



**HAL**  
open science

# Deep evidential fusion with uncertainty quantification and reliability learning for multimodal medical image segmentation

Ling Huang, Su Ruan, Pierre Decazes, Thierry Denœux

## ► To cite this version:

Ling Huang, Su Ruan, Pierre Decazes, Thierry Denœux. Deep evidential fusion with uncertainty quantification and reliability learning for multimodal medical image segmentation. *Information Fusion*, 2025, 113, pp.102648. <10.1016/j.inffus.2024.102648>. <hal-04681852>

**HAL Id: hal-04681852**

**<https://hal.science/hal-04681852v1>**

Submitted on 30 Aug 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

# Deep evidential fusion with uncertainty quantification and reliability learning for multimodal medical image segmentation

Ling Huang<sup>a</sup>, Su Ruan<sup>b</sup>, Pierre Decazes<sup>c</sup>, Thierry Denœux<sup>a,d</sup>

<sup>a</sup>*Université de technologie de Compiègne, CNRS, Heudiasyc, Compiègne, France*

<sup>b</sup>*Université de Rouen Normandie, Quantif, LITIS, Rouen, France*

<sup>c</sup>*Université de Rouen Normandie, Centre Henri Becquerel, Rouen, France*

<sup>d</sup>*Institut universitaire de France, Paris, France*

---

## Abstract

Single-modality medical images generally do not contain enough information to reach an accurate and reliable diagnosis. For this reason, physicians commonly rely on multimodal medical images for comprehensive diagnostic assessments. This study introduces a deep evidential fusion framework designed for segmenting multimodal medical images, leveraging the Dempster-Shafer theory of evidence in conjunction with deep neural networks. In this framework, features are first extracted from each imaging modality using a deep neural network, and features are mapped to Dempster-Shafer mass functions that describe the evidence of each modality at each voxel. The mass functions are then corrected by the contextual discounting operation, using learned coefficients quantifying the reliability of each source of information relative to each class. The discounted evidence from each modality is then combined using Dempster’s rule of combination. Experiments were carried out on a PET-CT dataset for lymphoma segmentation and a multi-MRI dataset for brain tumor segmentation. The results demonstrate the ability of the proposed fusion scheme to quantify segmentation uncertainty and improve segmentation accuracy. Moreover, the learned reliability coefficients provide some insight into the contribution of each modality to the segmentation process.

*Keywords:* Dempster-Shafer theory, Evidence theory, Medical image processing, Deep learning, Decision-level fusion

---

## 1. Introduction

Recent advances in medical imaging technologies have facilitated the acquisition of multimodal data such as Positron Emission Tomography (PET)/Computed Tomography (CT) and multi-sequence Magnetic Resonance Imaging (MRI). Images from a single modality provide partial insight into cancer and other abnormalities within the human body. Multimodal medical image analysis, which integrates information from diverse medical imaging modalities, significantly contributes to a comprehensive understanding of intricate medical conditions [90]. It encompasses factors such as the location, size, and extent of pathological

9 structures. Medical image segmentation based on the fusion of multimodal medical informa-  
10 tion allows clinicians to better delineate anatomical structures, lesions and abnormalities,  
11 thus enhancing the effectiveness of disease detection, diagnosis, and treatment planning.

12 Multimodal medical image fusion strategies can be implemented at different levels [84].  
13 At the lowest pixel level, multimodality images are concatenated as a single input. Alterna-  
14 tively, features can be extracted from different modalities and combined for further modeling  
15 and reasoning (feature-level fusion). Finally, in the decision-level approach, partial decisions  
16 are made independently based on each modality and aggregated to obtain a final decision.  
17 Though recent developments in multimodal medical image analysis have yielded promising  
18 experimental results, conventional multimodal medical image fusion strategies still suffer  
19 from some limitations. It is often difficult to explain why a given strategy works in a given  
20 context, and to quantify decision uncertainty in a reliable way. Moreover, most approaches  
21 are based on optimistic assumptions about data quality and, contrary to clinical knowledge,  
22 they treat images from different modalities as equally reliable when segmenting tumors,  
23 which may lead to biased or wrong decisions.

24 The success of information fusion depends on the relevance and complementarity of  
25 input information, the existence of prior knowledge about the information sources, and the  
26 expressive power of the uncertainty model employed [65, 24, 61]. Given that the quality  
27 of input information and prior knowledge is intricately tied to the data collection stage, a  
28 lot of work has been devoted to modeling uncertainties in a faithful way [39]. As a critical  
29 factor in the information fusion process [1, 38], accurate uncertainty quantification must be  
30 regarded as a primary objective to achieve precise multimodal medical image segmentation.

31 Early methods for quantifying uncertainty essentially relied on probabilistic models, of-  
32 ten integrated with Bayesian inference or sampling techniques to estimate uncertainty across  
33 various parameters or variables [34, 56]. The advent of deep neural networks has sparked  
34 renewed interest in uncertainty estimation [1], leading to the development of methods such  
35 as Monte-Carlo dropout [29] and deep ensembles [45]. However, it is important to note  
36 that these probabilistic models rely on assumptions about the underlying data distribution,  
37 and improper distributions can result in inaccurate uncertainty estimations. Furthermore,  
38 uncertainty quantification via inference or sampling algorithms heavily relies on computa-  
39 tional approximations and may lack rigorous theoretical justification [6, 7]. These and other  
40 limitations motivate the search for alternative approaches for uncertainty quantification for  
41 information fusion and decision-making applications.

42 Instead of making strong assumptions on actual data distribution, non-probabilistic  
43 methods use alternative mathematical frameworks or representations such as possibility  
44 theory [87, 21] and Dempster-Shafer theory (DST) [15, 69, 20] to quantify uncertainty. In  
45 particular, the latter formalism is an evidence-based information modeling, reasoning, and  
46 fusion framework that can be used with both supervised [16, 80, 79] and unsupervised learn-  
47 ing [48, 19], providing an effective way to handle imperfect (i.e., imprecise, uncertain, and  
48 conflicting) data. Compared to possibility theory, DST allows the quantification of both  
49 aleatory and epistemic uncertainty while providing a powerful mechanism for combining  
50 multiple unreliable pieces of information [61].

51 In multimodal medical image segmentation, effectively combining uncertain information

52 from diverse sources presents a significant challenge. Some learning-based approaches pro-  
 53 pose addressing conflicting decisions by introducing learnable weights [50, 4, 71]. The term  
 54 “weight” in those approaches usually refers to the importance of information. In contrast,  
 55 reliability pertains to the trustworthiness of the information and needs to be carefully an-  
 56 alyzed in different medical situations. Four major approaches have been used to provide  
 57 reliability coefficients: 1) modeling the reliability of sources using a degree of consensus [14];  
 58 2) modeling expert opinions using probability distributions [12]; 3) using external domain  
 59 knowledge or contextual information to model reliability coefficients [26]; 4) learning the  
 60 reliability coefficients from training data [25, 62], which is a very general approach that does  
 61 not require any prior domain knowledge or expert opinions. In this work, we consider an  
 62 even more flexible approach in which the reliability of each image modality is described by  
 63 several coefficients, one for each ground truth value. The reliability coefficient for source  $i$   
 64 and class  $k$  is then defined as one’s belief that the information from source  $i$  is reliable, if  
 65 the true class is  $k$ .

66 In this paper, we introduce a new approach to multimodal medical image segmentation  
 67 combining DST with deep neural networks<sup>1</sup>. The proposed fusion scheme comprises multiple  
 68 encoder-decoder-based feature extraction modules, DST-based evidence-mapping modules,  
 69 and a multimodality evidence fusion module. The evidence-mapping modules transform  
 70 the extracted features into mass functions representing the evidence from each imaging  
 71 modality about the class of each voxel. These mass functions are then corrected by a  
 72 contextual discounting operation, and the discounted pieces of evidence are combined by  
 73 Dempster’s rule of combination. The whole framework is trained end-to-end by minimizing  
 74 a loss function quantifying the errors before and after the fusion of information from each  
 75 modality. Our main contributions are, thus, the following:

- 76 1. We propose a new hybrid fusion architecture for multimodal medical images composed  
 77 of feature extraction, evidence-mapping, and combination modules.
- 78 2. Within this architecture, we integrate mechanisms for (i) quantifying segmentation un-  
 79 certainty using Dempster-Shafer mass functions, (ii) correcting these mass functions  
 80 to account for the relative reliability of each imaging modality using context discount-  
 81 ing, and (iii) combining corrected mass functions from different sources to reach final  
 82 segmentation decisions.
- 83 3. We introduce an improved two-part loss function making it possible to optimize the  
 84 segmentation performance of each individual source modality together with the overall  
 85 performance of the combined decisions.
- 86 4. Through extensive experiments with two real medical image datasets, we show that the  
 87 proposed decision-level fusion scheme improves segmentation reliability and quality as  
 88 compared to alternative pixel-level methods for exploiting different image modalities.

---

<sup>1</sup>This paper is an extended version of the short paper presented at the 25th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2022) [36]. This extended version includes a much more detailed description and explanation of the fusion framework, an improved optimization strategy with a two-part loss function, as well as extended results with a second dataset for lymphoma segmentation and an additional transformer-based feature-extraction module.

89 5. We show that the learned reliability coefficients provide some insight into the contri-  
 90 bution of each imaging modality in the segmentation process.

91 The rest of this paper is organized as follows. Background information and related work  
 92 are first recalled in Section 2. Our approach is then introduced in Section 3, and experimental  
 93 results are reported in Section 4. Finally, Section 5 concludes the paper and presents some  
 94 directions for further research.

## 95 2. Related work

96 The basic concepts of DST and its application to classification are first recalled in Section  
 97 2.1. The contextual discounting operation, which plays a central role in our approach, is  
 98 described separately in Section 2.2. The evidential neural network model used in this paper  
 99 is then introduced in Section 2.3, and related work on multimodal medical image fusion is  
 100 briefly reviewed in Section 2.4.

### 101 2.1. Dempster-Shafer theory

Let  $\Theta = \{\theta_1, \theta_2, \dots, \theta_K\}$  be the finite set of possible answers to some question, called the  
*frame of discernment*. Evidence about a variable taking values in  $\Theta$  can be represented by  
 a *mass function*  $m : 2^\Theta \rightarrow [0, 1]$ , such that

$$\sum_{A \subseteq \Theta} m(A) = 1 \quad \text{and} \quad m(\emptyset) = 0.$$

102 Each subset  $A \subseteq \Theta$  such that  $m(A) > 0$  is called a *focal set* of  $m$ . The mass  $m(A)$  represents  
 103 a share of a unit mass of belief allocated to focal set  $A$ , which cannot be allocated to any  
 104 strict subset of  $A$ . The mass  $m(\Theta)$  can be interpreted as a degree of ignorance. Full ignorance  
 105 is represented by the *vacuous* mass function  $m_\gamma$  verifying  $m_\gamma(\Theta) = 1$ . If all focal sets are  
 106 singletons, then  $m$  is said to be *Bayesian*; it is equivalent to a probability distribution.

*Belief and plausibility functions.* The information provided by a mass function  $m$  can also  
 be represented by a *belief function*  $Bel$  or a *plausibility function*  $Pl$  from  $2^\Theta$  to  $[0, 1]$  defined,  
 respectively, as:

$$Bel(A) = \sum_{B \subseteq A} m(B)$$

and

$$Pl(A) = \sum_{B \cap A \neq \emptyset} m(B) = 1 - Bel(\bar{A}),$$

for all  $A \subseteq \Theta$ , where  $\bar{A}$  denotes the complement of  $A$ . The quantity  $Bel(A)$  can be inter-  
 preted as a degree of support for  $A$ , while  $Pl(A)$  is a measure of lack of support against  $A$ .  
 The *contour function*  $pl$  associated to  $m$  is the function that maps each element  $\theta$  of  $\Theta$  to  
 its plausibility, i.e.,

$$pl(\theta) = Pl(\{\theta\}), \quad \forall \theta \in \Theta.$$

107 As shown below, this function can be easily computed when combining several pieces of  
 108 evidence; it plays an important role in decision-making.

109 *Dempster's rule.* In DST, the beliefs about a certain question are established by aggregat-  
 110 ing independent pieces of evidence represented by belief functions over the same frame of  
 111 discernment [69]. Given two mass functions  $m_1$  and  $m_2$  derived from two independent items  
 112 of evidence, the mass function  $m_1 \oplus m_2$  representing the pooled evidence is defined as

$$(m_1 \oplus m_2)(A) = \frac{1}{1 - \kappa} \sum_{B \cap C = A} m_1(B)m_2(C), \quad (1a)$$

113 for all  $A \subseteq \Theta, A \neq \emptyset$ , and  $(m_1 \oplus m_2)(\emptyset) = 0$ . The coefficient  $\kappa$  is the *degree of conflict*  
 114 between  $m_1$  and  $m_2$ ,

$$\kappa = \sum_{B \cap C = \emptyset} m_1(B)m_2(C). \quad (1b)$$

115 This operation is called *Dempster's rule of combination*. It is commutative and associa-  
 116 tive. The combined mass function  $m_1 \oplus m_2$  is called the *orthogonal sum* of  $m_1$  and  $m_2$ . Mass  
 117 functions  $m_1$  and  $m_2$  can be combined if and only if  $\kappa < 1$ . Let  $pl_1, pl_2$  and  $pl_1 \oplus pl_2$  denote  
 118 the contour functions associated with, respectively,  $m_1, m_2$  and  $m_1 \oplus m_2$ . The following  
 119 equation holds:

$$\forall \theta \in \Theta, \quad (pl_1 \oplus pl_2)(\theta) = \frac{pl_1(\theta)pl_2(\theta)}{1 - \kappa}. \quad (2)$$

120 The complexity of calculating the combined contour function using (2) is linear in the  
 121 cardinality of  $\Theta$ , whereas computing the combined mass function using (1) has, in the  
 122 worst-case, exponential complexity.

123 *Conditioning.* Given a mass function  $m$  and a nonempty subset  $A$  of  $\Theta$  such that  $Pl(A) > 0$ ,  
 124 the conditional mass function  $m(\cdot|A)$  is defined as the orthogonal sum of  $m$  and the mass  
 125 function  $m_A$  such that  $m(A) = 1$ . Conversely, given a conditional mass function  $m_0$  given  
 126  $A$  (expressing one's beliefs in a context where it is only known that the truth lies in  $A$ ), its  
 127 *conditional embedding* [73] is the least precise mass function  $m$  on  $\Theta$  such that  $m(\cdot|A) = m_0$ ;  
 128 it is obtained by transferring each mass  $m_0(C)$  to  $C \cup \bar{A}$ , for all  $C \subseteq A$ . Conditional  
 129 embedding is a form of "deconditioning", i.e., it performs the inverse of conditioning.

