An AI-Powered Approach for Medical Specialty Triage Using Natural Language Processing and Transformer Models

Anas Chahid, Ismail Chahid, Wafae Mrabti, Mohamed Emharraf, Mohammed Ghaouth Belkasmi
Mohamed First University, Oujda, 60000, Morocco

***Abstract*—Upon arrival at a hospital, patients require an initial assessment to determine the urgency of their condition and the appropriate medical specialty for their needs. This manual triage process, however, is often time-consuming and resource-intensive, leading to potential delays in care, patient dissatisfaction, and inefficient allocation of specialized medical staff. This study presents an AI-based solution to address this critical challenge. A model is introduced that automatically suggests a suitable medical specialty based on a textual description of a patient's symptoms, with the aim of improving the efficiency of the hospital's initial patient triage process. The proposed methodology involves preprocessing a large dataset of over 100,000 patient inquiries from online health forums and conducting a comparative analysis of multiple BERT-based models. Experimental results demonstrate that a domain-specific model, BiomedNLP-PubMedBERT, is particularly effective. To further enhance performance and address the inherent class imbalance in the dataset, a data augmentation strategy using synonym replacement and a weighted loss function was implemented. This combined approach achieved a final weighted F1-score of 92.91%, significantly outperforming the non-augmented baseline models. This work provides a practical path toward building effective automated triage tools that can streamline initial patient assessment and improve operational efficiency in hospital environments. The final model is publicly available for verification and further application.**

***Keywords*—Medical triage; natural language processing; BERT; deep learning; healthcare AI; text classification**

I. Introduction

A. Background

In the fast-paced environment of modern hospitals, the initial triage of incoming patients is a critical determinant of operational efficiency and patient outcomes[1]. Triage, derived from the French verb "trier" meaning to sort, is the process of prioritizing patients based on the severity of their condition to ensure that those with the most critical needs receive immediate attention[2].

This foundational step in emergency and general hospital intake involves a rapid assessment by trained medical personnel, typically nurses, who evaluate a patient's symptoms, vital signs, and medical history to assign a level of acuity[3], [4]. The most common triage systems, such as the Emergency Severity Index (ESI), categorize patients into different levels of urgency to guide the timeliness of care[5].

While established triage protocols are essential for managing patient flow, the process is not without its significant challenges. Manual triage is inherently dependent on the clinical experience and judgment of the triage nurse, which can introduce variability and potential for error[6]. Emergency departments are often crowded and chaotic environments where interruptions are common, which can lead to incomplete assessments and incorrect triage decisions[7]. These interruptions can delay the process, increasing wait times for all patients and potentially leading to adverse outcomes for those with time-sensitive conditions[8].

Furthermore, the increasing specialization within medicine adds another layer of complexity to the triage process. After determining the urgency of a patient's condition, the triage staff must also identify the most appropriate medical specialty to handle their care. This decision-making process can be particularly challenging when patients present with ambiguous or overlapping symptoms that could point to multiple specialties. In such cases, patients may be initially routed to a generalist or an incorrect specialist, leading to a phenomenon known as "doctor-shopping" within the hospital system. This can result in redundant diagnostic tests, delays in receiving the correct treatment, and increased healthcare costs.

The manual and often high-pressure nature of patient triage make it a resource-intensive process. It requires the dedicated time of skilled nursing staff who could otherwise be involved in direct patient care. In situations of high patient volume, the triage process can become a significant bottleneck, contributing to emergency department overcrowding and long waiting times. The proliferation of digital health information has empowered patients, but it can also lead to confusion and the spread of misinformation, further complicating the initial assessment. These challenges highlight a clear and pressing need for an intelligent, automated system to support and streamline the initial patient triage process in hospitals, ensuring that patients are directed to the right specialist in a timely and efficient manner.

B. Problem Statement

The core problem addressed in this research is the inefficiency inherent in the manual process of initial patient triage within hospitals, specifically the mapping of a patient's described symptoms to the correct medical specialty. This manual system is both time-consuming and resource-intensive, placing a significant burden on hospital staff and creating potential bottlenecks in patient flow. The process is prone to delays, which can negatively impact patient outcomes, especially for those with critical conditions. Furthermore, the complexity of accurately determining the appropriate specialty from a patient's initial complaint can lead to misdirection, resulting in multiple consultations, redundant testing, and increased

healthcare costs. This research aims to automate this process using a data-driven approach. The task is framed as a multi-class text classification problem; given a textual description of a patient's symptoms and medical history, the system must predict which of a predefined set of medical specialties is the most appropriate for consultation, thereby improving the speed and accuracy of the initial hospital triage.

C. Proposed Solution

This study proposes a solution based on a transformer model, Bidirectional Encoder Representations from Transformers (BERT)[9]. A large dataset of patient questions from online health forums is leveraged to train a model capable of understanding the nuanced language of symptom descriptions. The proposed approach begins with a rigorous data curation and analysis pipeline to understand the dataset's characteristics, including severe class imbalance and inter-specialty correlations. To find the most suitable architecture, a comparative evaluation of five distinct BERT-based models is conducted, highlighting the benefits of domain-specific pre-training. To address the data imbalance, a weighted loss function is implemented within a custom training framework. Finally, a successful data augmentation strategy using synonym replacement is demonstrated, which significantly boosts classification performance, particularly for underrepresented specialties.

D. Study Structure

The remainder of this study is organized as follows: Section II reviews related work in medical text classification and deep learning. Section III details the methodology, including data preprocessing, exploratory analysis, augmentation strategies, and model architecture. Section IV presents the experimental results and an in-depth analysis of the best-performing model. Section V discusses the limitations of the current study and outlines directions for future work. Finally, Section VI concludes the study.

II. Related Work

The application of machine learning in healthcare has become widespread, facilitating the analysis of vast datasets to predict disease outcomes, personalizing treatment plans, and enhancing diagnostic accuracy. Within this domain, Natural Language Processing (NLP) has been particularly impactful. Text classification, a key NLP task, enables the automatic categorization of clinical text, such as patient inquiries or clinical notes into predefined classes like diseases, symptoms, or, in the context of this study, the appropriate medical specialty. This capability is crucial for improving information retrieval, clinical decision support, and patient triage.

