valuable insights into the strengths and limitations of different approaches, contributing to a deeper understanding of their practical relevance in biomedical research.

3 Method

3.1 Binary Classificator

In binary text classification, the objective is to design a function $f : X \to \{0, 1\}$ that maps an input text X to a binary class label. Challenges such as limited and imbalanced datasets are common in this task, making robust text representation crucial. Large language models (LLMs), such as BERT [5], are particularly well-suited for this purpose. These models leverage attention mechanisms to capture contextual and semantic relationships, encoding text into dense numerical vectors.

Given an input sequence $X = [w_1, w_2, \ldots, w_N]$, BERT produces a representation $H(X) \in \mathbb{R}^{N \times d}$, where each token w_i is embedded as a d-dimensional vector (with $d = 768$ for standard BERT). Additionally, the output includes a special CLS token, which summarizes the global context of the sequence. For downstream tasks like classification, two primary strategies are common: using the CLS token as a global vector or averaging the token embeddings. Both methods yield a single dense vector $V \in \mathbb{R}^d$ representing the input text.

3.1.1 Handling Multiple Text Inputs. In biomedical contexts, documents often comprise distinct textual components such as titles (T) and abstracts (A). For our task, we focused on leveraging only these two components to create a practical system without relying on full-text articles. To integrate T and A into a unified representation, we experimented with several strategies:

(1) **Single Component Representations**:
- Use only the title: $V = H(T)$.
- Use only the abstract: $V = H(A)$.

(2) **Merged Representations**:
- Concatenate the title and abstract: $V = H([T; A])$.

(3) **Combined Representations**:
- Compute separate embeddings for the title and abstract, and then:
  (a) **Sum**: $V = H(T) + H(A)$,
  (b) **Concatenate**: $V = [H(T); H(A)]$,
  (c) **Weighted Average**: $V = \alpha H(T) + (1 - \alpha)H(A), \quad \alpha \in [0, 1]$.

These strategies ensure flexibility in integrating textual inputs, allowing us to assess the relative contribution of titles and abstracts to classification performance.

3.1.2 Binary Classification Pipeline. The binary classification task is structured into a pipeline, as illustrated in Figure 1. The pipeline comprises two main stages:

(1) **Middle-Lower Stage**: This stage encodes the input sequences (T, A) into a dense numeric vector $V = H(T, A)$ using the BERT model.
(2) **Upper Stage**: This stage maps the vector representation V to the binary class label ($y \in \{0, 1\}$) using a classifier $f(V) \to y$.

To evaluate this pipeline, we implemented a comprehensive evaluation framework:

- **Middle-Lower Stage**: We assessed 10 pre-trained BERT models from the Hugging Face library, resulting in 20 combinations for encoding T and A (10 models × 2 configurations). Each model was evaluated using a linear Support Vector Machine (SVM) classifier as a placeholder for the upper stage. SVM was selected due to its ability to identify optimal linear hyperplanes, maximizing the margin $M = \frac{2}{||w||}$ between classes through the optimization problem:

$$\text{minimize} \quad \frac{1}{2}||w||^2 \quad \text{subject to} \quad y_i \left(w^T x_i + b\right) \geq 1, \quad \forall i.$$

  This stage allowed us to compare the effectiveness of the text representations generated by different BERT models.
- **Upper Stage**: After selecting the best-performing BERT configuration, we fixed it as the encoder and evaluated 10 different binary classification algorithms from the Scikit-Learn library, including k-Nearest Neighbors, Random Forests, and SVMs. To improve robustness, we also tested ensemble methods, which aggregate predictions from M classifiers. The final decision y is computed as:

$$y = \text{mode}(y_1, y_2, \ldots, y_M),$$

  where $\text{mode}(y_1, y_2, \ldots, y_M)$ represents the most frequent class label.

3.1.3 Evaluation Metrics. To evaluate model performance, we adopted metrics tailored to the study's priorities. While traditional metrics like AUC-ROC and F1-score are commonly used, we focused on metrics that minimize critical risks in biomedical contexts. Specifically, we aimed to maximize the identification of relevant studies (class 1) while ensuring no studies of interest were overlooked.