130 *Plausibility-probability transformation.* Once a mass function representing the combined ev-  
 131 idence has been computed, it is often used to make a decision. Decision-making methods  
 132 in DST are reviewed in [18]. Here, we will use the simplest method [11], which consists in  
 133 computing a probability distribution on  $\Theta$  by normalizing the plausibilities of the singletons,

$$\forall \theta \in \Theta, \quad p(\theta) = \frac{pl(\theta)}{\sum_{k=1}^K pl(\theta_k)}. \quad (3)$$

134 Once probabilities have been computed, a decision can be made by maximizing the expected  
 135 utility. We note that this method fits well with Dempster's rule, as the plausibility of the  
 136 singletons can be easily computed from (2) without computing the whole combined mass  
 137 function.

138 *2.2. Modeling the reliability of evidence*

139 In the DST framework, the reliability of a source of information can be taken into  
 140 account using the *discounting* operation, which transforms a mass function into a weaker,  
 141 less informative one and thus allows us to combine information from unreliable sources [69].  
 142 Let  $m$  be a mass function on  $\Theta$  and  $\beta$  a real number in  $[0, 1]$  interpreted as the degree  
 143 of belief that the source mass function  $m$  is reliable. The discounting operation [69] with  
 144 discount rate  $1 - \beta$  transforms mass function  $m$  into a less informative one  ${}^\beta m$  defined as a  
 145 weighted sum of  $m$  and the vacuous mass function  $m_\gamma$ , with coefficients  $\beta$  and  $1 - \beta$ :

$${}^\beta m = \beta m + (1 - \beta) m_\gamma. \quad (4)$$

146 In the rest of this paper, we will refer to  $\beta$  as a *reliability coefficient*. When  $\beta = 1$ , we accept  
 147 the mass function  $m$  provided by the source and take it as a description of our knowledge;  
 148 when  $\beta = 0$ , we reject it and are left with the vacuous mass function  $m_\gamma$ .

149 The discounting operation plays an important role in many applications of DST, where  
 150 it makes it possible to take into account “meta-knowledge” about the reliability of a source  
 151 of information. It can be justified as follows [74]. Assume that  $m$  is provided by a source  
 152 that may be reliable ( $R$ ) or not ( $\neg R$ ). If the source is reliable, we adopt its opinion as ours,  
 153 i.e., we set  $m(\cdot|R) = m$ . If it is not reliable, then it leaves us in a state of total ignorance,  
 154 i.e.,  $m(\cdot|\neg R) = m_\gamma$ . Furthermore, assume that we have the following mass function on  
 155  $\mathcal{R} = \{R, \neg R\}$ :  $m_{\mathcal{R}}(\{R\}) = \beta$  and  $m_{\mathcal{R}}(\mathcal{R}) = 1 - \beta$ , i.e., our degree of belief that the source  
 156 is reliable is equal to  $\beta$ . Then, combining the conditional embedding of  $m(\cdot|R)$  with  $m_{\mathcal{R}}$   
 157 yields precisely  ${}^\beta m$  in (4), after marginalizing on  $\Theta$ .

158 *Contextual discounting.* In [58], the authors generalize the discounting operation using the  
 159 notion of *contextual discounting*, which makes it possible to account for richer metaknowl-  
 160 edge about the reliability of a source in different contexts, i.e., conditionally on different  
 161 hypotheses regarding the variable of interest. In the corresponding refined model,  $m(\cdot|R)$   
 162 and  $m(\cdot|\neg R)$  are defined as before, but our beliefs about the reliability of the source are  
 163 now defined by  $K$  coefficients  $\beta_1, \dots, \beta_K$ , one for each state in  $\Theta$ . More specifically, we have  
 164  $K$  conditional mass functions defined by  $m_{\mathcal{R}}(\{R\}|\theta_k) = \beta_k$  and  $m_{\mathcal{R}}(\mathcal{R}|\theta_k) = 1 - \beta_k$ , for  
 165  $k = 1, \dots, K$ . In this model,  $\beta_k$  is, thus, the degree of belief that the source of information  
 166 is reliable, given that the true state is  $\theta_k$ . As shown in [58], combining the conditional em-  
 167 beddings of  $m(\cdot|R)$  and  $m_{\mathcal{R}}(\cdot|\theta_k)$  for  $k = 1, \dots, K$  by Dempster’s rule yields the following  
 168 discounted mass function,

$${}^\beta m(A) = \sum_{B \subseteq A} m(B) \left( \prod_{\theta_k \in A \setminus B} (1 - \beta_k) \prod_{\theta_l \in \bar{A}} \beta_l \right) \quad (5)$$

169 for all  $A \subseteq \Theta$ , where  $\beta = (\beta_1, \dots, \beta_K)$  is the vector of all reliability coefficients, and a  
 170 product of terms is equal to 1 if the index set is empty. In many applications, we actually  
 171 do not need to compute the whole mass function (5): we can compute only the associated

172 contour function  ${}^\beta pl$ , which is all we need for decision-making. As shown in [58], this contour  
 173 function is equal to

$${}^\beta pl(\theta_k) = 1 - \beta_k + \beta_k pl(\theta_k), \quad k = 1, \dots, K. \quad (6)$$

174 It can be computed in linear time with respect to the size of  $\Theta$ , instead of exponential time  
 175 for  ${}^\beta m$ . An evidential  $k$  nearest neighbor rule based on the contextual discounting operation  
 176 was introduced in [22].

**Example 1.** Consider a simplified diagnostic problem in which a patient may have one of two diseases denoted by  $\theta_1$  and  $\theta_2$ . Assume that  $\theta_1$  is a heart disease while  $\theta_2$  is a lung disease. A cardiologist examines the patient and describes his opinion by the following mass function on  $\Theta = \{\theta_1, \theta_2\}$ :  $m(\{\theta_1\}) = 0.7$ ,  $m(\{\theta_2\}) = 0.2$ ,  $m(\Theta) = 0.1$ , i.e., his degrees of belief in  $\theta_1$  and  $\theta_2$  are, respectively, 0.7 and 0.2. Furthermore, suppose that the cardiologist is fully reliable to diagnose heart diseases ( $\beta_1 = 1$ ), i.e., if the true state of the patient is  $\theta_1$ , the physician's opinion can be fully trusted, whereas he is only 60% reliable to diagnose lung diseases ( $\beta_2 = 0.6$ ), i.e., if  $\theta_2$  is the true disease, there only is only 60% chance that the physician's diagnostic is relevant. Applying formula (5) to  $m$  gives the following discounted mass function:

$$\begin{aligned} {}^\beta m(\{\theta_1\}) &= \beta_2 m(\{\theta_1\}) = 0.42 \\ {}^\beta m(\{\theta_2\}) &= \beta_1 m(\{\theta_2\}) = 0.2 \\ {}^\beta m(\Theta) &= 1 - \beta_2 m(\{\theta_1\}) - \beta_1 m(\{\theta_2\}) = 0.38. \end{aligned}$$

The contour function of the original mass function  $m$  is

$$\begin{aligned} pl(\{\theta_1\}) &= 0.7 + 0.1 = 0.8 \\ pl(\{\theta_2\}) &= 0.2 + 0.1 = 0.3. \end{aligned}$$

After contextual discounting, we get

$$\begin{aligned} {}^\beta pl(\{\theta_1\}) &= 0.42 + 0.38 = 0.8 \\ {}^\beta pl(\{\theta_2\}) &= 0.2 + 0.38 = 0.58. \end{aligned}$$

177 We can check that  ${}^\beta pl(\{\theta_1\}) = 1 - 1 + 1 \times 0.8$  and  ${}^\beta pl(\{\theta_2\}) = 1 - 0.6 + 0.6 \times 0.3$ , which is  
 178 consistent with (6).

### 179 2.3. Evidential neural network

180 In [16], Dencœux proposed a DST-based evidential neural network (ENN) classifier in  
 181 which mass functions are computed based on distances between the input vector and pro-  
 182 totypes. As shown in Figure 1, the ENN model comprises a prototype activation layer, a  
 183 mass calculation layer, and a combination layer.

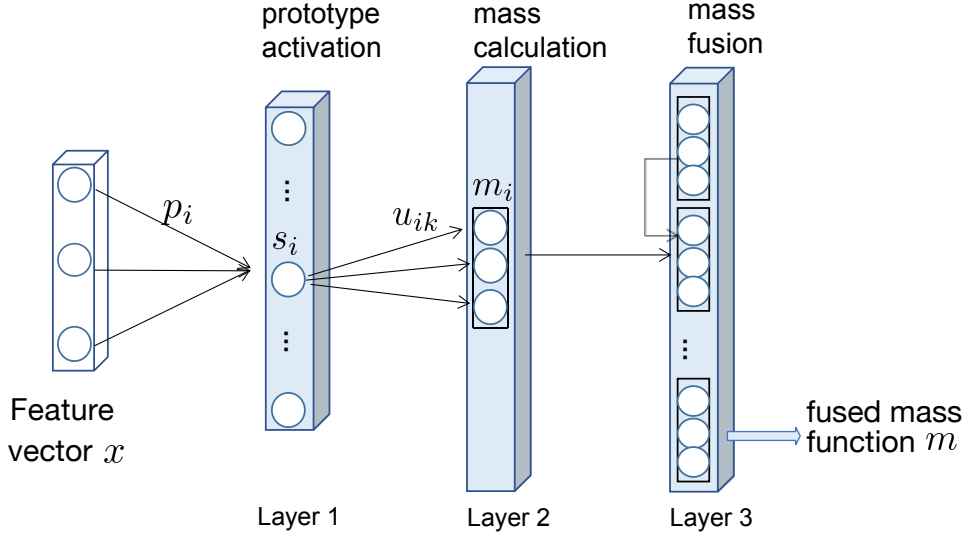


Figure 1: The evidential neural network model.

184 The prototype activation layer comprises  $I$  units, whose weight vectors are prototypes  
 185  $\mathbf{p}_1, \dots, \mathbf{p}_I$  in input space. The activation of unit  $i$  in the prototype layer is

$$s_i = \alpha_i \exp(-\gamma_i \|\mathbf{x} - \mathbf{p}_i\|^2), \quad (7)$$

186 where  $\gamma_i > 0$  and  $\alpha_i \in [0, 1]$  are two parameters. Each quantity  $s_i$  can be interpreted as a  
 187 degree of similarity between input vector  $\mathbf{x}$  and prototype  $\mathbf{p}_i$ .

The second hidden layer computes mass functions  $m_i$  representing the evidence of each prototype  $\mathbf{p}_i$ , using the following equations:

$$m_i(\{\theta_k\}) = u_{ik}s_i, \quad k = 1, \dots, K, \quad (8a)$$

$$m_i(\Theta) = 1 - s_i, \quad (8b)$$

188 where  $u_{ik}$  is the membership degree of prototype  $i$  to class  $\theta_k$ , and  $\sum_{k=1}^K u_{ik} = 1$ . The mass  
 189 function  $m_i$  can thus be seen as a discounted Bayesian mass function, with a discount rate  
 190  $1 - s_i$ ; its focal sets are singletons and  $\Theta$ . The mass assigned to  $\Theta$  increases with the distance  
 191 between  $\mathbf{x}$  and  $\mathbf{p}_i$ . Finally, the third layer combines the  $I$  mass functions  $m_1, \dots, m_I$  using  
 192 Dempster's rule (1). The output mass function  $m = \bigoplus_{i=1}^I m_i$  is a discounted Bayesian mass  
 193 function that summarizes the evidence of the  $I$  prototypes.

194 The idea of applying the above model to features extracted by a convolutional neural  
 195 network (CNN) was first proposed by Tong et al. in [79]. In this approach, the ENN module  
 196 becomes an “evidential layer”, which is plugged into the output of a CNN instead of the usual  
 197 softmax layer. The feature extraction and evidential modules are trained simultaneously.  
 198 Huang et al. applied the ENN model to medical image segmentation within a deep evidential  
 199 segmentation network [37].

200 **Remark 1.** The approach described in this section should not be confused with the “eviden-  
 201 tial deep learning” approach introduced in [68] and applied to brain tumor segmentation in

202 [93]. The latter approach is based on learning the parameters of a Dirichlet distribution that  
203 represents second-order uncertainty on the class probabilities. Although the parameters of  
204 the Dirichlet distribution can be formally identified to a mass function whose focal sets are  
205 the singletons  $\{\theta_k\}$  and the whole frame  $\Theta$ , this is actually a Bayesian approach that learns  
206 a probability distribution over the class probabilities through a suitable loss function.

#### 207 2.4. Multimodal medical image fusion

208 Multimodal medical image fusion can be performed at the pixel, feature or decision  
209 level. Pixel-level fusion is the traditional approach; it can be conducted directly in the  
210 spatial domain or indirectly through the application of transformations and representations.  
211 The fusion of high-level features is typically performed by a neural network learning a shared  
212 representation or a joint embedding space derived from multimodal features. Decision fusion  
213 consists in pooling decisions made independently from different image modalities; it can be  
214 performed with traditional or deep-learning approaches. In the following, we review previous  
215 work on multimodal medical image fusion, emphasizing the distinction between traditional  
216 and deep-learning approaches.

##### 217 2.4.1. Traditional approaches

218 Traditional fusion methods aim at combining relevant information (either pixels them-  
219 selves or low-level image features) from multiple images to produce a single fused image with  
220 enhanced features for further analysis. Four main approaches have been proposed: multi-  
221 scale transformation, sparse representation extraction, edge-preserving filters, and meta-  
222 heuristic optimization. The first three approaches focus on effective image representation,  
223 while the last one aims at combining the represented features efficiently.

224 The multi-scale transform approach decomposes images into different scales or frequency  
225 components using techniques such as wavelet transform [72], contourlet transforms [86],  
226 pyramid transforms [23] or curvelet transform [3], allowing relevant features from each source  
227 image to be combined. Sparse representation extraction assumes that multimodal images can  
228 be represented as a sparse linear combination of basis functions; search techniques such as  
229 dictionary learning [43] or sparse coding with dictionary learning [81] are used to obtain the  
230 sparse image representation and to merge images focusing on the most important features.  
231 Edge-preserving filters ensure the preservation of edges while smoothing images to ensure  
232 the fusion of critical features without blurring [75]. Commonly used filters include bilateral  
233 filters [47], guided filters [59], anisotropic diffusion [82], and total variation minimization  
234 [91]. The three above approaches can be used independently or in combination, which often  
235 yields better results. For example, in [35], Hu et al. propose a multimodal medical image  
236 fusion method based on separable dictionary learning and Gabor filtering; in [83], Wang et al.  
237 describe a multimodal medical image fusion method using Laplacian pyramid and adaptive  
238 sparse representations; in [51], Liu et al. introduce a general image fusion framework based  
239 on multi-scale transform and sparse representation.

240 In addition to studying effective image representations, a complementary research direc-  
241 tion has been to design meta-heuristic optimization algorithms allowing one to find the best  
242 fusion parameters for combining features obtained by different transform, sparse or fitting

243 algorithms. Many approaches use meta-heuristic optimization techniques such as genetic  
244 algorithms [3], particle swarm optimisation [77] or ant colony optimisation [70].