Previous research into medical specialty classification has explored a variety of machine learning techniques. Early and traditional approaches often relied on "shallow" learning models[10] like Support Vector Machines (SVMs). For instance, Weng, W. et al[11]. developed a robust NLP pipeline for classifying medical subdomains in clinical notes. Using a large dataset from MGH, their linear SVM model, which employed a hybrid feature representation of TF-IDF and UMLS concepts, achieved an impressive F1-score of 93.4%. They noted that while deep learning models showed a higher Area Under the Curve (AUC), the shallow learning classifiers yielded better F1 scores. Similarly, Faris et al[12]. tackled the challenge of classifying medical questions in the morphologically complex Arabic language. Their system, which used an ensemble of SVMs combined with Particle Swarm Optimization (PSO) for feature selection, achieved an 85% accuracy and a significant 95.9% feature reduction rate. However, such methods are often limited by their reliance on manual feature engineering and can struggle to capture the deep contextual nuances of natural language.

The advent of deep learning, particularly with the introduction of transformer-based models like Bidirectional Encoder Representations from Transformers (BERT), has significantly advanced the field. Kim, Y. et al[13]. demonstrated the superiority of a domain-specific, pre-trained BERT model for predicting one of 27 medical specialties. When fine-tuned on text from a medical question-answering website, their model outperformed other deep learning architectures like CNNs and LSTMs across accuracy, precision, recall, and F1-score. Their work also highlighted an important nuance: specialties with very specific symptoms yielded higher accuracy, while those covering broader organ systems were more challenging to predict.

Furthering the practical application of these models, Lee et al[14]. constructed an AI chatbot designed to recommend the appropriate medical specialty to patients via their smartphones. In their pipeline, a BERT model delivered the best performance with an F1-score of 0.768. However, due to computational resource constraints for deployment, an LSTM-based model was ultimately adopted for the live service, illustrating the practical trade-offs between performance and implementation.

Researchers have also explored hybrid architectures to push performance even further. Mao et al[15]. proposed a novel model, HyM, which automatically directs patients to one of eight specialties based on their symptom descriptions in Chinese. HyM combines features extracted from LSTM, TEXT-CNN, and BERT, along with TF-IDF, to capture both local and global semantic information. Evaluated on a dataset of over 40,000 items, HyM achieved 93.5% accuracy, outperforming its individual component models in both precision and stability. This body of work underscores a clear trend towards using sophisticated, pre-trained language models, often tailored to the biomedical domain, to effectively solve the challenge of automated medical specialty triage.

III. Methodology

The methodology is structured into several key stages: data acquisition and preprocessing, exploratory data analysis, data augmentation, model architecture design, and experimental setup.

A. Data Acquisition and Preprocessing

The foundation of the proposed model is a dataset of approximately 100,000 user-submitted questions from online health forums. Each entry consists of a patient's question and the corresponding medical specialty it was directed to.

The preprocessing pipeline was as follows:

- Specialty Filtering: To ensure each class had sufficient data for robust training, Gynecology and Oncology

were excluded. These specialties were severely underrepresented in the dataset, with sample sizes less than 50 entries being too small to allow the model to learn their features effectively.

- Data Cleaning: A multi-step cleaning process was applied:
  - Removal of null values and erroneous entries.
  - Deduplication of entries based on content to prevent data leakage between training and test sets.
  - Standard text normalization, including converting text to lowercase, removing URLs and HTML escape sequences, removing emojis, and normalizing whitespace.
- Length Filtering: To ensure meaningful content, entries with fewer than 50 characters were discarded. Additionally, to align with the input constraints of BERT-based models, entries with more than 512 tokens were removed.

After preprocessing, the final dataset size decreased to 97,309 entries, containing 13 unique medical specialties: Cardiology, Dentistry, Dermatology, Endocrinology, Gastroenterology, Infectious Disease, Nephrology, Neurology, Ophthalmology, Orthopedics, Otorhinolaryngology, Pulmonology, and Urology.

The cleaned dataset was split into a training set (80%) and a test set (20%) using stratified sampling to maintain the original class distribution. During model training, 10% of the training set was held out as a validation set for hyperparameter tuning, model selection, and early stopping.

B. Exploratory Data Analysis (EDA)

A thorough EDA was conducted to understand the dataset's characteristics and inform modeling decisions.

1) Specialty distribution analysis: The frequency of each medical specialty was visualized (see Fig. 1). The distribution revealed a significant class imbalance; Orthopedics being the most represented class, while others like Nephrology and Ophthalmology were far less frequent. This imbalance is a critical challenge that directly informed the decision to use a weighted loss function during training.

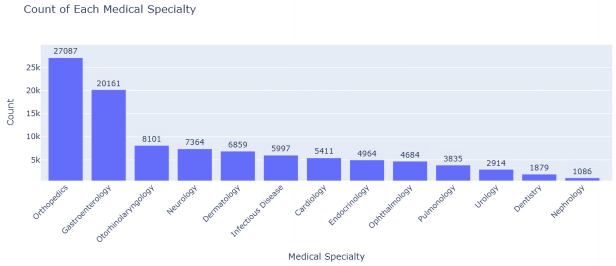

Fig. 1. Medical specialties data distribution.

2) Text length analysis: Histograms of character and token counts were generated, as shown in Fig. 2. This analysis confirmed that while most entries were well under the 512-token limit of standard BERT models, a non-trivial fraction exceeded this length. Setting max input tokens to 512 was therefore a necessary trade-off between computational feasibility and information retention.