- **True Positive Rate (TPR)**: Measures the proportion of correctly identified relevant studies:

$$\text{TPR} = \frac{\text{TP}_1}{\text{TP}_1 + \text{FN}_1}.$$

- **Positive Predictive Value (PPV)**: Evaluates the precision of identifying relevant studies:

$$\text{PPV} = \frac{\text{TP}_1}{\text{TP}_1 + \text{FP}_1}.$$

Here, TP_1, FN_1, and FP_1 represent the true positives, false negatives, and false positives for class 1, respectively. Our evaluation framework prioritizes high recall for relevant studies (TPR), ensuring the model effectively identifies all patient preference studies before optimizing precision (PPV).

By designing the evaluation in two stages, we mitigate the risk of overlooking relevant studies, ensuring the system prioritizes completeness before refining accuracy. This approach aligns with the ultimate goal of improving biomedical information retrieval without compromising critical patient-related outcomes.

3.2 Multilabel Classificator

PPS multilabels are analyzed to evaluate key dataset characteristics, such as imbalance and complexity. A portion of the dataset, consisting of unlabeled data, undergoes separate analysis through topic modeling techniques to identify additional potential categorizations or labels.

The full dataset, inclusive of the previously unclassified instances, is then used to train various multi-label classification models. Only

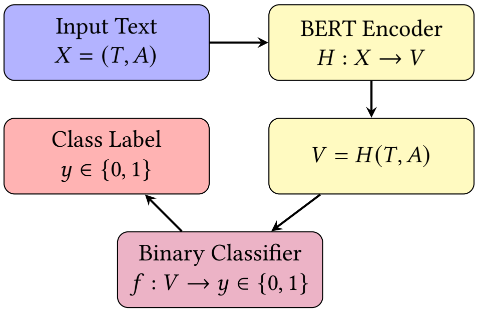


Figure 1: Binary Classification Pipeline

the models exhibiting optimal performance based on hyperparameter tuning are selected for further evaluation. These selected models are retrained using the entire dataset to ensure comprehensive learning and robustness. Finally, the performance of each model is assessed using a set of predefined classification metrics to determine their effectiveness in the multi-label classification task.

3.2.1 ML classificators. Five different models have been evaluated:

MLkNN[15] is an extension of the traditional k-Nearest Neighbors (kNN) algorithm, designed for multi-label classification. Unlike standard kNN, which assigns a single label to each instance, MLkNN predicts multiple labels by analyzing the label distribution among the k nearest neighbors in the training set. This model is simple and intuitive, effectively handling multi-label correlations and addressing class imbalances.

VW MLkNN[14] is an advanced version of the Multi-Label k-Nearest Neighbors (MLkNN) algorithm. VW MLkNN is designed to better account for instance similarity and label relevance in multi-label classification tasks. It introduces a weighting mechanism that refines the influence of the k-nearest neighbors based on their similarity to the query instance and the importance of their associated labels. This enhancement allows the algorithm to prioritize more relevant neighbors according to their similarity to the query. VW MLkNN is well-suited for handling complex relationships in high-dimensional datasets and is capable of managing noisy neighbors. However, the introduction of the weighting mechanism can increase computational complexity. Based on the article [14], a specific implementation has been created and fine-tuned using chatgpt.

Binary Relevance [6]: is a problem transformation technique that decomposes the task into independent binary classification problems, with one for each label. For a dataset with multiple labels, the Binary Relevance model trains several binary classifiers, each responsible for predicting the presence or absence of a specific label independently of the others. For this task, three classifiers are employed: k-Nearest Neighbors (kNN), Multinomial Naïve Bayes, and Support Vector Classifier (SVC)

RAkEL (Random k-Labelsets) [13] is a problem transformation approach for multi-label classification that builds upon the Label Powerset (LP) method. Instead of treating all labels simultaneously, as in LP, RAkEL reduces the complexity of the label space by dividing the labels into randomly selected subsets, known as k-labelsets. A separate classifier is trained for each subset, and the predictions from these classifiers are combined using an ensemble strategy. This method enhances robustness and accuracy by combining the predictions of multiple LP classifiers and capturing label correlations within each k-labelset. However, the quality and diversity of the k-labelsets, as well as careful selection of these subsets, can impact performance. Two classifiers have been investigated: Gaussian Naive Bayes and Random Forest.

