UIT_Concabietbay at MEDIQA-CORE-Task-1 2026: Multimodal Embedding-Based Ensemble Learning for Brain Tumor Subtype Classification

Notebook for the ImageCLEF Lab at CLEF 2026

Nguyen Minh Thoi[1,2], Luu Nhat Quang[1,2] and Thien B. Nguyen-Tat[1,2,*]

[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam

Abstract

This paper describes the UIT_Concabietbay system submitted to MEDIQA-CORE-Task-1 2026, the ImageCLEF medical task on brain tumor subtype classification. The task requires subject-level prediction of three related targets: Level 1 molecular grouping, low-grade versus high-grade glioma status, and WHO grade. Our system treats the task as multimodal tabular learning over precomputed imaging embeddings rather than as end-to-end training from raw MRI volumes or whole-slide pathology images. MRI embeddings are converted into fixed-length subject representations by mean and standard-deviation pooling. Pathology slide embeddings are summarized with mean, standard deviation, maximum pooling, norm-based attention pooling, and robust mean pooling. The resulting modality-specific feature tables are merged by subject identifier and augmented with modality-availability indicators; missing numeric features are handled inside model pipelines with median imputation and missingness indicators. For each target, we evaluate a classical machine-learning model zoo including regularized logistic regression, PCA-logistic regression, calibrated linear models, support-vector machines, random forests, extra trees, histogram gradient boosting, multilayer perceptrons, and optional gradient-boosting libraries. Model selection is based on out-of-fold macro-F1, with greedy probability blending, optional stacking, and conservative threshold tuning for the binary target. In the organizer post-verification spreadsheet, our strongest verified Test Set 1 average macro-F1 was 0.769 under the fully multimodal and MRI-plus-pathology conditions, while our strongest Test Set 2 average was 0.547 under the MRI-images-only and MRI-images-plus-reports conditions. A separate all-modality leaderboard score of 0.866 was obtained before organizer code verification and is reported only as pre-verification feedback.

Keywords

Brain tumor classification, CoRe-BT, MEDIQA-CORE, MRI embeddings, Pathology embeddings, Multimodal learning, Ensemble learning, Classical machine learning

1. Introduction

Brain tumor diagnosis increasingly depends on integrating evidence across radiology, histopathology, molecular testing, and clinical context. The 2021 WHO classification of central nervous system tumors emphasizes that tumor entities and grades are defined through integrated morphologic and molecular information rather than through imaging or histology alone [1]. This makes computational brain tumor classification a naturally multimodal problem: MRI captures macroscopic anatomy and tumor extent, while hematoxylin and eosin (H&E) pathology slides capture cellular morphology and tissue organization.

MEDIQA-CORE-Task-1 2026 focuses on brain tumor subtype classification within ImageCLEF 2026 and uses the CoRe-BT benchmark [2, 3, 4]. For each test subject, systems must predict three labels: Level 1 molecular grouping, LGG/HGG status, and WHO grade. The task is challenging because the number of labeled subjects is modest, class distributions are imbalanced, and modalities may be missing for some subjects. The benchmark provides raw MRI and pathology data, but also includes

CLEF 2026 Working Notes, 21–24 September 2026, Jena, Germany
*Corresponding author.
24521720@gm.uit.edu.vn (N. M. Thoi); 24521469@gm.uit.edu.vn (L. N. Quang); thiennt@uit.edu.vn (T. B. Nguyen-Tat)
0000-0002-4809-7126 (T. B. Nguyen-Tat)

precomputed embeddings from foundation models such as NeuroVFM for neuroimaging and GigaPath for pathology [5, 6]. These embeddings enable efficient experimentation without training large visual backbones from scratch.

Our team, UIT_Concabietbay, built a subject-level pipeline over the provided MRI and pathology embeddings. The system aggregates variable-size modality representations into fixed-length tabular features, handles missing modalities explicitly, trains target-specific classifiers, and combines complementary prediction candidates through out-of-fold (OOF) probability blending. This design was motivated by the small-data regime of the task. In such a setting, classical regularized models and carefully validated ensembles can be more stable than high-capacity end-to-end models, particularly when the available features are already generated by strong pretrained encoders.

The main contributions of our submitted system are:

- a reproducible subject-level pipeline for MRI and pathology embedding aggregation;
- modality-aware feature fusion with explicit availability indicators and imputation;
- target-specific model selection for three clinically related classification outputs;
- OOF-guided probability blending, optional stacking, and binary threshold tuning;
- organizer-verified post-code-execution results across modality conditions, with pre-verification leaderboard feedback reported separately.

