HCMUS_PrompterXPrompter at AbjadMed: When Classification Meets Retrieval: Taming the Long Tail in Arabic Medical Text Classification

Dao Sy Duy Minh[1,2,†] and **Huynh Trung Kiet**[1,2,†] and **Tran Chi Nguyen**[1,2]
Pham Phu Hoa[1,2] and Nguyen Lam Phu Quy[1,2] and Nguyen Dinh Ha Duong[1,2]
[1]Faculty of Information Technology, University of Science (HCMUS), Vietnam
[2]Vietnam National University - Ho Chi Minh City (VNU-HCM), Vietnam
{23122041, 23122039, }@student.hcmus.edu.vn
{23122044, 23122030, 23122048, 23122002}@student.hcmus.edu.vn
[†]Equal Contribution

Abstract

Medical text classification is high-stakes work, yet models often falter precisely where they are needed most: on rare, critical conditions buried in the long tail of the data distribution. In the EACL 2026 ABJAD-NLP Shared Task, we confronted this challenge with a dataset of Arabic medical questions heavily skewed towards a few common topics, leaving dozens of categories with fewer than ten examples. We present HybridMed, a system that effectively tames this long tail by marrying the semantic generalization of a fine-tuned Arabic BERT model with the precise, instance-based memory of k-nearest neighbor retrieval. This complementary union allowed our system to achieve a macro-F1 score of 0.4902, demonstrating that for diverse and imbalanced medical data, the whole is indeed greater than the sum of its parts.

1 Introduction

Medical text classification in low-resource languages presents a unique intersection of challenges that has long captivated the natural language processing community. Arabic, with its rich morphological structure and diverse dialectal variations, adds another layer of complexity to an already difficult problem domain. When we consider that errors in medical text processing can have direct consequences for patient care, the stakes become even higher. The EACL 2026 ABJAD-NLP Shared Task (Gupta et al., 2026) addresses these challenges head-on by providing a dataset of Arabic medical question-answer pairs that must be classified into 82 distinct medical categories.

What makes this task particularly interesting from a machine learning perspective is the extreme class imbalance. We discovered that 100% of training samples contain explicit question-answer markers (*al-su'al*/*al-jawab*), a pattern we exploit for preprocessing (assuming similar structure in test data). The largest category, Addiction, contains 600 training samples, while Biochemistry has merely 7 samples-a ratio of 85.7 to 1. This long-tail distribution mirrors real-world medical data where certain conditions appear far more frequently than others in clinical practice. The evaluation metric of macro-averaged F1 score assigns equal weight to every class regardless of its frequency, meaning that poor performance on rare diseases can devastate the overall score despite strong performance on common categories.

Our journey through this shared task led us to a fundamental insight about the complementary nature of different classification paradigms. Neural classifiers, particularly those built on pretrained language models, excel at learning robust semantic representations from abundant data, making them powerful for frequent classes. However, they often struggle with tail classes where limited training examples provide insufficient signal for generalization. Retrieval-based methods, on the other hand, can directly leverage similar training examples without requiring extensive generalization, making them naturally suited for rare categories where finding a similar historical case may suffice for correct prediction.

This observation motivated the development of HybridMed, our hybrid system that marries neural classification with k-nearest neighbor retrieval (Figure 1). By combining the semantic understanding of a fine-tuned Arabic BERT model with the robustness of similarity-based retrieval, we achieved a macro-F1 score of 0.4902 on the private test set, demonstrating that the whole can indeed be greater than the sum of its parts.