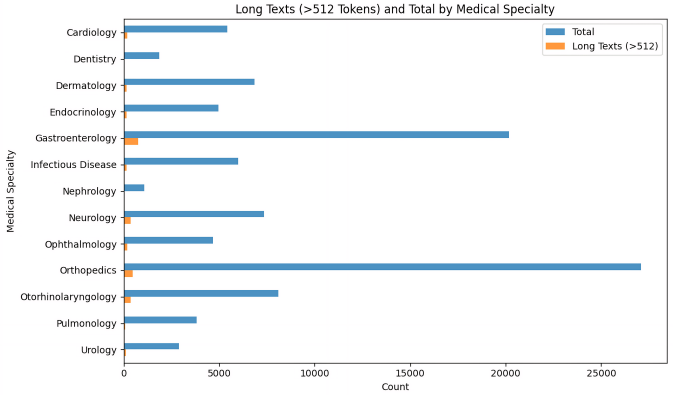

Fig. 2. Text length distribution.

3) Thematic analysis with word clouds: To visually inspect the defining terminology of each specialty, word clouds for each of the 13 classes were generated after filtering out a custom list of medical and general stop words (see Fig. 3). These visualizations confirmed that specialties were characterized by distinct vocabularies (e.g., "eye," "vision," "glasse" for Ophthalmology; "heart," "chest," "pressure" for Cardiology), validating the dataset's suitability for classification.

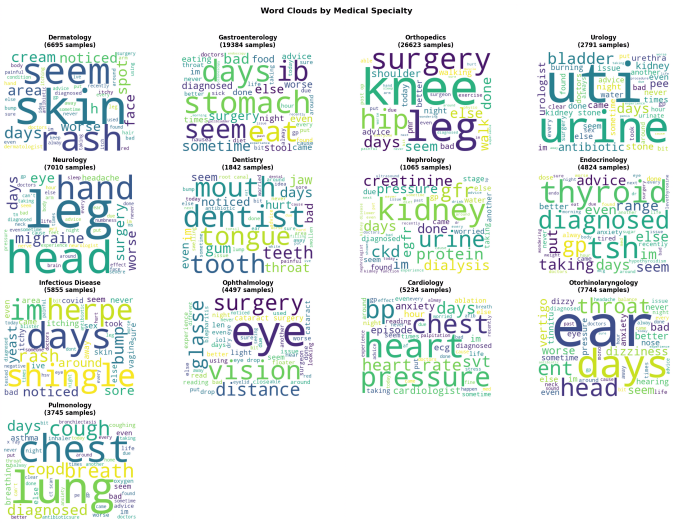

Fig. 3. WordCloud representation of medical specialties.

4) Embedding space visualization with UMAP: To understand the semantic relationships between patient queries, the paraphrase-multilingual-mpnet-base-v2[16] Sentence Transformer[17] model was used to generate embeddings for a 40,000-sample subset of the data. Uniform Manifold Approximation and Projection (UMAP)[18] was then used to project these high-dimensional embeddings into a 2D space.

The resulting plot (Fig. 4) showed clear clusters for some specialties, while others exhibited significant overlap, visually representing the difficulty of the classification task.

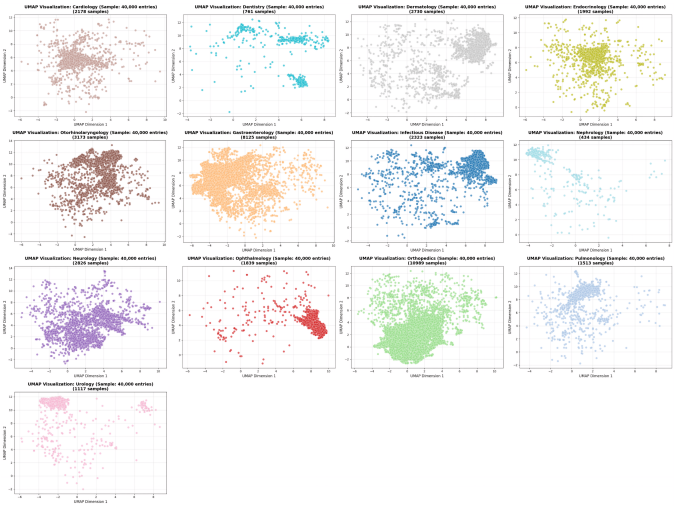

Fig. 4. Space embedding visualization of medical specialties.

5) Inter-specialty correlation analysis: To quantify the textual similarity between specialties, TF-IDF[19] vectorization followed by cosine similarity[20] calculation was employed. A heatmap of top similarity pairs was created (Fig. 5). This analysis highlighted intuitive correlations (e.g., high similarity between Dermatology and Infectious Disease) and less obvious ones (e.g., Dentistry and Otorhinolaryngology), likely due to shared words such as 'throat', 'ear pain', or 'swelling'. This reinforced the need for a model that can capture subtle contextual differences.

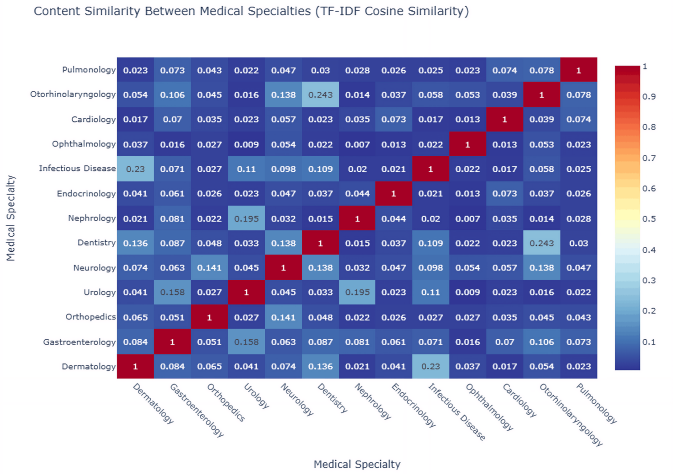

Fig. 5. Heatmap of similarities between medical specialties.

C. Data Augmentation

Based on the initial modeling results and the observed class imbalance, it was hypothesized that performance could be improved by augmenting the training data, especially for minority classes. Synonym replacement using the nlpaug[21] library was employed. For each sample in the original training set, a new sample was generated by randomly replacing words with their synonyms. This augmented data was added to the original training set, effectively doubling its size and diversity.