2. Related Work

2.1. Brain Tumor Classification and Multimodal Benchmarks

Glioma diagnosis and grading require combining histopathologic and molecular evidence, and the WHO classification provides the clinical framing for such integrated diagnosis [1]. CoRe-BT extends this direction computationally by providing paired MRI and pathology resources for brain tumor classification [4]. The MEDIQA-CORE task builds on this benchmark and asks participants to produce subject-level predictions for clinically meaningful labels [3]. Unlike segmentation-centered challenges, this task evaluates multimodal classification at the patient or subject level.

2.2. Foundation Embeddings for Medical Images

Foundation models are increasingly used as representation extractors for downstream medical learning. In computational pathology, whole-slide images are too large for direct processing in standard settings, making multiple-instance learning and pretrained slide embeddings attractive alternatives [7, 8]. GigaPath is a large whole-slide pathology foundation model trained from real-world digital pathology data and has shown strong transfer performance across pathology tasks [6]. For radiology, NeuroVFM has been proposed as a generalist visual foundation model for neuroimaging, supporting MRI and CT representations [5]. Our system does not fine-tune these encoders; instead, it uses the provided embeddings as fixed representations.

Related medical-imaging studies by Nguyen-Tat and collaborators have investigated MRI brain tumor segmentation, weakly supervised medical image segmentation, histopathology-based cancer classification, osteoarthritis severity diagnosis, and chest-X-ray-based COVID-19 pneumonia severity assessment [9, 10, 11, 12, 13]. These studies are complementary to the present work: they focus mainly on end-to-end or task-specific image-analysis pipelines, whereas our MEDIQA-CORE system uses organizer-provided embeddings and classical tabular ensembles for subject-level multimodal classification.

2.3. Classical Tabular Models and Ensembles

For small labeled datasets with high-dimensional embedding features, regularized linear models and classical ensemble methods remain strong baselines. Logistic regression and support-vector machines

provide variance-controlled decision boundaries, while random forests, extra trees, and gradient boosting offer nonlinear tabular alternatives implemented in scikit-learn and common boosting packages [14, 15, 16, 17, 18, 19, 20]. Stacking and probability averaging can improve robustness when model errors are complementary [21]. Our system follows this tradition: we emphasize OOF validation, calibrated probability outputs when possible, and probability-level fusion over a single highly tuned classifier.

3. Task and Data

3.1. Task Definition

The official prediction file for MEDIQA-CORE-Task-1 contains one row per test subject and four required columns:

- `subject_id`, the subject identifier;
- `level1_pred`, the Level 1 molecular grouping prediction;
- `lgghgg_pred`, the binary LGG/HGG prediction;
- `who_grade_pred`, the WHO grade prediction.

The three targets have different label structures. Level 1 and WHO grade are multiclass outputs with three integer labels, while LGG/HGG is binary. Consequently, our pipeline trains and selects models independently for each target rather than enforcing a single shared classifier.

3.2. Available Modalities

The CoRe-BT resources include MRI volumes, H&E whole-slide pathology images, pathology reports, segmentations, and precomputed embeddings [4]. The implementation verified for this paper uses the precomputed MRI and pathology embeddings as model inputs. Report text is present in local files and is used for subject ordering and manual audit context, but the submitted machine-learning pipeline does not train a text encoder or use report text as learned NLP features. We therefore describe the system as an MRI-pathology embedding system rather than a tri-modal image-text model.

Table 1 summarizes the main data sources used by the submitted pipeline.

Table 1: Main data sources used by the UIT_Concabietbay system.

Data source	Role	Usage in our system
Training labels	Supervised targets	Train target-specific classifiers for Level 1, LGG/HGG, and WHO grade
MRI embeddings	Radiology representation	Mean and standard-deviation pooling over subject-level embedding matrices
Pathology slide embeddings	Histopathology representation	Mean, standard deviation, maximum, attention, and robust-mean pooling over slide vectors
Test reports	Subject order and audit context	Not used as learned text features
Submission CSV	Official output	Four-column prediction file with integer labels

3.3. Local Data Observed in the Implementation

The local training label file contains 177 labeled subjects with columns `subject_id`, `level1_label`, `lgghgg_label`, and `who_grade_label`. The saved validation/evaluation split used by the main reproducible run contains 39 subjects, and the final local test-submission workflow contains 36 subjects.