#### 245 2.4.2. Deep learning approaches

246 Recent advances in deep learning have allowed breakthroughs in medical image fusion  
247 by making it to learn a joint embedding or a shared representation space from multiple  
248 features. Recent techniques include adversarial learning [67], co-training [89], multi-kernel  
249 learning [57], multi-task learning [52], etc. These methods exploit the ability of neural  
250 networks to extract meaningful representations and perform fusion in high-level feature  
251 spaces with learnable feature fusion rules. These approaches enable more sophisticated and  
252 robust image fusion, capable of handling complex relationships and producing high-quality  
253 fused image features. Here, we summarize three important models commonly used for multi-  
254 model medical image fusion.

255 *Convolutional Neural Networks.* Convolutional neural networks (CNNs) are widely used in  
256 image processing due to their strong feature representation capability. Within CNNs, various  
257 fusion operations can be used to effectively integrate information from different imaging  
258 modalities. Such operations include but are not limited to, concatenation, element-wise  
259 addition and multiplication, weighted sum, max pooling, etc. Fusion can occur at different  
260 stages of the network, i.e., early, middle, or late stages.

261 Early fusion stacks different modalities along a channel dimension and feeds into a single  
262 CNN [49]. This is the simplest operation but it requires high image registration quality. In  
263 the case of middle fusion, separate CNN branches are employed to extract features from each  
264 modality, which are subsequently concatenated at the feature level or fused in a particular  
265 common representation space. More recently, transformer-based CNN architectures, such  
266 as the Vision Transformer (ViT) [33], have also demonstrated considerable versatility in  
267 handling diverse types of data with the introduction of an attention mechanism [46]. CNNs  
268 can also be integrated with some traditional fusion ideas to obtain more robust fusion results  
269 using, e.g., the multiscale transformer [76] or multiscale residual pyramid attention network  
270 [28].

271 In contrast to the emphasis on image pixels or features in earlier fusion techniques, later  
272 fusion places greater importance on the aggregation of high-level decisions. It integrates in-  
273 formation derived from preliminary classifications with the application of appropriate fusion  
274 rules. Approaches can be classified into two main categories: 1) *hard fusion* methods, which  
275 merge logical information membership values, such as model ensembling with majority or  
276 average voting [42]; and 2) *soft fusion* methods, where classifiers assign numerical values to  
277 reflect their confidence in decisions, as exemplified by fuzzy voting [32, 27].

278 *Encoder-Decoder Networks.* Encoder-decoder networks are another type of convolutional  
279 neural network commonly used for image segmentation and reconstruction. Within the  
280 encoder-decoder network, multiple encoders are used to extract deep features from each  
281 modality. These features are subsequently integrated either through a straightforward con-  
282 catenation process or through a latent layer or learnt joint embedding space. The fused  
283 features are then passed to the decoder to produce the final image. Compared with CNNs,

284 the Encoder-decoder architecture offers a more structured and effective fusion framework  
285 with enhanced feature representation, precise spatial alignment, and flexible and effective fu-  
286 sion strategies. Multimodal Transformer (MMT) [85] is one of the most sophisticated forms  
287 of multimodal encoder-decoder networks that employ self-attention mechanisms to integrate  
288 and process multimodal data in an effective manner; nnFormer [92] has been identified as  
289 the most advanced model for multimodal MRI brain tumor segmentation.

290 *Generative Adversarial Networks.* Generative Adversarial Networks (GANs), composed of  
291 a generator and a discriminator, are capable of learning complex relationships between dis-  
292 parate modalities through the generation of highly realistic images via unsupervised adver-  
293 sarial training [30]. In the context of multimodal medical image fusion, the generator learns  
294 to generate a fused image that combines the semantic features of the inputs from different  
295 modalities. The discriminator guides the generator to produce high-quality fused images  
296 by distinguishing between the fused and the real images. GAN-based fusion methods are  
297 particularly useful for advanced medical image fusion tasks where the quality and realism  
298 of the fused image are of paramount importance, such as the combination of structural and  
299 functional imaging modalities. For example, in [88], the authors propose a conditional gen-  
300 erative adversarial network with a transformer for multimodal image fusion by introducing  
301 a wavelet fusion module to maintain long-distance dependencies across domains; in [67], the  
302 authors introduce an unsupervised medical fusion generative adversarial network to generate  
303 an image with CT bone structure and MRI soft tissue contrast by fusing CT and MRI image  
304 sequences.

305 Although a lot of research has been devoted to the study of multimodal medical image  
306 segmentation and promising experimental results have been obtained, modeling the relia-  
307 bility of each modality in a given context and quantifying the uncertainty on the outcome  
308 of the fusion process remain challenging research questions. In this paper, we address these  
309 questions using a deep evidential fusion framework combining deep learning with DST, and  
310 taking into account the reliability of each of the modalities being combined. The proposed  
311 decision-fusion framework is described in detail in the following section.

### 312 **3. Proposed framework**

313 The main idea of this paper is to hybridize a deep evidential fusion framework with un-  
314 certainty quantification and reliability learning for multimodal medical image segmentation  
315 under the framework of DST. The architecture of the system is described in Section 3.1, and  
316 the loss function used to train the whole framework end-to-end is presented in Section 3.2.

#### 317 *3.1. Architecture*

318 The proposed framework is depicted in Figure 2. Features are first extracted from differ-  
319 ent modalities using independent encoder-decoder feature-extraction (FE) modules. The fea-  
320 tures from each modality are then transformed into mass functions using evidence mapping  
321 (EM) modules. Finally, mass functions are discounted and combined in a multi-modality  
322 evidence fusion (MMEF) module. These modules are described in greater detail below.

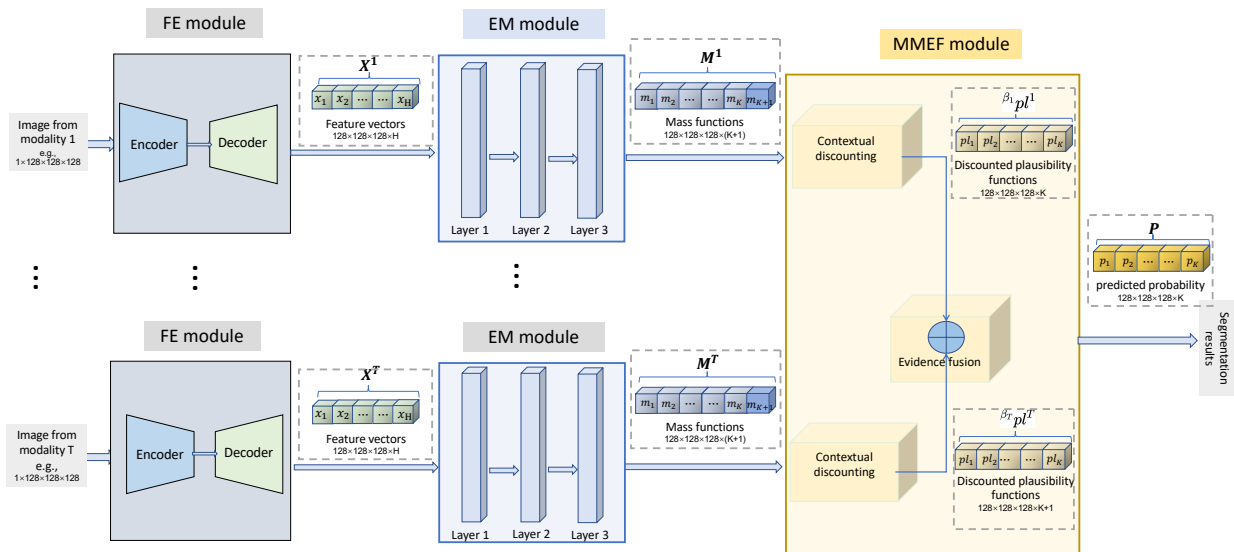


Figure 2: The proposed deep evidential fusion framework. It is composed of encoder-decoder *feature extraction* (FE) modules that represent images using deep features, *evidence mapping* (EM) modules that map deep features into mass functions, and a *multimodal evidence fusion* (MMEF) module that combines evidence from different modalities.

### 3.1.1. Feature-extraction (FE) module

Deep neural network architectures have been shown to be very powerful for extracting relevant information from high-dimensional data. Our approach is compatible with any deep FE architecture. The baseline model considered in this paper is UNet [41], a foundational medical image segmentation model. As illustrated in Figure 3, a UNet-based feature extraction module incorporates residual connections within each layer, following the same architecture as in [37]. Each layer of the module comprises encoding and decoding paths, connected by skip connections. In the encoding path (represented by blue blocks), the data undergoes downsampling through stride convolutions, while the decoding path (represented by green blocks) employs stride transpose convolutions for upsampling. The bottom layer, represented by the gray block, serves as the base connection without performing any down or up-sampling of the data. In Section 4.3, in addition to UNet, we will also consider the more recent nnUNet [40] and nnFormer [92] models as alternative FE modules. The settings of these modules will be described in Section 4.1.

### 3.1.2. Evidence mapping (EM) module

The EM module is based on the ENN architecture recalled in Section 2.3. It is identical to that described in [37]. As illustrated in Figure 2, we have one such module for each modality. The input to each module is a tensor containing the  $H$  features extracted for each voxel. The prototypes are, thus, vectors in the  $H$ -dimensional space of features extracted from modality  $t$  images by the FE module. As explained in Section 2.3, a prototype layer first computes the similarities between feature vectors and prototypes using (7). The next layer

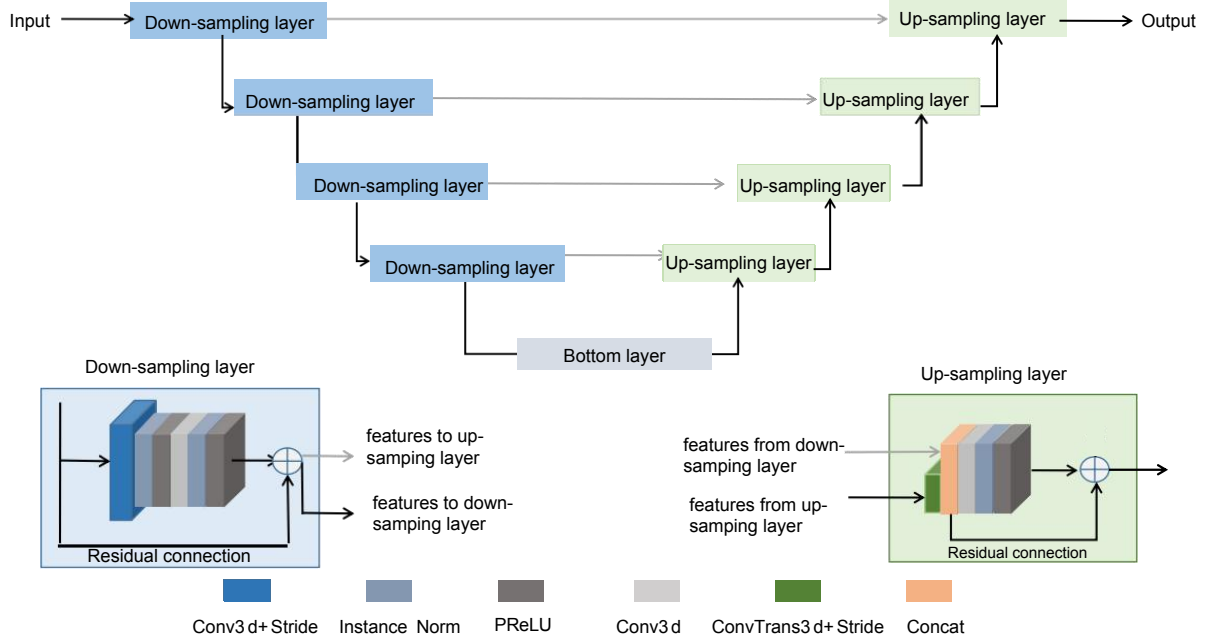


Figure 3: Schematic description of a UNet-based FE module. The network consists of a contracting path (down-sampling layers) and an expansive path (up-sampling layers), which gives it the u-shaped architecture. Reproduced based on [41].

344 computes mass functions for each prototype using (8) (see Figure 1). Finally, the prototype-  
 345 based mass functions are combined by Dempster’s rule (1) in a third layer. Denoting by  
 346  $\Theta = \{\theta_1, \dots, \theta_K\}$  the set of classes, the EM module thus computes, for each voxel  $n$  and  
 347 modality  $t$ , a mass function<sup>2</sup>  $m_n^t$  with focal sets  $\{\theta_k\}$ ,  $k = 1, \dots, K$  and  $\Theta$ . The mass  $m_n^t(\Theta)$   
 348 is a measure of the segmentation uncertainty for classifying voxel  $n$  in the image of modality  
 349  $t$ .

### 350 3.1.3. Multi-modality evidence fusion (MMEF) module

This module first transforms the contour functions from the EM modules using the contextual discounting operation recalled in Section 2.2. The contour function for voxel  $n$  and modality  $t$  is obtained from mass function  $m_n^t$  as

$$pl_n^t(\theta_k) = m_n^t(\{\theta_k\}) + m_n^t(\Theta), \quad k = 1, \dots, K.$$

351 Using (6), the discounted contour function is given by

$$\beta^t pl_n^t(\theta_k) = 1 - \beta_k^t + \beta_k^t pl_n^t(\theta_k), \quad k = 1, \dots, K, \quad (9)$$

<sup>2</sup>Throughout this paper, we use an upper index  $t$  to denote modalities, and lower indices  $n$  and  $k$  to denote, respectively, voxels and classes.

