MELBA Journal Machine Learning for Biomedical Imaging

Towards Trustworthy AI for Glioma Diagnosis: A Task-Aware Evaluation of Uncertainty Quantification

Gonzalo Esteban **Mosquera Rojas** [1], Sebastian R. **van der Voort** [2], Carolin M. **Pirkl** [3], Sandeep **Kaushik** [4], Marion **Smits** [1,5,6], Stefan **Klein** [1]

1 Department of Radiology and Nuclear Medicine, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands
2 Department of Medical Informatics, Amsterdam UMC, University of Amsterdam, Amsterdam, the Netherlands
3 GE HealthCare, Munich, Germany
4 GE HealthCare, USA
5 Brain tumor Centre, Erasmus MC Cancer Institute, Rotterdam, the Netherlands
6 Medical Delta, Delft, the Netherlands

Abstract

Uncertainty Quantification (UQ) is a key requirement for trustworthy AI in high-stakes medical image analysis. In this work, we evaluate UQ within a multi-task Deep Learning (DL) framework for MRI-based glioma diagnosis that performs tumor segmentation and predicts Isocitrate Dehydrogenase (IDH) mutation status, 1p/19q co-deletion status and tumor grade. We use Monte Carlo Dropout (MCD) as the primary sampling-based UQ method for the detailed task-aware analysis, obtaining predictive uncertainty and its aleatoric and epistemic components. We assess uncertainty along complementary axes: (i) MC sample convergence of uncertainty estimates and their decomposition into aleatoric and epistemic components, (ii) calibration of predictive probabilities, and (iii) operational utility of uncertainty estimates, including error detection, selective prediction, and associations with tumor segmentation performance. We also compare MCD with Deep Ensembles (DE) and Monte Carlo Deep Ensembles (MCDE) to assess whether the observed operational utility of uncertainty estimates extends beyond a single UQ method. For the tumor segmentation task, we further examine how different voxel-wise uncertainty aggregation strategies influence case-level reliability assessment, thereby explicitly accounting for task-specific uncertainty representation. Additionally, we study task interactions to quantify how tumor segmentation quality and uncertainty relate to the prediction of the tumor features. Finally, we explore whether a composite trust score integrating tumor segmentation and classification uncertainty improves error detection. Across tasks, uncertainty estimates supported meaningful error detection, while calibration depended on the dropout rate, with moderate rates yielding the most reliable predictive probabilities. Decomposition provided task-dependent interpretability but did not consistently improve error detection over predictive uncertainty alone. The comparison with DE and MCDE showed that ensemble-based uncertainty estimates provided comparable operational utility, although no UQ method consistently dominated across all tasks and metrics. Moreover, the proposed trust score did not consistently outperform classification uncertainty for selective prediction, indicating that task-specific predictive uncertainty remains the most informative operational indicator of trust. Overall, our results provide a task-aware evaluation strategy and practical guidance towards the development of trustworthy AI for glioma diagnosis.

Keywords

Glioma, Magnetic Resonance Imaging, Deep Learning, Uncertainty Quantification, Monte Carlo Dropout, Deep Ensembles

Article informations

https://doi.org/10.59275/j.melba.2026-456d

Volume 2026, Received: 2026-03-09, Published 2026-09-28

Corresponding author: g.mosquerarojas@erasmusmc.nl
Special issue: Uncertainty for Safe Utilization of Machine Learning in Medical Imaging (UNSURE) 2025
Guest editors: Mobarak Hoque, Raghav Mehta, Cheng Ouyang, Chen Qin, Marianne Rakic, Sandy Wells

1. Introduction

Gliomas are among the most aggressive primary brain tumors and are associated with substantial morbidity and mortality worldwide (Ostrom et al., 2019). Advances in molecular neuropathology have transformed glioma classification, shifting diagnostic standards from purely histological grading towards integrated molecular definitions. The current World Health Organization (WHO) classification incorporates molecular markers such as Isocitrate Dehydrogenase (IDH) mutation status and 1p/19q co-deletion status, alongside tumor grade, to stratify gliomas into biologically and clinically meaningful subtypes (Louis et al., 2021). These characteristics directly influence therapeutic strategy and patient prognosis, making accurate characterization essential for clinical decision-making.

Magnetic resonance imaging (MRI) is the primary non-invasive imaging modality used for diagnosis, treatment planning, and longitudinal monitoring of patients with glioma (Abdalla et al., 2020). In clinical practice, MRI is often assessed using qualitative criteria or simple measures based on tumor diameter, providing only a rudimentary view of tumor burden. Automated or semi-automated segmentation can produce more detailed and reproducible measurements of tumor subregions, supporting improved diagnosis, treatment planning, and longitudinal monitoring. However, developing robust algorithms remains challenging due to variability in tumor appearance and subtle intensity differences (Menze et al., 2014). Beyond spatial delineation, Deep Learning (DL) approaches have demonstrated promising capability in predicting molecular markers such as IDH mutation and 1p/19q co-deletion status directly from imaging data, a paradigm often referred to as radiogenomics (Chang et al., 2018; Kickingereder et al., 2019).

To exploit shared imaging representations, multi-task learning frameworks have been proposed that jointly perform tumor segmentation and genetic classification within a single model architecture. Such models leverage shared feature representations and may improve efficiency and generalization compared to isolated single-task approaches (Kordnoori et al., 2024; Zhou et al., 2019; van der Voort et al., 2023). However, while predictive performance has advanced substantially, translation of these systems into clinical workflows remains limited due to the black-box nature of Neural Networks (NNs), which lack reliable confidence scores about their predictions (Lambert et al., 2024; Guo et al., 2017).

In high-stakes medical applications, predictive accuracy alone is insufficient. AI systems must communicate the reliability of their predictions to support safe clinical integration (Ojha et al., 2025; Abdar et al., 2022; He et al., 2025). Uncertainty Quantification (UQ) has therefore emerged as a key component of trustworthy medical AI, aiming to offer rationale behind the DL models' predictions and increase their interpretability (Lambert et al., 2024). Uncertainty estimates can enable selective automation by identifying cases requiring human review, and thus improving transparency in decision-support systems.

Despite this growing interest, uncertainty in multi-task medical AI systems remains insufficiently characterized. In multi-task settings, segmentation and classification branches share intermediate representations but produce heterogeneous outputs. Uncertainty may propagate differently across tasks, and segmentation quality may influence subtype prediction reliability. Furthermore, predictive uncertainty can be decomposed into epistemic uncertainty, reflecting model uncertainty due to limited knowledge, and aleatoric uncertainty, reflecting intrinsic data ambiguity (Kendall and Gal, 2017). While theoretically appealing, recent work has questioned whether disentangling uncertainty sources consistently yields practical benefits in real-world settings, emphasizing that the usefulness of such decomposition for segmentation tasks may depend on dataset characteristics and downstream evaluation criteria (Kahl et al., 2024). Systematic empirical evaluation of these aspects in clinically relevant multi-task systems remains limited.

These considerations highlight the need for a comprehensive and task-aware assessment of AI-based diagnostic systems. In this work, we present a systematic evaluation of UQ within a multi-task DL framework for MRI-based glioma diagnosis. The pipeline jointly performs tumor segmentation and predicts IDH mutation status, 1p/19q co-deletion status, and tumor grade. The DL architecture, which was initially proposed by van der Voort et al. (2023), showed promising performance but lacked UQ and analysis of the interaction between the different tasks. Here, we use Monte Carlo Dropout (MCD) (Gal and Ghahramani, 2016) as the primary sampling-based UQ method to obtain predictive uncertainty and its aleatoric and epistemic components. In addition, we compare MCD with Deep Ensembles (DE) and Monte Carlo Deep Ensembles (MCDE) to assess whether the observed operational utility of uncertainty estimates extends beyond a single UQ method.

We evaluate uncertainty along multiple complementary axes: Monte Carlo (MC) sample convergence of uncertainty estimates, uncertainty decomposition, calibration of predictive probabilities, operational utility of uncertainty estimates including a comparison across different UQ methods, and cross-task interactions between tumor segmentation and the reliability of tumor feature prediction. We further analyze how different voxel-wise aggregation strategies affect segmentation-level uncertainty assessment. Through this comprehensive evaluation, we aim to clarify when and how uncertainty estimates meaningfully contribute to reliability in a multi-task glioma diagnosis AI system. By grounding uncertainty analysis in clinically relevant tasks and opera-

tional metrics, our study provides practical guidance towards the development of trustworthy AI for this application.

2. Related Works

2.1 MRI-based radiogenomics for glioma characterization

Radiogenomics has emerged as a promising paradigm for the non-invasive characterization of gliomas by linking quantitative imaging features to molecular and genomic tumor profiles. Several comprehensive reviews highlight the growing role of radiomics and radiogenomics in supporting precision medicine for glioma management, emphasizing applications ranging from molecular subtyping and risk stratification to prognosis estimation and treatment monitoring (Mitra, 2021; Singh et al., 2021; Fathi Kazerooni et al., 2021).

MRI constitutes one of the central imaging modalities in these efforts. Structural sequences, including T1-weighted (T1w), Contrast-Enhanced T1-weighted (T1wCE), T2-weighted (T2w), and Fluid-Attenuated Inversion Recovery (FLAIR), provide rich morphological information that can be exploited by radiomics pipelines for tumor characterization (Singh et al., 2021). These approaches have demonstrated potential for predicting molecular subtypes, distinguishing true progression from pseudoprogression, estimating overall survival and progression-free survival, and identifying recurrence patterns (Fathi Kazerooni et al., 2021).

In parallel, DL–based radiogenomics has gained substantial momentum. Convolutional Neural Networks (CNNs) have been used to automatically extract hierarchical imaging features directly from MRI data, often outperforming classical radiomics approaches in molecular subtype prediction (Li et al., 2022). For instance, Buda et al. (2020) demonstrated that CNNs can predict genomic subtypes of lower-grade gliomas directly from MRI, exploring both training from scratch and transfer learning strategies. Comparative studies between radiomics and DL models suggest that learned representations may capture complementary or superior discriminative information relative to handcrafted features (Li et al., 2022).

Despite promising performance, translation of radiogenomic models into routine clinical practice remains limited. Reviews consistently identify challenges related to reproducibility, generalizability across institutions, and robustness under distribution shifts (Fathi Kazerooni et al., 2021; Singh et al., 2021). In addition, uncertainty and interpretability of DL models are increasingly recognized as critical factors for clinical acceptance (Singh et al., 2021).

2.2 Multi-task Deep Learning

Multi-task DL has gained increasing attention in medical imaging as a strategy to leverage shared representations across related tasks, often leading to improved generalization, regularization, and computational efficiency compared to isolated single-task models. By jointly optimizing multiple objectives, multi-task frameworks can exploit complementary supervision signals and encourage more robust feature learning. Multi-task DL has been successfully implemented in different medical imaging applications. For instance, Zhou et al. (2019) demonstrated improved breast tumor classification by combining segmentation and classification within a unified framework for ultrasound imaging.

Kaushik et al. (2023) proposed a multi-task network combining segmentation and regression to generate synthetic CT images from MRI for radiation therapy planning, where segmentation was used to localize bone regions of interest and guide accurate density prediction during image translation. Kordnoori et al. (2024) reported enhanced diagnostic performance for brain tumor characterization through joint segmentation and classification of gliomas, meningiomas, and pituitary tumors.

Within neuro-oncology, multi-task approaches have been explored both for segmentation enhancement and for joint segmentation–classification pipelines. Several studies have focused on improving tumor delineation through auxiliary tasks. For instance, Ngo et al. (2020) proposed a multi-task framework for small brain tumor segmentation in MRI, incorporating an auxiliary feature reconstruction task to preserve fine-grained tumor characteristics. Similarly, Huang et al. (2021) introduced a DL multi-task framework based on a V-Net architecture with dual decoders, where a distance transform prediction task regularized mask prediction and led to improved segmentation contour accuracy.

More recent efforts have extended multi-task paradigms towards integrated diagnostic pipelines in glioma research. Li et al. (2023) developed a transformer-based multi-task model for simultaneous glioma segmentation and identification of infiltrated brain regions, demonstrating that shared boundary information from segmentation can enhance classification-related tasks. In a clinically oriented setting, Chakrabarty et al. (2023) proposed a 2.5D hybrid multi-task Convolutional Neural Network (CNN) to jointly localize, segment, and predict IDH mutation status and 1p/19q co-deletion status, integrating imaging features with prior clinical knowledge through feature fusion mechanisms.

In the work of van der Voort et al. (2023), a single 3D CNN capable of jointly performing tumor segmentation and predicting IDH mutation status, 1p/19q co-deletion status, and tumor grade from pre-operative MRI was proposed. Their framework achieved good performance on an independent test set, representing one of the first unified models providing the necessary components for WHO subtype derivation non-invasively. However, UQ was not incorporated, and interactions between tumor segmentation quality, reliability of tumor features prediction, and

task-specific uncertainty were not systematically analyzed.

2.3 Uncertainty Quantification and Decomposition

UQ in DL for Medical Image Analysis encompasses a broad range of approaches, including Bayesian Neural Networks (BNNs), MC sampling methods, DE, Test-time Augmentation (TTA), Evidential Learning (EL), Generative Modelling (GM), and Conformal Prediction (CP) frameworks (Lambert et al., 2024). These methods differ in their ability to model aleatoric and epistemic uncertainty and in their computational requirements, particularly for high-dimensional tasks such as 3D segmentation. The same review by Lambert et al. (2024) highlights that no single method consistently dominates across clinical applications, and emphasizes the importance of task-specific and application-driven evaluation of uncertainty estimates. MCD remained the most used method for UQ among 218 papers analyzed in this review, followed by DE.

Several approaches have been proposed to decompose predictive uncertainty into aleatoric (data-related) and epistemic (model-related) components. In a Bayesian framework, the components are obtained via Predictive Entropy and Mutual Information (MI) (Smith and Gal, 2018; Mukhoti et al., 2023). More recent methods derive uncertainty components directly from network outputs, for example within 3D U-Net architectures (Jones et al., 2022) or single-model diffusion-based frameworks (Chan et al., 2024), demonstrating potential clinical relevance for identifying ambiguous or out-of-distribution cases.

However, the practical value of disentanglement remains contested. Kahl et al. (2024) showed that successful separation in controlled settings for segmentation tasks does not necessarily translate to real-world medical data, and that its downstream benefit strongly depends on the task, dataset properties, and aggregation strategy. Similarly, Mukhoti et al. (2023) emphasized that predictive entropy summarizes total uncertainty and therefore confounds aleatoric and epistemic uncertainty. They showed that predictive entropy can still be effective when one uncertainty source is low or dominant, but may become less informative when ambiguous in-distribution samples and out-of-distribution samples both lead to high predictive entropy. These findings suggest that epistemic–aleatoric decomposition may be useful for interpreting the source of uncertainty, but that its operational benefit should not be assumed and instead needs to be validated within the specific application context.

2.4 Uncertainty Quantification in Multi-Task Settings

While UQ has been extensively studied for single-task segmentation or classification, its role in multi-task learning remains comparatively underexplored. In particular, few works explicitly analyze how uncertainty behaves across interacting tasks or how it propagates within shared representations.

Ruan et al. (2020) proposed Mt-UcGAN, a unified framework for joint tumor segmentation, quantification, and uncertainty estimation in renal CT imaging, incorporating an uncertainty-constrained adversarial mechanism within a multi-task architecture. Similarly, Mehta et al. (2021) investigated the propagation of voxel-level uncertainty across cascaded tasks, demonstrating that uncertainty estimates from one stage can improve performance in downstream segmentation tasks. However, these studies focus primarily on segmentation or cascaded processing pipelines rather than on tightly integrated frameworks that jointly perform different tasks.

To the best of our knowledge, the interaction between tumor segmentation uncertainty and molecular subtype prediction uncertainty has not been systematically investigated in multi-task AI frameworks focused on glioma diagnosis. Although segmentation has been used in multi-task approaches under the theoretical assumption that tumor shape information can guide AI models in learning relevant features for the tumor subtyping, it remains unclear whether uncertainty in the shared segmentation backbone influences reliability in the prediction of relevant tumor features, and whether cross-task aggregation improves operational trust.

3. Methods

3.1 Multi-task Deep Learning Glioma Subtyping and Uncertainty Quantification Framework

As shown in Figure 1, the pipeline operates on four preoperative structural MRI sequences: T1w, T1wCE, T2w, and FLAIR. The end goal is to accurately predict three key glioma characteristics: IDH mutation status (wildtype or mutated), 1p/19q co-deletion status (intact or co-deleted), and tumor grade (grade 2, 3, or 4).

The framework includes a pre-processing module, and an encoder-decoder network (van der Voort et al., 2023; Ronneberger et al., 2015) that performs tumor segmentation as an auxiliary task to extract imaging features at different resolution levels both in the encoder and decoder pathways. These features are concatenated and fed to a classification branch composed of fully connected layers for the prediction of each of the tumor features. Details on the pre-processing and architecture can be found in Appendix A.1 and A.2, respectively.