MRI embeddings are available for most training subjects and all validation/test subjects used in the saved runs, while pathology slide availability is incomplete. The code therefore treats missing modalities as part of the problem instead of dropping subjects with incomplete modality coverage.

The organizer-provided results spreadsheet reports a broader official data distribution of 183 training subjects, 42 validation subjects, 36 Test Set 1 subjects, and 18 Test Set 2 subjects. Because these official counts differ from the subset used in the saved local run, we report local validation and organizer post-verification results as separate evidence sources. Specifically, the official distribution contains 183 training subjects (35 with both MRI and histopathology, 16 histopathology-only, and 132 MRI-only), 42 validation subjects (8 both-present, 3 histopathology-only, and 31 MRI-only), 36 Test Set 1 subjects with both modalities present, and 18 Test Set 2 subjects with MRI only; the full-dataset totals are 79 both-present, 19 histopathology-only, and 181 MRI-only subjects, for 279 subjects in total.

4. System Overview

Figure 1 shows the main workflow. The system first loads MRI and pathology embedding files for each subject. It then applies modality-specific pooling operations to obtain fixed-length feature vectors. Feature tables are merged by `subject_id`, availability masks are retained, and candidate models are evaluated separately for each target. OOF probabilities are used both for model selection and for ensemble construction. Final predictions are assembled in the required CSV format.