352 where  $\boldsymbol{\beta}^t = (\beta_1^t, \dots, \beta_K^t)$  is the vector of discounting (reliability) coefficients for modality  $t$ .  
 353 We recall that  $\beta_k^t$  represents our degree of belief that the modality  $t$  is reliable when it is  
 354 known that the actual class of voxel  $n$  is  $\theta_k$ . From (2), the combined contour function at  
 355 voxel  $n$  can then be computed up to a multiplicative constant by multiplying the contour  
 356 functions for the  $T$  modalities as

$$\beta p_n(\theta_k) \propto \prod_{t=1}^T \beta^t p_n^t(\theta_k), \quad k = 1, \dots, K,$$

357 where  $\boldsymbol{\beta} = (\boldsymbol{\beta}^1, \dots, \boldsymbol{\beta}^T)$  is the vector of  $KT$  reliability coefficients for the the  $K$  classes and  
 358  $T$  modalities. Finally, the predicted probability distribution  $P_n$  for voxel  $n$  after combining  
 359 evidence from the  $T$  modalities is obtained from (3) as

$$\beta p_n(\theta_k) = \frac{\beta p_n(\theta_k)}{\sum_{l=1}^K \beta p_n(\theta_l)} = \frac{\prod_{t=1}^T (1 - \beta_k^t + \beta_k^t p_n^t(\theta_k))}{\sum_{l=1}^K \prod_{t=1}^T (1 - \beta_l^t + \beta_l^t p_n^t(\theta_l))}, \quad k = 1, \dots, K. \quad (10)$$

360 The learnable parameters in this module are the  $KT$  reliability coefficients in vector  $\boldsymbol{\beta}$ .

### 361 3.2. Loss function

The whole framework is optimized by minimizing the following loss function,

$$\text{loss} = \text{loss}_s + \text{loss}_f,$$

362 where

- 363 • The term  $\text{loss}_s$  is the Dice loss quantifying the segmentation performance of each source  
 364 modality independently, with

$$\text{loss}_s = \sum_{t=1}^T \left[ 1 - \frac{2 \sum_{n=1}^N \sum_{k=1}^K m_n^t(\{\theta_k\}) \times G_{kn}}{\sum_{n=1}^N \sum_{k=1}^K (m_n^t(\{\theta_k\}) + G_{kn})} \right], \quad (11)$$

365 where  $N$  is the number of voxels, and  $G_{kn} = 1$  if voxel  $n$  belongs to class  $\theta_k$ , and  
 366  $G_{kn} = 0$  otherwise;

- 367 • The term  $\text{loss}_f$  quantifies the segmentation performance after combination:

$$\text{loss}_f = 1 - \frac{2 \sum_{n=1}^N \sum_{k=1}^K \beta p_n(\theta_k) \times G_{kn}}{\sum_{n=1}^N \sum_{k=1}^K \beta p_n(\theta_k) + G_{kn}}, \quad (12)$$

368 where  $\beta p_n$  is the predicted probability distribution for voxel  $n$  given by (10).

369 The learnable parameters are the weights of the FE module, the prototypes and associ-  
 370 ated parameters  $\alpha_i$ ,  $\gamma_i$  and  $u_{ik}$  of the EM module, and the reliability coefficients  $\beta_k^t$  in the  
 371 MMEF module. Learning the reliability coefficients is an original feature of our approach.  
 372 As shown in Sections 4.2 and 4.3, these coefficients can allow us to gain some insight into  
 373 the multi-modality segmentation process.

## 374 4. Experiments and results

375 In this section, the proposed framework described in Section 3 is applied to two real  
376 multimodal medical image datasets. The experimental settings are first described in Section  
377 4.1. The results on the two datasets are then reported in Sections 4.2 and 4.3.

### 378 4.1. Experimental settings

379 *Datasets.* The proposed framework was tested on two multimodal medical image datasets.

380 The *PET-CT lymphoma dataset* contains 3D images from 173 patients who were diag-  
381 nosed with large B-cell lymphomas and underwent PET-CT examination<sup>3</sup>. For lymphoma  
382 segmentation, PET imaging helps identify active tumor sites by highlighting areas of in-  
383 creased metabolic activity. In contrast, CT imaging provides anatomical information about  
384 the size, shape, location, and surrounding structures of lymphoma tumors. While PET  
385 image makes it possible to obtain functional information about the tumor and surrounding  
386 tissues, CT images provide complementary anatomical details allowing for more accurate  
387 segmentation. The lymphomas in mask images were delineated manually by experts and  
388 considered as ground truth. Figure 4 shows an example of PET and CT images of a patient  
389 with lymphomas. The PET and CT images and the corresponding mask images have differ-  
390 ent sizes and spatial resolutions due to the use of different imaging machines and operations.  
391 For CT images, the size varies from  $267 \times 512 \times 512$  to  $478 \times 512 \times 512$ . For PET images,  
392 the size varies from  $276 \times 144 \times 144$  to  $407 \times 256 \times 256$ .

393 The *multi-MRI brain tumor* dataset was made available for the BraTS2021 challenge  
394 [5]. The original BraTS2021 dataset comprises training, validation, and test sets with,  
395 respectively, 1251, 219, and 570 cases. There are four modalities: FLAIR, T1Gd, T1, and  
396 T2 with  $240 \times 240 \times 155$  voxels. Figure 5 shows examples of four-modality MRI slices for  
397 one patient. The appearance of brain tumors varies in different modalities [5]. T1Gd MRI  
398 images are obtained following the administration of a gadolinium-based contrast agent that  
399 enhances areas with disrupted blood-brain barrier such as tumor regions, making tumors  
400 appear hyperintense (bright) and improving the visibility of tumor margins. FLAIR MRI  
401 images suppress the signal from cerebrospinal fluid (CSF), highlighting pathological changes  
402 while suppressing the CSF signal. T2 MRI images are sensitive to tissue water content  
403 and provide good contrast between soft tissues. Tumors with increased water content often  
404 appear hyperintense (bright) on T2 images. T1 MRI images are crucial for identifying tumor  
405 location and structural details by their excellent anatomical detail. Annotations of scans  
406 comprise gadolinium (GD)-enhancing tumor (ET), necrotic and non-enhancing tumor core  
407 (NRC/NET), and peritumoral edema (ED). The task of the BraTS2021 challenge was to  
408 segment the images into three overlapping regions: ET, tumor core (TC, the union of ET  
409 and NRC/NET), and whole tumor (WT, the union of ET, NRC/NET, and ED). In this  
410 work, we evaluated the segmentation performances with respect to these three overlapping  
411 regions to allow a fair comparison with other state-of-the-art methods. Additionally, we also

---

<sup>3</sup>The study was approved as a retrospective study by the Henri Becquerel Center Institutional Review Board.

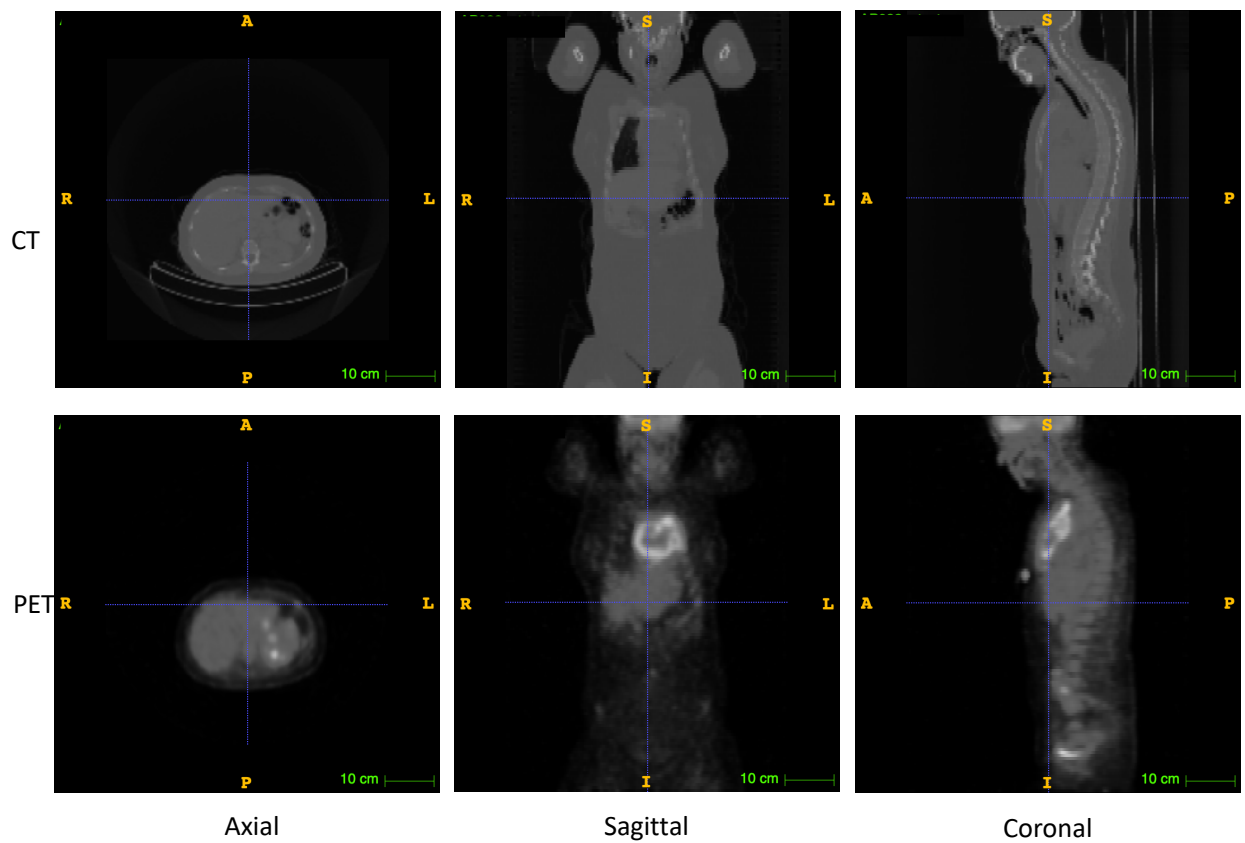


Figure 4: Example of a patient with lymphomas. The first and second rows showcase, respectively, CT and PET slices, depicting axial, sagittal, and coronal views. The lymphomas correspond to the bright regions in PET slices.

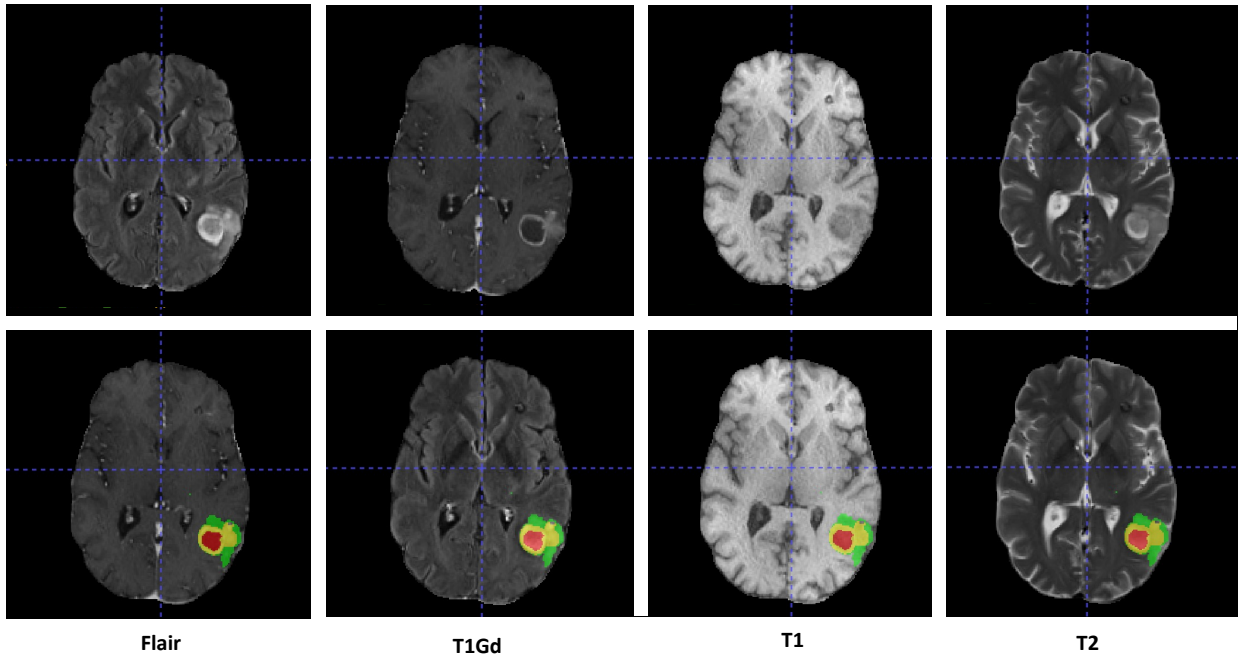


Figure 5: Examples of a patient with brain tumors in four MRI modalities: FLAIR, T1Gd, T1, and T2. The first and second rows show, respectively, the original images and the images with tumor masks for the three classes: peritumoral edema (ED, green), enhancing tumor (ET, yellow), and necrotic tumor core or non-enhancing tumor (NCR/NET, red).

412 compared the results with respect to the three original non-overlapping tumor regions to  
 413 highlight the impact of contextual discounting on subregion segmentation.

414 *Pre-processing.* For the PET-CT dataset, we first normalized the PET, CT and mask images:  
 415 (1) for PET images, we applied a random intensity shift and scale to each channel with a shift  
 416 value of 0 and scale value of 0.1; (2) for CT images, the shift and scale values were set to 1000  
 417 and 1/2000; (3) for mask images, the intensity value was normalized into the  $[0, 1]$  interval by  
 418 replacing the outside value by 1. We then resized the PET and CT images to  $256 \times 256 \times 128$   
 419 by linear interpolation and mask images to  $256 \times 256 \times 128$  by nearest neighbor interpolation.  
 420 Lastly, CT and PET images were registered using B-spline interpolation. Following [37], we  
 421 randomly divided the 173 scans into subsets of size 138, 17, and 18 for, respectively, training,  
 422 validation, and test. The training process was then repeated five times to test the stability  
 423 of our framework, with different data used exactly once as the validation and test data.

424 For the BraTS2021 dataset, we used the same pre-processing operation as in [60]. We  
 425 first performed a min-max scaling operation and clipped intensity values to standardize  
 426 all volumes; we then cropped/padded the volumes to a fixed size of  $128 \times 128 \times 128$  by  
 427 removing the unnecessary background (the cropping/padding operation was only applied  
 428 to training data). No data augmentation technique was applied, and no additional data  
 429 was used in this study. Since the ground truth labels are unavailable for the validation  
 430 and test sets, we trained and tested our framework with the training set. Following [60],  
 431 we randomly divided the 1251 training scans into subsets of 834, 208, and 209 cases for

432 training, validation, and testing, respectively. The process was repeated five times to test  
433 the stability of our framework. All the preprocessing methods mentioned in this paper can  
434 be found in the SimpleITK [53] toolkit.

435 All the compared methods used the same dataset composition and pre-processing opera-  
436 tions. They were implemented in Python with the PyTorch-based medical image framework  
437 MONAI<sup>4</sup>.

438 *Parameter initialization and learning.* At the FE stage, the number of filters in UNet was  
439 set to (8, 16, 32, 64, 128) with kernel size equal to five and convolutional strides equal to  
440 (2, 2, 2, 2) for layers from left to right. For nnUNet used in Section 4.3, the kernel size  
441 was set to (3, (1, 1, 3), 3, 3) and the upsample kernel size was set to (2, 2, 1) with strides  
442 ((1, 1, 1), 2, 2, 1). For nnFormer used in Section 4.3, the crop size was set to (128, 128, 128)  
443 with embedding dimension set to 96 and the number of heads was set to (3, 6, 12, 24). The  
444 number of extracted features was  $H = 2$  for the PET-CT lymphoma dataset and  $H = 4$  for  
445 the multi-MRI BraTS2021 dataset.