3.2 Data

In compliance with the expected input of the model, the dataset was composed of different subsets containing the four structural MRI sequences, accompanied by its corre-

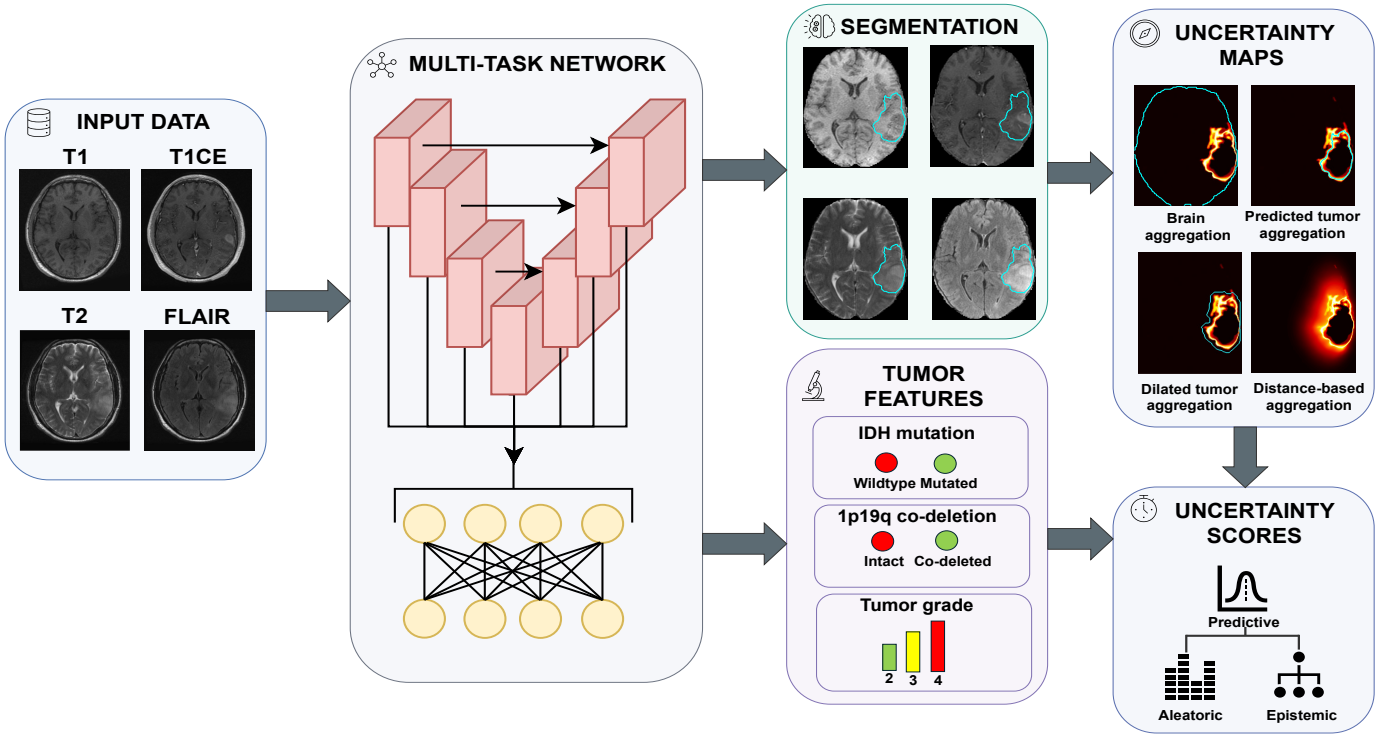

Figure 1: Overview of the multi-task glioma subtyping framework with uncertainty quantification. The input data consist of four structural MRI sequences per patient, namely T1-weighted (T1w), contrast-enhanced T1-weighted (T1wCE), T2-weighted (T2w), and fluid-attenuated inversion recovery (FLAIR). The sequences are preprocessed using registration, bias field correction, and skull stripping. Subsequently, the data are used to train a Deep Learning model that jointly performs tumor segmentation and prediction of IDH mutation status, 1p/19q co-deletion status, and tumor grade. For each task, predictive, aleatoric, and epistemic uncertainty are computed. For tumor segmentation, voxel-wise uncertainty maps are converted into case-level uncertainty scores using four aggregation strategies: brain-level aggregation, predicted-tumor aggregation, dilated-tumor aggregation, and distance-weighted aggregation around the tumor region. In the tumor features panel, the colors of the icons vary from tumors associated with poorer prognosis (red) to those with better prognosis (green).

sponding tumor segmentation and labels (when available), regarding IDH mutation status, 1p/19q co-deletion status and tumor grade.

The train and test sets are comprised of a collection of both in-house and publicly available datasets of patients with adult-type glioma, including BraTS (Bakas et al., 2017, 2018; Menze et al., 2014), Brain tumor Progression (Schmainda and Prah, 2018), CPTAC-GBM (CPTAC, 2018), Erasmus Glioma Database (EGD) (van der Voort et al., 2021), IvyGAP (Shah et al., 2016) and REMBRANDT (Scarpace et al., 2019) for the training set; and TCGA-GBM (Scarpace et al., 2016) and TCGA-LGG (Pedano et al., 2016) for the test set. The distribution of the data is presented in Table 1.

Table 1: Train and test data distribution

Set	Dataset	Number of cases	Total
Train	BraTS	156	1466
	Brain tumor Progression	20	
	CPTAC-GBM	45	
	EGD	775	
	IvyGAP	39	
	In-house	322	
	REMBRANDT	109	
Test	TCGA-GBM	133	236
	TCGA-LGG	103	

Since not all data samples had information regarding molecular features and tumor grade, this was handled with a masked loss during training. Details on this procedure are given in Appendix A.2. Table 10 in Appendix B provides additional information on the data distribution for each of the tumor features, evidencing the highly-imbalanced nature of all the tasks across datasets.

3.3 Uncertainty Quantification

3.3.1 Bayesian Modeling

Given a training dataset consisting of input $\mathcal{X} = \{x_1, \ldots, x_n\}$ and corresponding output $\mathcal{Y} = \{y_1, \ldots, y_n\}$, Bayesian modeling aims to infer parameters $\boldsymbol{\omega}$ of a function $y = f^{\boldsymbol{\omega}}(x)$ that are likely to have generated the observed output.

Before observing any data, an initial belief about plausible parameter values is set. In the Bayesian framework, this belief is encoded through a prior distribution over the parameter space, denoted as $p(\boldsymbol{\omega})$. The prior reflects assumptions about the model parameters before incorporating evidence from the data.

Once data are observed, the prior belief is updated. To perform this update a likelihood function is required. This function specifies the probabilistic mechanism by which outputs are generated given inputs and parameters, thus describing how well a particular parameter configuration explains the observed data. It is defined as:

$$p(\mathcal{Y} \mid \mathcal{X}, \boldsymbol{\omega}). \tag{1}$$

Applying Bayes' theorem, the posterior distribution over the parameters is given by

$$p(\boldsymbol{\omega} \mid \mathcal{X}, \mathcal{Y}) = \frac{p(\mathcal{Y} \mid \mathcal{X}, \boldsymbol{\omega})\, p(\boldsymbol{\omega})}{p(\mathcal{Y} \mid \mathcal{X})}, \tag{2}$$

where $p(\mathcal{Y} \mid \mathcal{X})$ is the marginal likelihood (or evidence), and acts as a normalization constant ensuring that the posterior integrates to one.

The posterior distribution characterizes the updated belief over model parameters after observing the data. Given a new input x^*, predictions are obtained by marginalizing over all possible parameter configurations, weighted by their posterior probability:

$$p(y^* \mid x^*, \mathcal{X}, \mathcal{Y}) = \int p(y^* \mid x^*, \boldsymbol{\omega})\, p(\boldsymbol{\omega} \mid \mathcal{X}, \mathcal{Y})\, d\boldsymbol{\omega}. \tag{3}$$

This integral expresses the Bayesian predictive distribution, which accounts for uncertainty in the model parameters by averaging predictions over the posterior distribution.

3.3.2 Uncertainty Decomposition and MCD

Let N be a Neural Network (NN) with weights $\mathbf{W}$. The theoretical definition of the predictive distribution in Equation 3 is analytically intractable for N given its high-dimensional weight space. MCD (Gal and Ghahramani, 2016) was then proposed as an approximation to Bayesian inference.

In MCD, Bayesian inference is approximated by applying dropout to N with rate δ at train time, and then performing T stochastic forward passes with the same dropout rate enabled at test time. Given some training data $\mathcal{D}$ and a new input x^*, each forward pass t samples a different set of network weights $\mathbf{W}_t$, yielding a collection of predictions $\{p(y^*|x^*, \mathbf{W}_t)\}_{t=1}^T$. The predictive distribution is then approximated by:

$$p(y^*|x^*, \mathcal{D}) \approx \frac{1}{T} \sum_{t=1}^{T} p(y^*|x^*, \mathbf{W}_t). \tag{4}$$

The total predictive uncertainty can be quantified by computing the entropy of this predictive distribution (Smith and Gal, 2018; Depeweg et al., 2018):

$$U_{\text{predictive}} = -\sum_c \left(\frac{1}{T} \sum_{t=1}^{T} p(y^* = c \mid x^*, \mathbf{W}_t) \right) \times \log \left(\frac{1}{T} \sum_{t=1}^{T} p(y^* = c \mid x^*, \mathbf{W}_t) \right), \tag{5}$$

where x^* denotes a test input sample and y^* the corresponding predicted label. $\mathbf{W}_t$ represents the network weights obtained at MC sample t when dropout is applied at test time, and T is the total number of MC forward passes (MC samples hereinafter). The term $p(y^* = c \mid x^*, \mathbf{W}_t)$ denotes the predicted probability for class c produced by the model in the t-th MC sample. The inner average $\frac{1}{T}\sum_{t=1}^{T} p(y^* = c \mid x^*, \mathbf{W}_t)$ approximates the predictive distribution by averaging the class probabilities across all MC samples. Finally, c indexes the set of possible classes.

The predictive uncertainty can be decomposed into two components. The aleatoric uncertainty, representing the inherent noise in the data, is estimated by averaging the entropy of the predictive distribution from all T MC samples (Smith and Gal, 2018; Depeweg et al., 2018):

$$U_{\text{aleatoric}} = \frac{1}{T} \sum_{t=1}^{T} \left(-\sum_c p(y^* = c \mid x^*, \mathbf{W}_t) \times \log p(y^* = c \mid x^*, \mathbf{W}_t) \right), \tag{6}$$

where $p(y^* = c \mid x^*, \mathbf{W}_t)$ represents the predicted probability for class c at inference under weight realization $\mathbf{W}_t$. In this case, the entropy is computed for each MC sample t and subsequently averaged across the T samples.

The epistemic uncertainty describes the uncertainty coming from the model parameters. It is computed as the difference between the predictive and aleatoric components, yielding the MI (Smith and Gal, 2018; Depeweg et al., 2018):

$$U_{\text{epistemic}} = \left(\frac{1}{T} \sum_{t=1}^{T} \sum_c p_{t,c} \log p_{t,c} \right) - \left(\sum_c \bar{p}_c \log \bar{p}_c \right), \tag{7}$$

where $\bar{p}_c$ represents the mean predicted probability for class c across all T MC samples, and $p_{t,c}$ describes the class c probability at MC sample t.

3.4 Evaluation Experiments:

3.4.1 MC Sample Convergence and Decomposition of Uncertainty Estimates

We performed a systematic evaluation of MCD parameters, varying the dropout rate δ and the number of MC samples T and observing the magnitudes for predictive, aleatoric and epistemic uncertainty, which were computed as shown in equations 5, 6 and 7. Specifically, we explored values of δ between 0.2 and 0.6 in increments of 0.05, and values of T from 10 to 100 in steps of 10. These ranges were selected to analyze the effect of both parameters in a wide spectrum, covering both moderate and higher values, and considering theoretical and empirical insights from the literature (Gal and Ghahramani, 2016; Kendall et al., 2015; Srivastava et al., 2014).

We define the dropout rate as the probability of dropping out units, meaning that higher rates reduce the network predictive capacity more aggressively. Too low a dropout rate may lead to underestimation of uncertainty, while excessively high values can degrade predictive performance or lead to underfitting. The upper bound of 0.6 is chosen such that even with aggressive dropout, sufficient model capacity is retained without requiring network width scaling. Likewise, the number of MC samples influences the fidelity of uncertainty estimates: too few may yield noisy approximations, but the more samples, the longer the computation time at inference.

A sufficient number of MC samples was defined as the point at which the average uncertainty estimates showed limited additional change with increasing sampling.

3.4.2 Calibration and Quality Analysis

Model calibration in the classification tasks was assessed using Expected Calibration Error (ECE) and Negative Log-Likelihood (NLL). ECE quantifies the discrepancy between predicted confidence and empirical accuracy. Given predicted labels $\hat{y}_i$ and predicted confidence values $\hat{p}_i$ for a case i from the total number of cases N, predictions are partitioned into B confidence bins. Here, $\hat{p}_i$ denotes the maximum softmax probability assigned to the predicted class $\hat{y}_i$. For bin b, let $\mathrm{acc}(b)$ denote the empirical accuracy and $\mathrm{conf}(b)$ the mean confidence. ECE is computed as:

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{|B_b|}{N} \left|\mathrm{acc}(b) - \mathrm{conf}(b)\right|, \tag{8}$$

where $|B_b|$ is the number of samples in bin b. We used $B = 5$ equally spaced confidence bins across all classification tasks and dropout configurations. This was a practical choice to balance calibration-curve resolution with the reliability of bin-wise accuracy estimates, since using more bins reduces the number of cases per bin and can make ECE more sensitive to small changes in case assignment.

NLL evaluates both correctness and confidence by penalizing low probability assigned to the true class:

$$\mathrm{NLL} = -\frac{1}{N} \sum_{i=1}^{N} \log \hat{p}_{i,y_i}, \tag{9}$$

where $\hat{p}_{i,y_i}$ denotes the predicted probability assigned to the true class y_i for case i.

Importantly, both ECE and NLL were computed exclusively from the softmax output probabilities of the MCD ensemble and therefore reflect calibration of the model probability outputs, rather than calibration of the uncertainty estimates themselves.

Additionally, we computed the Receiver Operating Characteristic Area Under the Curve (AUC) of the MCD ensemble model at each configuration as a general measure of classification performance using the softmax probabilities. In case of the tumor grade task, which consisted of three classes, it was calculated using the one-vs-rest strategy.

For the tumor segmentation task, the quality of uncertainty estimates was assessed by computing the Pearson correlation coefficient ρ between case-level uncertainty scores and the Dice Similarity Coefficient (DSC) between the predicted segmentation and the ground truth for all cases in the test set, under the assumption that informative uncertainty estimates should correlate negatively with tumor segmentation accuracy. To assess whether the observed relationships were robust to non-linear but monotonic trends, we additionally computed Spearman rank correlation ρ_s.

Case-level uncertainty is derived by aggregating entropy-based voxel-wise uncertainty estimates. This aggregation was explored in four different regions. First, full-brain aggregation was computed by averaging uncertainty over the brain mask obtained using HD-BET (Isensee et al., 2019). Second, predicted tumor aggregation was computed by averaging uncertainty within the predicted tumor mask obtained from the MCD ensemble. Third, to include uncertainty immediately adjacent to the predicted tumor boundary, we computed a dilated predicted-tumor aggregation, where the predicted tumor mask was expanded using a 5-voxel binary dilation. Finally, we evaluated a boundary-weighted aggregation in which all brain voxels contributed to the case-level score with weights decreasing as a function of their distance to the predicted tumor boundary. Specifically, for voxel v, the weight was defined as

$$w(v) = \exp\left(-\frac{d(v)^2}{2\sigma^2}\right),$$

where $d(v)$ is the Euclidean distance from voxel v to the predicted tumor boundary for a given case, and $\sigma = 5$ voxels. The boundary-weighted uncertainty score was then computed as

$$U^{\mathrm{bw}} = \frac{\sum_{v\in\Omega^{\mathrm{brain}}} w(v)u_v}{\sum_{v\in\Omega^{\mathrm{brain}}} w(v)},$$

with u_v denoting the uncertainty value at voxel v. This allowed us to compare whole-brain, tumor-local, peri-tumor, and boundary-aware aggregation strategies.

3.4.3 Operational Utility of Uncertainty Estimates

To evaluate the operational utility of uncertainty estimates, we quantified their ability to identify erroneous predictions across tasks and to support selective prediction by retaining the most certain cases. Table 2 provides a summary of the metrics used, their interpretation, and expected ranges indicative of strong performance.