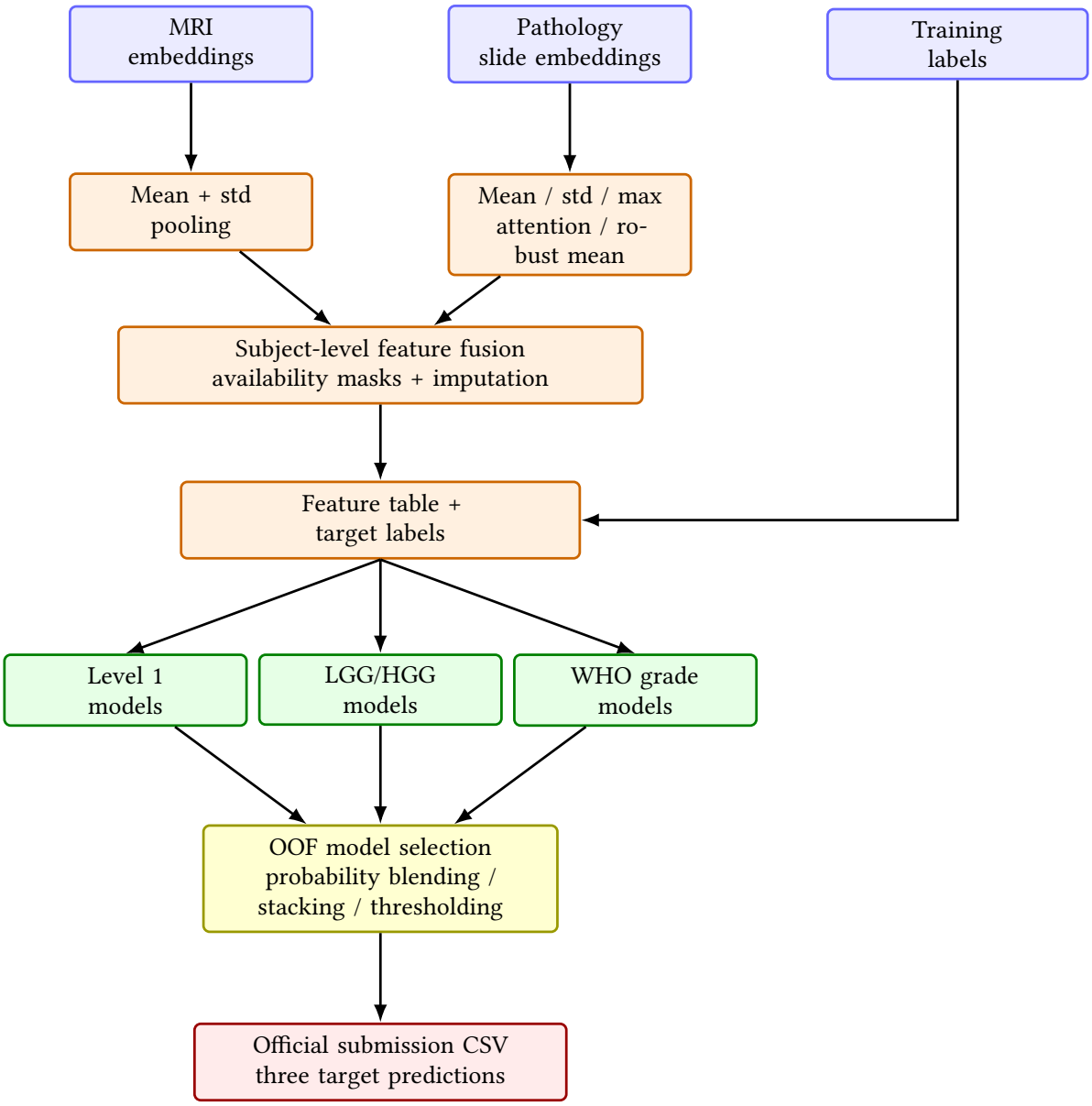

Figure 1: Overview of the UIT_Concabietbay pipeline. MRI and pathology embeddings are pooled into subject-level features, fused by identifier, and used to train target-specific classifiers. Out-of-fold probabilities guide model selection and ensemble construction.

5. Feature Extraction

5.1. MRI Embedding Features

Let $E_i \in \mathbb{R}^{n_i \times d}$ denote the MRI embedding matrix for subject i, where n_i is the number of embedding vectors and d is the embedding dimension. The implementation accepts one-dimensional, two-dimensional, or higher-rank tensor representations and reshapes them into a matrix when needed. In the main configuration, MRI features are derived using mean and standard-deviation pooling:

$$\mu_i = \frac{1}{n_i} \sum_{j=1}^{n_i} E_{ij}, \qquad \sigma_i = \sqrt{\frac{1}{n_i} \sum_{j=1}^{n_i} (E_{ij} - \mu_i)^2}.$$

The final MRI feature vector is the concatenation of these pooled vectors, plus metadata such as modality availability, file count, token count, embedding dimension, and the norm of the mean

embedding. If multiple compatible MRI files exist for a subject, they are vertically concatenated before pooling. Files with incompatible dimensions or invalid numeric values are skipped and logged.

5.2. Pathology Slide Embedding Features

Pathology slide embeddings are loaded from slide-level files using the configured `last_layer_embed` representation. For a subject with slide vectors $S_i \in \mathbb{R}^{m_i \times d}$, the main pipeline extracts mean, standard deviation, maximum pooling, norm-based attention pooling, and robust mean pooling.

The attention vector is computed from embedding norms. For slide j, let $r_j = \|S_{ij}\|_2$. If more than one slide is available and the norm variance is nonzero, the implementation standardizes the norms, clips the standardized values, applies a softmax, and forms a weighted sum:

$$\alpha_j = \frac{\exp(z_j)}{\sum_{k=1}^{m_i} \exp(z_k)}, \qquad a_i = \sum_{j=1}^{m_i} \alpha_j S_{ij}.$$

Robust mean pooling removes norm outliers using a median absolute deviation rule when enough slides are available; otherwise it falls back to the arithmetic mean. Tile-level pathology features are supported in the source code through HDF5 summaries, but tile features are disabled in the saved main run described here. This distinction is important for reproducibility: the submitted system should be described primarily as using slide-level pathology embeddings.

5.3. Fusion and Missing Modalities

MRI and pathology feature tables are left-merged by `subject_id`. The feature table retains explicit modality masks, including `mri_available`, `path_slide_available`, and `path_tile_available`. Missing numeric values are not filled during feature construction for the main model path. Instead, model pipelines use median imputation with missingness indicators, allowing classifiers to exploit both observed embedding values and the fact that a modality is absent.

The implementation defines several feature configurations:

- `mri`: MRI-derived features only;
- `path_slide`: pathology slide features only;
- `mri_path_slide`: early fusion of MRI and pathology slide features;
- `all`: all available numeric features in the saved configuration.

6. Model Training

6.1. Target-Specific Classifiers

The three labels are trained independently. This choice avoids forcing one model to share a representation across targets with different class distributions and error costs. For Level 1 and WHO grade, predictions are obtained with the maximum-probability class. For LGG/HGG, the pipeline may tune a positive-class threshold on OOF probabilities; the tuned threshold is used only when it improves OOF macro-F1 relative to the default decision rule.

6.2. Model Zoo

The candidate model families include:

- dummy prior baselines for sanity checking;
- logistic regression with and without class balancing;
- PCA followed by logistic regression;
- calibrated ridge classifiers and calibrated linear SVMs;

- RBF-kernel SVMs with probability output;
- random forests and extra trees;
- histogram gradient boosting;
- multilayer perceptrons with early stopping;
- optional LightGBM, XGBoost, and CatBoost candidates when installed.