446 To train our fusion framework, we proceeded in three steps. First, FE modules (i.e.,  
447 UNet, nnUNet, or nnFormer) were pre-trained independently for each modality during 50  
448 epochs. Then, the weights of the FE modules were fixed, and the parameters of the EM  
449 and MMEF modules were optimized. Finally, the whole framework was fine-tuned for a  
450 few epochs. The initial values of parameters  $\alpha_i$  and  $\gamma_i$  in the EM modules were set to 0.5  
451 and 0.01, and the membership degrees  $u_{ik}$  were initialized randomly by drawing uniform  
452 random numbers, and normalizing. We used,  $I = 10$  prototypes for the PET-CT lymphoma  
453 dataset, and  $I = 20$  prototypes for the more complex multi-MRI BraTS2021 dataset. These  
454 prototypes were randomly initialized from a normal distribution with zero mean and an  
455 identity covariance matrix. Details about the initialization of the EM module can be found  
456 in [37]. The reliability coefficients  $\beta_k^t$  in the MMEF module were initialized at 0.5.

457 For both datasets, we used the Adam optimization algorithm with an early stopping  
458 strategy: training was stopped when there was no improvement in performance on the  
459 validation set during ten epochs. The initial learning rate was set to 0.01. The batch size  
460 was set to 4. For all the compared methods, the model with the best performance on the  
461 validation set was saved as the final model for testing<sup>5</sup>.

462 *Evaluation criteria.* Although many authors have shown that segmentation performance can  
463 be improved by merging multimodal medical images into deep neural networks [63, 2], the  
464 reliability of information sources and the quality of uncertainty quantification have rarely  
465 been investigated. Here, the former issue will be addressed by analyzing the reliability  
466 coefficients  $\beta_k^t$  defined in Section 3.1.3. To assess the quality of uncertainty quantification,  
467 we will use three metrics: the *Brier score* [9], the *negative log-likelihood* (NLL), and *Expected*  
468 *Calibration Error* (ECE) [31]. These metrics provide a robust evaluation framework for the

---

<sup>4</sup>More details about how to use those models can be found in MONAI core tutorials <https://monai.io/started.html##monaicore>.

<sup>5</sup>The code is available at <https://github.com/iWeisskohl/Deep-evidential-fusion>.

469 uncertainty of the segmentation results, with smaller values indicating better performance.  
 470 Their definitions are recalled below.

The Brier Score and NLL are defined, respectively, as

$$\text{BS} = \frac{1}{N} \sum_{n=1}^N (P_n - G_n)^2,$$

and

$$\text{NLL} = - \sum_{n=1}^N G_n \log P_n + (1 - G_n) \log(1 - P_n),$$

471 where  $G_n$  is the ground truth of voxel  $n$ ,  $P_n$  is the predicted probability of voxel  $n$ , and  $N$   
 472 is the number of voxels.

The ECE measures the correspondence between predicted probabilities and ground truth. The output normalized plausibilities of the model are first discretized into equally spaced bins  $E_b$ ,  $b \in [1, B]$  ( $B = 10$  in this paper). The accuracy of bin  $E_b$  is defined as

$$\text{acc}(E_b) = \frac{1}{|E_b|} \sum_{n \in E_b} \mathbf{1}(S_n = G_n),$$

where  $S_n$  is the predicted class label for voxel  $n$  and  $\mathbf{1}(\cdot)$  is the indicator function. The average confidence of bin  $E_b$  is defined as

$$\text{conf}(E_b) = \frac{1}{|E_b|} \sum_{n \in E_b} P_n.$$

The ECE is the weighted average of the difference in accuracy and confidence of the bins:

$$\text{ECE} = \sum_{b=1}^B \frac{|E_b|}{N} | \text{acc}(E_b) - \text{conf}(E_b) |.$$

473 A model is perfectly calibrated when  $\text{acc}(E_b) = \text{conf}(E_b)$  for all  $b \in \{1, \dots, B\}$ , in which case  
 474  $\text{ECE} = 0$ .

475 Since our dataset has imbalanced foreground and background proportions, we only con-  
 476 sidered voxels belonging to the foreground or tumor region to calculate the above three  
 477 indices. For the PET-CT lymphoma dataset, focusing only on the tumor region is not easy  
 478 since the lymphomas are scattered throughout the whole body. Thus, we focused on the  
 479 foreground region for this dataset. For the BraTS2021 dataset, we followed the suggestion  
 480 from [66] to focus on the tumor region for the reliability evaluation. For each patient in the  
 481 test set, we defined a bounding box covering the foreground or tumor region and calculated  
 482 the corresponding values in this bounding box. For all segmentation performance criteria,  
 483 the reported results were obtained by calculating the criteria for each test 3D scan and then  
 484 averaging over the patients.

In addition to evaluating segmentation reliability, we also measured segmentation accuracy using the *Dice score*. In a segmentation task, the Dice score measures the volume of the overlapping region of the predicted object and the ground truth object as

$$\text{Dice} = \frac{2TP}{FP + 2TP + FN},$$

where  $TP$ ,  $FP$ , and  $FN$  denote, respectively, the numbers of true positive, false positive, and false negative voxels.

#### 4.2. Segmentation results on the PET-CT lymphoma dataset

*Segmentation uncertainty.* The results concerning uncertainty estimation are reported in Table 1. Our model (MMEF-UNet) was compared to

1. UNet with a softmax decision layer (the baseline);
2. UNet with Monte-Carlo (MC) dropout [29] and deep ensemble [45], two popular techniques for improving the uncertainty quantification capabilities of probabilistic deep neural networks;
3. ENN-UNet, composed of UNet as the FE module and the EM module in place of the softmax layer; this is the architecture studied in [37];
4. RBF-UNet, an alternative model composed of UNet and a radial-basis function (RBF) module in place of the softmax layer; as shown in [37], this model makes it possible to compute output belief functions that are similar to those computed by ENN-UNet.

We can remark that approaches 1 to 4 above implement pixel-level fusion, whereas our approach is based on decision-level fusion. As for uncertainty quantification, UNet, UNet-MC and UNet-Ensemble are probabilistic methods. UNet only computes point estimates of class probabilities without taking into account second-order uncertainty. UNet-MC applies dropout during both training and inference, sampling multiple forward passes to estimate uncertainty by averaging the predictions. UNet-Ensemble quantifies uncertainty by averaging the predictions obtained from multiple independently-trained models. In contrast, ENN-UNet and RBF-UNet are evidential methods: they both calculate belief functions to represent segmentation evidence and uncertainty under the DST framework. For UNet-MC, the dropout rate was set to 0.2 and the number of samples was set to five; we averaged the five output probabilities at each voxel as the final output of the model. For UNet-ensembles, the number of samples was set to five; the five output probabilities were then averaged at each voxel as the final output of the model. The settings of ENN-UNet and RBF-ENN are the same as those reported in [37].

From Table 1, we can see that Monte-Carlo dropout and deep ensembles do not significantly improve the segmentation reliability as compared to the baseline UNet model, as shown, e.g., by the higher NLL values. In contrast, the addition of the EM module to the FE module, as implemented in ENN-UNet, brings a significant improvement, particularly according to NLL; the RBF-UNet model yields similar results. The decision-fusion framework MMEF-UNet brings an additional improvement according to all three criteria (ECE,

Table 1: Means and standard errors of segmentation quality and reliability measures for MMEF-UNet and the referenced uncertainty quantification methods on the lymphoma dataset. The best results are in bold and the second best are underlined.

Model	ECE↓	Brier score ↓	NLL↓	Dice score ↑
UNet	$0.056 \pm 3.6 \times 10^{-3}$	$0.065 \pm 3.9 \times 10^{-3}$	$0.310 \pm 8.8 \times 10^{-2}$	$0.770 \pm 3.2 \times 10^{-2}$
UNet-MC	$0.053 \pm 4.6 \times 10^{-3}$	$0.062 \pm 4.9 \times 10^{-3}$	$0.400 \pm 8.7 \times 10^{-2}$	$0.801 \pm 1.1 \times 10^{-2}$
UNet-Ensemble	$0.063 \pm 7.6 \times 10^{-2}$	$0.064 \pm 4.0 \times 10^{-3}$	$0.343 \pm 7.2 \times 10^{-2}$	$0.802 \pm 6.7 \times 10^{-3}$
ENN-UNet	<u><math>0.050 \pm 3.5 \times 10^{-3}</math></u>	$0.062 \pm 3.9 \times 10^{-3}$	<u><math>0.191 \pm 1.4 \times 10^{-2}</math></u>	<u><math>0.805 \pm 7.1 \times 10^{-3}</math></u>
RBF-UNet	$0.051 \pm 3.3 \times 10^{-3}$	<u><math>0.061 \pm 0.9 \times 10^{-3}</math></u>	$0.193 \pm 1.3 \times 10^{-2}$	$0.802 \pm 6.9 \times 10^{-3}$
MMEF-UNet (ours)	<b><math>0.045 \pm 1.3 \times 10^{-3}</math></b>	<b><u><math>0.056 \pm 2.7 \times 10^{-3}</math></u></b>	<b><math>0.180 \pm 1.3 \times 10^{-2}</math></b>	<b><math>0.811 \pm 3.4 \times 10^{-2}</math></b>

520 Brier score, NLL) and outperforms the other models: specifically, we observe decreases of  
521 1.1%, 0.9%, and 13% in ECE, Brier score, and NLL, respectively, as compared to UNet.  
522 We can conclude that, compared to the baseline model, both the EM and MMEF modules  
523 contribute to a higher segmentation reliability.

524 These findings are, to some extent, confirmed by Figure 6, which shows the calibra-  
525 tion plots (also known as reliability diagrams) for the compared methods on the lymphoma  
526 dataset. Calibration plots are graphical representations showing how well the probabilistic  
527 predictions of a segmentation model are calibrated, i.e., how well confidence matches accu-  
528 racy. In the left graph of Figure 6, we can see that the curve corresponding to UNet-Ensemble  
529 is closer to the diagonal than those of UNet and UNet-MC, which indicates better calibra-  
530 tion. Looking at the right graph in Figure 6, we can see that the three DST-based models,  
531 ENN-UNet, RBF-UNet, and MMEF-UNet, have better calibration performance than the  
532 probabilistic ones, as shown by their calibration curves closer to the diagonal. Among them,  
533 MMEF-UNet shows the best calibration performance as ENN-UNet is slightly overconfident,  
534 while RBF-UNet is slightly underconfident.

535 *Segmentation accuracy.* The segmentation accuracy was measured by the Dice score, as  
536 shown in Table 1. Compared with the baseline model UNet, our proposal MMEF-UNet  
537 significantly increases segmentation performance, as shown by a 4.1% increase in the Dice  
538 score. Compared with the two DST-based deep evidential segmentation methods, MMEF-  
539 UNet has a higher Dice score (although the difference with ENN-UNet is not statistically  
540 significant). Figure 7 shows an example of visualized segmentation results obtained by UNet,  
541 ENN-UNet, RBF-UNet, and MMEF-UNet. We can see that UNet and RBF-UNet are more  
542 conservative (they correctly detect only a subset of the tumor voxels), while ENN-UNet is  
543 more radical (some of the voxels that do not belong to tumors are predicted as tumors).  
544 In contrast, the tumor regions predicted by MMEF-UNet better overlap the ground-truth  
545 tumor region, especially for the isolated lymphomas, which is also reflected by the promising  
546 Dice score value. These conclusions are consistent with the calibration trends displayed in  
547 Figure 6.

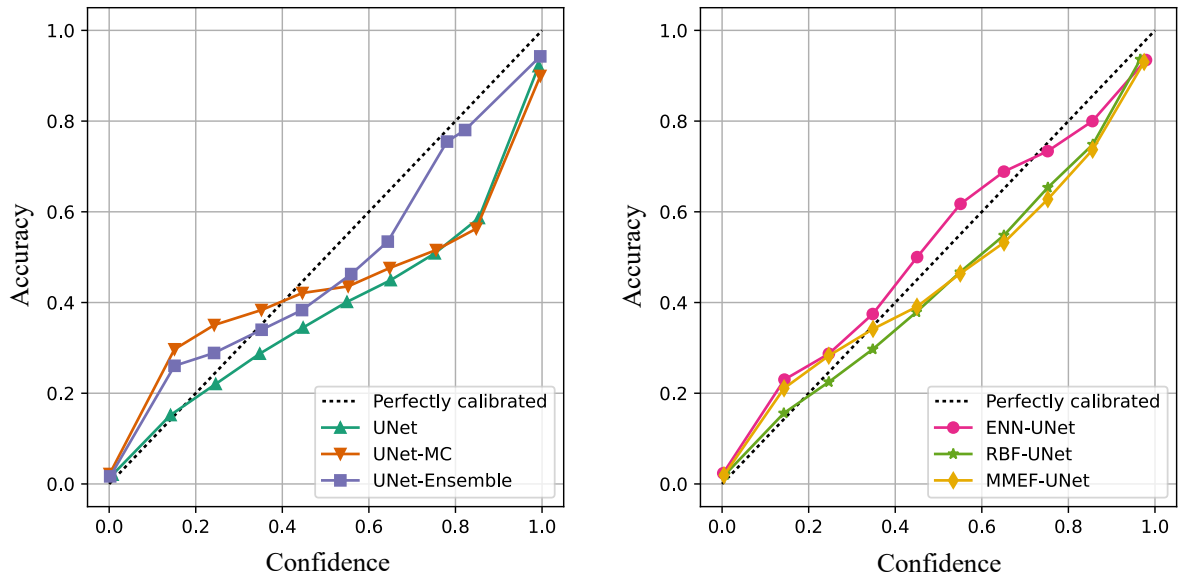


Figure 6: Calibration plots for probabilistic (left) and evidential (right) deep segmentation models.

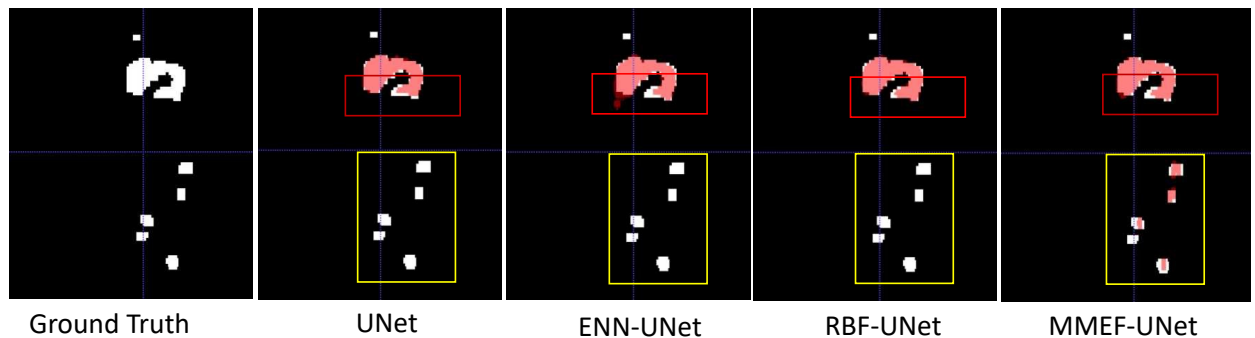


Figure 7: Examples of visualized segmentation results: from left to right, ground truth, and segmentation results obtained by UNet, ENN-UNet, RBF-UNet, and MMEF-UNet. The white and red regions represent, respectively, the ground truth and the segmentation result. Red and yellow boxes highlight the main differences in segmenting large and small isolated tumors.