For each classification task, a binary error indicator for case i was defined as:

$$e_i = \begin{cases} 1, & \text{if prediction } i \text{ is incorrect} \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

For the tumor segmentation, the binary error indicator was defined as:

$$e_i^{\mathrm{seg}} = \begin{cases} 1, & \text{if } DSC_i \le \tau \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

Since this binary definition of segmentation error depends on the selected DSC threshold τ, we additionally performed a threshold-sensitivity analysis for $\tau \in \{0.60, 0.70, 0.80\}$. We computed the segmentation error rate to assess whether the operational utility of segmentation uncertainty depended heavily on the value of τ, or whether the conclusions remained consistent across more lenient and stricter definitions of error.

U-AUC: In this case, we measured the Uncertainty-based Receiver Operating Characteristic Area Under the Curve (U-AUC): MCD-derived uncertainty scores U were treated as continuous predictors of error e_i, and the area under this curve was computed. In this setting, this metric evaluates the probability that a randomly chosen erroneous case receives a higher uncertainty score than a randomly chosen correct case:

$$U\text{-}AUC = \mathbb{P}(U_{e=1} > U_{e=0}). \tag{12}$$

While U-AUC is threshold-independent, it may be optimistic when errors are rare. It aims to answer the question: *does uncertainty rank errors above successes overall?*

Average Precision (AP): Given the typically imbalanced nature of medical imaging tasks, we additionally computed the AP, corresponding to the area under the precision-recall curve:

$$\mathrm{AP} = \sum_k (R_k - R_{k-1})P_k, \tag{13}$$

where P_k and R_k denote precision and recall at threshold k.

The baseline AP of a random classifier equals the error rate:

$$\epsilon = \frac{1}{N}\sum_{i=1}^{N} e_i. \tag{14}$$

Thus, AP values above ϵ indicate better-than-random identification of errors. It aims to answer the question: *Is uncertainty useful when errors are rare?*

Error-Rate Lift: To quantify error identification relative to baseline prevalence, we computed Lift:

$$\mathrm{Lift} = \frac{\mathrm{AP}}{\epsilon}. \tag{15}$$

A Lift of 1 indicates performance equivalent to the baseline error rate, as expected under random ranking. A Lift greater than 1 indicates that erroneous cases tend to receive higher uncertainty scores than non-erroneous cases.

Area Under the Risk–Coverage Curve (AURC): To additionally evaluate uncertainty estimates in a selective prediction setting, we computed the AURC (El-Yaniv et al., 2010; Jaeger et al., 2022; Zenk et al., 2025). For our application, we computed an empirical definition obtained by progressively retaining cases from lowest to highest uncertainty. Let U_i denote the uncertainty score for case i, with larger values indicating lower confidence. The N cases were ordered from most certain to least certain, such that $U_{(1)} \le U_{(2)} \le \cdots \le U_{(N)}$. The case-wise risk r_i was defined as:

$$r_i = \begin{cases} e_i, & \text{for classification tasks,} \\ 1 - \mathrm{DSC}_i, & \text{for tumor segmentation,} \end{cases} \tag{16}$$

where $e_i = 1$ for an incorrect classification and $e_i = 0$ otherwise.

For a retained set containing the k most certain cases, coverage was defined as:

$$C_k = \frac{k}{N}, \tag{17}$$

and selective risk was computed as the mean risk among the retained cases:

$$R(C_k) = \frac{1}{k}\sum_{j=1}^{k} r_{(j)}. \tag{18}$$

The AURC was then computed by averaging the selective risk over all retained-set sizes:

$$\mathrm{AURC} = \frac{1}{N}\sum_{k=1}^{N} R(C_k). \tag{19}$$

Table 2: Summary of the metrics used to assess operational utility of uncertainty estimates, namely, Uncertainty-based Receiver Operating Characteristic Area Under the Curve (U-AUC), Average Precision (AP), Lift, and Area Under the Risk–Coverage Curve (AURC). For each metric, the purpose, possible values, and interpretation are presented.

Metric	Purpose	Possible Values	Interpretation
U-AUC	Measures the probability that a randomly chosen error case has higher uncertainty than a correct case.	0 – 1	Values close to 1 indicate uncertainty correctly ranks errors above successes.
AP	Evaluates precision-recall trade-off for predicting errors, suitable for imbalanced tasks.	0 – 1	Values above the baseline error rate (ϵ) indicate better-than-random error identification.
Lift	Quantifies error-identification performance relative to baseline prevalence.	$0 - \frac{1}{\epsilon}$	Values > 1 indicate that uncertainty successfully prioritizes error cases. A value of 1 indicates random selection.
AURC	Evaluates the mean residual risk among retained cases as coverage varies from most certain to all cases.	0 – 1	Lower values indicate that uncertainty successfully retains lower-risk cases at reduced coverage.

This corresponds to a discrete retained-set approximation of the risk–coverage curve area. Lower AURC values indicate that low-uncertainty retained cases have lower residual risk across coverage levels. This metric therefore aims to answer the question: *how much prediction risk remains when retaining only the most certain cases?*

3.4.4 Comparison of MCD with other UQ methods

Although MCD is widely used for uncertainty estimation in DL, relying on a single UQ method may limit the interpretation of a task-aware uncertainty analysis. After selecting the final MCD configuration, we therefore compared it with two ensemble-based strategies, DE and MCDE. These methods were selected because they are compatible with the multi-task architecture used in this study. Ensembles can also be interpreted as Bayesian Model Averaging (He et al., 2020; Wilson and Izmailov, 2020; Mukhoti et al., 2023). Therefore, the same framework for disentangling epistemic and aleatoric uncertainty described in 3.3.2 was applied.

For DE, we trained $M = 5$ independently initialized models using the same architecture, data split, preprocessing, loss functions, and training procedure (including dropout) as in the main experiments. The ensemble members differed only in their fixed random seeds, which affected initialization, mini-batch ordering, and stochastic data augmentation. At inference, dropout was disabled and the final predictive probability vector was obtained by averaging the deterministic probability vectors across ensemble members:

$$\bar{\mathbf{p}}_i^{\mathrm{DE}} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{p}_{i,m},$$

where $\mathbf{p}_{i,m}$ denotes the class-probability vector for case i predicted by ensemble member m.

For MCDE, we combined deep ensembles with MCD by enabling dropout at inference for each independently trained ensemble member. For each model, we used the selected MCD configuration identified in the MCD analyses. The final MCDE predictive probability vector was obtained by averaging predictions across both ensemble members and MC samples:

$$\bar{\mathbf{p}}_i^{\mathrm{MCDE}} = \frac{1}{MT} \sum_{m=1}^{M} \sum_{t=1}^{T} \mathbf{p}_{i,m,t},$$

where $\mathbf{p}_{i,m,t}$ denotes the probability vector for case i from the t-th MC sample of ensemble member m.

This additional comparison was included to provide a broader evaluation of UQ beyond MCD and to assess whether the operational utility of uncertainty estimates was consistent across dropout-based and ensemble-based sampling strategies. It was performed after selecting the final MCD configuration and using the same operational-utility metrics defined in Section 3.4.3: U-AUC, AP, Lift, and AURC. This allowed us to assess whether the operational utility observed for MCD was specific to dropout-based stochastic inference, or whether comparable behavior was also observed for ensemble-based uncertainty estimates.

3.4.5 Interaction Between Tumor Segmentation and Classification

To investigate cross-task dependencies, we analyzed the relationship between tumor segmentation performance and uncertainty, as well as classification correctness.

Tumor Segmentation Quality and Classification Correctness: We used DSC scores as a measure of tumor segmentation quality and compared its values between correctly and incorrectly classified cases for each of the tasks. Differences between groups were evaluated using the Mann–Whitney U test to assess whether segmentation performance systematically differed between classification outcomes.

Uncertainty Correlation Across Tasks: To assess shared uncertainty patterns, we computed Pearson correlation coefficients between tumor segmentation uncertainty U_{seg} and classification uncertainty U_{cls}. This analysis was performed separately for predictive, aleatoric, and epistemic components.

Tumor Segmentation Uncertainty as Predictor of Classification Error: Finally, we evaluated whether segmentation uncertainty alone was able to predict classification errors by computing U-AUC, AP, Lift and AURC using U_{seg} as the predictor and e_{cls} as the outcome. This analysis quantified whether uncertainty in the shared segmentation backbone propagated to prediction reliability of the tumor features.

Composite Trust Score:

To investigate whether incorporating tumor segmentation uncertainty improved the identification of classification errors beyond classification uncertainty alone, we defined a composite trust score that integrated uncertainty from both tasks. The trust score for task t and case i was defined as:

$$\text{Trust}_{i,t} = 1 - \left(\alpha_t \tilde{U}_{i,t} + (1-\alpha_t)\,\tilde{U}_{i,\text{seg}}\right), \qquad (20)$$

where $\tilde{U}_{i,t}$ and $\tilde{U}_{i,\text{seg}}$ denote the normalized predictive uncertainty for the classification and the segmentation task, respectively, and $\alpha_t \in [0,1]$ represents the weighting factor for task t.

The weight α_t was determined based on the relative ability of each uncertainty source to identify classification errors, quantified using Lift. Specifically, let Lift_t denote the Lift obtained when using classification uncertainty to predict classification errors on task t, and $\text{Lift}_{\text{seg}\to t}$ the Lift obtained when using tumor segmentation uncertainty for the same purpose. The task-specific weight was defined as:

$$\alpha_t = \frac{\text{Lift}_t}{\text{Lift}_t + \text{Lift}_{\text{seg}\to t}}. \qquad (21)$$

This formulation assigns greater weight to the uncertainty source with greater error-identification performance relative to baseline prevalence. Intuitively, if classification uncertainty is highly predictive of errors, α_t approaches 1 and the trust score relies primarily on classification uncertainty. Conversely, if segmentation uncertainty provides additional predictive value, its contribution increases accordingly.

The subtraction from 1 ensures that higher trust scores correspond to more reliable predictions, facilitating interpretation and enabling direct comparison with confidence-based referral strategies. The proposed trust score was evaluated using selective prediction analysis to determine whether combining uncertainty sources improved error identification compared to classification uncertainty alone.

4. Results

4.1 MC-Sample Convergence and Decomposition of Uncertainty Estimates

As shown in Figure 2, the average uncertainty values reached a plateau across tasks once the number of MC samples was approximately 20–30, depending on the task. Increasing dropout rates generally led to proportional increases in predictive, aleatoric and epistemic uncertainties.

For all the tumor features prediction tasks, namely, IDH mutation status, 1p/19q co-deletion status, and tumor grade, the aleatoric uncertainty component was the main contributor to the predictive uncertainty. Conversely, epistemic uncertainty dominated in the tumor segmentation task.

4.2 Calibration and Quality Analysis

Based on the convergence patterns in Figure 2, we selected $T = 30$ for the remaining analyses, as this was the point at which average uncertainty values had generally plateaued across tasks. Figure 3 presents the ECE, NLL and AUC as a function of dropout rate δ for the classification tasks. For the IDH mutation status prediction, both ECE and NLL reached their lowest values at dropout 0.25, with subtle fluctuations between 0.2 and 0.4 and then a steady increase thereafter. AUC was highest at 0.2 and generally declined as dropout increased. In the case of 1p/19q co-deletion status prediction, ECE and NLL were minimized at 0.2 and generally rose with higher dropout rates. AUC followed a similar decreasing pattern. For the tumor grade prediction, ECE was lowest at 0.25 and gradually increased with dropout, while NLL rose consistently. AUC also showed a decreasing trend for dropout rates higher than 0.35.

Although ECE, NLL, and AUC showed broadly related trends, the curves were not identical across dropout rates. For IDH mutation status prediction, ECE showed a local decrease between dropout rates 0.35 and 0.40, whereas NLL did not show a corresponding decrease and AUC remained below its value at lower dropout rates. For 1p/19q co-deletion status prediction, ECE and NLL both increased after dropout 0.20, but ECE showed smaller local fluctuations at higher dropout rates while NLL and AUC indicated clearer degradation. For tumor grade prediction, the difference between the metrics was most apparent: ECE showed local decreases at intermediate dropout values, particularly around dropout rates 0.25 and 0.45, whereas NLL increased more steadily and AUC decreased at higher dropout rates.

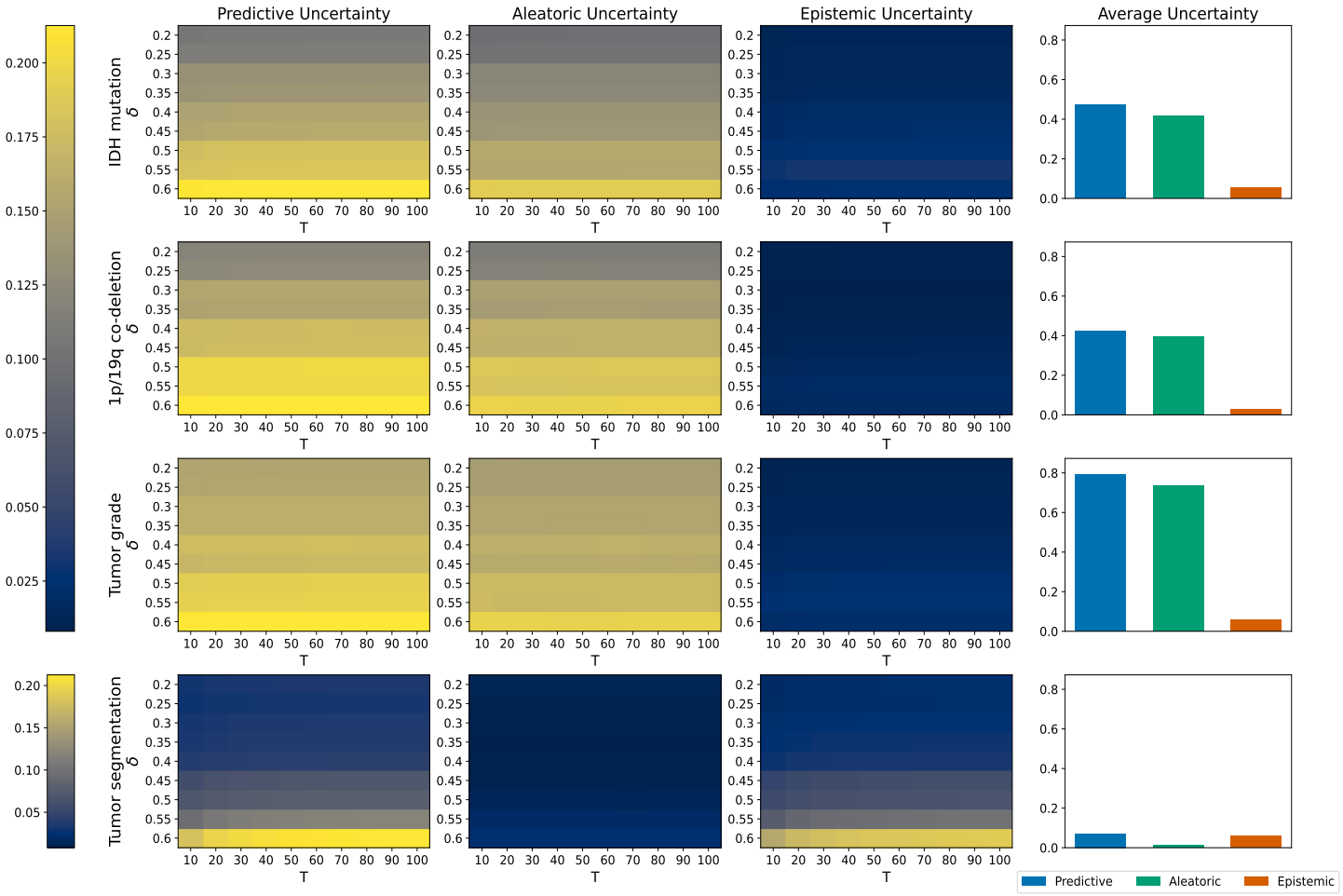

Figure 2: First three columns from left to right: predictive, aleatoric and epistemic uncertainty value heatmaps across all tasks as a function of the number of MCD samples T and the dropout rate δ. The value at each cell of the heatmap represents the mean uncertainty across all cases of the test set. The last column presents the average value for each type of uncertainty over all experiments with the blue, green and red bars representing predictive, aleatoric and epistemic uncertainty, respectively. Uncertainty values for classification tasks were normalized at the same scale to ease the analysis of the patterns. For the segmentation task the heatmaps were kept in the original scale.

Figure 4 presents the uncertainty quality analysis for the tumor segmentation task. Brain mask-based aggregation showed weak correlations with DSC across dropout rates, with values fluctuating around zero and no consistent negative association. In contrast, predicted tumor aggregation showed the strongest and most stable negative correlations across the evaluated dropout range. Dilated tumor aggregation followed a similar trend at low and moderate dropout rates, but the correlation became progressively weaker at higher dropout values. Boundary-weighted aggregation also showed a negative association with DSC at low and moderate dropout rates, although this association weakened more markedly as dropout increased. The lower panel shows that the mean segmentation DSC was highest at moderate dropout rates, peaking at 0.25. It generally decreased with stronger dropout. Based on its consistently strong negative association with DSC and its lack of additional aggregation hyperparameters, predicted tumor aggregation was selected for the downstream segmentation uncertainty analyses.