Multilabel Twin SVM [4]: is an extension of Twin Support Vector Machines (TWSVM) designed for multi-label classification. This model formulates the problem as multiple binary classification tasks, one for each label, utilizing the efficiency of the TWSVM framework. The Multilabel Twin SVM constructs two non-parallel hyperplanes for each label and aims to optimize the separation of instances belonging to each label from those that do not. It does not consider label correlation and is sensitive to hyperparameter tuning.

Topic Modeling Analysis [8]: For this investigation, which was conducted on unlabeled data, the Bertopic model has been used. Bertopic is a topic modeling framework that leverages advanced natural language processing (NLP) techniques, including pre-trained transformers like BERT, combined with dimensionality reduction and clustering algorithms. This approach provides interpretable topic modeling while preserving the semantic relationships in textual data.

4 Experiments

4.1 Managing Binary Dataset and Class Imbalance

Our experimentation commenced with the dataset provided by the University of Turin, which represents PubMed's search results for patient preference studies conducted in 2023. This dataset consists of 1,215 entries, each corresponding to a scientific article, alongside 17 columns. Key columns include **Title**, **Abstract**, and **Class Label**, with the latter distinguishing studies into two categories: relevant (class 1) and non-relevant (class 0). The dataset exhibits a pronounced class imbalance, with 986 articles labeled as non-relevant and only 229 as relevant.

A word frequency analysis revealed that the term *"patient preferences"* is frequently mentioned in both classes, emphasizing its thematic relevance across the dataset. Notably, the text length of both titles and abstracts falls within BERT's 512-token limit, eliminating the need for additional pre-processing steps.

Given the dataset's class imbalance, we adopted a cross-validation strategy to ensure an equitable distribution of samples across training and testing subsets. This approach maintained proportional class representation across all folds, providing a more robust and reliable performance evaluation framework. Despite the availability of advanced text augmentation models, their limitations in the clinical and biomedical domains prompted us to focus on balancing strategies without resorting to synthetic data generation [16].

4.2 Evaluation of BERT Models

We initiated our experimentation by evaluating pre-trained BERT-based models across two input formats: **Title (T)** and **Abstract (A)**. Despite freezing the models' weights to reduce computational demands, the complexity of these models required several hours of

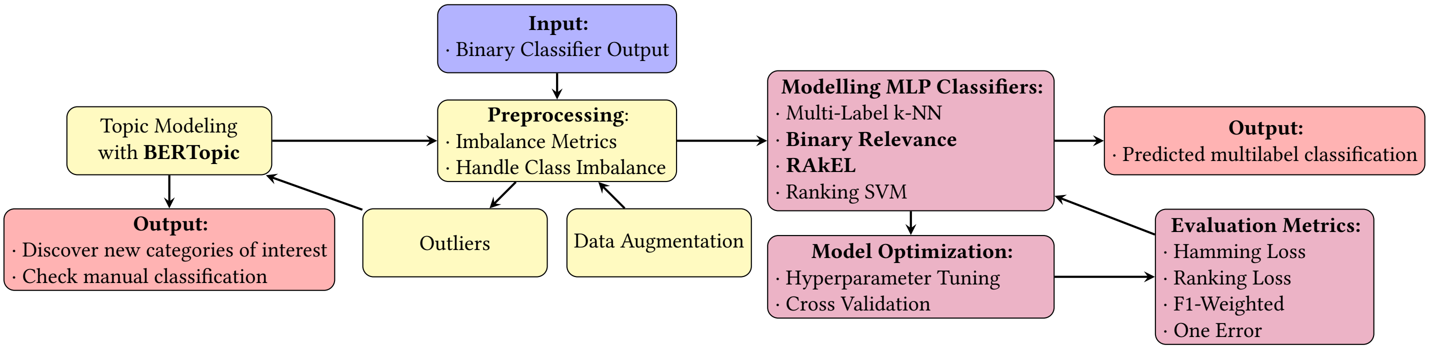


Figure 2: Multilabel Classification Investigation Pipeline

GPU-accelerated processing on a Tesla T4 (2,560 CUDA cores, 16 GB GDDR6).