While effective at increasing data volume, naive synonym replacement has limitations. It can occasionally introduce semantic drift or contextually inappropriate substitutions, such as replacing a specific clinical descriptor with a more generic or incongruous term. Although a manual inspection of a subset (50 entries per specialty) confirmed the general preservation of clinical meaning for the vast majority of samples, the augmentation process did not utilize a strict medical ontology filter. Consequently, rare instances of semantic noise may have been introduced. Future implementations could benefit from a more sophisticated filtering step, such as restricting replacements to clinically equivalent terms using a medical ontology like UMLS (Unified Medical Language System), to ensure the absolute integrity of the clinical narrative.

Table I shows representative examples of this process, demonstrating how synonyms can create diverse yet clinically equivalent training samples.

D. Model Architecture: BERT for Sequence Classification

The core of the proposed solution is the BERT architecture, a transformer-based model pre-trained to learn deep bidirectional language representations. Its ability to capture the context of words from both left and right directions makes it exceptionally powerful for understanding nuanced text, such as patient symptom descriptions. Recognizing that not all pre-trained models are equally suited for specialized domains, the methodology involved a comparative study of multiple BERT variants to identify the most effective architecture for medical text. This approach allowed for the establishment of a robust baseline, ensuring the final model was optimized for the specific vocabulary and semantics of the healthcare domain. The specific models evaluated in the study are detailed in Section III-E1.

1) Theoretical background: The power of BERT and other transformers lies in the self-attention mechanism[22]. This mechanism allows the model to weigh the importance of different words in the input text when producing a representation for each word. The attention score is calculated as:

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where, Q (Query), K (Key), and V (Value) are linear projections of the input embeddings; the dot product of Q and K determines the similarity between words, which is then scaled by the square root of the key dimension, d_k. A softmax function converts these scores into weights, which are applied to the V matrix to produce the final output. This allows the model to capture long-range dependencies and contextual relationships effectively.

BERT is pre-trained on two unsupervised tasks:

- Masked Language Model (MLM)[23]: 15% of the input tokens are masked, and the model must predict them based on the surrounding context.

TABLE I. COMPARISON OF ORIGINAL SENTENCES AND THEIR AUGMENTED COUNTERPARTS, WITH REPLACED WORDS IN BOLD

Specialty	Original Text	Augmented Text
Dentistry	"i have a decaying tooth on the right side of my mouth... i get a throbbing sensation when i lay down. my head also feels like it has pressure. any advice is appreciated"	"i have a **rotten** tooth on the right side of my mouth... i get a **hurting feeling** when i lay down. my head also feels like it has **tightness**. any **aid** is appreciated"
Otorhinolaryngology	"after the huge pop my nose felt clear, but i still have a headache and now eye twitching."	"after the **colossal snap** my nose felt **fair**, but i still have a headache and now eye **spasms**."
Neurology	"I have been experiencing severe headaches and dizziness for the past week. The pain is getting worse."	"I have been experiencing **intense** headaches and dizziness for the past **workweek**. The **ache** is getting worse"

- Next Sentence Prediction (NSP)[24]: The model receives pairs of sentences and must predict if the second sentence logically follows the first.

2) Fine-tuning for classification: For the classification task, each pre-trained BERT model was adapted through a process known as fine-tuning[25]. This involves adding a single linear classification layer on top of the final hidden state of the special classify token (CLS). The CLS token's output is designed to be an aggregate representation of the entire input sequence, making it suitable for classification tasks.

The entire model, including the original pre-trained BERT layers and the new classification layer, is then trained end-to-end on the labeled dataset. This allows the model to adjust its learned representations to the specific task of medical specialty triage. The output of the classification layer is passed through a softmax function to produce a probability distribution over the 13 medical specialties.

E. Experimental Setup

1) Model variants: A comprehensive comparative study of 5 different BERT models was conducted to establish a strong baseline and select the best candidate architecture. These included:

- BERT-base-uncased
- BiomedNLP-PubMedBERT[26], [27]
- BlueBERT[28], [29]
- BioMedBERT[30]
- ModernBERT[31], [32]

2) Training configuration: All the models were trained with the following configuration:

- Max Sequence Length: 512 tokens
- Batch Size: 16
- Learning Rate: 5×10^{-6} (with a cosine annealing scheduler)
- Number of Epochs: 3
- Warmup Steps: 200
- Optimizer: AdamW

3) Handling class imbalance: To address the class imbalance identified in the EDA, a custom Trainer class that modifies the standard CrossEntropyLoss function was implemented. Class weights were calculated to be inversely proportional to their frequency. These weights were then passed to the loss function, compelling the model to pay more attention to misclassifications in minority classes.

4) Evaluation metrics: Model performance was evaluated using a comprehensive set of metrics suitable for multi-class and potentially imbalanced classification tasks. These metrics are derived from the per-class counts of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

Accuracy: Represents the proportion of total correct predictions. While intuitive, it can be misleading on imbalanced datasets.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

To gain deeper insight beyond accuracy, Precision, Recall, and F1-Score were employed; these metrics evaluate performance on a per-class basis.

- Precision measures the accuracy of positive predictions for a class c.

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c} \tag{3}$$

- Recall measures the model's ability to identify all relevant instances of a class c.

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c} \tag{4}$$

- F1-Score is the harmonic mean of Precision and Recall.

$$\text{F1}_c = 2 \cdot \frac{\text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \tag{5}$$

To summarize performance across all classes, two types of averages for these metrics were computed:

- Weighted Average: Calculates the metric for each class and finds the average, weighted by the number of true instances for each class n_c. For a metric M over a set of classes C:

$$M_{\text{weighted}} = \frac{\sum_{c \in C} n_c \cdot M_c}{\sum_{c \in C} n_c} \tag{6}$$

- Macro Average: Calculates the metric independently for each class and takes the unweighted average. This treats all classes as equally important.

$$M_{\text{macro}} = \frac{1}{|C|} \sum_{c \in C} M_c \tag{7}$$

For the experiments, the Weighted F1-Score on the validation set was chosen as the primary metric for model selection and early stopping, as it offers the best balance between precision and recall.