Linear models are useful because the embedding dimensionality is high relative to the number of labeled subjects. Tree ensembles and boosting models provide nonlinear alternatives, while calibrated linear models and SVMs add complementary decision boundaries. The final selected ensemble is not a CatBoost-only model; CatBoost is one optional member of the candidate zoo.

6.3. Cross-Validation

The main saved run uses stratified cross-validation with five folds and a fixed random seed. Stratification uses the Level 1 label by default and can fall back to a combined target label when every combined stratum has enough examples. Each candidate model produces OOF predictions, fold-level metrics, and saved fold models. The recorded metrics include accuracy, balanced accuracy, macro-F1, weighted-F1, and log loss. Macro-F1 is the primary selection metric because the targets are imbalanced and the official score is macro-F1 based.

7. Ensembling and Submission Assembly

7.1. Greedy Probability Blending

The main selected strategy is greedy probability blending. For each target, the pipeline starts with candidate OOF probability matrices and iteratively adds the model whose inclusion yields the best macro-F1. Selected probability matrices are averaged with normalized weights. This simple fusion strategy is robust in small-data settings because it uses validation evidence while avoiding a high-capacity meta-model unless there is enough OOF signal.

7.2. Stacking

The code also supports stacking with a logistic-regression meta-classifier trained over concatenated OOF probabilities. Stacking is treated as an optional candidate variant rather than as the default final strategy. This is deliberate: with only 177 labeled training subjects, a meta-model can overfit if base predictions are highly correlated.

7.3. Consistency Checks

The implementation evaluates consistency rules between WHO grade and LGG/HGG status. For example, one candidate post-processing rule maps WHO grade 1/2 to LGG and grades 3/4 to HGG. Such rules are not applied blindly; they are retained only if they improve OOF macro-F1 or are supported by candidate analysis. This conservative approach avoids improving one output at the expense of another in the averaged official score.

7.4. Final Submission

The submitted prediction files follow the required format and contain integer predictions. The 36-subject Run ID 728 leaderboard feedback is retained in Table 6 as all-modality pre-verification feedback generated before organizer code verification. The organizer post-verification results in Table 5 are taken from the organizer-provided spreadsheet after code verification and execution under specified modality conditions. Because the hidden test labels are not available locally, the local workspace cannot independently recompute these official scores. The workspace also contains later candidate-assembly

Table 2
Team-conducted local validation metrics from the main saved CoReBT run. These values are not organizer-verified hidden-test results.

Target	Accuracy	Balanced Acc.	Macro-F1	Weighted-F1
Level 1	0.7910	0.6653	0.6953	0.7849
LGG/HGG	0.8757	0.7902	0.7902	0.8757
WHO grade	0.7401	0.5890	0.5995	0.7323
Mean	0.8023	0.6815	0.6950	0.7976

artifacts; therefore, exact reproduction of a specific submitted run requires preserving the archived prediction file, submission ZIP, artifact manifest, and candidate-selection note rather than relying on a mutable root-level prediction file.

8. Experimental Setup

8.1. Implementation

The system is implemented in Python using NumPy and pandas for data handling, scikit-learn for preprocessing and most model families, joblib for model serialization, and optional boosting libraries such as XGBoost, LightGBM, and CatBoost. The reproducible pipeline stores feature tables, model rankings, fold metrics, OOF probabilities, selected metrics, trained fold models, and submission variants. Manifest-based inference reloads saved fold models and averages their predicted probabilities.

8.2. Local Validation Protocol

The main reproducible run used:

- 177 training subjects;
- 39 local evaluation subjects;
- feature sets `mri`, `path_slide`, `mri_path_slide`, and `all`;
- five-fold stratified CV with seed 42;
- macro-F1 as the primary model-selection score.

The saved feature build produced train and evaluation tables with 177 and 39 rows, respectively, and no invalid feature files were recorded in the main run.

9. Results

9.1. Local Validation

Table 2 reports team-conducted local validation metrics from the main saved run. These values were used for model selection and are not organizer-verified hidden-test results. LGG/HGG obtained the highest local macro-F1, while WHO grade was the most difficult target. This pattern is plausible because WHO grade is a three-class prediction problem and may require finer morphologic and molecular evidence.