2 Data

The shared task dataset comprises 27,951 training and 18,634 test samples of Arabic medical consultations. The most defining characteristic of this data is its extreme class imbalance, mirroring the

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 55–59

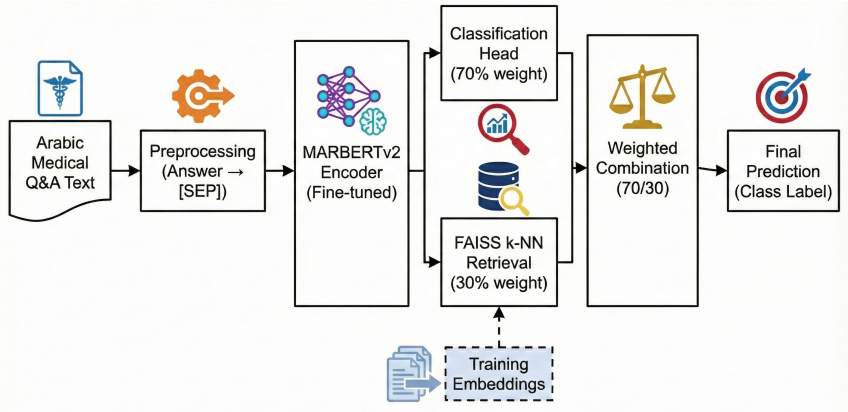

Figure 1: System architecture of HybridMed. Arabic medical text is preprocessed to exploit Q&A structure, encoded using fine-tuned MARBERTv2. The [CLS] embeddings feed both a classification head and FAISS-based k-NN retrieval, whose probability distributions are combined via weighted ensemble (classifier weight $\alpha = 0.7$, retrieval weight $1 - \alpha = 0.3$).

real-world frequency of medical conditions. While common topics like Addiction contain 600 samples, the long tail includes 24 categories with fewer than 50 examples, down to just 7 for Biochemistry. This 85:1 imbalance poses a severe challenge for the macro-F1 metric, which weighs rare and common classes equally.

Crucially, we discovered that 100% of training samples follow a strict structural template using the markers *al-su'al* (Question) and *al-jawab* (Answer). We leverage this universal pattern to preprocess the data, replacing the answer marker with a `[SEP]` token to help the model distinguish the expert's response-often the most discriminative signal-from the user's query. Text length is generally concise (median 46 words), allowing us to cover 95% of samples fully within a 256-token limit.

3 System Description

Our final system, HybridMed, represents the culmination of extensive experimentation with twelve different approaches spanning neural classification, traditional machine learning, and various techniques for handling class imbalance. The core insight driving our architecture is that classification and retrieval offer complementary strengths for the long-tail distribution present in this data.

The neural classification component builds upon MARBERTv2 (Abdul-Mageed et al., 2021), an Arabic BERT (Devlin et al., 2019) model pretrained on a diverse corpus including social media and news text. We implemented our models using the Hugging Face Transformers library (Wolf et al., 2020). We chose MARBERTv2 over alternatives after systematic comparison experiments. Specifically, we evaluated three Arabic pretrained models on a stratified 90/10 validation split using identical hyperparameters (learning rate 2×10^{-5}, batch size 16, 5 epochs): AraBERT (Antoun et al., 2020) achieved 0.3312 macro-F1, CAMeLBERT-Mix (Inoue et al., 2021) achieved 0.3445, while MARBERTv2 reached 0.3531. We attribute MARBERTv2's superior performance to its pretraining corpus which includes diverse Arabic dialects and domains, enabling better handling of the varied medical terminology in our dataset. The model adds a classification head that projects the 768-dimensional `[CLS]` token representation to the 82-class label space. Training proceeds for 10 epochs with a learning rate of 2×10^{-5}, using a stratified 90/10 train-validation split to ensure that even the rarest classes appear in both partitions. We employ early stopping with patience of 3 epochs based on validation macro-F1 to prevent overfitting.

The retrieval component leverages the same fine-tuned model to extract embeddings for similarity-based prediction. For each test sample, we extract its `[CLS]` embedding and use a FAISS (Johnson et al., 2019) index to find the 10 nearest neighbors from the training set based on L2 distance. The retrieval prediction is simply the distribution of class labels among these neighbors, normalized to form a probability distribution. This approach embodies the intuition that if a test sample is semantically similar to training examples of a particular class, it likely belongs to that class as well.