IV. Results and Discussion

A. Model Performance Comparison

The performance of the 5 baseline models, trained on the non-augmented dataset, is detailed in Table II. The analysis reveals two primary findings:

Domain-Specific Pre-Training is Crucial for Performance: There is a performance divide between general-purpose models and those pre-trained on biomedical text. The general-purpose models, BERT-base-uncased and ModernBERT, achieved accuracies of 86.80% and 82.88%, respectively. Notably, ModernBERT, a newer architecture proposed as a potential replacement for BERT, was the lowest-performing model in the evaluation, underscoring the limitations of general-domain models on this specialized task. In contrast, all models pre-trained on biomedical corpora delivered superior results, with the top three BlueBERT (88.26%), BiomedNLP-PubMedBERT (88.23%), and PubMedBERT (88.20%) all having been pre-trained on PubMed. This not only validates the hypothesis regarding the value of specialized pre-training but also highlights that nuances in pre-training methodology can lead to meaningful performance differences.

BiomedNLP-PubMedBERT Emerges as the Most Balanced Model: While BlueBERT achieved the highest accuracy (88.26%), BiomedNLP-PubMedBERT demonstrated the most robust and well-rounded performance. It secured the highest Weighted F1 score (88.41%), indicating the best balance between precision and recall across all classes, and the highest Macro Recall (86.96%), signaling superior performance on minority classes. This strong performance, combined with a joint-lowest validation loss of 0.47, led to its selection as the optimal baseline model for subsequent experiments.

TABLE II. Performance Comparison of Baseline BERT Models

Model	Acc.	W. F1	M. F1	W. Recall	Val. Loss
ModernBERT	82.88	83.00	80.12	82.88	0.62
BERT-base	86.80	86.89	84.31	86.79	0.53
PubMedBERT	88.20	88.28	85.94	88.19	0.48
BlueBERT	88.26	88.32	85.84	88.26	0.51
BiomedNLP	88.23	88.41	85.93	88.23	0.47

B. Impact of Data Augmentation

The best baseline model (BiomedNLP-PubMedBERT) was then retrained from scratch on the training set enhanced with synonym replacement. This single change resulted in a significant improvement across all key metrics, as detailed in Table III.

TABLE III. Impact of Data Augmentation on Tuned BiomedNLP-PubMedBERT

Model Config.	Acc.	W. F1	M. F1	Val. Loss
Baseline Data	88.23	88.41	85.93	0.47
Augmented Data	92.88	92.91	91.89	0.29

Fig. 6 compares the confusion matrices of the model before and after augmentation. The baseline model (left) shows notable confusion between clinically related specialties. In contrast, the augmented model (right) shows much stronger diagonal values, indicating a significant boost in recall and a reduction in misclassifications across the board. This visually confirms the effectiveness of the data augmentation strategy.

The augmented model achieved a final Weighted F1-Score of 92.91% and an accuracy of 92.88%, demonstrating the clear effectiveness of this strategy for improving model generalization and robustness.

C. In-Depth Analysis of the Best Model (Augmented BiomedNLP-PubMedBERT)

The following analysis was conducted on the final model's performance on the held-out test set, providing an unbiased evaluation of its generalization capabilities.

1) Training dynamics: Analysis of the training history plots (Fig. 7 and Fig. 8) for the final model showed stable convergence. The training loss consistently decreased while the validation loss flattened, indicating that the model was learning effectively without significant overfitting. Validation accuracy and F1-scores progressively improved and stabilized, further supporting the choice of hyperparameters and the use of early stopping.

2) Confusion matrix analysis: The confusion matrix for the final augmented model (see Fig. 6) provides a clear view of its performance. The normalized matrix shows high recall (diagonal values) for most classes. Areas of confusion often align with the specialty correlations observed in the EDA. For instance, a small percentage of Nephrology cases were misclassified as Urology, which is clinically possible given the symptomatic overlap related to the urinary system. Similarly, confusion between Cardiology and Pulmonology was observed.

For instance, an entry reading 'hi guys, I've been previously diagnosed with kidney stones and I currently have constant pain when urinating...' was labeled 'Nephrology' but predicted as 'Urology'. This highlights the model's challenge in distinguishing between issues of organ function (Nephrology) versus the urinary tract (Urology) when keywords overlap.

3) Per-class performance: The detailed classification report, presented in Table IV, further breaks down the model's strengths and weaknesses. Classes with distinct terminology, such as Dermatology and Ophthalmology, achieved very high F1-scores. The most challenging classes tended to be those with fewer samples or those with less localized symptoms, such as Dentistry and Neurology. Despite this, even the worst-performing class achieved a respectable F1-score, validating the model's overall utility.