9.2. Feature Ablation

Table 3 summarizes a team-conducted feature-set ablation from the saved run. These ablation experiments were not executed or verified by the task organizers. MRI-only features were stronger than pathology slide-only features. Early fusion of MRI and pathology slide embeddings improved over

Table 3
Team-conducted feature-set ablation summary from the saved run. These experiments were not executed or verified by the task organizers.

Feature setting	Mean best-task macro-F1
MRI only	0.6245
Pathology slide only	0.4305
MRI + pathology slide	0.6671
All available features	0.6699

Table 4
High-level summary of selected local ensemble strategies.

Target	Selected strategy	Feature usage
Level 1	Greedy probability blend	All-feature and MRI-only variants
LGG/HGG	Greedy probability blend	MRI + pathology slide and all-feature variants
WHO grade	Greedy probability blend	MRI + pathology slide and all-feature variants

Table 5
Organizer post-verification results for UIT_Concabietbay after code verification and execution under the organizer-specified modality conditions.

Test set	Organizer condition	Rank	Level 1	WHO grade	LGG/HGG	Average
Test 1	Fully multimodal (MRI images + MRI reports + pathology images)	4	0.878	0.729	0.700	0.769
Test 1	MRI images + MRI reports	3	0.666	0.611	0.807	0.695
Test 1	MRI images + pathology images	3	0.878	0.729	0.700	0.769
Test 1	Pathology images + MRI reports	5	0.515	0.756	0.700	0.657
Test 2	MRI images + MRI reports	2	0.402	0.508	0.730	0.547
Test 2	MRI images only	1	0.402	0.508	0.730	0.547
Test 2	MRI reports only	3	0.127	0.231	0.433	0.264

either modality alone, and the all-feature setting obtained the best mean best-task macro-F1. These results suggest that pathology embeddings provide complementary signal, but their standalone utility is limited by incomplete availability and by simple slide-level aggregation.

9.3. Selected Ensemble Components

The final selected local ensemble used different feature sets and base learners by target. Table 4 summarizes the saved selection at a high level. Level 1 combined an all-feature logistic-regression variant with an MRI-only logistic-regression variant. LGG/HGG combined an MRI-pathology PCA-logistic model with an all-feature random forest. WHO grade combined three logistic-regression variants over MRI-pathology and all-feature settings.

9.4. Organizer Post-Verification Results

Table 5 reports the UIT_Concabietbay rows from the organizer post-verification spreadsheet. The average is the arithmetic mean of the three macro-F1 components. The spreadsheet reports separate Test Set 1 and Test Set 2 results under modality conditions executed after code verification. Rows that include MRI reports should be interpreted in light of our implementation: report files were accepted for subject ordering and audit context, but report text was not used as a learned NLP feature. Consequently, conditions that add or remove reports can be identical for our system.

Table 6
Pre-verification all-modality MEDIQA-CORE-Task-1 leaderboard feedback. These scores were generated before organizer code verification and are not the final post-verification results.

Rank	Team	Run ID	Level 1	WHO grade	LGG/HGG	Average
1	catcd	1066	1.000	0.956	1.000	0.985
2	cnvntq	1049	0.918	0.886	0.911	0.905
3	**UIT_Concabietbay**	**728**	**0.900**	**0.858**	**0.839**	**0.866**
4	htruong47	949	0.915	0.697	0.859	0.824
5	anubhav27	1039	0.540	0.772	0.958	0.757
6	cclark339	986	0.792	0.679	0.765	0.745
–	*MEDIQA Organizers*	1185	0.539	0.726	0.920	0.728

9.5. Pre-Verification Leaderboard Feedback

Table 6 reports the all-modality leaderboard feedback available before organizer code verification. These scores were generated before the task organizers executed and verified participant code under the modality-specific conditions in Table 5; therefore, they are retained only as historical leaderboard feedback and not as the final organizer-verified results.

10. Discussion

10.1. Why Embedding-Based Classical Learning Worked

The organizer post-verification results show that a carefully validated embedding-based pipeline can remain competitive for multimodal brain tumor classification, although the verified scores are lower than the earlier all-modality leaderboard feedback. Precomputed embeddings reduce the need for large GPU resources and avoid unstable end-to-end fine-tuning on a small dataset. Classical models then operate on compact representations with strong inductive biases: linear models control variance in high dimensions, while tree ensembles and boosting models capture nonlinear interactions.