Table 2: Estimated reliability coefficient  $\beta_k^t$  (means and standard errors) after training for the background and lymphoma classes and the two modalities. Higher values correspond to greater contribution to the segmentation.

$\beta_k^t$	background	lymphomas
PET	$0.999 \pm 8.9 \times 10^{-3}$	$0.996 \pm 4.5 \times 10^{-3}$
CT	$0.863 \pm 1.8 \times 10^{-2}$	$0.975 \pm 8.9 \times 10^{-3}$

Table 3: Segmentation quality and reliability of UNet and ENN-UNet applied to the lymphoma dataset with a single modality (CT or PET).

Model	ECE↓	Brier score ↓	NLL↓	Dice score ↑
UNet (CT)	$0.133 \pm 4.9 \times 10^{-3}$	$0.157 \pm 9.8 \times 10^{-3}$	$0.571 \pm 3.6 \times 10^{-2}$	$0.544 \pm 2.8 \times 10^{-2}$
UNet (PET)	$0.060 \pm 4.0 \times 10^{-3}$	$0.068 \pm 4.0 \times 10^{-3}$	$0.348 \pm 8.2 \times 10^{-2}$	$0.764 \pm 2.9 \times 10^{-2}$
ENN-UNet (CT)	$0.131 \pm 8.5 \times 10^{-3}$	$0.156 \pm 1.0 \times 10^{-2}$	$0.521 \pm 3.2 \times 10^{-2}$	$0.543 \pm 2.7 \times 10^{-2}$
ENN-UNet (PET)	$0.050 \pm 4.9 \times 10^{-3}$	$0.064 \pm 5.4 \times 10^{-3}$	$0.195 \pm 2.2 \times 10^{-2}$	$0.781 \pm 3.5 \times 10^{-2}$

548 *Analysis of reliability coefficients.* Table 2 reports the learned reliability coefficients. We can  
549 see that they are higher for the PET modality. This is consistent with domain knowledge, as  
550 mentioned in Section 4.1: PET images provide functional information about tumor activity  
551 and make it possible to identify active tumor sites, whereas CT images essentially provide  
552 detailed anatomical information (e.g., size, shape, and location) about lymph nodes and  
553 surrounding tissues and are used as a complement to PET images. This is also confirmed  
554 by the results presented in Table 3, showing that the performance of UNet and ENN-UNet  
555 with either CT alone or PET alone, the latter configuration yielding better results.

#### 556 4.3. Segmentation results on the multi-MRI BraTS2021 dataset

557 *Segmentation uncertainty.* For the BraTS2021 dataset, we tested the segmentation perfor-  
558 mance of our fusion framework with UNet as well as two alternative FE modules: nnUNet  
559 and nnFormer. The nnUNet model was reported to have the best performance in the  
560 BraTS2021 challenge [54] and nnFormer is now one of the state-of-the-art brain tumor  
561 segmentation models. The complete frameworks with nnUNet and nnFormer as a feature ex-  
562 tractor are referred to, respectively, as MMEF-nnUNet and MMEF-nnFormer. We compared  
563 our results with three baseline models: UNet, nnUNet, and nnFormer, and three Monte  
564 Carlo-based uncertainty segmentation models: UNet-MC, nnUNet-MC, and nnFormer-MC.  
565 Since the results obtained in Section 4.2, as well as those reported in [37] have shown that  
566 ENN-UNet and RBF-UNet yield similar results, here we only compared the performance of  
567 the ENN-based models, i.e., ENN-UNet, ENN-nnUNet and ENN-nnFormer. Moreover, we  
568 did not test the performance of deep ensemble models because applying them to larger-scale  
569 datasets exceeds our computation resources.

570 As with the lymphoma dataset, we used the ECE, Brier score, and NLL metrics to  
571 assess segmentation uncertainty. The results with UNet, nnUNet and nnFormer in the  
572 FE module are presented, respectively, in Tables 4, 5 and 6. We can see that our fusion  
573 model consistently outperforms the baseline models with all three FE models and across all  
574 uncertainty evaluation metrics, although the differences are more significant when UNet is  
575 used as a feature extractor. Indeed, the fusion mechanism can be expected to have a smaller  
576 impact when information sources are more informative. Overall, MMEF-nnUNet achieves  
577 the highest segmentation reliability with the lowest ECE, Brier score, and NLL values, and  
578 MMEF-nnFormer yields the second-best results.

Table 4: Reliability measures (means and standard errors) for MMEF-UNet and the reference methods based on UNet on the BraTS2021 dataset. The best results are in bold and the second bests are underlined.

Model	ECE↓	Brier score ↓	NLL↓
UNet	$0.071 \pm 1.8 \times 10^{-3}$	$0.141 \pm 1.8 \times 10^{-3}$	$2.475 \pm 2.2 \times 10^{-3}$
UNet-MC	$0.067 \pm 1.3 \times 10^{-3}$	$0.135 \pm 4.5 \times 10^{-3}$	$2.264 \pm 7.3 \times 10^{-2}$
ENN-UNet	<u><math>0.065 \pm 1.3 \times 10^{-3}</math></u>	<u><math>0.130 \pm 4.5 \times 10^{-3}</math></u>	<u><math>2.250 \pm 3.6 \times 10^{-2}</math></u>
MMEF-UNet (ours)	<b><math>0.060 \pm 1.3 \times 10^{-3}</math></b>	<b><math>0.115 \pm 2.2 \times 10^{-3}</math></b>	<b><math>2.189 \pm 4.1 \times 10^{-2}</math></b>

Table 5: Reliability measures (means and standard errors) for MMEF-nnUNet and the reference methods based on nnUNet on the BraTS2021 dataset. The best results are in bold. The best results are in bold, and the second-best results are underlined.

Model	ECE↓	Brier score ↓	NLL↓
nnUNet	<u><math>0.053 \pm 2.2 \times 10^{-3}</math></u>	$0.109 \pm 4.5 \times 10^{-3}$	$1.823 \pm 7.3 \times 10^{-2}$
nnUNet-MC	<b><math>0.051 \pm 1.8 \times 10^{-3}</math></b>	<u><math>0.107 \pm 4.5 \times 10^{-3}</math></u>	$1.810 \pm 5.8 \times 10^{-2}$
ENN-nnUNet	<u><math>0.053 \pm 1.8 \times 10^{-3}</math></u>	$0.109 \pm 4.9 \times 10^{-3}$	<u><math>1.804 \pm 8.2 \times 10^{-2}</math></u>
MMEF-nnUNet (ours)	<b><math>0.051 \pm 1.3 \times 10^{-3}</math></b>	<b><math>0.102 \pm 2.7 \times 10^{-3}</math></b>	<b><math>1.748 \pm 5.9 \times 10^{-2}</math></b>

Table 6: Reliability measures (means and standard errors) for MMEF-nnFormer and the reference methods based on nnUNet on the BraTS2021 dataset. The best results are in bold and the second-best are underlined.

Model	ECE↓	Brier score ↓	NLL↓
nnFormer	$0.055 \pm 1.6 \times 10^{-3}$	$0.111 \pm 3.2 \times 10^{-3}$	$1.917 \pm 5.5 \times 10^{-2}$
nnFormer-MC	<u><math>0.053 \pm 1.8 \times 10^{-3}</math></u>	<u><math>0.107 \pm 3.6 \times 10^{-3}</math></u>	<b><math>1.756 \pm 6.1 \times 10^{-2}</math></b>
ENN-nnFormer	$0.055 \pm 1.4 \times 10^{-3}$	$0.110 \pm 3.6 \times 10^{-3}$	$1.907 \pm 7.0 \times 10^{-2}$
MMEF-nnFormer (ours)	<b><math>0.052 \pm 0.6 \times 10^{-3}</math></b>	<b><math>0.103 \pm 1.2 \times 10^{-3}</math></b>	<u><math>1.787 \pm 2.2 \times 10^{-2}</math></u>

Table 7: Dice score (means and standard errors) for MMEF-UNet and the reference methods based on UNet on the BraTS2021 dataset. The best results are in bold and the second bests are underlined.

Model	ET	TC	WT	Mean
UNet	$0.807 \pm 9.4 \times 10^{-3}$	$0.825 \pm 8.5 \times 10^{-3}$	$0.881 \pm 6.7 \times 10^{-3}$	$0.837 \pm 7.2 \times 10^{-3}$
UNet-MC	$0.812 \pm 1.3 \times 10^{-2}$	$0.832 \pm 1.1 \times 10^{-2}$	$0.886 \pm 6.3 \times 10^{-3}$	$0.843 \pm 8.9 \times 10^{-3}$
ENN-UNet	<u><math>0.810 \pm 1.3 \times 10^{-2}</math></u>	<u><math>0.842 \pm 1.1 \times 10^{-2}</math></u>	<u><math>0.896 \pm 5.4 \times 10^{-3}</math></u>	<u><math>0.849 \pm 9.4 \times 10^{-3}</math></u>
MMEF-UNet (ours)	<b><math>0.833 \pm 1.2 \times 10^{-2}</math></b>	<b><math>0.854 \pm 7.2 \times 10^{-3}</math></b>	<b><math>0.907 \pm 4.9 \times 10^{-3}</math></b>	<b><math>0.864 \pm 5.8 \times 10^{-3}</math></b>

Table 8: Dice score (means and standard errors) for MMEF-nnUNet and the reference methods based on nnUNet on the BraTS2021 dataset. The best results are in bold and the second bests are underlined.

Model	ET	TC	WT	Mean
nnUNet	$0.791 \pm 4.9 \times 10^{-3}$	$0.850 \pm 5.8 \times 10^{-3}$	$0.912 \pm 3.5 \times 10^{-3}$	$0.851 \pm 4.4 \times 10^{-3}$
nnUNet-MC	$0.802 \pm 4.4 \times 10^{-3}$	$0.860 \pm 4.9 \times 10^{-3}$	<u><math>0.916 \pm 4.4 \times 10^{-3}</math></u>	$0.859 \pm 4.9 \times 10^{-3}$
ENN-nnUNet	<u><math>0.807 \pm 9.8 \times 10^{-3}</math></u>	<u><math>0.869 \pm 1.9 \times 10^{-2}</math></u>	$0.915 \pm 5.4 \times 10^{-3}$	<u><math>0.863 \pm 9.8 \times 10^{-3}</math></u>
MMEF-nnUNet (ours)	<b><math>0.832 \pm 9.8 \times 10^{-3}</math></b>	<b><math>0.873 \pm 2.6 \times 10^{-3}</math></b>	<b><math>0.918 \pm 1.3 \times 10^{-3}</math></b>	<b><math>0.875 \pm 4.4 \times 10^{-3}</math></b>

579 *Segmentation accuracy.* Segmentation accuracy was evaluated by the Dice score for the three  
580 overlapping regions, ET, TC, and WT, as well as by the mean Dice score. The results with  
581 UNet, nnUNet and nnFormer as feature extractors are reported, respectively, in Tables 7,  
582 8 and 9. Again, we can see that our fusion strategy improves segmentation accuracy for  
583 all three FE models. Overall, the highest segmentation accuracy was achieved by MMEF-  
584 nnFormer, with an increase of 1.5 % in the mean Dice score compared with the second-best  
585 method, ENN-nnFormer.

586 We also report the Dice score for the segmentation of the three original tumor regions:  
587 ED, ET, and NRC/NET in Table 10. As we can see, the baseline nnFormer shows good  
588 performance for segmenting ED and ET, while it does not perform as well for segmenting  
589 NRC/NET. Indeed, the lack of clear contrast, the similar signal intensities to normal brain  
590 tissue, the infiltrative growth patterns, and the need for multi-modal data make the seg-  
591 mentation of NRC/NET inherently more challenging compared to ED and ET. When the  
592 MMEF-nnFormer approach was applied, the Dice scores for the ED, ET, and NRC/NET im-  
593 proved by 0.6%, 1.6%, and 6.5%, respectively. The substantial improvement in NRC/NET  
594 segmentation is particularly encouraging, as it demonstrates the effectiveness of the proposed  
595 fusion method for delineating fuzzy tumor boundaries and solving challenging segmentation  
596 tasks.

597 Figures 8 and 9 show two segmentation cases when using nnFormer as the feature extrac-  
598 tor. Figure 8 shows an easy segmentation case where only one tumor type is present. Both  
599 the Flair and T1Gd images exhibit good segmentation performance with only a few misla-  
600 beled voxels. It is surprising to see that concatenating multimodal medical images as the  
601 input for nnFormer resulted in worse outcomes, with the most mislabeled voxels. This might

Table 9: Dice score (means and standard errors) for MMEF-nnFormer and the reference methods based on nnFormer on the BraTS2021 dataset. The best results are in bold and the second bests are underlined.

Model	ET	TC	WT	Mean
nnFormer	<u>0.839</u> $\pm 3.8 \times 10^{-3}$	0.878 $\pm 2.9 \times 10^{-3}$	<b>0.915</b> $\pm 2.4 \times 10^{-3}$	0.877 $\pm 1.7 \times 10^{-3}$
nnFormer-MC	0.837 $\pm 3.7 \times 10^{-3}$	0.877 $\pm 4.5 \times 10^{-3}$	<u>0.914</u> $\pm 2.9 \times 10^{-3}$	0.876 $\pm 2.3 \times 10^{-3}$
ENN-nnFormer	0.836 $\pm 9.8 \times 10^{-3}$	<u>0.882</u> $\pm 5.6 \times 10^{-2}$	<u>0.914</u> $\pm 5.2 \times 10^{-3}$	<u>0.878</u> $\pm 3.2 \times 10^{-3}$
MMEF-nnFormer (ours)	<b>0.854</b> $\pm 7.5 \times 10^{-3}$	<b>0.911</b> $\pm 5.4 \times 10^{-3}$	<u>0.914</u> $\pm 2.3 \times 10^{-3}$	<b>0.893</b> $\pm 4.8 \times 10^{-3}$

Table 10: Dice score (means and standard errors) for MMEF-nnFormer and nnFormer on the BraTS2021 dataset in segmenting detailed tumor class.