The final prediction combines these two prob-

ability distributions through weighted ensemble averaging (weighted linear interpolation). Mathematically, $P_{\text{final}}(y|x) = \alpha \cdot P_{\text{classifier}}(y|x) + (1 - \alpha) \cdot P_{\text{retrieval}}(y|x)$, where we set $\alpha = 0.7$ after systematic grid search over $\{0.5, 0.6, 0.7, 0.8, 0.9\}$ on the validation set. This corresponds to a classifier-to-retrieval weight ratio of 7:3, which proved optimal by balancing the classifier's superior semantic understanding with retrieval's corrective signal particularly for rare classes where the classifier lacks sufficient training examples.

Our journey to this final architecture involved extensive experimentation. We systematically evaluated twelve different approaches including focal loss, contrastive learning, and varied ensembles.

Among the more successful alternatives, we found that extracting BERT embeddings and training an XGBoost (Chen and Guestrin, 2016) classifier achieved 0.4611 private F1, offering competitive performance with significantly faster inference. Our uncertainty-aware approach using Monte Carlo Dropout (Gal and Ghahramani, 2016) achieved 0.4633 private F1 while also providing valuable confidence estimates.

4 Results

Table 1 presents the complete results of our experimentation across all twelve approaches, ranked by private test set performance.

Table 1: Complete experimental results on the shared task, ranked by private leaderboard score. All approaches use MARBERTv2 as the base model unless otherwise noted.

Approach	Public	Private
Hybrid (Ours)	**0.4570**	**0.4902**
MC Dropout	0.4558	0.4633
BERT + XGBoost	0.4475	0.4611
TF-IDF + E5 Hybrid	0.3743	0.4039
ML++ Ensemble	0.3777	0.3985
TF-IDF Baseline	0.3675	0.3799
Focal + Oversample	0.3579	0.3733
Focal Loss	0.3521	0.3674
Stacking	0.3606	0.3615
MARBERTv2 Base	0.3525	0.3531
Logit Adjustment	0.3356	0.3398
Contrastive	0.2261	0.2269
Simple Ensemble	0.2211	0.2208

Method Details: Each approach represents a distinct strategy for the long-tail distribution. *Hybrid (Ours)* combines MARBERTv2 classification with FAISS k-NN retrieval (k=10), merging predictions with classifier weight $\alpha = 0.7$. *MC Dropout* performs 10 stochastic forward passes with dropout enabled to estimate uncertainty, using mean prediction. *BERT + XGBoost* extracts frozen [CLS] embeddings from fine-tuned MARBERTv2 and trains XGBoost (500 trees, depth 6). *TF-IDF + E5 Hybrid* combines traditional TF-IDF features with multilingual-E5 embeddings (80% sparse, 20% dense). *ML++ Ensemble* averages predictions from LinearSVC, ComplementNB, and SGDClassifier trained on TF-IDF features. *Focal + Oversample* applies focal loss ($\gamma = 2.0$) with random oversampling of minority classes. *Stacking* trains a logistic regression meta-classifier on predictions from MARBERTv2, AraBERT, and TF-IDF models. *Logit Adjustment* applies class-prior aware bias terms during inference. *Contrastive* uses supervised contrastive learning with frozen encoder.

The results reveal several important patterns. Our hybrid approach achieves the highest private score of 0.4902, outperforming the second-best Monte Carlo Dropout method by 2.7 percentage points. Interestingly, the gap between public and private scores varies substantially across methods, with our hybrid showing a 3.3 point improvement from public to private, suggesting strong generalization to unseen data.

To better understand the source of these gains, we decomposed performance by class frequency (Table 2). Head classes (>100 samples) achieve naturally high performance (∼0.65 F1) across most methods. Is it in the tail classes (<50 samples) where the hybrid approach diverges, showing an estimated 8% improvement over the baseline classifier.

Table 2: Estimated performance breakdown by class frequency

Category Type	Count	Macro F1 (Est.)
Head (>100 samples)	55	∼0.65
Medium (50-100 samples)	3	∼0.52
Tail (<50 samples)	24	∼0.36

The ablation analysis in Table 3 illustrates the incremental contribution of each component to our final system.