Our evaluation across tasks and metrics indicated that the MCD model with a dropout rate of 0.25 provided a favorable trade-off between predictive performance and uncertainty calibration, and was therefore selected for the remaining evaluation.

To further assess whether the reported Pearson correlations reflected the case-wise uncertainty and performance relationship, we visualized predictive uncertainty against tumor segmentation DSC at the selected configuration ($T = 30$, $\delta = 0.25$). The corresponding scatter plots are provided in Appendix D. For predicted tumor and dilated tumor aggregation, the scatter plots showed a clear decreasing trend, consistent with the strong negative Pearson and Spearman correlations. Boundary-weighted aggregation also showed a negative monotonic trend, although

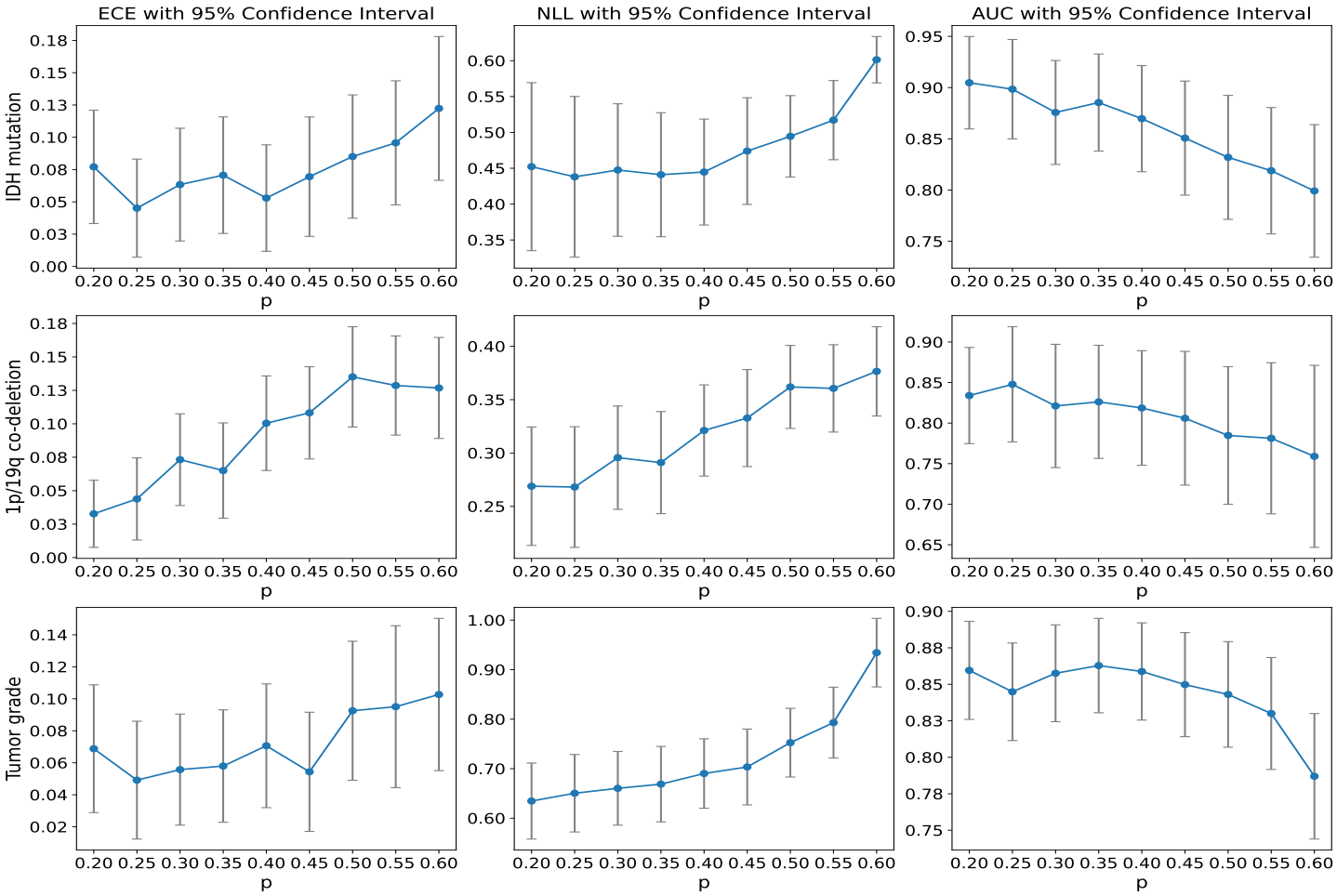

Figure 3: Expected Calibration Error (ECE), Negative Log-Likelihood (NLL) and Receiver Operating Characteristic Area Under the Curve (AUC) values as a function of dropout rate δ. For each point in every graph, a 95% confidence interval is presented. This interval was obtained by $1000\times$ bootstrap resampling of the test set.

with a weaker linear association. In contrast, brain mask aggregation showed no clear relationship between predictive uncertainty and DSC. Overall, Pearson and Spearman correlation values were similar across the aggregation strategies, indicating that the main findings were not dependent on the use of a strictly linear association measure.

Figure 5 further highlights the ability of uncertainty scores to discriminate between high- and low-quality predictions, with low-quality predictions generally exhibiting higher uncertainty across tasks. The difference between the distributions for all tasks was statistically significant according to the Mann–Whitney U test.

In order to contextualize the uncertainty-based error detection analyses, we additionally reported the predictive performance of the selected MCD configuration used throughout the remaining experiments ($T = 30$, $\delta = 0.25$). Table 3 reports the task-specific class distribution in the test set, along with different performance metrics, including accuracy, balanced accuracy, macro-averaged F1-score and AUC. The class distributions highlight substantial imbalance, particularly for 1p/19q co-deletion status and tumor grade prediction tasks.

4.3 Operational Utility of Uncertainty Estimates

For tumor segmentation, the primary error definition used $\tau_{\text{DSC}} = 0.70$. A sensitivity analysis using different values of $\tau_{\text{DSC}} \in \{0.60, 0.70, 0.80\}$ is reported in Appendix C.

Table 4 summarizes the ability of predictive, aleatoric, and epistemic uncertainty estimates to identify erroneous predictions across all tasks, evaluated using U-AUC, AP, Lift, and AURC. For each metric, $1000\times$ bootstrap confidence intervals are reported in the table to indicate the variability of the estimates. The observed error rates of 0.18, 0.10, 0.29, and 0.17 for IDH mutation status, 1p/19q co-deletion status, tumor grade and tumor segmentation, respectively, correspond to the baseline AP for a random classifier.

For IDH mutation status prediction, predictive uncertainty achieved an AP of 0.42, corresponding to a Lift of 2.37 relative to the baseline error rate. Aleatoric and

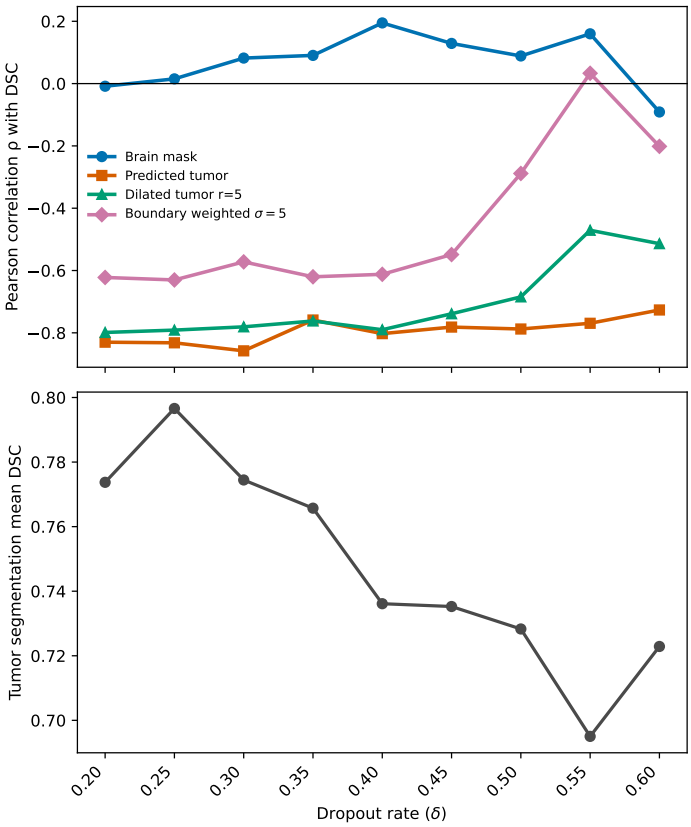

Figure 4: Tumor segmentation uncertainty quality analysis. The top panel presents the Pearson correlation coefficient between case-level predictive uncertainty and Dice Similarity Coefficient (DSC) as a function of the dropout rate δ used for the MCD ensemble. Correlations are shown for four spatial aggregation strategies: brain-mask aggregation, predicted-tumor aggregation, 5-voxel dilated predicted-tumor aggregation, and boundary-weighted aggregation using $\sigma = 5$ voxels. More negative correlations indicate that higher case-level uncertainty is associated with lower segmentation accuracy. The bottom panel presents the mean DSC between the predicted tumor segmentation and the ground truth in the test set as a function of the dropout rate δ.

epistemic uncertainty yielded lower AP values of 0.32 and 0.34, corresponding to Lifts of 1.79 and 1.89, respectively. AURC values were similar across uncertainty components, with all three components achieving an AURC of 0.10. Despite aleatoric uncertainty being dominant in magnitude (as shown in Figure 2), predictive uncertainty provided the numerically highest Lift.

For 1p/19q co-deletion status prediction, predictive and aleatoric uncertainty achieved AP values of 0.37 and 0.39, corresponding to Lifts of 3.55 and 3.70, respectively, while epistemic uncertainty yielded a lower AP of 0.16 (Lift 1.54). Predictive and aleatoric uncertainty also achieved lower

Table 3: Predictive performance at the selected MCD configuration ($T = 30$, $\delta = 0.25$) for the three classification tasks. For each task, the class distribution is also reported. Values are shown with 95% confidence intervals obtained by $1000\times$ bootstrap resampling of the test set. AUC is reported at task level: for tumor grade, it corresponds to the macro-averaged one-vs-rest AUC across grade classes. Total test set size differs across tasks because molecular and histopathological annotations were not available for all patients. Arrows in metric column headers indicate that higher values are preferable (↑).

Task	Class distribution	Accuracy ↑	Balanced accuracy ↑	Macro F1 score ↑	AUC ↑
IDH	Wildtype: 129 Mutated: 85	0.82 [0.77,0.87]	0.79 [0.73,0.84]	0.80 [0.74,0.85]	0.90 [0.85,0.94]
1p/19q	Intact: 205 Co-deleted: 25	0.90 [0.85,0.93]	0.56 [0.50,0.63]	0.57 [0.47,0.67]	0.85 [0.77,0.91]
Grade	Grade 2: 45 Grade 3: 58 Grade 4: 132	0.71 [0.64,0.77]	0.60 [0.57,0.64]	0.51 [0.47,0.54]	0.85 [0.81,0.88]

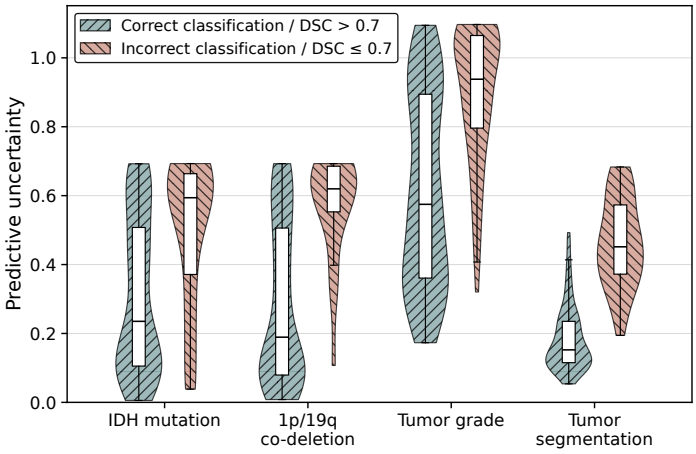

Figure 5: Violin plots of the predictive uncertainty across different tasks, comparing the distributions of high quality and low quality predictions, obtained with the optimal configuration of the MCD model at $T = 30$ and $\delta = 0.25$. For the classification tasks, high and low quality predictions correspond to correctly and incorrectly classified cases, respectively. For the tumor segmentation task, high and low quality predictions represent cases with $DSC > 0.7$ and $DSC \leq 0.7$, respectively. For all tasks, the difference between the distributions for both groups was statistically significant (Mann–Whitney U test).

AURC values of 0.03, compared with 0.05 for epistemic uncertainty. In this task, the dominant aleatoric component also achieved the numerically highest Lift.

For tumor grade prediction, predictive and aleatoric uncertainty achieved identical AP values of 0.59, corre-

sponding to Lifts of 2.00, whereas epistemic uncertainty achieved a lower AP of 0.45 (Lift 1.52). Similarly, predictive and aleatoric uncertainty achieved lower AURC values of 0.12, compared with 0.17 for epistemic uncertainty.

For tumor segmentation, aleatoric uncertainty achieved the highest AP of 0.83 (Lift 5.01), followed closely by predictive uncertainty with an AP of 0.82 (Lift 4.95) and epistemic uncertainty with an AP of 0.81 (Lift 4.90). All three uncertainty components achieved high U-AUC values, ranging from 0.95 to 0.96, and similar AURC values of 0.13. Although epistemic uncertainty was dominant in magnitude (as shown in Figure 2), aleatoric and predictive uncertainty still achieved comparable or higher Lift.

Overall, while all uncertainty types identified errors better than expected from the baseline error rates, the component with the highest magnitude (as shown in Figure 2) did not consistently correspond to the highest Lift or lowest AURC. For IDH mutation status prediction, predictive uncertainty had the highest point estimates for AP and Lift, but the corresponding confidence intervals overlapped with those of aleatoric and epistemic uncertainty. For 1p/19q co-deletion status prediction and tumor grade prediction, predictive and aleatoric uncertainty showed very similar performance, whereas epistemic uncertainty generally showed lower AP, lower Lift, and higher AURC. For tumor segmentation, all three uncertainty components showed comparable performance, with U-AUC values between 0.95 and 0.96, AP values between 0.81 and 0.83, Lift values between 4.90 and 5.01, and identical AURC values of 0.13. Therefore, it can be noted that predictive uncertainty achieved performance comparable to that of individual decomposed components across all tasks given the high overlap between the confidence intervals.

4.4 Comparison of MCD with other UQ methods

After selecting the final MCD configuration, we compared MCD with two ensemble-based UQ methods, DE and MCDE, using the same operational-utility framework. Table 5 reports the performance of predictive uncertainty for error detection and selective prediction across the three classification tasks and tumor segmentation using predicted-tumor aggregation.

For the classification tasks, all three methods produced uncertainty estimates with Lift values above 1, indicating better-than-baseline error identification. For IDH mutation status prediction, MCDE had the lowest AURC, while MCD achieved the highest AP and Lift. For 1p/19q co-deletion status prediction, DE and MCD showed the same U-AUC values, with DE achieving the highest AP and Lift. For tumor grade prediction, DE achieved the highest U-AUC, AP, and Lift, whereas MCDE achieved the lowest error rate and AURC.

For tumor segmentation, predictive uncertainty was aggregated within the predicted tumor region. All three methods achieved high U-AUC values, indicating that predicted-tumor uncertainty was informative for identifying low-quality segmentations. MCDE achieved the highest U-AUC and Lift, whereas MCD achieved the highest AP. The lowest AURC values were observed for DE and MCDE.

Overall, ensemble-based uncertainty estimates provided operational utility comparable to MCD, but no method consistently dominated across all tasks and metrics. In particular, MCDE tended to provide favorable AURC values, suggesting slightly improved selective-risk behavior, whereas DE or MCD more often achieved higher AP or Lift. Because confidence intervals overlapped in several comparisons, these differences were interpreted as task- and metric-dependent trends rather than as evidence of consistent superiority of one uncertainty method. Therefore, the subsequent analyses were based on MCD.

The corresponding results for aleatoric and epistemic uncertainty are reported in Appendix E, Table 11. For aleatoric uncertainty, the trends were broadly consistent with the predictive-uncertainty analysis: all three methods showed operational utility above the baseline error rate across the classification tasks, and tumor segmentation uncertainty aggregated within the predicted tumor region remained highly informative. For epistemic uncertainty, classification error-detection performance was weaker and less consistent than for predictive or aleatoric uncertainty. This was most evident for 1p/19q co-deletion status prediction, where epistemic uncertainty showed lower U-AUC, AP, and Lift than the corresponding predictive and aleatoric uncertainty estimates. Across the classification tasks, MCD-based epistemic uncertainty tended to retain the most consistent operational utility, whereas the ensemble-based epistemic estimates showed more metric-dependent behavior.