Model	ED	ET	NRC/NET	Mean
nnFormer	0.817 $\pm 5.0 \times 10^{-3}$	0.839 $\pm 3.8 \times 10^{-3}$	0.740 $\pm 7.2 \times 10^{-3}$	0.799 $\pm 2.5 \times 10^{-3}$
MMEF-nnFormer (ours)	<b>0.823</b> $\pm 3.3 \times 10^{-3}$	<b>0.855</b> $\pm 7.5 \times 10^{-3}$	<b>0.805</b> $\pm 7.3 \times 10^{-3}$	<b>0.828</b> $\pm 4.7 \times 10^{-3}$

602 be due to the hard fusion strategy of nnFormer, i.e., image concatenation, which cannot mit-  
603 igate the impact of noisy information. Consequently, the fused results are sometimes not  
604 as good as those from single-modality inputs. The proposed MMEF-nnFormer approach  
605 achieves the best performance, with fewer mislabeled voxels compared to other methods.  
606 Figure 9 illustrates a challenging segmentation scenario involving a tumor with ED, ET,  
607 and NRC/NET components. We can remark that the FLAIR image alone provides suffi-  
608 cient information to accurately segment ED, which is consistent with domain knowledge.  
609 Overall, the MMEF-nnFormer model yields the best results in this case. This example  
610 illustrates the ability of our method to improve segmentation accuracy by appropriately  
611 weighting and combining information from different modalities.

612 *Analysis of reliability coefficients.* We first recall some clinical domain knowledge of MRI  
613 images in segmenting brain tumors:

- 614 1. T1Gd images are particularly useful for delineating tumor boundaries by making tumor  
615 regions hyperintense (bright);
- 616 2. FLAIR images help delineate tumor boundaries, assess tumor infiltration into sur-  
617 rounding brain tissue, and are particularly sensitive to peritumoral edema, which ap-  
618 pears hyperintense (bright) on FLAIR sequences;
- 619 3. T2 images help delineate tumor extent, identify peritumoral edema, and assess the  
620 relationship between the tumor and surrounding brain structures;
- 621 4. Tumors typically appear hypointense (dark) on T1 images, while the contrast between  
622 the tumor and surrounding normal brain tissue may not always be sufficient for accu-  
623 rate segmentation.

624 Figure 10 shows the learned reliability coefficients  $\beta_k^t$  estimated by MMEF-nnFormer,  
625 for the four modalities and the three tumor classes. It can be seen that the evidence from

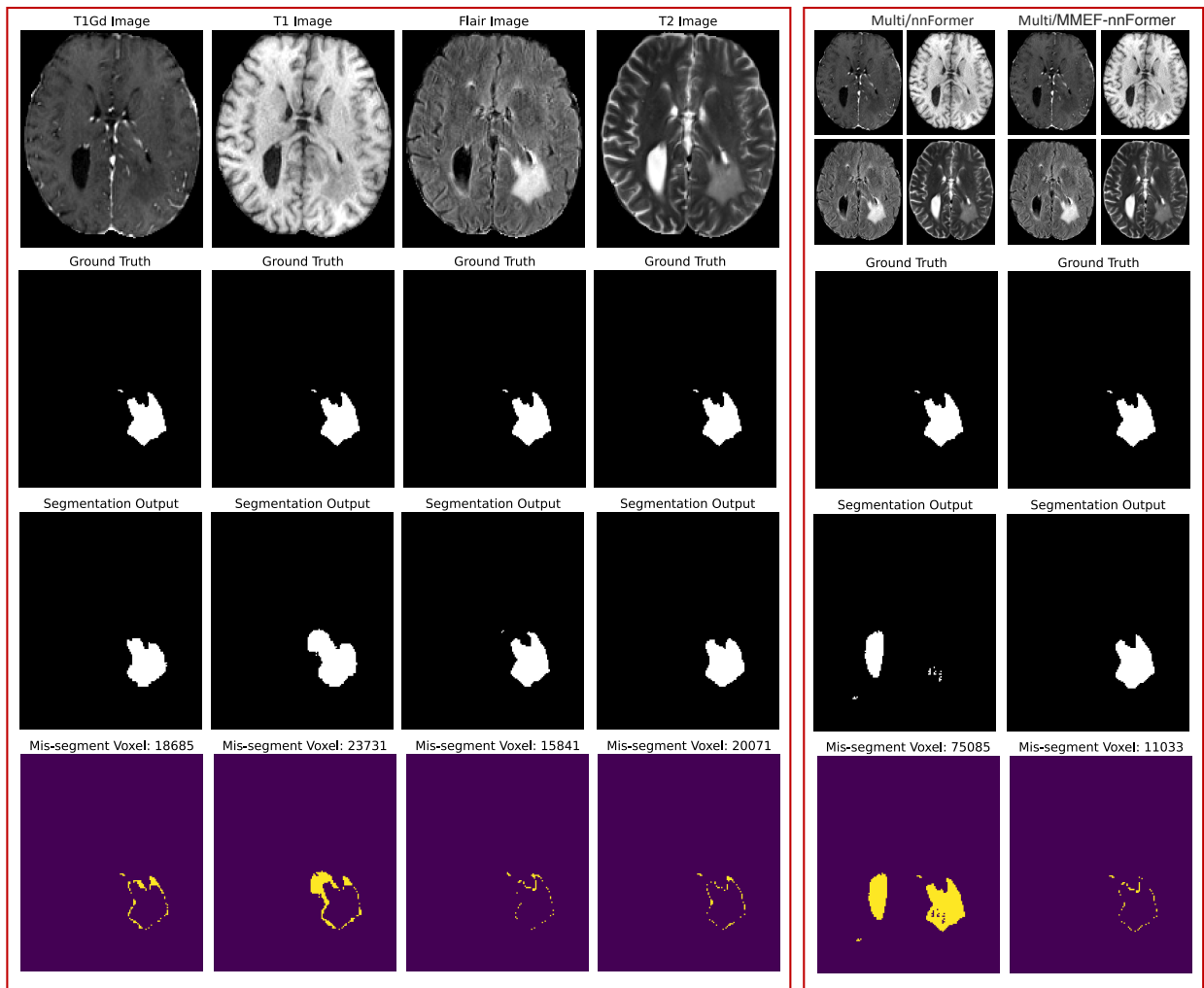


Figure 8: Examples of an easy tumor segmentation case. The first and second rows display the input modalities and the tumor ground truth, respectively. The third and last rows present the segmentation output and the mis-segmented voxels (highlighted in yellow). The left red block shows results from single-modality input using nnFormer, while the right red block compares results from multimodal input using nnFormer (left column) and MMEF-nnFormer (right column).

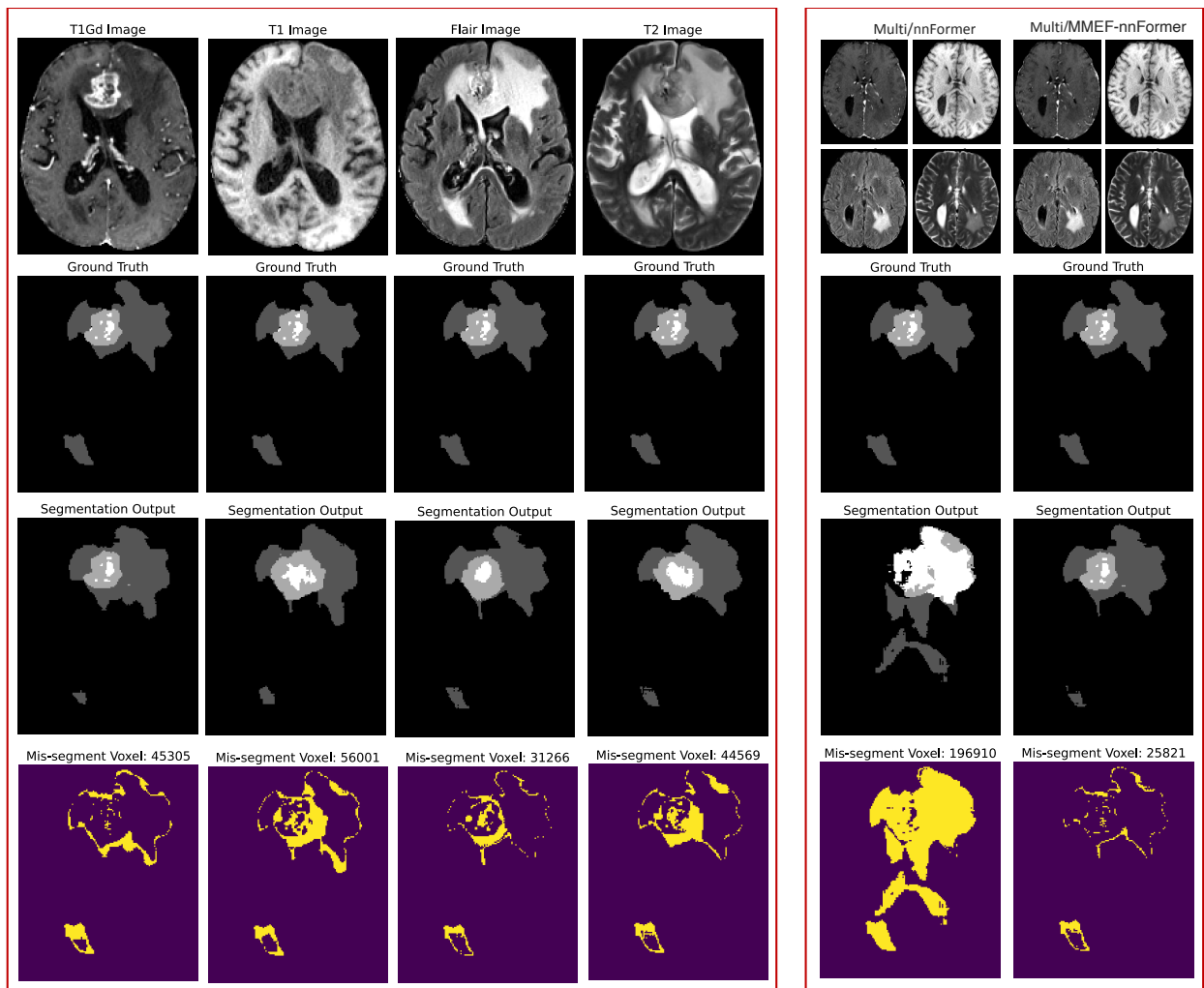


Figure 9: Examples of a challenging tumor segmentation case. The first and second rows display the input modalities and the tumor ground truth, respectively. The third and last rows present the segmentation output and the mis-segmented voxels (highlighted in yellow). The left red block shows results from single-modality input using nnFormer, while the right red block compares results from multimodal input using nnFormer (left column) and MMEF-nnFormer (right column).

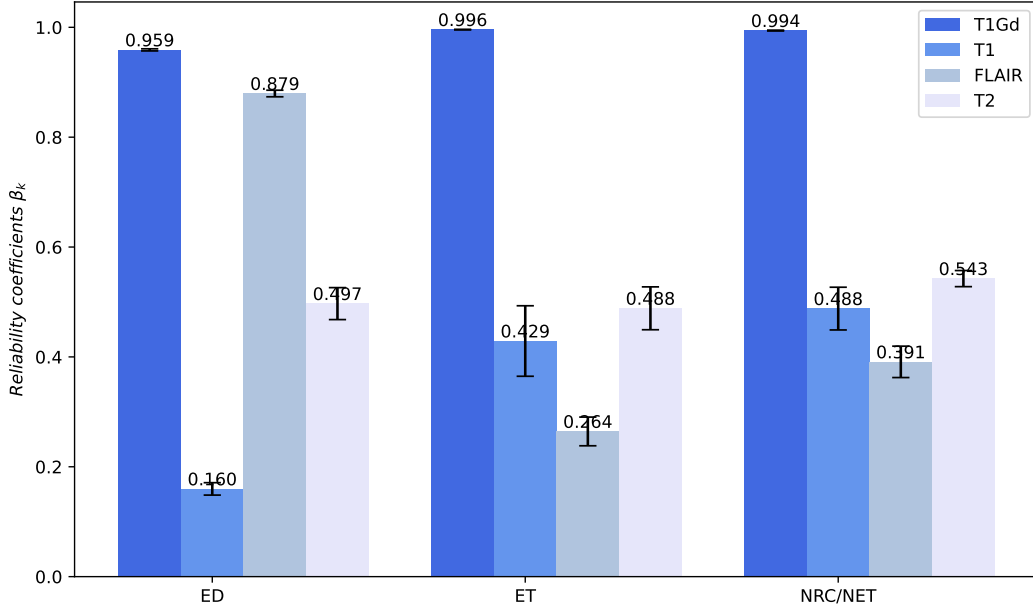


Figure 10: Estimated reliability coefficients  $\beta_k$  (means and standard errors) after training of MMEF-nnFormer for classes ED, ET, and NRC/NET in the four modalities. Higher values correspond to greater contribution to the segmentation.

626 the T1Gd modality is reliable when the true class is ED, ET, or NRC/NET, with all the  
627 reliability values greater than 0.9. In contrast, the evidence from the FLAIR modality is  
628 more reliable for the ED class with a high-reliability coefficient of 0.879 against, respectively,  
629 0.26 and 0.39 for ET and NRC/NET. The evidence from the T2 modality shows similar  
630 reliability in segmenting the three classes with a reliability coefficient of around 0.5. The  
631 evidence from the T1 modality is the least reliable one, compared with the other three MRI  
632 modalities. These results are consistent with domain knowledge about these modalities as  
633 reported in [5] and recalled at the beginning of this section, i.e., T1Gd images are useful  
634 for delineating tumor boundaries, FLAIR images are sensitive to ED, and T1 images are  
635 not sufficient for accurate tumor segmentation. This transparency and explainability of the  
636 decision-making process can be expected to enhance end-users’ trust and can be seen as  
637 significant advantages of the proposed multimodal evidence fusion approach, as opposed to  
638 the “black box” nature of conventional deep learning segmentation models.

#### 639 4.4. Discussion

640 In the following, we provide some discussion about the generalizability, computational  
641 complexity, and limitations of our approach.

642 *Generalizability.* The main advantage of our framework is its ability to model and learn the  
643 reliability of each image modality, which can be crucial when dealing with diverse, potentially

644 noisy, or low-quality data. While multimodal medical image segmentation tasks are the focus  
645 of this paper, the proposed deep evidential fusion framework can be applied to a broader  
646 range of challenging medical tasks involving heterogeneous data sources. For instance, in  
647 medical tasks such as diagnosing dementia or Alzheimer’s disease, various heterogeneous  
648 medical data are available [10]. These data can include lower-quality brain MRI images due  
649 to brain degeneration, textual data on disease history and progression, time-series data on  
650 blood-brain-barrier integrity, cerebrovascular information, and other relevant physiological  
651 measures. Traditional models struggle to effectively address this heterogeneous data within  
652 a single neural network [13], and recent work also proposed to address data heterogeneity  
653 with model ensembles and hard decision fusion [78]. Our deep evidential fusion framework  
654 could be well-suited to analyze such heterogeneous medical tasks. By learning the reliability  
655 coefficients for each of the modalities, our model can effectively combine the evidence from  
656 heterogeneous sources to reach a more informed and explainable diagnostic decision.