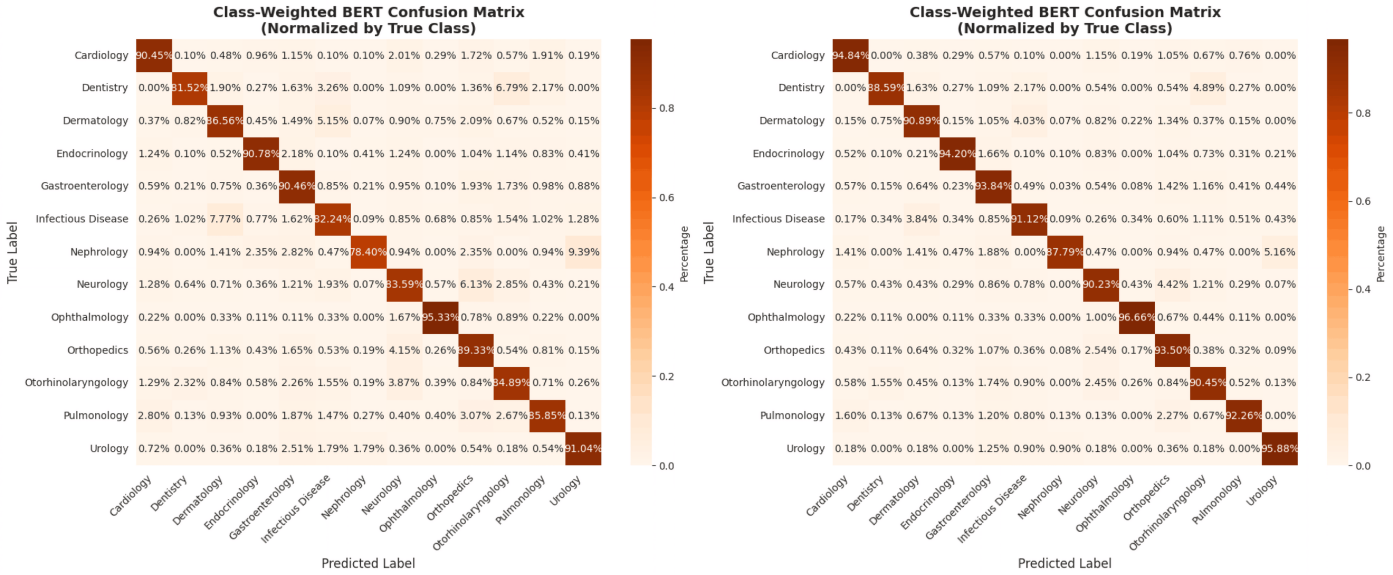

Fig. 6. Visualizing the impact of data augmentation. The confusion matrix on the left shows the performance of the baseline BiomedNLP-PubMedBERT model. The matrix on the right shows the final model after being retrained on the augmented dataset.

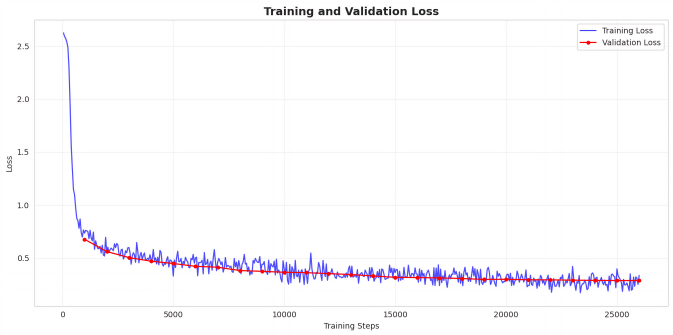

Fig. 7. Training and validation loss during training.

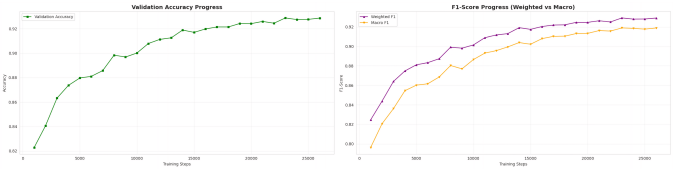

Fig. 8. Progress of accuracy, F1, and weighted F1 scores during training.

4) Qualitative analysis: To evaluate the model's ability to generalize beyond its training data, a qualitative analysis was conducted. The model was tested on a set of phrases intentionally structured to be more direct and concise than the conversational forum posts it was trained on. This allowed testing different aspects of its performance, from simple keyword recognition to its ability to handle clinical ambiguity.

Below are some examples:

TABLE IV. PER-CLASS PRECISION, RECALL, AND MACRO F1-SCORE ON THE TEST SET

Class	Precision	Recall	F1-Score
Cardiology	0.9177	0.9484	0.9328
Dentistry	0.8468	0.8859	0.8659
Dermatology	0.8982	0.9089	0.9035
Endocrinology	0.9528	0.9420	0.9474
Gastroenterology	0.9556	0.9384	0.9469
Infectious Disease	0.8833	0.9112	0.8970
Nephrology	0.9303	0.8779	0.9034
Neurology	0.8394	0.9023	0.8697
Ophthalmology	0.9656	0.9666	0.9661
Orthopedics	0.9605	0.9350	0.9476
Otorhinolaryngology	0.9074	0.9045	0.9059
Pulmonology	0.9128	0.9226	0.9177
Urology	0.9256	0.9588	0.9419

Case 1: Unambiguous Keyword Association.

- Input: "Patient has a broken leg from a fall."
- Predicted Specialty: Orthopedics
- Analysis: Correct

Case 2: Differentiating Regional Specialties.

- Input: "Complaining of dizziness, ringing in the ears, and hearing loss."
- Predicted Specialty: Otorhinolaryngology
- Analysis: Correct

Case 3: Handling Symptom Ambiguity (Chest Pain).

- Input: "Severe chest pain and shortness of breath during exertion."

- Predicted Specialty: Pulmonology
- Analysis: While both Pulmonology and Cardiology are clinically valid specialties for these symptoms, the model's preference for Pulmonology is likely driven by the strong association of 'shortness of breath' with respiratory conditions in the training data. Clinically, this presentation is highly ambiguous and could indicate a pulmonary embolism, but it is equally indicative of a severe cardiac event such as ischemia or angina. The prediction of Pulmonology alone is therefore incomplete from a clinical safety perspective. This highlights a critical limitation in single-label classification for ambiguous presentations, reinforcing the necessity for a multi-label approach or a top-k recommendation system to ensure comprehensive and clinically defensible triage.

Case 4: Inferring from Plain English Language.

- Input: "A red, itchy rash has been spreading on my arm for a week."
- Predicted Specialty: Dermatology
- Analysis: Correct

Case 5: Distinguishing Overlapping Urinary System Specialties.

- Input: "Patient reports a burning sensation during urination and bladder pain."
- Predicted Specialty: Urology
- Analysis: Correct

This qualitative assessment provides valuable insight into the model's practical utility, highlighting its strengths in contextual understanding and identifying potential areas for improvement in handling ambiguity.