4.5 Interaction Between Tumor Segmentation and Classification

Figure 6 shows the distribution of tumor segmentation DSC scores stratified by classification correctness for each classification task. For IDH mutation status, correctly classified cases exhibited higher segmentation performance compared to misclassified cases, and this difference was statistically significant (Mann–Whitney U test). For 1p/19q co-deletion status, correctly classified cases tended to present slightly higher DSC scores; however, the difference did not reach statistical significance. For tumor grade, correctly classified cases again showed higher DSC scores, with a statistically significant difference between groups (Mann–Whitney U test).

Table 6 presents the correlations between tumor segmentation uncertainty and classification uncertainty. Overall,

Table 4: Operational utility of uncertainty estimates per task. Assessment is made using Uncertainty-based Receiver Operating Characteristic Area Under the Curve (U-AUC), Average Precision (AP), Lift, and Area Under the Risk–Coverage Curve (AURC), derived from using uncertainty as a predictor of errors. Values are reported with 95% confidence intervals obtained by $1000\times$ bootstrap resampling of the test set. The numerically highest values for U-AUC, AP, and Lift, and the numerically lowest values for AURC, are indicated in bold for each task. Arrows in metric column headers indicate whether higher (↑) or lower (↓) values are preferable.

Task	Error rate ↓	Uncertainty											
		Predictive				Aleatoric				Epistemic			
		U-AUC ↑	AP ↑	Lift ↑	AURC ↓	U-AUC ↑	AP ↑	Lift ↑	AURC ↓	U-AUC ↑	AP ↑	Lift ↑	AURC ↓
IDH mutation status prediction	0.18	**0.73** [0.62,0.82]	**0.42** [0.28,0.57]	**2.37** [1.72,3.34]	**0.10** [0.05,0.15]	0.71 [0.61,0.80]	0.32 [0.22,0.46]	1.79 [1.39,2.54]	**0.10** [0.05,0.15]	0.71 [0.61,0.80]	0.34 [0.23,0.50]	1.89 [1.45,2.80]	**0.10** [0.05,0.16]
1p/19q co-deletion status prediction	0.10	**0.84** [0.76,0.91]	0.37 [0.22,0.58]	3.55 [2.37,6.07]	**0.03** [0.01,0.05]	**0.84** [0.77,0.91]	**0.39** [0.22,0.58]	**3.70** [2.51,6.19]	**0.03** [0.01,0.04]	0.68 [0.58,0.77]	0.16 [0.10,0.29]	1.54 [1.22,2.59]	0.05 [0.03,0.08]
Tumor grade prediction	0.29	0.78 [0.72,0.84]	**0.59** [0.47,0.71]	**2.00** [1.66,2.47]	**0.12** [0.09,0.17]	**0.79** [0.73,0.85]	**0.59** [0.47,0.72]	**2.00** [1.68,2.50]	**0.12** [0.09,0.17]	0.70 [0.62,0.77]	0.45 [0.35,0.58]	1.52 [1.28,1.94]	0.17 [0.12,0.24]
Tumor segmentation	0.17	0.95 [0.92,0.98]	0.82 [0.72,0.91]	4.95 [3.89,6.67]	**0.13** [0.12,0.14]	**0.96** [0.94,0.98]	**0.83** [0.72,0.91]	**5.01** [3.93,6.80]	**0.13** [0.12,0.14]	0.95 [0.92,0.98]	0.81 [0.71,0.90]	4.90 [3.85,6.64]	**0.13** [0.12,0.14]

Table 5: Comparison of uncertainty quantification methods using predictive uncertainty. Assessment is made using Uncertainty-based Receiver Operating Characteristic Area Under the Curve (U-AUC), Average Precision (AP), Lift, and Area Under the Risk–Coverage Curve (AURC), derived from using predictive uncertainty as a predictor of errors. Values are reported with 95% confidence intervals obtained by $1000\times$ bootstrap resampling of the test set. For tumor segmentation, predictive uncertainty was aggregated within the predicted tumor region. The numerically highest values for U-AUC, AP, and Lift, and the numerically lowest values for AURC, are indicated in bold for each task. Arrows in metric column headers indicate whether higher (↑) or lower (↓) values are preferable.

Task	Method	Error ↓	U-AUC ↑	AP ↑	Lift ↑	AURC ↓
IDH mutation	DE	**0.17**	0.72 [0.62,0.82]	0.37 [0.24,0.56]	2.15 [1.57,3.23]	0.10 [0.05,0.15]
	MCD	0.18	0.73 [0.62,0.82]	**0.42** [0.28,0.57]	**2.37** [1.72,3.34]	0.10 [0.05,0.15]
	MCDE	**0.17**	**0.73** [0.63,0.83]	0.34 [0.23,0.50]	2.03 [1.53,2.98]	**0.09** [0.05,0.14]
1p/19q co-deletion	DE	0.10	**0.84** [0.76,0.91]	**0.39** [0.22,0.59]	**3.90** [2.53,6.62]	**0.03** [0.01,0.04]
	MCD	0.10	**0.84** [0.76,0.91]	0.37 [0.22,0.58]	3.55 [2.37,6.07]	**0.03** [0.01,0.05]
	MCDE	**0.09**	0.81 [0.74,0.89]	0.26 [0.16,0.45]	2.84 [2.06,4.91]	**0.03** [0.01,0.04]
Tumor grade	DE	0.30	**0.80** [0.73,0.86]	**0.61** [0.50,0.73]	**2.04** [1.70,2.53]	0.13 [0.09,0.18]
	MCD	0.29	0.78 [0.72,0.84]	0.59 [0.47,0.71]	2.00 [1.66,2.47]	0.12 [0.09,0.17]
	MCDE	**0.28**	0.78 [0.72,0.84]	0.51 [0.40,0.65]	1.82 [1.53,2.30]	**0.12** [0.08,0.17]
Tumor segmentation	DE	0.13	0.94 [0.90,0.96]	0.65 [0.48,0.82]	4.91 [3.70,7.30]	**0.12** [0.11,0.13]
	MCD	0.17	0.95 [0.92,0.98]	**0.82** [0.72,0.91]	4.95 [3.89,6.67]	0.13 [0.12,0.14]
	MCDE	**0.12**	**0.96** [0.92,0.98]	0.79 [0.65,0.89]	**6.37** [4.82,9.16]	**0.12** [0.12,0.13]

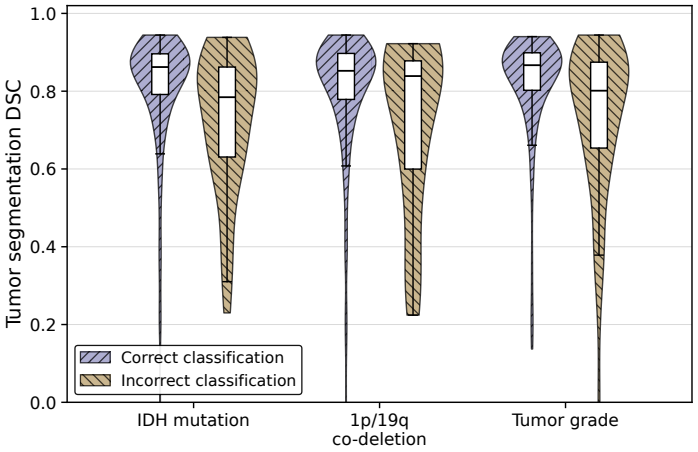

Figure 6: Distribution of tumor segmentation Dice Similarity Coefficient (DSC) scores for correctly and incorrectly classified cases for the IDH mutation status, 1p/19q co-deletion status and tumor grade. Cases are stratified according to whether the corresponding prediction was correct or incorrect. For each task, higher DSC values indicate better agreement between predicted and reference tumor segmentations. Statistical significance between groups was found for the IDH mutation status and tumor grade prediction (Mann–Whitney U test).

predictive and aleatoric uncertainty components exhibited positive correlations across the classification tasks, indicating that cases with higher tumor segmentation uncertainty also tended to present higher classification uncertainty. These positive associations were supported by confidence intervals that remained above zero for the predictive and aleatoric components across all three classification tasks.

For IDH mutation status prediction, tumor segmentation uncertainty showed moderate correlations with predictive and aleatoric classification uncertainty, with correlation coefficients of 0.48 and 0.51, respectively. The association with epistemic uncertainty was weaker, with a correlation coefficient of 0.09, and its confidence interval included zero, indicating no clear positive association for this component.

For 1p/19q co-deletion status prediction, correlations were weaker overall. Predictive and aleatoric uncertainty

Table 6: Pearson correlation (ρ) between tumor segmentation uncertainty and classification uncertainty for each task. Correlations were computed separately for predictive, aleatoric, and epistemic uncertainty components. Values are reported with 95% confidence intervals obtained by $1000\times$ bootstrap resampling of the test set. The numerically highest correlation coefficient per task is indicated in bold.

Task	Uncertainty		
	Predictive	Aleatoric	Epistemic
IDH mutation status prediction	0.48 [0.38, 0.57]	$\mathbf{0.51}$ [0.42, 0.58]	0.09 [-0.04, 0.21]
1p/19q co-deletion status prediction	0.22 [0.12, 0.33]	$\mathbf{0.24}$ [0.15, 0.33]	-0.15 [-0.25, -0.05]
Tumor grade prediction	$\mathbf{0.53}$ [0.45, 0.61]	0.50 [0.42, 0.57]	0.10 [-0.03, 0.26]

showed modest positive associations, with correlation coefficients of 0.22 and 0.24, respectively. In contrast, epistemic uncertainty showed a weak negative correlation of -0.15, and its confidence interval remained below zero. This indicates that, for 1p/19q co-deletion status prediction, higher tumor segmentation epistemic uncertainty was not associated with higher classification epistemic uncertainty.

For tumor grade prediction, tumor segmentation uncertainty was again moderately correlated with predictive and aleatoric classification uncertainty, with correlation coefficients of 0.53 and 0.50, respectively. The association with epistemic uncertainty remained weak, with a correlation coefficient of 0.10, and its confidence interval included zero. Overall, predictive and aleatoric uncertainty demonstrated consistent positive cross-task associations, whereas epistemic uncertainty showed weaker and less consistent relationships across tasks.

Table 7 summarizes the ability of tumor segmentation uncertainty to identify classification errors. Overall, tumor segmentation uncertainty provided moderate discrimination of classification errors, with U-AUC values ranging from 0.61 to 0.69. In terms of AP, predictive segmentation uncertainty achieved values of 0.30, 0.23, and 0.44 for IDH, 1p/19q, and tumor grade prediction, corresponding to Lifts of 1.68, 2.16, and 1.52, respectively. Similar Lift values were observed for aleatoric and epistemic segmentation uncertainty, ranging from 1.45 to 2.55 depending on the task and component. For all tasks and uncertainty components, the lower bounds of the Lift confidence intervals remained above 1, indicating that high segmentation uncertainty identified subsets of cases with higher classification error rates than expected under random selection. AURC values ranged from 0.06 to 0.22, with lower values indicating lower residual risk among retained cases.

Across tasks, no single segmentation uncertainty component was consistently best across all metrics. Aleatoric segmentation uncertainty achieved the highest U-AUC in all three classification tasks. However, the highest AP and Lift were obtained by epistemic uncertainty for IDH mutation status prediction, aleatoric uncertainty for 1p/19q co-deletion status prediction, and epistemic uncertainty for tumor grade prediction. In each task, the confidence intervals of the segmentation uncertainty components overlapped for U-AUC, AP, and Lift.

Compared to classification uncertainty (Table 4), tumor segmentation uncertainty showed lower error-detection and selective-prediction performance when comparing the predictive uncertainty scores directly. For IDH mutation status prediction, predictive segmentation uncertainty achieved a Lift of 1.68 and an AURC of 0.15, whereas predictive classification uncertainty achieved a Lift of 2.37 and an AURC of 0.10. For 1p/19q co-deletion status prediction, predictive segmentation uncertainty achieved a Lift of 2.16 and an AURC of 0.06, compared with a Lift of 3.55 and an AURC of 0.03 for predictive classification uncertainty. For tumor grade prediction, predictive segmentation uncertainty achieved a Lift of 1.52 and an AURC of 0.22, compared with a Lift of 2.00 and an AURC of 0.12 for predictive classification uncertainty. Thus, for all three classification tasks, predictive classification uncertainty achieved higher Lift and lower AURC than predictive tumor segmentation uncertainty.

The same pattern was observed for aleatoric uncertainty, where classification uncertainty achieved higher Lift and lower AURC than tumor segmentation uncertainty for all three classification tasks. For epistemic uncertainty, the comparison was less uniform for Lift: classification uncertainty achieved a higher Lift for IDH mutation status prediction, whereas tumor segmentation uncertainty had a higher Lift for 1p/19q co-deletion status and tumor grade prediction. However, classification epistemic uncertainty still achieved lower AURC than tumor segmentation epistemic uncertainty for all three tasks. Overall, tumor segmentation uncertainty was informative for identifying classification errors, but classification-specific uncertainty remained the more direct and generally stronger error predictor.

The proposed trust score demonstrated consistent ability to identify classification errors across tasks (Table 8). For IDH mutation status prediction, the trust score achieved a U-AUC of 0.73 and an AP of 0.38, corresponding to a Lift of 2.12 relative to the baseline error rate of 0.18, with an AURC of 0.09. For 1p/19q co-deletion status prediction, the trust score yielded a U-AUC of 0.85 and an AP of 0.42, representing a Lift of 3.98 over the baseline error rate of 0.10, with an AURC of 0.02. For tumor grade prediction, the trust score achieved a U-AUC of 0.77 and an AP of 0.51, corresponding to a Lift of 1.77 relative to the baseline error rate of 0.29, with an AURC of 0.13. For all three tasks, the lower bounds of the Lift confidence intervals remained

above 1, indicating that low-trust cases contained more classification errors than expected under random selection.

Compared with predictive classification uncertainty alone (Table 4), the trust score did not show a consistent improvement across tasks. For IDH mutation status prediction, the trust score achieved the same U-AUC as predictive classification uncertainty, lower AP and Lift, and slightly lower AURC. For 1p/19q co-deletion status prediction, the trust score achieved higher point estimates for U-AUC, AP, and Lift, and a lower AURC than predictive classification uncertainty alone. In contrast, for tumor grade prediction, the trust score achieved lower U-AUC, AP, and Lift, and higher AURC than predictive classification uncertainty alone. The confidence intervals for the trust score overlapped with those of predictive classification uncertainty across these comparisons. Therefore, although the trust score slightly improved the point estimates for 1p/19q co-deletion status prediction, the results do not show a consistent improvement over predictive classification uncertainty alone across all classification tasks.

5. Discussion

UQ is increasingly recognized as a key requirement for trustworthy artificial intelligence in medical imaging, yet its practical value remains highly dependent on the task and application context. In this work, we performed a task-aware evaluation of uncertainty within a multi-task DL framework for glioma diagnosis, analyzing its MC sample convergence, magnitude and decomposition, calibration, and operational utility, as well as its interactions between tumor segmentation and tumor molecular subtyping features. Collectively, our results provide insights into both the capabilities and limitations of uncertainty estimates in this clinical application.

We first assessed the convergence of average uncertainty estimates as a function of the number of MC samples and dropout rate. Across tasks, mean uncertainty values generally reached a plateau after approximately 20–30 MC samples, indicating that further sampling produced limited changes in average uncertainty magnitude. This supports the use of a computationally practical number of MC samples for the subsequent analyses.

Increasing the dropout rate generally led to higher uncertainty magnitudes, reflecting broader posterior approximations and increased predictive variability. However, excessive dropout resulted in reduced predictive performance and degraded uncertainty quality, suggesting that overly strong stochastic perturbations can degrade predictive outputs and reduce the practical value of uncertainty estimates. This highlights the importance of balancing stochasticity and predictive fidelity when computing uncertainty through MCD.

The analysis of uncertainty decomposition revealed consistent task-dependent patterns. Aleatoric uncertainty dominated in classification tasks, whereas epistemic uncertainty was more prominent in tumor segmentation, which likely reflects fundamental differences in the nature of the tasks. Molecular subtype prediction from MRI is inherently limited by the indirect relationship between imaging appearance and genetic status, introducing irreducible uncertainty possibly due to biological variability. In contrast, tumor segmentation is more directly informed by image appearance, tumor boundaries, and local texture patterns. Therefore, higher epistemic uncertainty in segmentation may reflect model-related limitations in representing atypical morphologies, ambiguous boundaries, or underrepresented imaging patterns in MRI.