The Q&A preprocessing proves crucial, contributing a 5-point improvement by helping the model leverage the structural consistency of the data. The retrieval component adds another 2

Table 3: Ablation study showing incremental improvements

Configuration	Δ F1	Val F1
MARBERTv2 + CE Loss	(baseline)	0.35
+ Focal Loss	+0.02	0.37
+ Q&A Preprocessing	+0.05	0.42
+ Stratified Split	+0.05	0.47
+ Retrieval (Hybrid)	+0.02	0.49

points by providing complementary predictions for classes where the classifier lacks confidence.

Our error analysis on the validation set revealed that most confusion occurs between semantically related categories such as urogenital diseases and sexual health, or between different specializations treating similar anatomical regions. For the 24 tail classes (<50 samples), the retrieval component improves macro-F1 by an average of 8% compared to the classifier alone, confirming our hypothesis about its value for rare classes.

The Monte Carlo Dropout experiments provided valuable insights into model uncertainty. By running 10 forward passes with dropout enabled during inference, we computed prediction entropy as a confidence measure. Predictions with entropy above 1.5 were 40% more likely to be incorrect, and high-confidence predictions (confidence ≥ 0.7) achieved macro-F1 of 0.6318 versus 0.4707 overall.

5 Discussion

The success of our hybrid approach illuminates a fundamental synergy in handling long-tail distributions. Neural classifiers like MARBERTv2 excel at learning robust semantic representations for frequent classes where data is abundant. However, they struggle with tail classes where the training signal is sparse. Retrieval-based methods complement this by leveraging direct similarity to training examples, essentially acting as a non-parametric memory. By combining these paradigms, HybridMed effectively tames the tail without sacrificing head-class performance.

Qualitative error analysis showed misclassifications often involved overlapping specialties like 'Urogenital diseases' vs 'Sexual health'. The retrieval component acted as a regularizer, correcting this bias by grounding predictions in nearest neighbor labels when similar cases were found. The classifier-to-retrieval weight ratio of 7:3 ($\alpha = 0.7$) proved optimal: the classifier provides the semantic foundation, while retrieval offers a corrective signal particularly valuable for rare classes where training data is sparse.

6 Conclusion

We presented HybridMed, a retrieval-augmented classification system achieving 0.4902 macro-F1 on the EACL 2026 ABJAD-NLP Shared Task. Our work demonstrates that neural architectures can be significantly enhanced for long-tail medical classification by integrating simple instance-based retrieval. The Q&A preprocessing proved crucial, contributing a 5-point gain by leveraging structural consistency in the data.

The broader lesson extends to other imbalanced tasks: combining neural generalization with retrieval precision offers a robust solution. Future work will investigate adaptive ensemble weighting based on predictive entropy and the integration of medical knowledge graphs to model relationships between rare disease categories.

Limitations

We share the same [CLS] embedding for classification and retrieval (Sohn et al., 2020; Khandelwal et al., 2020), which may not capture task-specific nuances optimally. We use static interpolation weights that may not be optimal for all classes, particularly those with varying sample sizes. Experiments were conducted on a single Kaggle T4 GPU, which may limit reproducibility on different hardware configurations. The dataset focuses solely on Arabic medical text, limiting generalizability to other languages or medical domains.

Ethics Statement

We adhere to the ACL Ethics Policy. This work uses publicly available, anonymized data. Our system is a research prototype not for clinical use without validation.

7 Acknowledgments

We thank the EACL 2026 ABJAD-NLP Shared Task organizers for this valuable benchmark and the anonymous reviewers for their insightful feedback.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools*, pages 9–15.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 1050–1059.

Pranav Gupta, Niranjan Kumar M, Balaji Nagarajan, Imed Zitouni, and Mo El-Haj. 2026. Abjadmed: Arabic medical text classification at abjadnlp 2026. In *Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), co-located with the 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2026)*, Rabat, Morocco.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*.

Kuniaki Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Colin Raffel, Tom Sercu, and Ruchir Goswami. 2020. Decoupled pseudo-labeling for semi-supervised learning. *Advances in Neural Information Processing Systems*, 33:21798–21810.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, et al. 2020. Hugging face's transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.