Table 1 summarizes the performance of the top three models—**PubMed-BERT**, **BioMed-BERT**, and **PubMed-BERT-MS-MARCO**—in terms of True Positive Rate (TPR), Positive Predictive Value (PPV), and Area Under the Precision-Recall Curve (AUC-PR).

BERT-Base Model	TPR	PPV	AUC-PR
PubMed-BERT	0.831	0.844	0.915
BioMed-BERT	0.822	0.825	0.893
PubMed-BERT-MS-MARCO	0.819	0.803	0.889

Table 1: Performance metrics of top-performing BERT-based models.

PubMed-BERT emerged as the leading model, delivering superior performance across all metrics. BioMed-BERT, while slightly behind, demonstrated competitive results and was also selected for further exploration to ensure comprehensive evaluation.

4.3 High-Performance Classifier Chains

Building on the BERT models' embeddings, we tested various traditional machine learning classifiers for binary classification. Through iterative hyperparameter tuning, two classifier chains demonstrated exceptional performance:

(1) **PubMed-BERT with k-Nearest Neighbors (k-NN)**
(2) **BioMed-BERT with Support Vector Machine (SVM) using an RBF kernel**

As detailed in Table 2, these chains addressed the dataset's requirements effectively:

Encoder	Classifier	TPR	PPV
PubMed-BERT	k-NN	**0.953**	0.719
BioMed-BERT	SVM (RBF)	0.903	**0.918**

Table 2: Results of high-performance classifier chains.

The PubMed-BERT and k-NN chain achieved an impressive TPR of 0.953, minimizing the risk of overlooking relevant studies. Conversely, the BioMed-BERT and SVM chain excelled in precision, achieving a PPV of 0.918, aligning with the goal of filtering irrelevant documents.

4.4 Ensemble Method Design

To leverage the complementary strengths of these classifier chains, we developed a soft majority voting ensemble. By testing numerous weight and threshold combinations, we identified the optimal parameters shown in Table 3.

α (PubMed)	β (BioMed)	Threshold
0.4375	0.5625	0.3875

Table 3: Optimal parameters for the ensemble method.

This ensemble achieved a harmonious balance between TPR and PPV, as evidenced by the confusion matrix in Table 4.

Actual	Predicted	
	Non-PPS	PPS
Non-PPS	96.68% (TN)	3.32% (FP)
PPS	4.08% (FN)	95.92% (TP)

Table 4: Confusion matrix showing the relationship between actual and predicted classes.

This configuration achieved minimal false negatives (FN: 4.08%) and false positives (FP: 3.32%), striking an effective balance between the high TPR of the k-NN chain and the high PPV of the SVM chain. These results align with the dual objectives of this research: retaining all relevant studies while excluding irrelevant ones with high precision.

4.5 ML (Multilabel) Datasets

From the identified PPS-related papers, two distinct categorizations were established: Clinical Areas and Interventions. These categorizations are registered manually and are represented as multilabel datasets, which characteristics [1] [12] are investigated in the table 5. The Clinical Areas dataset contains 12 labels and 88 unique labelsets (distinct combinations of labels). Furthermore, its complexity, measured at 13.19, exceeds that of the Interventions dataset, which