Importantly, however, uncertainty magnitude alone did not directly indicate operational utility. Components that dominated in magnitude did not necessarily perform best in the evaluation metrics. Decomposing predictive uncertainty into aleatoric and epistemic components did not consistently improve error detection performance either. This finding is consistent with the mathematical definition of the decomposition used in this work: predictive uncertainty can be high because of aleatoric uncertainty, epistemic uncertainty, or both. Therefore, when one uncertainty source dominates, predictive uncertainty may already provide a strong operational signal for error detection, even if it does not distinguish why the model is uncertain. The bootstrap confidence intervals supported this interpretation, showing that the operational utility of predictive uncertainty was comparable to that of its components across tasks.

These findings are in line with a recent work by Kahl et al. (2024), which questions the practical utility of uncertainty decomposition in real-world settings by demonstrating that uncertainty decomposition in segmentation works in synthetic settings but does not necessarily translate to real-world data, and therefore the benefit or feasibility of uncertainty decomposition cannot be assumed a priori. Instead, its value must be evaluated empirically for each specific task and dataset. Our results extend these observations to a multi-task setting including three classification tasks and demonstrate that uncertainty decomposition provides insight into the origin of predictive uncertainty but its operational utility is task-limited.

Calibration and quality analysis further demonstrated that the quality of uncertainty estimates was dependent on the amount of dropout applied. Moderate dropout levels achieved the best performance, while excessive dropout degraded calibration and predictive likelihood. This likely reflects the trade-off between model confidence and predictive accuracy: insufficient stochasticity leads to overconfident predictions, whereas excessive stochasticity reduces predictive consistency. These results highlight that uncertainty

Table 7: Ability of tumor segmentation uncertainty to predict classification errors. Assessment is performed using Uncertainty-based Receiver Operating Characteristic Area Under the Curve (U-AUC), Average Precision (AP), Lift, and Area Under the Risk–Coverage Curve (AURC). Values are reported with 95% confidence intervals obtained by $1000\times$ bootstrap resampling of the test set. The numerically highest values for U-AUC, AP, and Lift, and the numerically lowest values for AURC, are indicated in bold for each task. Arrows in metric column headers indicate whether higher (↑) or lower (↓) values are preferable.

Task	Error rate ↓	Tumor segmentation uncertainty as a predictor of classification errors											
		Predictive				Aleatoric				Epistemic			
		U-AUC ↑	AP ↑	Lift ↑	AURC ↓	U-AUC ↑	AP ↑	Lift ↑	AURC ↓	U-AUC ↑	AP ↑	Lift ↑	AURC ↓
IDH mutation status prediction	0.18	0.62 [0.51,0.72]	0.30 [0.21,0.46]	1.68 [1.27,2.60]	0.15 [0.08,0.22]	**0.65** [0.55,0.75]	0.28 [0.20,0.40]	1.54 [1.22,2.23]	**0.13** [0.07,0.20]	0.61 [0.50,0.72]	**0.32** [0.22,0.46]	**1.79** [1.30,2.72]	0.15 [0.09,0.23]
1p/19q co-deletion status prediction	0.10	0.68 [0.57,0.78]	0.23 [0.12,0.41]	2.16 [1.38,4.15]	**0.06** [0.03,0.11]	**0.69** [0.57,0.79]	**0.27** [0.14,0.43]	**2.55** [1.50,4.56]	**0.06** [0.03,0.11]	0.67 [0.56,0.78]	0.24 [0.12,0.40]	2.26 [1.36,4.19]	**0.06** [0.03,0.11]
Tumor grade prediction	0.29	0.65 [0.57,0.73]	0.44 [0.35,0.58]	1.52 [1.25,1.94]	0.22 [0.15,0.30]	**0.67** [0.59,0.74]	0.42 [0.34,0.55]	1.45 [1.23,1.82]	**0.19** [0.14,0.26]	0.65 [0.56,0.73]	**0.45** [0.35,0.58]	**1.54** [1.27,1.95]	0.22 [0.16,0.30]

Table 8: Performance of the proposed trust score for predicting classification errors. Assessment is performed using Uncertainty-based Receiver Operating Characteristic Area Under the Curve (U-AUC), Average Precision (AP), Lift, and Area Under the Risk–Coverage Curve (AURC). Values are reported with 95% confidence intervals obtained by $1000\times$ bootstrap resampling of the test set. Arrows in metric column headers indicate whether higher (↑) or lower (↓) values are preferable.

Task	Error rate ↓	Trust score			
		U-AUC ↑	AP ↑	Lift ↑	AURC ↓
IDH mutation status prediction	0.18	0.73 [0.64,0.81]	0.38 [0.26,0.53]	2.12 [1.58,3.10]	0.09 [0.05,0.14]
1p/19q co-deletion status prediction	0.10	0.85 [0.77,0.91]	0.42 [0.25,0.58]	3.98 [2.71,6.39]	0.02 [0.01,0.04]
Tumor grade prediction	0.29	0.77 [0.71,0.83]	0.51 [0.41,0.65]	1.77 [1.48,2.23]	0.13 [0.09,0.17]

quality cannot be optimized independently of predictive performance, reinforcing the need for joint evaluation of both aspects when developing diagnostic AI systems.

The behavior of ECE, NLL, and AUC across dropout rates further illustrates that calibration and predictive performance capture related but distinct properties of the model probability outputs. ECE evaluates calibration by comparing predicted confidence with empirical accuracy, and therefore reflects whether the numerical confidence values are aligned with the observed correctness rate. In contrast, NLL evaluates the probability assigned to the true class, averaged over all cases, and therefore depends on both confidence and correctness. It penalizes probability mass assigned away from the true class, especially for confident incorrect predictions. Conversely, AUC evaluates discriminative performance by measuring how well the model separates cases across different classification thresholds, without directly assessing whether the predicted probabilities are calibrated.

These differences explain why the metrics did not always change in parallel across dropout rates. A decrease in ECE at a specific dropout configuration can indicate improved agreement between confidence and empirical accuracy, without necessarily implying that the model assigns higher probabilities to the true classes or improves class discrimination. Conversely, an increase in NLL can occur when the average probability assigned to the true class decreases, even if the confidence values remain reasonably aligned with empirical accuracy. Similarly, AUC may decrease when the separation of classes worsens, even if the calibration error is unchanged or reduced. These observations support evaluating ECE, NLL, and AUC, together, rather than interpreting any single metric as a complete summary of the model predictive performance.

In the tumor segmentation task, case-level uncertainty was negatively correlated with segmentation accuracy, indicating that uncertainty estimates reflected meaningful variations in prediction quality. The strength of this relationship depended strongly on the spatial aggregation strategy. Full-brain aggregation showed weak associations with DSC, suggesting that averaging uncertainty over the entire brain is not well suited for case-level segmentation reliability assessment. This is likely because most brain voxels correspond to confidently predicted background and contribute limited information about tumor delineation quality, thereby diluting the uncertainty signal arising from the lesion region. In contrast, aggregation within the predicted tumor region showed the strongest and most stable negative association with segmentation accuracy.

The dilated tumor and boundary-weighted analyses further addressed whether uncertainty immediately outside the predicted boundary could provide additional information. These strategies also showed negative associations with DSC, supporting the relevance of spatially localized uncertainty around the tumor, although they did not provide a clearer advantage over predicted tumor aggregation and introduced additional hyperparameters such as dilation radius or distance-weighting scale.

A central objective of this study was to evaluate the operational utility of uncertainty estimates by quantifying their ability to detect erroneous predictions. Across tasks, uncertainty estimates identified errors better than expected

from the baseline error rates, with Lift values above 1 and AURC values indicating reduced residual risk among retained high-confidence cases. These results indicate that uncertainty estimates can serve as practical indicators of prediction reliability in the studied AI-based glioma diagnosis framework.

The comparison with DE and MCDE further showed that this operational utility was not specific to MCD. Across tasks, ensemble-based uncertainty estimates achieved performance comparable to MCD, although no method consistently dominated across all metrics. This supports the generality of the task-aware evaluation framework, while also showing that the preferred UQ method may depend on the intended use case: ranking errors, identifying rare failures, or reducing residual risk among retained cases.

For tumor segmentation, the binary definition of segmentation error depended on the selected DSC threshold. The threshold-sensitivity analysis showed that the segmentation error rate increased, as expected, when stricter DSC thresholds were used. Nevertheless, segmentation uncertainty remained informative across the evaluated thresholds, with high error-detection performance and Lift values above 1. This indicates that the operational conclusion that segmentation uncertainty identifies low-quality segmentations was not driven by the primary threshold of $\mathrm{DSC} \leq 0.70$, although the absolute values of AP and Lift varied with the induced error prevalence.

Because the classification tasks rely on shared representations learned jointly with segmentation, we further investigated their interactions. Tumor segmentation quality was significantly associated with classification correctness for IDH mutation status and tumor grade prediction, indicating that errors in structural representation learning may propagate to classification tasks. In contrast, the association between segmentation quality and 1p/19q co-deletion status prediction did not reach statistical significance. This suggests that, while accurate structural representation is important for certain molecular predictions, its influence may vary depending on the specific biological feature being predicted. One possible explanation is that imaging correlates of 1p/19q co-deletion status are less strongly linked to tumor morphology captured by segmentation, and may instead depend more on subtler texture or intensity patterns. Overall, these findings support the hypothesis that tumor segmentation performance partially reflects the quality of shared latent features used for molecular prediction, but that this dependency is task-specific rather than universal.

Correlation analysis revealed moderate associations between tumor segmentation and classification uncertainty for predictive and aleatoric components, suggesting that some uncertainty patterns are shared across tasks. The bootstrap confidence intervals supported positive associations for predictive and aleatoric uncertainty across all three classification tasks. In contrast, epistemic uncertainty showed weaker and less consistent relationships: its confidence intervals included zero for IDH mutation status and tumor grade prediction, while for 1p/19q co-deletion status prediction the association was weakly negative. Thus, shared cross-task uncertainty patterns were mainly observed for predictive and aleatoric uncertainty, rather than for the epistemic component.

Tumor segmentation uncertainty alone demonstrated moderate ability to predict classification errors, with Lift values above 1 across all three classification tasks, indicating that it contained information about classification reliability. However, when comparing predictive uncertainty scores directly, classification-specific uncertainty achieved higher Lift and lower AURC than tumor segmentation uncertainty for all three classification tasks. This suggests that although segmentation contributes to classification reliability, classification-specific predictive uncertainty remains the most informative predictor of classification error.

To investigate whether integrating information across tasks could improve reliability assessment, we proposed a composite trust score combining classification and tumor segmentation uncertainty. The trust score identified low-trust cases with higher classification error rates than expected under random selection across tasks, with Lift values above 1. However, compared with predictive classification uncertainty alone, it did not provide a consistent improvement. For 1p/19q co-deletion status prediction, the trust score improved the point estimates for U-AUC, AP, Lift, and AURC. This improvement was not observed for IDH mutation status or tumor grade prediction, where predictive classification uncertainty performed comparably or better depending on the metric. This result suggests that classification uncertainty already captures the most relevant information about classification reliability in this setting.

The previous finding highlights an important principle for trustworthy AI development: additional complexity does not necessarily improve reliability estimation. Instead, task-specific predictive uncertainty measures may provide the most direct and informative indicators of prediction confidence. At the same time, the trust score provides a useful conceptual framework for integrating uncertainty across tasks. Although its current formulation did not consistently improve selective risk reduction, it establishes a basis for future developments. Future trust-score formulations may require task-specific calibration, non-linear integration strategies, or explicit modeling of when tumor segmentation uncertainty is expected to contribute additional information beyond classification uncertainty alone.

This study has some limitations. First, uncertainty was evaluated using structural MRI modalities only. While these sequences form the backbone of current radiological assess-

ment, advanced imaging techniques such as diffusion- and perfusion-weighted MRI provide complementary information that may be of utility to strengthen interpretability and robustness of the uncertainty estimates.

Second, our convergence analysis focused on average uncertainty magnitudes as a function of the number of MC samples and dropout rate. Therefore, it should not be interpreted as a complete stability assessment. We did not evaluate case-wise ranking stability, repeatability across independently trained models or random seeds, or the variability of downstream error-detection metrics across repeated MC realizations. These aspects represent important directions for future work, particularly when uncertainty estimates are used for patient-level referral or prioritization.

Beyond the specific methods evaluated in this study, the task-aware evaluation framework proposed here provides a general basis for assessing uncertainty estimates in clinically oriented DL models. Because the framework evaluates convergence, calibration, uncertainty decomposition, error-detection performance, and selective prediction, its principles can be applied to newer UQ methods beyond MCD, DE, and MCDE. Future work may therefore extend this comparison to additional UQ approaches and assess method-specific trade-offs in calibration, computational cost, scalability, and clinical deployment feasibility.

6. Conclusion

This study presents a comprehensive, task-aware evaluation of UQ in a multi-task DL framework for glioma diagnosis. Our findings show that average MCD uncertainty estimates converged with a practical number of MC samples, that calibration and uncertainty quality depended on the dropout configuration, and that uncertainty estimates provided meaningful information about model reliability. Across tasks, uncertainty estimates enabled identification of prediction errors and supported selective-prediction analyses, although their utility depended on the task, uncertainty component, and operational metric considered.

Uncertainty decomposition offered insight into task-specific uncertainty sources, with aleatoric uncertainty being more prominent in classification and epistemic uncertainty being more prominent in tumor segmentation. However, decomposition did not provide a consistent operational advantage over predictive uncertainty alone. For tumor segmentation, uncertainty estimates were informative for identifying low-quality segmentations, particularly when aggregated within tumor-focused regions rather than across the full brain.

Although tumor segmentation performance and uncertainty were associated with classification reliability, classification predictive uncertainty remained the most direct and informative indicator of classification error, and the proposed composite trust score did not consistently improve over it. The comparison of MCD with DE and MCDE further showed that the operational value of uncertainty estimates was not restricted to a single UQ method. Overall, these findings emphasize that UQ should be evaluated in relation to the intended clinical task and use case. The task-aware evaluation framework proposed here provides practical guidance for assessing uncertainty in DL-based diagnostic tools and represents a promising step towards trustworthy AI for glioma diagnosis.

Acknowledgments

This work is part of the "Trustworthy AI for MRI" ICAI lab within the project ROBUST: Trustworthy AI-based Systems for Sustainable Growth with project number KICH3.LTP.20.006, financed by the Dutch Research Council (NWO), GE HealthCare, and the Dutch Ministry of Economic Affairs and Climate Policy (EZK) under the program LTP KIC 2020-2023.

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding treatment of animals or human subjects.

Conflicts of Interest

Authors GEMR, MS, and SKlein are part of the Erasmus MC ICAI lab "Trustworthy AI for MRI", a public-private research program partially funded by GE HealthCare (payment to institution). CP and SKaushik are employees of GE HealthCare. SvdV declares no competing interests.

Data availability

The public datasets included in this study are available online:

- BraTS: `http://braintumorsegmentation.org/`
- Brain Tumor Progression: `https://doi.org/10.7937/K9/TCIA.2018.15quzvnb`
- CPTAC-GBM: `https://doi.org/10.7937/K9/TCIA.2018.3rje41q1`
- EGD: `https://xnat.bmia.nl/REST/projects/egd`
- IvyGAP: `https://doi.org/10.7937/K9/TCIA.2016.XLwaN6nL`
- REMBRANDT: `https://doi.org/10.7937/K9/TCI`

A.2015.588OZUZB

- TCGA-GBM: https://doi.org/10.7937/K9/TCIA.2016.RNYFUYE9
- TCGA-LGG: https://doi.org/10.7937/K9/TCIA.2016.L4LTD3TK

The code used for this paper is publicly available in the following repository: https://github.com/ErasmusMC-NeuroOnco/PrognosAIs-UQ

References

Gehad Abdalla, Ahmed Hammam, Mustafa Anjari, Dr Felice D'Arco, and Dr Sotirios Bisdas. Glioma surveillance imaging: current strategies, shortcomings, challenges and outlook. *BJR— Open*, 2(1):20200009, 2020.

Moloud Abdar, Abbas Khosravi, Sheikh Mohammed Shariful Islam, U Rajendra Acharya, and Athanasios V Vasilakos. The need for quantification of uncertainty in artificial intelligence for clinical data analysis: increasing the level of trust in the decision-making process. *IEEE Systems, Man, and Cybernetics Magazine*, 8(3):28–40, 2022.

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific Data*, 4(1):1–13, 2017.

Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv preprint arXiv:1811.02629*, 2018.

Mateusz Buda, Ehab A AlBadawy, Ashirbani Saha, and Maciej A Mazurowski. Deep radiogenomics of lower-grade gliomas: convolutional neural networks predict tumor genomic subtypes using MR images. *Radiology: Artificial Intelligence*, 2(1):e180050, 2020.

Satrajit Chakrabarty, Pamela LaMontagne, Joshua Shimony, Daniel S Marcus, and Aristeidis Sotiras. MRI-based classification of IDH mutation and 1p/19q codeletion status of gliomas using a 2.5 D hybrid multi-task convolutional neural network. *Neuro-Oncology Advances*, 5(1):vdad023, 2023.