10.2. MRI and Pathology Complementarity

The team-conducted local ablation indicates that MRI embeddings carry the strongest standalone signal in the saved run. Pathology slide features alone perform worse, but fusion improves over MRI alone. These ablation experiments were not verified or executed by the organizers, so they should be interpreted as local diagnostic evidence rather than official modality results. There are several possible explanations. First, pathology availability is incomplete. Second, slide-level pooling may discard spatial or cellular details that more advanced multiple-instance learning methods could exploit. Third, the target labels include molecular and grade information that may be reflected differently across imaging modalities.

10.3. Local Validation Versus Official Evaluation

The revised result reporting separates three evidence sources: local CV and ablation results produced by our team, pre-verification all-modality leaderboard feedback, and organizer post-verification results after code execution under modality conditions. The gap among these sources highlights the instability of small medical benchmarks. With only 177 labeled subjects in the saved local training file, fold composition and class imbalance can strongly affect local estimates. OOF validation remains necessary, but it should not be interpreted as a precise estimate of hidden-test performance. The submitted system therefore used target-specific selection rather than a single global model choice.

10.4. Reproducibility Considerations

The codebase supports reproducible feature construction, CV training, manifest-based inference, and generation of multiple submission variants. However, a code audit also found that later local artifacts can combine predictions from different candidate submissions. For camera-ready reproducibility, the exact Run-728 prediction file, ZIP file, artifact manifest, and candidate-selection note should be archived together. Future versions should include a single command that reconstructs the exact official submission from immutable inputs.

10.5. Limitations

The system has four main limitations. First, it does not train directly from raw MRI volumes or raw whole-slide images, so it cannot adapt the visual encoders to the task. Second, it does not use pathology reports as learned text features, even though report text may contain useful clinical evidence. Third, pathology aggregation is based on simple slide-level pooling rather than an end-to-end multiple-instance learning model. Fourth, the final submission process involves candidate selection beyond the main CV training loop, which should be documented more rigorously in future releases.

11. Conclusion

We presented the UIT_Concabietbay system for MEDIQA-CORE-Task-1 2026. The system formulates brain tumor subtype prediction as subject-level multimodal tabular learning over precomputed MRI and pathology embeddings. MRI features are derived with mean and standard-deviation pooling, pathology slide features use multiple complementary pooling operations, and missing modalities are handled with availability masks and median imputation. Target-specific classical classifiers are selected with OOF macro-F1 and combined through greedy probability blending, with optional stacking and binary threshold tuning. In the organizer post-verification spreadsheet, our strongest verified average macro-F1 was 0.769 on Test Set 1 and 0.547 on Test Set 2 under the reported modality conditions. The earlier 0.866 all-modality leaderboard score is retained only as pre-verification feedback rather than as a final organizer-verified result.

Future work should focus on stronger pathology aggregation, incorporation of report text, calibration-aware ensembling, and a fully scripted final-submission reconstruction path. More advanced multimodal architectures may also improve performance if regularized carefully for the small-data setting.

Acknowledgments

We thank the MEDIQA-CORE and ImageCLEF organizers for providing the CoRe-BT benchmark, the evaluation platform, and the official task feedback.

This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT to assist with language polishing, manuscript structuring, LaTeX formatting, and source-code auditing. The authors reviewed and edited the manuscript and take full responsibility for its content.

References

[1] D. N. Louis, A. Perry, P. Wesseling, D. J. Brat, I. A. Cree, D. Figarella-Branger, C. Hawkins, H. K. Ng, S. M. Pfister, G. Reifenberger, R. Soffietti, A. von Deimling, D. W. Ellison, The 2021 WHO

classification of tumors of the central nervous system: a summary, Neuro-Oncology 23 (2021) 1231–1251. doi:10.1093/neuonc/noab106.

[2] B. Ionescu, H. Müller, D. Stanciu, A. Radu, R. Bolborici, M. Negru, A. Ene, V. Vasilescu, A.-A. Nicolae, L. Ştefan, M. Constantin, M. Dogariu, A. Andrei, H. Damm, T. M. G. Pakull, A. Ben Abacha, A. García Seco de Herrera, C. M. Friedrich, R. Brüngel, L. Reinartz, H. Schäfer, C. S. Schmidt, B. Bracke, P. Nath, B. Eryılmaz, M. Hjuler, D. Fabre, C. Lemaire, B. Lecouteux, D. Schwab, D. Dimitrov, M. S. Hee, M. Ahsan, S. Ahmad, D. Zlatkova, G. Pachov, Z. Xie, P. Nakov, I. Koychev, J. E. Heras Rivera, D. K. Low, W. Yim, J. Ruzevick, D. Child, M. Kurt, Z. Sun, F. Xia, M. Yetisgen, A. Radzhabov, Y. Prokopchuk, V. Kovalev, D. Karpenka, S. A. Hicks, S. Gautam, M. A. Riegler, V. Thambawita, P. Halvorsen, M. El Sakka, J. Mothe, A. Băicoianu, C. Florea, M. Ivanovici, Overview of ImageCLEF 2026: Multimodal challenges in medicine, science, agritech, and security, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Springer Lecture Notes in Computer Science LNCS, Jena, Germany, 2026.