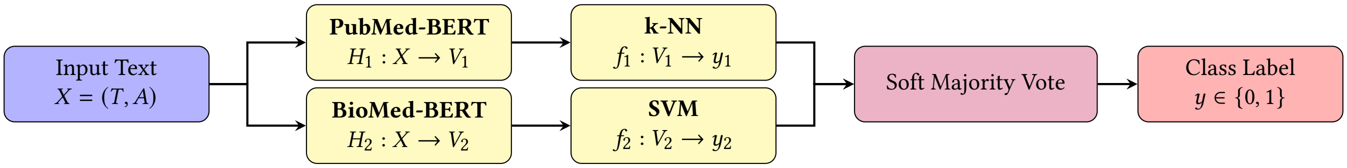


Figure 3: Ensemble Model

ML dataset	Labels	Label-sets	Card	Dens	TCS
Clinical Areas	12	88	0.87	0.07	13.19
Interventions	6	23	0.63	0.11	11.12

Table 5: Parameters of multilabel datasets

has a complexity of 11.12. The datasets were further examined for imbalance using various imbalance metrics [12]. Notably, both datasets exhibit a similar meanIR (mean imbalance ratio), with a slightly higher value observed for the Interventions dataset (table 6) as well as CVIR (coefficient of variation of imbalance ratio), a metric akin to the standard deviation. To conclude the analysis of the data set, the correlation between labels was assessed. Understanding label correlations is critical for selecting an appropriate multilabel classification (MLC) model. Certain models, such as Binary Relevance, do not inherently consider label dependencies, while others, like MLkNN, preserve such correlations if they exist. In clinical domains, the correlation between labels is very low (less than 0.1). Similarly, for the intervention dataset, the correlation remains below 0.1. However, it is slightly higher for imaging and screening, reaching 0.29. This suggests to use correlation-conservative models with this categorization.

ML dataset	mean IR	Max IR	CVIR	Scumble
Clinical Areas	4.53	12.20	1.91	0.63
Interventions	**4.63**	8.82	2.37	0.44

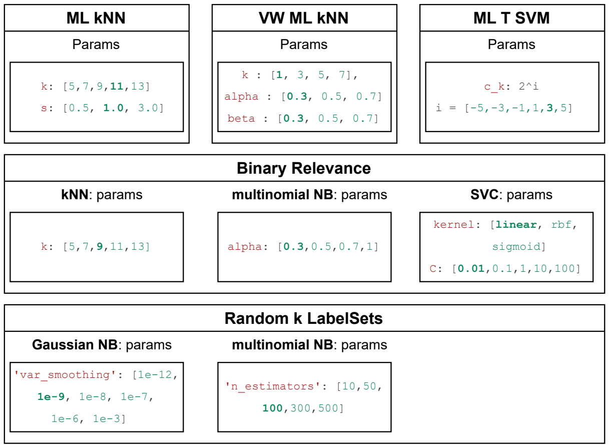
Table 6: Imbalance analysis of multilabel datasets

4.6 MLC Experiment Configuration

4.6.1 Multilabel classificator hyperparams finetuning & datasets subsets. The primary goal of the multilabel classification component is to evaluate the performance of the models described in chapter 3.2.1. For this task their optimal hyperparameters are studied. The complete test set is described in the scheme 4. The performance of the models is primarily evaluated using the Hamming Loss, a metric extensively applied in multilabel classification due to its straightforward interpretation and its suitability for handling imbalanced datasets.

As per the binary classifier, to ensure an imbalanced approach and a robust evaluation, cross-validation with three, five, and seven folds has been incorporated into the analysis.

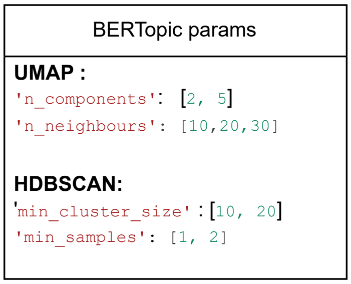
Once the optimal hyperparameters are determined, the models are retrained using the complete dataset under the identified configurations.



```
k: [5,7,9,11,13]
s: [0.5, 1.0, 3.0]

k : [1, 3, 5, 7],
alpha : [0.3, 0.5, 0.7]
beta : [0.3, 0.5, 0.7]

c_k: 2^i
i = [-5,-3,-1,1,3,5]

k: [5,7,9,11,13]

alpha: [0.3,0.5,0.7,1]

kernel: [linear, rbf,
sigmoid]
C: [0.01,0.1,1,10,100]

'var_smoothing': [1e-12,
1e-9, 1e-8, 1e-7,
1e-6, 1e-3]

'n_estimators': [10,50,
100,300,500]
```