V. Limitations and Future Work

While the proposed model demonstrates high performance, several limitations are acknowledged:

- Data Source Bias: The data is sourced entirely from online health forums, which may differ from clinical patient intake language in vocabulary, symptom specificity, and formality. Consequently, the model's generalizability to real hospital environments remains uncertain. A quantitative evaluation on a held-out dataset of actual, de-identified Electronic Health Records (EHRs) is necessary to assess this domain gap effectively.
- Limited Scope: Several major specialties were excluded. A production system would need to be more comprehensive.
- Single Label Classification: The model assigns only one specialty per query. Many real-world cases are complex and may require consultations from multiple specialties. This can be handled by showing the top 2 predicted specialties based on the confidence score.

Future work will explore addressing these limitations:

- Incorporate Diverse Data: Exploring training on more formal data sources, such as de-identified Electronic Health Records (EHRs), to improve clinical validity.
- Expand Specialty Coverage: The model will be extended to include a wider range of medical and surgical specialties.
- Change the Augmentation Technique: Using a different approach to data augmentation by leveraging large language models such as MedGemma[33]. This would allow for more sophisticated augmentation through paraphrasing and rephrasing entire sentences rather than simple synonym replacement.

VI. Conclusion

This research presented an automated system leveraging Natural Language Processing and transformer models to predict the appropriate medical specialty from patient symptom descriptions. The comparative analysis established that domain-specific pre-training, particularly with the BiomedNLP-PubMedBERT architecture, provides a superior foundation for this clinical text classification task. By implementing a weighted loss function and a synonym replacement data augmentation strategy, the inherent class imbalance was effectively mitigated, yielding a highly accurate and robust classification model. This study contributes empirical evidence supporting the efficacy of tailored deep learning pipelines for medical triage. The findings demonstrate a viable pathway toward integrating intelligent decision-support tools into initial patient assessment workflows. Such automated systems have the potential to expedite diagnostic routing, optimize resource allocation, and ultimately enhance the operational efficiency of healthcare facilities.

Statements and Declarations

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Funding

The authors did not receive support from any organization for the submitted work.

Data Availability

The dataset analyzed during the current study was curated from publicly available online health forums. Due to the sensitive nature of the data and to protect user privacy, the raw dataset is not publicly available. However, aggregated data and processing scripts can be made available from the corresponding author upon reasonable request.

The trained model, inference code, and model card are openly available in the Hugging Face Model Hub at https://huggingface.co/anaschahid/medical-specialty-classifier. The model is also deployed as an interactive web application for demonstration purposes at https://medclass.anaschahid.work. These resources ensure the reproducibility and practical accessibility of the proposed solution.

Ethical Considerations

This study is based on publicly available data collected from online health forums. All data was fully anonymized prior to analysis. No personally identifiable information (PII), such as usernames, locations, or specific dates, was used in this research to ensure the privacy of the original posters. As the research was conducted using publicly accessible, anonymized data, formal ethics approval from an institutional review board was not required. The study was performed in accordance with the ethical standards as laid down in the 1964 Declaration of Helsinki and its later amendments.

REFERENCES

[1] G. FitzGerald, G. A. Jelinek, D. Scott, and M. F. Gerdtz, "Emergency department triage revisited," *Emergency Medicine Journal*, vol. 27, no. 2, pp. 86–92, 2010.

[2] C. C. Yancey and M. C. O'Rourke, *Emergency Department Triage*. Treasure Island (FL): StatPearls Publishing, 2025, pMID: 32491515. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK557583/

[3] M. Christ, F. Grossmann, D. Winter, R. Bingisser, and E. Platz, "Modern triage in the emergency department," *Deutsches Ärzteblatt International*, vol. 107, no. 50, pp. 892–898, 2010.

[4] N. Farrohknia, M. Castrén, A. Ehrenberg, L. Lind, S. Oredsson, H. Jonsson, K. Asplund, and K. E. Göransson, "Emergency department triage scales and their components: a systematic review of the scientific evidence," *Scandinavian journal of trauma, resuscitation and emergency medicine*, vol. 19, no. 1, pp. 1–13, 2011.

[5] R. C. Wuerz, L. W. Milne, D. R. Eitel, D. Travers, and N. Gilboy, "Reliability and validity of a new five-level triage instrument," *Academic Emergency Medicine*, vol. 7, no. 3, pp. 236–242, 2000.

[6] B. M. Porto, "Improving triage performance in emergency departments using machine learning and natural language processing: a systematic review," *BMC Emergency Medicine*, vol. 24, p. 219, nov 2024, pMID: 39558255 PMCID: PMC11575054.

[7] K. D. Johnson, G. L. Gillespie, and K. Vance, "Effects of Interruptions on Triage Process in Emergency Department: A Prospective, Observational Study," *Journal of nursing care quality*, vol. 33, no. 4, pp. 375–381, 2018, pMID: 29319593 PMCID: PMC6037611.

[8] G. Savioli, I. F. Ceresa, M. A. Bressan, G. B. Piccini, A. Varesi, V. Novelli, A. Muzzi, S. Cutti, G. Ricevuti, C. Esposito, A. Voza, A. Desai, Y. Longhitano, A. Saviano, A. Piccioni, F. Piccolella, A. Bellou, C. Zanza, and E. Oddone, "Five Level Triage vs. Four Level Triage in a Quaternary Emergency Department: National Analysis on Waiting Time, Validity, and Crowding—The CREONTE (Crowding and REOrganization National TriagE) Study Group," *Medicina*, vol. 59, no. 4, p. 781, apr 2023, pMID: 37109739 PMCID: PMC10143416.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding," may 2019, arXiv:1810.04805 [cs]. [Online]. Available: http://arxiv.org/abs/1810.04805

[10] C. Friedman, P. O. Alderson, J. H. Austin, J. J. Cimino, and S. B. Johnson, "A general natural-language text processor for clinical radiology." *Journal of the American Medical Informatics Association*, vol. 1, no. 2, pp. 161–174, 1994, pMID: 7719797 PMCID: PMC116194.

[11] W.-H. Weng, K. B. Wagholikar, A. T. McCray, P. Szolovits, and H. C. Chueh, "Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach," *BMC Medical Informatics and Decision Making*, vol. 17, no. 1, p. 155, dec 2017.