[3] A. Ben Abacha, J. E. Heras Rivera, D. K. Low, W.-w. Yim, J. Ruzevick, D. Child, M. Kurt, Overview of the MEDIQA-CORE 2026 task 1 on brain tumor subtype classification, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.

[4] J. E. Heras Rivera, D. K. Low, X. Xiong, J. J. Ruzevick, D. D. Child, W.-w. Yim, M. Kurt, A. Ben Abacha, CoRe-BT: A multimodal radiology-pathology-text benchmark for robust brain tumor typing, volume abs/2603.03618, 2026. URL: https://arxiv.org/abs/2603.03618.

[5] A. Kondepudi, et al., Health system learning achieves generalist neuroimaging models, 2025. arXiv:2511.18640.

[6] H. Xu, et al., A whole-slide foundation model for digital pathology from real-world data, Nature 630 (2024) 181–188. doi:10.1038/s41586-024-07441-w.

[7] M. Ilse, J. M. Tomczak, M. Welling, Attention-based deep multiple instance learning, in: Proceedings of the 35th International Conference on Machine Learning, volume 80 of *Proceedings of Machine Learning Research*, PMLR, 2018, pp. 2127–2136.

[8] M. Y. Lu, D. F. K. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, F. Mahmood, Data-efficient and weakly supervised computational pathology on whole-slide images, Nature Biomedical Engineering 5 (2021) 555–570. doi:10.1038/s41551-020-00682-w.

[9] T. B. Nguyen-Tat, T.-Q. T. Nguyen, H.-N. Nguyen, V. M. Ngo, Enhancing brain tumor segmentation in MRI images: A hybrid approach using UNet, attention mechanisms, and transformers, Egyptian Informatics Journal 27 (2024) 100528. doi:10.1016/j.eij.2024.100528.

[10] T. B. Nguyen-Tat, H.-A. Vo, P.-S. Dang, QMaxViT-Unet+: A query-based MaxViT-Unet with edge enhancement for scribble-supervised segmentation of medical images, Computers in Biology and Medicine 187 (2025) 109762. doi:10.1016/j.compbiomed.2025.109762.

[11] T. B. Nguyen-Tat, A. T. Vu-Xuan, V. M. Ngo, Design of a novel fuzzy ensemble CNN framework for ovarian cancer classification using tissue microarray images, Image and Vision Computing 161 (2025) 105604. doi:10.1016/j.imavis.2025.105604.

[12] T. B. Nguyen-Tat, T.-P. Nguyen-Duong, Optimizing knee osteoarthritis severity diagnostics: A GA-enhanced deep ensemble approach in medical imaging, Ain Shams Engineering Journal 16 (2025) 103524. doi:10.1016/j.asej.2025.103524.

[13] T. B. Nguyen-Tat, V.-T. Tran-Thi, V. M. Ngo, Predicting the severity of COVID-19 pneumonia from chest X-Ray images: A convolutional neural network approach, EAI Endorsed Transactions on Industrial Networks and Intelligent Systems 12 (2024). doi:10.4108/eetinis.v12i1.6240.

[14] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning 20 (1995) 273–297. doi:10.1007/BF00994018.

[15] L. Breiman, Random forests, Machine Learning 45 (2001) 5–32. doi:10.1023/A:1010933404324.

[16] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Machine Learning 63 (2006) 3–42. doi:10.1007/s10994-006-6226-1.

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay,

Scikit-learn: Machine learning in python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[18] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794. doi:10.1145/2939672.2939785.

[19] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, LightGBM: A highly efficient gradient boosting decision tree, in: Advances in Neural Information Processing Systems, volume 30, 2017.

[20] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, A. Gulin, CatBoost: unbiased boosting with categorical features, in: Advances in Neural Information Processing Systems, volume 31, 2018.

[21] D. H. Wolpert, Stacked generalization, Neural Networks 5 (1992) 241–259. doi:10.1016/S0893-6080(05)80023-1.