Figure 4: Hyperparams optimization grid

The unlabeled data, comprising 25.5% and 40.8% of the instances in the clinical areas and interventions datasets, respectively, was utilized to train the **BERTopic** model. Figure 5 presents the experiments conducted to optimize the model configuration.



```
'n_components': [2, 5]
'n_neighbours': [10,20,30]

'min_cluster_size':[10, 20]
'min_samples': [1, 2]
```

Figure 5: BERTopic model blocks finetuning

4.7 MLC Evaluation

4.7.1 Multilabel metrics. The main metric analyzed is the **Hamming Loss**, which is defined as the fraction of incorrectly predicted labels (false positives and false negatives) over the total number of labels across all instances. The formula is given by:

$$\text{Hamming Loss} = \frac{1}{n \cdot q} \sum_{i=1}^{n} \sum_{j=1}^{q} \mathbb{I}\left(y_{ij} \neq \hat{y}_{ij}\right)$$

where:

- n: Total number of instances in the dataset.
- q: Total number of labels.
- y_{ij}: True label for the j-th label of the i-th instance ($y_{ij} \in \{0, 1\}$).
- $\hat{y}_{ij}$: Predicted label for the j-th label of the i-th instance ($\hat{y}_{ij} \in \{0, 1\}$).
- $\mathbb{I}(\cdot)$: Indicator function, which equals 1 if the condition inside is true ($y_{ij} \neq \hat{y}_{ij}$) and 0 otherwise.

In addition to Hamming Loss, other metrics are utilized to evaluate the performance of the models. These are briefly explained below:

Ranking Loss: This metric assesses the proportion of label pairs that are incorrectly ordered by the model. A label pair is considered incorrectly ordered if the model assigns a higher score to an irrelevant label than to a relevant one. Lower values of ranking loss indicate better model performance.

F1-Score (Micro & Weighted): The F1-score is used to balance precision and recall in the evaluation of classification models. The micro F1-score computes the unweighted average of F1-scores across all labels, treating all labels equally, whereas the weighted F1-score accounts for label frequency, giving more weight to labels with higher occurrences.

Coverage Error: This metric evaluates how many steps are required in the ranked list of predicted labels to include all the true labels for a given sample. Lower values indicate superior performance.

One Error: One error measures the frequency at which the top-ranked label, according to the model's predictions, is not among the true labels. A lower one error value reflects better performance.

These metrics collectively provide a comprehensive evaluation of the models, addressing various aspects of their performance such as label ranking, precision-recall balance, and prediction accuracy.

For the ***BERTopic*** model, the evaluation considers the *diversity* metrics: it measures the uniqueness and variety of words across the topics. Higher diversity indicates that topics are more distinct from one another, which is desirable in topic modeling to ensure minimal overlap between topics.

4.7.2 Multilabel evaluations. Results are presented for the clinical area categorization (similar insights can be derived from the interventions dataset) in table 7. The Binary Relevance (SVC) models achieve the best performance in terms of Hamming loss (0.022) and F1-score (0.81), while the lowest values for ranking loss, coverage error, and one error are observed with the BR method using the multinomial Naive Bayes classifier. It is important to note that the Binary Relevance approach does not account for correlations between labels. However, we have observed that the correlation is minimal for the clinical dataset, while it is slightly higher for the interventions dataset. Therefore, the performance of other classifiers, such as MLkNN (which ranks second for Hamming loss after BR) and RAkEL (which performs best for F1-score), should also be highlighted.

With respect to the BERTopic model, the configuration yielding favorable diversity and a suitable number of topics corresponds to the HDBSCAN settings (*min cluster size* = 10, *min samples* = 1) combined with UMAP parameters (*n neighbors* = 10, *n components* = 5). This combination aligns with model index 3 in Figure 6. Such settings promote higher topic density across the generated topics. It is recommended to visually inspect all the topics produced to ensure meaningful clustering.