Matthew Chan, Maria Molina, and Chris Metzler. Estimating epistemic and aleatoric uncertainty with a single model. *Advances in Neural Information Processing Systems*, 37:109845–109870, 2024.

Peter Chang, Jack Grinband, Brian D Weinberg, Marios Bardis, Vahe Khy, Guillermo Cadena, Min Su, Soonmee Cha, Christopher G Filippi, Daniela Bota, et al. Deep-learning convolutional neural networks accurately classify genetic mutations in gliomas. *AJNR American Journal of Neuroradiology*, 39(7):1201–1207, 2018.

CPTAC. The Clinical Proteomic Tumor Analysis Consortium Glioblastoma Multiforme Collection (CPTAC-GBM) (Version 16) [dataset]. *The Cancer Imaging Archive*, 2018.

Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In *International Conference on Machine Learning*, pages 1184–1193. PMLR, 2018.

Ran El-Yaniv et al. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11 (5), 2010.

Anahita Fathi Kazerooni, Stephen J Bagley, Hamed Akbari, Sanjay Saxena, Sina Bagheri, Jun Guo, Sanjeev Chawla, Ali Nabavizadeh, Suyash Mohan, Spyridon Bakas, et al. Applications of radiomics and radiogenomics in high-grade gliomas in the era of precision medicine. *Cancers*, 13(23):5921, 2021.

Vladimir Fonov, Alan C Evans, Kelly Botteron, C Robert Almli, Robert C McKinstry, D Louis Collins, Brain Development Cooperative Group, et al. Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage*, 54(1):313–327, 2011.

Vladimir S Fonov, Alan C Evans, Robert C McKinstry, C Robert Almli, and DL Collins. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, 47:S102, 2009.

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

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

Bobby He, Balaji Lakshminarayanan, and Yee Whye Teh. Bayesian deep ensembles via the neural tangent kernel. *Advances in neural information processing systems*, 33: 1010–1022, 2020.

Wenchong He, Zhe Jiang, Tingsong Xiao, Zelin Xu, and Yukun Li. A survey on uncertainty quantification methods for deep learning. *ACM Computing Surveys*, 2025.

He Huang, Guang Yang, Wenbo Zhang, Xiaomei Xu, Weiji Yang, Weiwei Jiang, and Xiaobo Lai. A deep multi-task learning framework for brain tumor segmentation. *Frontiers in Oncology*, 11:690244, 2021.

Fabian Isensee, Marianne Schell, Irada Pflueger, Gianluca Brugnara, David Bonekamp, Ulf Neuberger, Antje Wick, Heinz-Peter Schlemmer, Sabine Heiland, Wolfgang Wick, et al. Automated brain extraction of multisequence MRI using artificial neural networks. *Human Brain Mapping*, 40(17):4952–4964, 2019.

Paul F Jaeger, Carsten T Lüth, Lukas Klein, and Till J Bungert. A call to reflect on evaluation practices for failure detection in image classification. *arXiv preprint arXiv:2211.15259*, 2022.

Craig K Jones, Guoqing Wang, Vivek Yedavalli, and Haris Sair. Direct quantification of epistemic and aleatoric uncertainty in 3D U-net segmentation. *Journal of Medical Imaging*, 9(3):034002–034002, 2022.

Kim-Celine Kahl, Carsten T Lüth, Maximilian Zenk, Klaus Maier-Hein, and Paul F Jaeger. VALUES: A framework for systematic validation of uncertainty estimation in semantic segmentation. *arXiv preprint arXiv:2401.08501*, 2024.

Sandeep S Kaushik, Mikael Bylund, Cristina Cozzini, Dattesh Shanbhag, Steven F Petit, Jonathan J Wyatt, Marion I Menzel, Carolin Pirkl, Bhairav Mehta, Vikas Chauhan, et al. Region of interest focused MRI to synthetic CT translation using regression and segmentation multi-task network. *Physics in Medicine & Biology*, 68 (19):195003, 2023.

Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 30, 2017.

Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*, 2015.

Philipp Kickingereder, David Bonekamp, Marta Nowosielski, Andreas Kratz, Martin Sill, Stefanie Burth, Wolfgang Wick, Heinz-Peter Schlemmer, Alexander Radbruch, Martin Bendszus, et al. Radiogenomics of glioblastoma: machine learning–based classification of molecular characteristics by using multiparametric and multiregional MR imaging features. *Radiology*, 290(3):674–682, 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Shirin Kordnoori, Maliheh Sabeti, Mohammad Hossein Shakoor, and Ehsan Moradi. Deep multi-task learning structure for segmentation and classification of supratentorial brain tumors in MR images. *Interdisciplinary Neurosurgery*, 36:101931, 2024.

Benjamin Lambert, Florence Forbes, Senan Doyle, Harmonie Dehaene, and Michel Dojat. Trustworthy clinical AI solutions: a unified review of uncertainty quantification in deep learning models for medical image analysis. *Artificial Intelligence in Medicine*, 150:102830, 2024.

Yiming Li, Dong Wei, Xing Liu, Xing Fan, Kai Wang, Shaowu Li, Zhong Zhang, Kai Ma, Tianyi Qian, Tao Jiang, et al. Molecular subtyping of diffuse gliomas using magnetic resonance imaging: comparison and correlation between radiomics and deep learning. *European Radiology*, 32(2):747–758, 2022.

Yin Li, Kaiyi Zheng, Shuang Li, Yongju Yi, Min Li, Yufan Ren, Congyue Guo, Liming Zhong, Wei Yang, Xinming Li, et al. A transformer-based multi-task deep learning model for simultaneous infiltrated brain area identification and segmentation of gliomas. *Cancer Imaging*, 23(1):105, 2023.

David N Louis, Arie Perry, Pieter Wesseling, Daniel J Brat, Ian A Cree, Dominique Figarella-Branger, Cynthia Hawkins, Ho-Keung Ng, Stefan M Pfister, Guido Reifenberger, et al. The 2021 WHO classification of tumors of the central nervous system: a summary. *Acta Neuropathologica*, 143(1):11–28, 2021.

Raghav Mehta, Thomas Christinck, Tanya Nair, Aurelie Bussy, Swapna Premasiri, Manuela Costantino, M Mallar Chakravarthy, Douglas L Arnold, Yarin Gal, and Tal Arbel. Propagating uncertainty across cascaded medical imaging tasks for improved deep learning inference. *IEEE Transactions on Medical Imaging*, 41(2):360–373, 2021.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2014.

Sushmita Mitra. Deep learning with radiogenomics towards personalized management of gliomas. *IEEE Reviews in Biomedical Engineering*, 16:579–593, 2021.

Jishnu Mukhoti, Andreas Kirsch, Joost Van Amersfoort, Philip HS Torr, and Yarin Gal. Deep deterministic uncertainty: A new simple baseline. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24384–24394, 2023.

Duc-Ky Ngo, Minh-Trieu Tran, Soo-Hyung Kim, Hyung-Jeong Yang, and Guee-Sang Lee. Multi-task learning for small brain tumor segmentation from MRI. *Applied Sciences*, 10(21):7790, 2020.

Jaya Ojha, Oriana Presacan, Pedro G. Lind, Eric Monteiro, and Anis Yazidi. Navigating uncertainty: A user-perspective survey of trustworthiness of AI in healthcare. *ACM Transactions on Computing for Healthcare*, 6(3): 1–32, 2025.

Quinn T Ostrom, Gino Cioffi, Haley Gittleman, Nirav Patil, Kristin Waite, Carol Kruchko, and Jill S Barnholtz-Sloan. CBTRUS statistical report: primary brain and other central nervous system tumors diagnosed in the United States in 2012–2016. *Neuro-Oncology*, 21(Supplement_5):v1–v100, 2019.

Nancy Pedano, Adam E Flanders, Lisa Scarpace, Tom Mikkelsen, Jennifer M Eschbacher, Beth Hermes, Victor Sisneros, Jill Barnholtz-Sloan, and Quinn Ostrom. The Cancer Genome Atlas Low Grade Glioma Collection (TCGA-LGG). *The Cancer Imaging Archive*, 2016.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III*, pages 234–241. Springer, 2015.

Yanan Ruan, Dengwang Li, Harry Marshall, Timothy Miao, Tyler Cossetto, Ian Chan, Omar Daher, Fabio Accorsi, Aashish Goela, and Shuo Li. Mt-UcGAN: multi-task uncertainty-constrained GAN for joint segmentation, quantification and uncertainty estimation of renal tumors on CT. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 439–449. Springer, 2020.

L Scarpace, AE Flanders, R Jain, T Mikkelsen, and DW Andrews. Data from REMBRANDT [Data set]. The Cancer Imaging Archive, 2019.

Lisa Scarpace, Tom Mikkelsen, Soonmee Cha, Sujaya Rao, Sangeeta Tekchandani, David Gutman, Joel H Saltz, Bradley J Erickson, Nancy Pedano, Adam E Flanders, et al. The Cancer Genome Atlas Glioblastoma Multiforme Collection (TCGA-GBM). *The Cancer Imaging Archive*, 2016.

Kathleen Schmainda and Melissa Prah. Data from brain-tumor-progression. *The Cancer Imaging Archive*, 21, 2018.

N Shah, X Feng, M Lankerovich, RB Puchalski, and B Keogh. Data from Ivy Glioblastoma Atlas Project (IvyGAP) [Data set]. *The Cancer Imaging Archive*, 2016.

Gagandeep Singh, Sunil Manjila, Nicole Sakla, Alan True, Amr H Wardeh, Niha Beig, Anatoliy Vaysberg, John Matthews, Prateek Prasanna, and Vadim Spektor. Radiomics and radiogenomics in gliomas: a contemporary update. *British Journal of Cancer*, 125(5):641–657, 2021.

Lewis Smith and Yarin Gal. Understanding measures of uncertainty for adversarial example detection. *arXiv preprint arXiv:1803.08533*, 2018.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

Sebastian R van der Voort, Fatih Incekara, Maarten MJ Wijnenga, Georgios Kapsas, Renske Gahrmann, Joost W Schouten, Hendrikus J Dubbink, Arnaud JPE Vincent, Martin J van den Bent, Pim J French, et al. The Erasmus Glioma Database (EGD): Structural MRI scans, WHO 2016 subtypes, and segmentations of 774 patients with glioma. *Data in Brief*, 37:107191, 2021.

Sebastian R van der Voort, Fatih Incekara, Maarten MJ Wijnenga, Georgios Kapsas, Renske Gahrmann, Joost W Schouten, Rishi Nandoe Tewarie, Geert J Lycklama, Philip C De Witt Hamer, Roelant S Eijgelaar, et al. Combined molecular subtyping, grading, and segmentation of glioma using multi-task deep learning. *Neuro-Oncology*, 25(2):279–289, 2023.

Andrew G Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *Advances in neural information processing systems*, 33:4697–4708, 2020.

Maximilian Zenk, David Zimmerer, Fabian Isensee, Jeremias Traub, Tobias Norajitra, Paul F Jäger, and Klaus Maier-Hein. Comparative benchmarking of failure detection methods in medical image segmentation: unveiling the role of confidence aggregation. *Medical image analysis*, 101:103392, 2025.

Tianlang Zhou, Su Ruan, and Stephane Canu. Multi-task learning for segmentation and classification of tumors in 3D medical images. In *Machine Learning in Medical Imaging*, pages 339–347. Springer, 2019.

Appendix A. Implementation details

A.1 Pre-processing

Prior to model training, all MRI scans underwent a standardized pre-processing pipeline consistent with common practices in Medical Image Analysis. These steps were designed to reduce inter-subject variability, improve anatomical alignment across patients, and ensure numerical stability during network optimization.

First, all scans were spatially registered to the MNI152 atlas space (Fonov et al., 2009, 2011) to establish a common anatomical reference frame. This atlas has a resolution of $1\times1\times1$ mm^3 and a size of $197\times233\times189$ voxels. Modality-specific registration was performed to preserve anatomical correspondence: T1w and T1wCE images were registered to the T1w MNI152 atlas, while T2w and FLAIR images were registered to the T2w atlas template. This modality-aware registration strategy ensures optimal alignment of tissue contrasts while minimizing interpolation artifacts across sequences. Standard-space alignment reduces anatomical variability unrelated to pathology and facilitates consistent learning of spatial patterns by the NN.

Following registration, bias field correction was applied to each scan to mitigate low-frequency intensity inhomogeneities caused by magnetic field non-uniformities. Correcting these intensity distortions is particularly important in multi-center datasets, like the one used in this work, as scanner-dependent intensity variations can otherwise introduce spurious features that degrade model generalization.

Subsequently, skull-stripping was performed using the HD-BET tool (Isensee et al., 2019) to generate an accurate brain mask for the MNI152 atlas. Removing non-brain tissues (e.g., skull, fat, and extracranial structures) serves two purposes: (i) it restricts the learning process to anatomically relevant regions and (ii) reduces the risk of the model exploiting non-biological artifacts. The resulting brain mask was further used to compute a tight bounding box around the intracranial volume, allowing spatial cropping of the scans, and resulting in a uniform scan size of 145x182x152 voxels. This step reduces computational burden, decreases background redundancy, and increases the effective proportion of informative voxels presented to the model.

Finally, intensity normalization was performed via z-score standardization within the brain mask region. For each scan, voxel intensities were centered and scaled using the mean and standard deviation computed over brain tissue only. Masked normalization avoids contamination from background voxels and ensures comparable intensity distributions across subjects and modalities, thereby promoting stable training dynamics and improving convergence.

A.2 Architecture and training procedure

The architecture (van der Voort et al., 2023; Ronneberger et al., 2015) depicted in Figure 7 processes full 3D volumes. It was trained for up to 100 epochs with ADAM (Kingma and Ba, 2014) optimizer and minimizing the following weighted loss function (van der Voort et al., 2023):

$$\begin{aligned}\mathcal{L} = W_{\text{IDH}} \cdot \mathcal{L}_{\text{IDH}} + W_{\text{1p/19q}} \cdot \mathcal{L}_{\text{1p/19q}} \\ + W_{\text{grade}} \cdot \mathcal{L}_{\text{grade}} + W_{\text{seg}} \cdot \mathcal{L}_{\text{seg}}.\end{aligned} \tag{22}$$

$$W_t = \begin{cases} \frac{N_{\text{train}}}{N_t} & \text{for } t \in \{\text{IDH}, \text{1p/19q}, \text{grade}\} \\ 1 & \text{for } t = \text{seg} \end{cases}, \tag{23}$$

where N_{train} corresponds to the number of training subjects, and N_t to the amount of available labeled samples for each task t.

To account for missing data and intrinsic class imbalance in each task, a masked, class-weighted categorical cross-entropy loss was used. This formulation ensures that only samples with available labels contribute to the loss, and that underrepresented classes are not neglected during optimization. Let a batch contain N samples and C classes. For each sample $i \in \{1, \ldots, N\}$, let $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,C})$ denote the one-hot encoded ground truth label, and $\hat{\mathbf{y}}_i = (\hat{y}_{i,1}, \ldots, \hat{y}_{i,C})$ the predicted class probabilities. Let $m_i \in \{0, 1\}$ be a binary mask indicating whether sample i has a valid label. A fixed weight w_c is assigned to each class c, computed from the full training dataset as:

$$w_c = \frac{N_t}{C \cdot n_c}, \tag{24}$$

where N_t corresponds to the amount of labeled training samples for the task, and n_c is the number of occurrences of class c among those training samples. The masked, class-weighted loss for a batch is then defined as (van der Voort et al., 2023):

$$\mathcal{L}_{\text{masked}} = \begin{cases} \frac{1}{\sum_{i=1}^{N} m_i} \sum_{i=1}^{N} m_i \left(-\sum_{c=1}^{C} w_c \, y_{i,c} \log \hat{y}_{i,c} \right), \\ \qquad \text{if } \sum_{i=1}^{N} m_i > 0, \\ 0, \quad \text{otherwise.} \end{cases} \tag{25}$$

For the tumor segmentation task, we use the DSC loss, which directly optimizes for the overlap between the predicted segmentation and the ground truth. Let $\hat{p}_j \in [0, 1]$ and $g_j \in [0, 1]$ denote the predicted and ground truth values for voxel j over a total of M voxels. The DSC loss is defined as:

$$\mathcal{L}_{\text{DSC}} = 1 - \frac{2 \sum_{j=1}^{M} \hat{p}_j g_j}{\sum_{j=1}^{M} \hat{p}_j^2 + \sum_{j=1}^{M} g_j^2 + \epsilon},$$

where ϵ is a small constant added for numerical stability.

Appendix B. Data

Table 10 provides a detailed overview of the dataset used in this study. The table reports the number of cases in the training and test sets, together with the distribution of available labels for IDH mutation status, 1p/19q co-deletion status, and tumor grade. For each task, cases with missing annotations are reported as N/A, reflecting incomplete molecular or histopathological information. This breakdown highlights both the class imbalance across tasks and the presence of missing labels, which motivated the use of task-specific masking during training.