[12] H. Faris, M. Habib, M. Faris, M. Alomari, and A. Alomari, "Medical speciality classification system based on binary particle swarms and ensemble of one *vs.* rest support vector machines," *Journal of Biomedical Informatics*, vol. 109, p. 103525, sep 2020.

[13] Y. Kim, J.-H. Kim, Y.-M. Kim, S. Song, and H. J. Joo, "Predicting medical specialty from text based on a domain-specific pre-trained BERT," *International Journal of Medical Informatics*, vol. 170, p. 104956, feb 2023.

[14] H. Lee, J. Kang, and J. Yeo, "Medical Specialty Recommendations by an Artificial Intelligence Chatbot on a Smartphone: Development and Deployment," *Journal of Medical Internet Research*, vol. 23, no. 5, p. e27460, may 2021, pMID: 33882012 PMCID: PMC8104000.

[15] C. Mao, Q. Zhu, R. Chen, and W. Su, "Automatic medical specialty classification based on patients' description of their symptoms," *BMC Medical Informatics and Decision Making*, vol. 23, no. 1, p. 15, jan 2023.

[16] N. Reimers and I. Gurevych, "paraphrase-multilingual-mpnet-base-v2 model," 2020. [Online]. Available: https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

[17] I. Gurevyc and N. Reimers, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," aug 2019, arXiv:1908.10084 [cs]. [Online]. Available: http://arxiv.org/abs/1908.10084

[18] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform Manifold Approximation and Projection for Dimension Reduction," sep 2020, arXiv:1802.03426 [stat]. [Online]. Available: http://arxiv.org/abs/1802.03426

[19] K. SPARCK JONES, "A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL," *Journal of Documentation*, vol. 28, no. 1, pp. 11–21, jan 1972.

[20] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, nov 1975.

[21] E. Ma, "nlpaug: Data augmentation for nlp," 2020. [Online]. Available: https://github.com/makcedward/nlpaug

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, vol. 30, 2017, pp. 5998–6008. [Online]. Available: https://doi.org/10.48550/arXiv.1706.03762

[23] W. Liang and Y. Liang, "Bpdec: Unveiling the potential of masked language modeling decoder in bert model pretraining," jan 2024, arXiv:2401.15861 [cs]. [Online]. Available: http://arxiv.org/abs/2401.15861

[24] Y. Sun, Y. Zheng, C. Hao, and H. Qiu, "Nsp-BERT: A Prompt-based Few-Shot Learner through an Original Pre-training Task -Next Sentence Prediction," in *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, oct 2022, pp. 3233–3250. [Online]. Available: https://aclanthology.org/2022.coling-1.286/

[25] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune bert for text classification?" in *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18*. Springer, 2019, pp. 194–206. [Online]. Available: https://doi.org/10.48550/arXiv.1905.05583

[26] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Biomednlp-biomedbert-base-uncased-abstract model," 2020. [Online]. Available: https://huggingface.co/microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract

[27] R. Tin, Y. Gu, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing," *ACM Transactions on Computing for Healthcare*, vol. 3, no. 1, pp. 1–23, jan 2022, arXiv:2007.15779 [cs].

[28] Y. Peng, S. Yan, and Z. Lu, "bluebert_pubmed_mimic_uncased_l-24_h-1024_a-16 model," 2019. [Online]. Available: https://huggingface.co/bionlp/bluebert_pubmed_mimic_uncased_L-24_H-1024_A-16

[29] S. Ya, Y. Peng, and Z. Lu, "Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets," jun 2019, arXiv:1906.05474 [cs]. [Online]. Available: http://arxiv.org/abs/1906.05474

[30] S. Chakraborty, E. Bisong, S. Bhatt, T. Wagner, R. Elliott, and F. Mosconi, "Biomedbert: A Pre-trained Biomedical Language Model for QA and IR," in *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain: International Committee on Computational Linguistics, dec 2020, pp. 669–679. [Online]. Available: https://aclanthology.org/2020.coling-main.59/

[31] B. Warner, A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini, A. Gallagher, R. Biswas, F. Ladhak, T. Aarsen, N. Cooper, G. Adams, J. Howard, and I. Poli, "Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference," dec 2024, arXiv:2412.13663 [cs]. [Online]. Available: http://arxiv.org/abs/2412.13663

[32] B. Warner, A. Chaffin, B. Clavié, and O. Weller, "Modernbert-base model," 2024. [Online]. Available: https://huggingface.co/answerdotai/ModernBERT-base

[33] A. Sellergren, S. Kazemzadeh, T. Jaroensri, A. Kiraly, M. Traverse, T. Kohlberger, S. Xu, F. Jamil, C. Hughes, C. Lau, J. Chen, F. Mahvar, L. Yatziv, T. Chen, B. Sterling, S. A. Baby, S. M. Baby, J. Lai, S. Schmidgall, L. Yang, K. Chen, P. Bjornsson, S. Reddy, R. Brush, K. Philbrick, M. Asiedu, I. Mezerreg, H. Hu, H. Yang, R. Tiwari, S. Jansen, P. Singh, Y. Liu, S. Azizi, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Riviere, L. Rouillard, T. Mesnard, G. Cideron, J.-b. Grill, S. Ramos, E. Yvinec, M. Casbon, E. Buchatskaya, J.-B. Alayrac, D. Lepikhin, V. Feinberg, S. Borgeaud, A. Andreev, C. Hardin, R. Dadashi, L. Hussenot, A. Joulin, O. Bachem, Y. Matias, K. Chou, A. Hassidim, K. Goel, C. Farabet, J. Barral, T. Warkentin, J. Shlens, D. Fleet, V. Cotruta, O. Sanseviero, G. Martins, P. Kirk, A. Rao, S. Shetty, D. F. Steiner, C. Kirmizibayrak, R. Pilgrim, D. Golden, and L. Yang, "Medgemma Technical Report," jul 2025, arXiv:2507.05201 [cs]. [Online]. Available: http://arxiv.org/abs/2507.05201