In the categorization of clinical areas, three potential new labels were identified: skin diseases (Psoriasis - 36 papers), visual system diseases (Cataract - 21 papers) and diseases of the urinary system (Overactive Bladder - 16 papers). Diseases classification according to WHO ICD-11 statistics [11] have been used.

Identifying topics related to interventions presents greater challenges due to the nature of the data. Topics tend to emerge naturally based on medical domains rather than specific intervention types. To address this, a semi-guided approach could be employed to steer the model toward identifying intervention-related topics. Nevertheless, some papers have been identified that may potentially be classified under interventions, such as those related to drugs (e.g., contraceptive counseling, 23 papers) and screening (e.g., testing/diagnostics, 14 papers).

Model	f1-m	f1-w	h-loss	r-loss	c-err	one-err
BR (SVC)	**0.82**	**0.81**	**0.022**	0.16	2.73	0.44
BR (kNN)	0.78	0.77	0.027	0.20	3.22	0.48
ML kNN	0.76	0.75	*0.030*	0.08	1.71	0.61
RAkEL(gNB)	*0.77*	*0.77*	0.033	0.14	2.54	0.46
VW-MLkNN	0.75	0.75	0.034	0.17	2.88	0.47
BR (mNB)	0.70	0.72	0.047	**0.32**	**1.24**	**0.41**
RAkEL(RF)	0.45	0.39	0.049	0.50	6.51	0.73
MLT SVM	0.48	0.48	0.056	0.43	5.71	0.69

Table 7: Models performance. Clinical areas results

5 Conclusions

This study aimed to develop a robust multilabel classification system capable of identifying and categorizing relevant research papers using only their titles and abstracts. The proposed framework consists of a binary classification model that filters Patient Preference Studies (PPS) from the main dataset, followed by a multilabel classifier that assigns relevant labels to PPS.

The binary classification model, implemented as an ensemble of PubmedBert-kNN and BioMedBERT-SVM classifiers, minimizes the loss of relevant papers through majority voting. For the multilabel classification task, Binary Relevance (BR) with an SVC classifier demonstrated the best performance, achieving the lowest Hamming loss and highest F1-score. The RAkEL model with an mnNB classifier also emerged as a strong candidate, effectively maintaining label correlation. Additionally, the BERTopic model facilitated the discovery of new and meaningful labels, identifying three for clinical areas and two for intervention datasets, significantly reducing manual effort and domain expertise requirements.

Despite these advancements, the study highlights the limited availability of embedding models specifically designed for medical

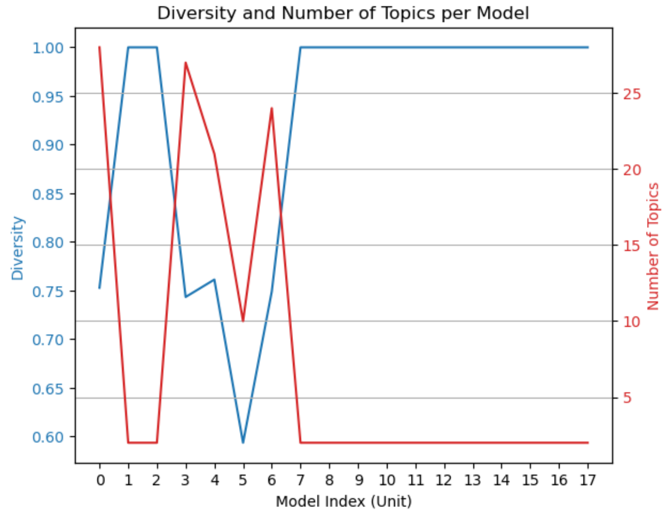


Figure 6: Diversity and # of topics for clinical areas investigation

research literature. Expanding public datasets and developing new embedding models could further enhance the performance of classification systems. With larger datasets, training domain-specific large language models (LLMs) also becomes a viable path forward. Future work should focus on refining the MLC framework as the label set is finalized and the scope of investigation becomes more defined. The BERTopic model has proven particularly useful for categorizing clinical areas, while a semi-guided approach may better suit intervention-related tasks. These findings underscore the potential of automated classification systems to streamline literature reviews and support medical research.