Appendix C. Segmentation error threshold sensitivity analysis

To assess whether the operational utility of segmentation uncertainty depended on the binary definition of segmentation error, we repeated the segmentation error-detection analysis for three DSC thresholds: $\tau_{\mathrm{DSC}} = 0.60$, $\tau_{\mathrm{DSC}} = 0.70$, and $\tau_{\mathrm{DSC}} = 0.80$. These thresholds were chosen to represent a more lenient definition focused on severe failures, the primary intermediate operating point used in the main analysis, and a stricter definition of low-quality segmentation. For each threshold, segmentation error was defined as $\mathrm{DSC} \leq \tau_{\mathrm{DSC}}$, and the threshold-specific error rate, U-AUC, AP, and Lift were recomputed for predictive, aleatoric and epistemic uncertainty. This analysis was not used to optimize the segmentation error threshold, but to assess whether the conclusion that segmentation uncertainty identifies low-quality segmentations was robust to the selected binarization.

As expected, increasing the DSC threshold resulted in a higher segmentation error rate, from 0.1 at $\tau_{\mathrm{DSC}} = 0.60$, to 0.17 at $\tau_{\mathrm{DSC}} = 0.70$, and 0.31 at $\tau_{\mathrm{DSC}} = 0.80$. Despite this change in error prevalence, segmentation uncertainty remained strongly informative across thresholds. For predictive uncertainty, U-AUC remained high across all thresholds, ranging from 0.96 at $\tau_{\mathrm{DSC}} = 0.60$ to 0.93 at $\tau_{\mathrm{DSC}} = 0.80$. AP values were consistently above the corresponding threshold-specific baseline error rates, and Lift values remained greater than 1 for all thresholds. Similar patterns were observed for the aleatoric and epistemic uncertainty components. These results indicate that the operational conclusion that segmentation uncertainty identifies low-quality segmentations was not driven by the specific primary threshold of $\mathrm{DSC} \leq 0.70$, although the absolute metric values varied with the prevalence of segmentation errors induced by each threshold.

Table 9: Sensitivity of segmentation uncertainty error detection to the DSC threshold used to define segmentation error. For each threshold, segmentation error was defined as $\mathrm{DSC} \leq \tau_{\mathrm{DSC}}$. Error rate is reported as the proportion of cases classified as segmentation errors. U-AUC, AP, and Lift are reported for predictive, aleatoric, and epistemic uncertainty, with 95% confidence intervals obtained by $1000\times$ bootstrap resampling of the test set. Arrows in metric column headers indicate whether higher (↑) or lower (↓) values are preferable.

τ_{DSC}	Error rate ↓	Uncertainty	U-AUC ↑	AP ↑	Lift ↑
0.60	0.10	Predictive	0.96 [0.93,0.98]	0.78 [0.64,0.90]	7.96 [5.73,12.41]
		Aleatoric	0.97 [0.95,0.99]	0.82 [0.67,0.92]	8.33 [5.99,13.09]
		Epistemic	0.96 [0.93,0.98]	0.76 [0.61,0.88]	7.77 [5.61,12.26]
0.70	0.17	Predictive	0.95 [0.92,0.98]	0.82 [0.72,0.91]	4.95 [3.89,6.67]
		Aleatoric	0.96 [0.94,0.98]	0.83 [0.72,0.91]	5.01 [3.93,6.80]
		Epistemic	0.95 [0.92,0.98]	0.81 [0.71,0.90]	4.90 [3.85,6.64]
0.80	0.31	Predictive	0.93 [0.89,0.96]	0.88 [0.81,0.93]	2.83 [2.41,3.46]
		Aleatoric	0.93 [0.89,0.97]	0.89 [0.83,0.95]	2.87 [2.45,3.53]
		Epistemic	0.92 [0.88,0.95]	0.87 [0.80,0.92]	2.80 [2.38,3.42]

Appendix D. Correlation patterns between segmentation uncertainty estimates and tumor segmentation DSC

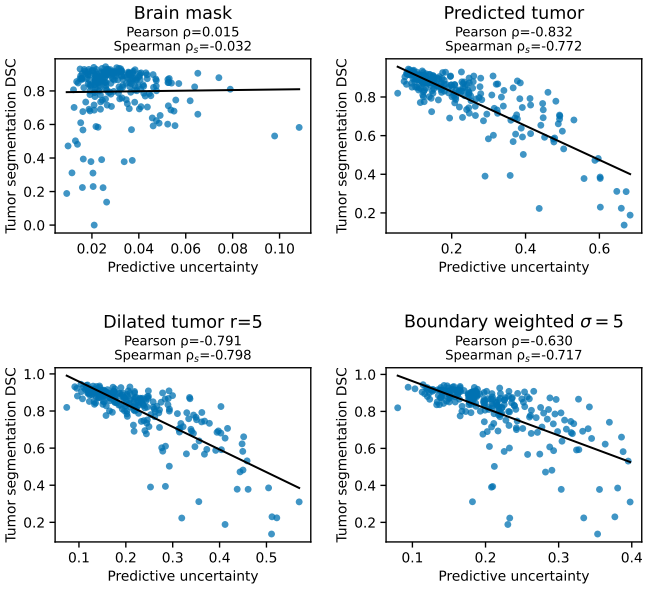

Figure 8: Case-wise association between predictive segmentation uncertainty and tumor segmentation DSC at the selected MCD configuration ($T = 30$, $\delta = 0.25$). Results are shown for brain mask, predicted tumor, 5-voxel dilated tumor, and boundary-weighted aggregation ($\sigma = 5$ voxels). Each point represents one test case; Pearson ρ and Spearman ρ_s are reported for each strategy.

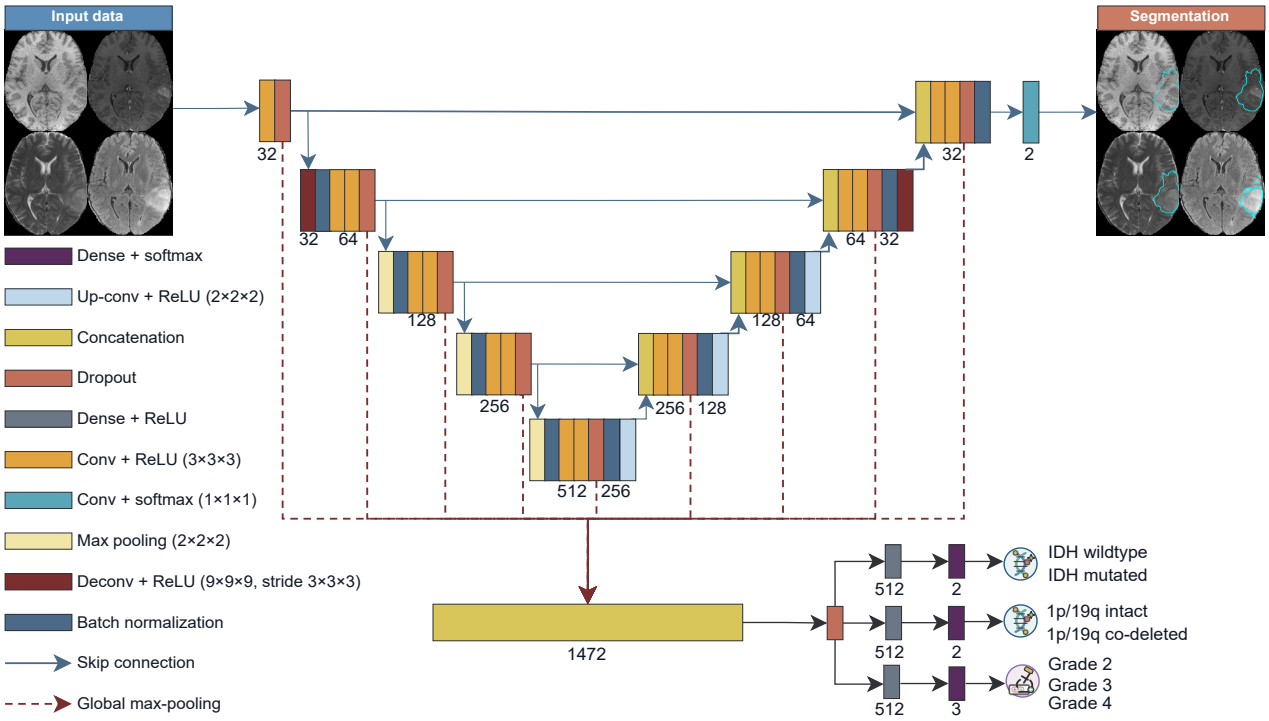

Figure 7: Overview of the multi-task DL network used for the tumor segmentation and prediction of IDH mutation status, 1p/19q co-deletion status and tumor grade. The architecture processes full 3D volumes. The labels in the figure define the layer type, and the numbers indicate either the number of filters, dense units or features. Figure based on the one presented in the work of van der Voort et al. (2023).

Figure 8 shows the case-wise relationship between predictive segmentation uncertainty and tumor segmentation DSC at the selected MCD configuration. This analysis was included to assess whether the uncertainty–DSC associations reported using Pearson correlation were consistent with the corresponding rank-based Spearman correlation.

Across the tumor-localized uncertainty scores, Pearson and Spearman correlations showed the same overall interpretation: higher predictive uncertainty was associated with lower tumor segmentation DSC. For predicted tumor aggregation, both coefficients indicated a strong negative association, with Pearson $\rho = -0.832$ and Spearman $\rho_s = -0.772$. A similar pattern was observed for dilated tumor aggregation, with Pearson $\rho = -0.791$ and Spearman $\rho_s = -0.798$. Boundary-weighted aggregation also showed a negative relationship, with Pearson $\rho = -0.630$ and Spearman $\rho_s = -0.717$.

For brain mask aggregation, both coefficients were close to zero, with Pearson $\rho = 0.015$ and Spearman $\rho_s = -0.032$, indicating no meaningful association between brain-level predictive uncertainty and tumor segmentation DSC. Thus, the conclusions were consistent across Pearson and Spearman correlation: tumor-localized uncertainty showed a negative uncertainty–performance relationship, whereas brain-level aggregation did not. This supports the use of Pearson correlation in the main analysis, while showing that the observed findings were not dependent on assuming a strictly linear relationship.

Appendix E. Comparison of UQ Methods

Table 11 provides the component-wise comparison of MCD, DE, and MCDE using aleatoric and epistemic uncertainty. These analyses complement the main predictive-uncertainty comparison by showing how the decomposed uncertainty components behave across tasks and methods. For tumor segmentation, uncertainty was aggregated within the predicted tumor region, matching the strategy used in the main method-comparison analysis.

Overall, aleatoric uncertainty showed trends broadly consistent with predictive uncertainty, with useful error-detection performance across classification tasks and high performance for tumor segmentation. Epistemic uncertainty was more variable across methods and tasks. It remained informative for tumor segmentation with all three methods, with the strongest U-AUC, AP, and Lift observed for MCD, but was generally weaker for classification error detection.

Table 10: Detailed distribution of the data used for training and testing the Deep Learning (DL) architecture for each one of the classification tasks. N/A (not available) represents missing data. Tumor segmentation ground truth was available for all cases.

Dataset	Subset	IDH mutation status			1p/19q co-deletion status			Tumor grade			
		Wildtype	Mutated	N/A	Intact	Co-deleted	N/A	2	3	4	N/A
Train	BraTS	0	0	156	0	0	156	0	0	0	156
	Brain tumor Progression	0	0	20	0	0	20	0	0	0	20
	CPTAC-GBM	0	0	45	0	0	45	0	0	0	45
	EGD	312	155	308	186	73	516	135	80	502	58
	IvyGAP	32	6	1	27	3	9	0	1	36	2
	In-house	86	53	183	107	22	193	24	6	191	101
	REMBRANDT	0	0	109	0	0	109	37	21	32	19
Test	TCGA-GBM	107	5	21	127	0	6	0	0	132	1
	TCGA-LGG	22	80	1	78	25	0	45	58	0	0

Table 11: Comparison of Uncertainty Quantification methods using aleatoric and epistemic uncertainty. Assessment is made using Uncertainty-based Receiver Operating Characteristic Area Under the Curve (U-AUC), Average Precision (AP), Lift, and Area Under the Risk–Coverage Curve (AURC), derived from using each uncertainty component as a predictor of errors. Values are reported with 95% confidence intervals obtained by $1000\times$ bootstrap resampling of the test set. The error column denotes the method-specific classification error rate for classification tasks and the proportion of cases with DSC ≤ 0.70 for tumor segmentation. For the latter, uncertainty was aggregated within the predicted tumor region. The highest displayed values for U-AUC, AP, and Lift, and the lowest displayed values for AURC, are indicated in bold for each task and uncertainty component. Arrows in metric column headers indicate whether higher (↑) or lower (↓) values are preferable.

Task	Method	Error ↓	Aleatoric uncertainty				Epistemic uncertainty			
			U-AUC ↑	AP ↑	Lift ↑	AURC ↓	U-AUC ↑	AP ↑	Lift ↑	AURC ↓
IDH mutation	DE	0.17	**0.73** [0.62,0.82]	**0.39** [0.26,0.56]	**2.26** [1.66,3.31]	**0.09** [0.05,0.15]	0.64 [0.54,0.74]	0.28 [0.17,0.43]	1.60 [1.20,2.45]	0.12 [0.07,0.18]
	MCD	0.18	0.71 [0.61,0.80]	0.32 [0.22,0.46]	1.79 [1.39,2.54]	0.10 [0.05,0.15]	**0.71** [0.61,0.80]	**0.34** [0.23,0.50]	**1.89** [1.45,2.80]	**0.10** [0.05,0.16]
	MCDE	0.17	**0.73** [0.62,0.82]	0.36 [0.24,0.53]	2.12 [1.55,3.13]	**0.09** [0.05,0.14]	0.64 [0.54,0.73]	0.30 [0.20,0.47]	1.81 [1.31,2.83]	0.12 [0.07,0.19]
1p/19q co-deletion	DE	0.10	0.83 [0.75,0.90]	0.29 [0.18,0.49]	2.95 [2.12,5.25]	0.03 [0.01,0.04]	0.66 [0.55,0.76]	**0.18** [0.10,0.34]	**1.79** [1.19,3.65]	**0.05** [0.03,0.08]
	MCD	0.10	**0.84** [0.77,0.91]	**0.39** [0.22,0.58]	**3.70** [2.51,6.19]	0.03 [0.01,0.04]	**0.68** [0.58,0.77]	0.16 [0.10,0.29]	1.54 [1.22,2.59]	**0.05** [0.03,0.08]
	MCDE	0.09	0.81 [0.74,0.89]	0.28 [0.17,0.47]	3.02 [2.11,5.55]	**0.02** [0.01,0.04]	0.60 [0.48,0.72]	0.12 [0.08,0.22]	1.33 [1.04,2.27]	0.06 [0.03,0.09]
Tumor grade	DE	0.30	**0.80** [0.73,0.85]	**0.59** [0.48,0.71]	1.99 [1.65,2.47]	0.13 [0.09,0.18]	0.64 [0.56,0.72]	**0.47** [0.36,0.60]	**1.58** [1.31,2.02]	0.22 [0.16,0.29]
	MCD	0.29	0.79 [0.73,0.85]	**0.59** [0.47,0.72]	**2.00** [1.68,2.50]	**0.12** [0.09,0.17]	**0.70** [0.62,0.77]	0.45 [0.35,0.58]	1.52 [1.28,1.94]	**0.17** [0.12,0.24]
	MCDE	0.28	0.78 [0.71,0.84]	0.53 [0.42,0.66]	1.90 [1.58,2.40]	**0.12** [0.08,0.17]	0.63 [0.55,0.71]	0.41 [0.32,0.54]	1.45 [1.21,1.90]	0.21 [0.14,0.28]
Tumor segmentation	DE	0.13	**0.96** [0.93,0.98]	0.81 [0.66,0.90]	6.11 [4.58,8.68]	0.13 [0.12,0.13]	0.93 [0.89,0.96]	0.63 [0.46,0.80]	4.75 [3.56,7.10]	**0.12** [0.12,0.13]
	MCD	0.17	**0.96** [0.94,0.98]	**0.83** [0.72,0.91]	5.01 [3.93,6.80]	0.13 [0.12,0.14]	**0.95** [0.92,0.98]	**0.81** [0.71,0.90]	**4.90** [3.85,6.64]	0.13 [0.12,0.14]
	MCDE	0.12	**0.96** [0.92,0.98]	0.79 [0.65,0.90]	**6.43** [4.83,9.28]	**0.12** [0.12,0.13]	0.83 [0.76,0.89]	0.40 [0.25,0.59]	3.25 [2.26,4.96]	0.13 [0.12,0.14]