Finally, our work is open-source and available at the following link: https://github.com/adsp-polito/2024-P8-PPS

References

[1] Flavia Cristina Bernardini, Rodrigo Barbosa da Silva, Rodrigo Magalhães Rodovalho, and Edwin Benito Mitacc Meza. 2014. Cardinality and Density Measures and Their Influence to Multi-Label Learning Methods. *Learning and nonlinear models* 12, 1 (2014), 53–71.

[2] Katherine E. Brown, Chao Yan, Zhuohang Li, Xinmeng Zhang, Benjamin X. Collins, You Chen, Ellen Wright Clayton, Murat Kantarcioglu, Yevgeniy Vorobeychik, and Bradley A. Malin. 2024. Not the Models You Are Looking For: Traditional ML Outperforms LLMs in Clinical Prediction Tasks. *medRxiv* (2024). doi:10.1101/2024.12.03.24318400 arXiv:https://www.medrxiv.org/content/early/2024/12/05/2024.12.03.24318400.full.pdf

[3] Canyu Chen, Jian Yu, Shan Chen, Che Liu, Zhongwei Wan, Danielle Bitterman, Fei Wang, and Kai Shu. 2024. ClinicalBench: Can LLMs Beat Traditional ML Models in Clinical Prediction? arXiv:2411.06469 [cs.CL] https://arxiv.org/abs/2411.06469

[4] Wei-Jie Chen, Yuan-Hai Shao, Chun-Na Li, and Nai-Yang Deng. 2016. MLTSVM: a novel twin support vector machine to multi-label learning. *Pattern Recognition* 52 (2016), 61–74.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL] https://arxiv.org/abs/1810.04805

[6] Ioannis Vlahavas Eleftherios Spyromitros, Grigorios Tsoumakas. 2008. An Empirical Study of Lazy Multilabel Classification Algorithms. In *Proc. 5th Hellenic Conference on Artificial Intelligence (SETN 2008)* (Syros, Greece).

[7] Ioannis Katakis Grigorios Tsoumakas and Ioannis Vlahavas. 2006. A review of multi-label classification models. (2006).

[8] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794* (2022).

[9] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. arXiv:arXiv:2007.15779

[10] Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing. 2017. Learning from class-imbalanced data: Review of methods and applications. *Expert systems with applications* 73 (2017), 220–239.

[11] World Health Organization. n.d.. ICD-11 for Mortality and Morbidity Statistics (Version: 11). https://icd.who.int/dev11/l-m/en. Accessed: 2025-01-25.

[12] Adane Nega Tarekegn, Mario Giacobini, and Krzysztof Michalak. 2021. A review of methods for imbalanced multi-label classification. *Pattern Recognition* 118 (2021), 107965. doi:10.1016/j.patcog.2021.107965

[13] G. Tsoumakas, I. Katakis, and I. Vlahavas. 2011. Random k-Labelsets for Multilabel Classification. *IEEE Transactions on Knowledge and Data Engineering* 23, 7 (July 2011), 1079–1089. doi:10.1109/TKDE.2010.164

[14] Zhe Wang, Hao Xu, Pan Zhou, and Gang Xiao. 2023. An Improved Multilabel k-Nearest Neighbor Algorithm Based on Value and Weight. *Computation* 11, 2 (2023), 32–.

[15] Min-Ling Zhang and Zhi-Hua Zhou. 2007. ML-KNN: A lazy learning approach to multi-label learning. *Pattern recognition* 40, 7 (2007), 2038–2048.

[16] Huanhuan Zhao, Haihua Chen, Thomas A. Ruggles, Yunhe Feng, Debjani Singh, and Hong-Jun Yoon. 2024. Improving Text Classification with Large Language Model-Based Data Augmentation. *Electronics* 13, 13 (2024). doi:10.3390/electronics13132535