657 Beyond medical image processing, our approach could be applied to multimodal data  
658 fusion in other domains, such as reviewed in [44] and [8]. As examples of potential appli-  
659 cation domains where heterogeneous data need to be processed to make decisions, we can  
660 mention remote sensing and earth observations, in which light detection and ranging (Li-  
661 DAR), synthetic aperture radar (SAR), and hyperspectral images need to be combined for,  
662 e.g., improved classification of objects. As noted in [44], SAR and LiDAR use different elec-  
663 tromagnetic frequencies and thus interact differently with materials and surfaces. It would  
664 thus be beneficial to apply different discounting (reliability) coefficients to these sensor data  
665 depending on the nature of the objects of interest. This conjecture needs, of course, to be  
666 validated experimentally, which goes beyond the scope of this paper.

667 *Computational complexity.* Although the operations of DST have, in the worst case, expo-  
668 nential complexity, the mass functions computed in the EM module have only  $K$  focal sets,  
669 where  $K$  is the number of classes, and the contextual discounting operation computed in  
670 the MMEF is applied to the contour function, as explained in Section 3.1.3. Consequently,  
671 the number of operations performed in the EM and MMEF modules is only linear in the  
672 number of classes. More precisely, as shown in [16], each forward and backward propagation  
673 for one voxel and one modality in the EM module has complexity  $O(I(H + K))$ , where  $I$  is  
674 the number of prototypes,  $H$  is the number of features extracted by the FE module, and  $K$   
675 is the number of classes. In the MMEF module, the discounting of the  $T$  mass functions for  
676 each voxel using (9) and their combination using (10) can be performed in  $O(KT)$  opera-  
677 tions, and the backward pass (gradient calculation) requires the same computational effort.  
678 Overall, the complexity of our model is, thus, similar to that performed in standard neural  
679 network architectures based on weighted sums. In terms of computing times, pre-training  
680 each of the FE modules with the nnFormer architecture took approximately one hour on  
681 our machine<sup>6</sup> for the BraTS2021 dataset, and training the whole system end-to-end took 2.3  
682 hours. The total training time (6.4 hours) is slightly less than that of nnFormer with the  
683 four modalities (7.8 hours). As far as state-of-the-art uncertainty quantification techniques

---

<sup>6</sup>All models were trained on an NVIDIA A100-SXM4 graphics card with 40 GB GPU memory.

684 are concerned, Monte Carlo dropout does not significantly impact training time, while the  
685 deep ensemble method is notoriously time-consuming because it implies training several  
686 models. Overall, our framework based on DST and decision fusion is at least as efficient as  
687 alternative uncertainty quantification approaches.

688 *Limitations.* Our approach is based on combining high-level information extracted from  
689 each modality by the FE and EM modules in the form of mass functions. It, thus, has  
690 all the advantages and limitations of decision-level fusion approaches. On the plus side, it  
691 is highly modular and can still provide sensible results when only some of the modalities  
692 are available. This advantage is not crucial in multimodal image segmentation applications  
693 because all modalities are usually available, but it can matter in other potential applications  
694 such as remote sensing, as mentioned above. Another advantage of decision fusion is that  
695 the fusion process is simple and transparent, as already discussed in Sections 4.2 and 4.3.  
696 On the minus side, decision-level fusion is, at least in principle, suboptimal because it does  
697 not consider all input data globally: we can always construct a classification task in which  
698 a single classifier trained with a set of features will perform better than a combination  
699 of classifiers trained with each of the features. The good performances of our approach  
700 reported in Sections 4.2 and 4.3 show that this potential suboptimality is not an issue in the  
701 considered medical image segmentation applications, but it could be in other applications.  
702 Another limitation of our approach is that, to keep computations simple, we do not combine  
703 the whole discounted mass functions in the MMEF module, but only the contour functions.  
704 As a result, the output at each voxel is not a full mass function (with  $2^K - 1$  focal sets),  
705 which prevents us from harnessing the full power of DST, such as some of the decision rules  
706 reviewed in [18]. This and other limitations will be addressed in future work.

## 707 5. Conclusion

708 We have proposed a deep decision-level fusion architecture for multi-modality medical  
709 image segmentation. In this approach, features are first extracted from each modality using  
710 a deep neural network such as UNet. An evidence-mapping module based on prototypes in  
711 feature space then computes a Dempster-Shafer mass function at each voxel. To account  
712 for the varying reliability of different information sources in different contexts, the mass  
713 functions are transformed using the contextual discounting operation before being combined  
714 by Dempster’s rule. The whole framework is trained end-to-end by minimizing a loss function  
715 that quantifies prediction error both at the modality level and after fusion.

716 This model has been evaluated using two real-world datasets for lymphoma segmentation  
717 in PET-CT images and brain tumor segmentation in multi-MRI images. In both cases, our  
718 approach has been shown to allow for better uncertainty quantification and image segmenta-  
719 tion as compared to various alternative schemes based on pixel-level fusion. In particular, as  
720 compared to UNet, nnUNet or nnFormer alone with a softmax layer, the introduction of the  
721 evidential mapping module (computing the mass functions) improves the results, and the  
722 decision-level fusion scheme with contextual discounting brings an additional improvement.  
723 Furthermore, the values found for the reliability coefficients are consistent with domain

724 knowledge, which suggests that these coefficients can provide useful insight into the fusion  
725 process.

726 This work can be extended in many directions. First, as discussed in Section 4.4, our  
727 DST-based fusion approach can be applied to a variety of learning tasks in which several  
728 sources of information must be combined. In the biomedical domain, it could be applied  
729 to fuse heterogeneous data such as signals, personal information, biomarkers, gene infor-  
730 mation, etc. In remote sensing, a potential application could be, e.g., the fusion of Lidar,  
731 SAR and hyperspectral data. References [44] and [8] mention many other applications in  
732 which multimodal data fusion plays an important role, including human-machine interac-  
733 tion, meteorological monitoring using weather radar and satellite data, or concrete structural  
734 monitoring through fusing ultrasonic, impact echo, capacitance, and radar. From a theo-  
735 retical point of view, our approach could be extended in several directions. As mentioned  
736 in Section 4.4, we could combine not only the contour functions from the EM module but  
737 the whole mass functions, which would allow us to compute richer outputs that could be  
738 exploited within more sophisticated decision strategies such as partial classification [55], or  
739 further combined with other data. We could also consider other mass-function correction  
740 methods making it possible to account for more diverse meta-knowledge about information  
741 sources such as proposed, e.g. in [62], and/or other combination rules such as the cautions  
742 rule [17] or variants with learnable parameters as used in [64].

## 743 Acknowledgements

744 This work was supported by the China Scholarship Council (No. 201808331005). It  
745 was carried out in the framework of the Labex MS2T, which was funded by the French  
746 Government, through the program “Investments for the Future” managed by the National  
747 Agency for Research (Reference ANR-11-IDEX-0004-02).

## 748 References

- 749 [1] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao,  
750 A. Khosravi, U. R. Acharya, et al. A review of uncertainty quantification in deep learning: Techniques,  
751 applications and challenges. *Information fusion*, 76:243–297, 2021.
- 752 [2] G. Andrade-Miranda, V. Jaouen, O. Tankyevych, C. C. Le Rest, D. Visvikis, and P.-H. Conze. Multi-  
753 modal medical transformers: A meta-analysis for medical image segmentation in oncology. *Computer-  
754 ized Medical Imaging and Graphics*, 110:102308, 2023.
- 755 [3] M. Arif and G. Wang. Fast curvelet transform through genetic algorithm for multimodal medical image  
756 fusion. *Soft Computing*, 24(3):1815–1836, 2020.
- 757 [4] C. Asha, S. Lal, V. P. Gurupur, and P. P. Saxena. Multi-modal medical image fusion with adaptive  
758 weighted combination of NSST bands using chaotic grey wolf optimization. *IEEE Access*, 7:40782–  
759 40796, 2019.
- 760 [5] U. Baid, S. Ghodasara, S. Mohan, M. Bilello, E. Calabrese, E. Colak, K. Farahani, J. Kalpathy-Cramer,  
761 F. C. Kitamura, S. Pati, et al. The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor  
762 segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021.
- 763 [6] M. Bauer, M. Van der Wilk, and C. E. Rasmussen. Understanding probabilistic sparse gaussian process  
764 approximations. *Advances in neural information processing systems*, 29, 2016.

- 765 [7] Y. Bengio, I. Goodfellow, and A. Courville. *Deep learning*, volume 1. MIT press Cambridge, MA, USA,  
766 2017.
- 767 [8] R. Bokade, A. Navato, R. Ouyang, X. Jin, C.-A. Chou, S. Ostadabbas, and A. V. Mueller. A cross-  
768 disciplinary comparison of multimodal data fusion approaches and applications: Accelerating learning  
769 through trans-disciplinary information sharing. *Expert Systems with Applications*, 165:113885, 2021.
- 770 [9] G. W. Brier et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*,  
771 78(1):1–3, 1950.
- 772 [10] C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic,  
773 O. Delaneau, J. OConnell, et al. The uk biobank resource with deep phenotyping and genomic data.  
774 *Nature*, 562(7726):203–209, 2018.
- 775 [11] B. R. Cobb and P. P. Shenoy. On the plausibility transformation method for translating belief function  
776 models to probability models. *International Journal of Approximate Reasoning*, 41(3):314–330, 2006.
- 777 [12] R. Cooke et al. *Experts in uncertainty: opinion and subjective probability in science*. Oxford University  
778 Press, 1991.
- 779 [13] Y. Dai, D. Qiu, Y. Wang, S. Dong, and H.-L. Wang. Research on computer-aided diagnosis of Alzheimers  
780 disease based on heterogeneous medical data fusion. *International Journal of Pattern Recognition and*  
781 *Artificial Intelligence*, 33(05):1957001, 2019.
- 782 [14] F. Delmotte, L. Dubois, and P. Borne. Context-dependent trust in data fusion within the possibil-  
783 ity theory. In *1996 IEEE International Conference on Systems, Man and Cybernetics. Information*  
784 *Intelligence and Systems*, volume 1, pages 538–543. IEEE, 1996.
- 785 [15] A. P. Dempster. Upper and lower probability inferences based on a sample from a finite univariate  
786 population. *Biometrika*, 54(3-4):515–528, 1967.
- 787 [16] T. Denœux. A neural network classifier based on Dempster-Shafer theory. *IEEE Transactions on*  
788 *Systems, Man, and Cybernetics-Part A: Systems and Humans*, 30(2):131–150, 2000.
- 789 [17] T. Denœux. Conjunctive and disjunctive combination of belief functions induced by non distinct bodies  
790 of evidence. *Artificial Intelligence*, 172:234–264, 2008.
- 791 [18] T. Denœux. Decision-making with belief functions: A review. *International Journal of Approximate*  
792 *Reasoning*, 109:87–110, 2019.
- 793 [19] T. Denœux. NN-EVCLUS: neural network-based evidential clustering. *Information Sciences*, 572:297–  
794 330, 2021.
- 795 [20] T. Denœux, D. Dubois, and H. Prade. Representations of uncertainty in artificial intelligence: Beyond  
796 probability and possibility. In P. Marquis, O. Papini, and H. Prade, editors, *A Guided Tour of Artificial*  
797 *Intelligence Research*, volume 1, chapter 4, pages 119–150. Springer Verlag, 2020.
- 798 [21] T. Denœux, D. Dubois, and H. Prade. Representations of uncertainty in artificial intelligence: Prob-  
799 ability and possibility. In P. Marquis, O. Papini, and H. Prade, editors, *A Guided Tour of Artificial*  
800 *Intelligence Research*, volume 1, chapter 3, pages 69–117. Springer Verlag, 2020.
- 801 [22] T. Denœux, O. Kanjanatarakul, and S. Sriboonchitta. A new evidential k-nearest neighbor rule based  
802 on contextual discounting with partially supervised learning. *International Journal of Approximate*  
803 *Reasoning*, 113:287–302, 2019.
- 804 [23] J. Du, W. Li, B. Xiao, and Q. Nawaz. Union laplacian pyramid with multiple features for medical  
805 image fusion. *Neurocomputing*, 194:326–339, 2016.
- 806 [24] D. Dubois, W. Liu, J. Ma, and H. Prade. The basic principles of uncertain information fusion. an  
807 organised review of merging rules in different representation frameworks. *Information Fusion*, 32:12–  
808 39, 2016.
- 809 [25] Z. Elouedi, K. Mellouli, and P. Smets. Assessing sensor reliability for multisensor data fusion within the  
810 transferable belief model. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*,  
811 34(1):782–787, 2004.
- 812 [26] S. Fabre, A. Appriou, and X. Briottet. Presentation and description of two classification methods using  
813 data fusion based on sensor management. *Information Fusion*, 2(1):49–71, 2001.
- 814 [27] P. H. Foo and G. W. Ng. High-level information fusion: An overview. *J. Adv. Inf. Fusion*, 8(1):33–72,  
815 2013.

- 816 [28] J. Fu, W. Li, J. Du, and Y. Huang. A multiscale residual pyramid attention network for medical image  
817 fusion. *Biomedical Signal Processing and Control*, 66:102488, 2021.
- 818 [29] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in  
819 deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.
- 820 [30] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and  
821 Y. Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence,  
822 and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran  
823 Associates, Inc., 2014.
- 824 [31] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In  
825 *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.
- 826 [32] D. L. Hall, M. McNeese, J. Llinas, and T. Mullen. A framework for dynamic hard/soft fusion. In *2008*  
827 *11th International Conference on Information Fusion*, pages 1–8. IEEE, 2008.
- 828 [33] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, et al. A survey  
829 on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110,  
830 2022.
- 831 [34] G. E. Hinton and D. Van Camp. Keeping the neural networks simple by minimizing the description  
832 length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory*,  
833 pages 5–13, 1993.
- 834 [35] Q. Hu, S. Hu, and F. Zhang. Multi-modality medical image fusion based on separable dictionary  
835 learning and gabor filtering. *Signal Processing: Image Communication*, 83:115758, 2020.
- 836 [36] L. Huang, T. Denoeux, P. Vera, and S. Ruan. Evidence fusion with contextual discounting for  
837 multi-modality medical image segmentation. In *Medical Image Computing and Computer Assisted*  
838 *Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Pro-*  
839 *ceedings, Part V*, pages 401–411. Springer, 2022.
- 840 [37] L. Huang, S. Ruan, P. Decazes, and T. Denoeux. Lymphoma segmentation from 3d PET-CT images  
841 using a deep evidential network. *International Journal of Approximate Reasoning*, 149:39–60, 2022.
- 842 [38] L. Huang, S. Ruan, and T. Denoeux. Application of belief functions to medical image segmentation:  
843 A review. *Information fusion*, 91:737–756, 2023.
- 844 [39] L. Huang, S. Ruan, Y. Xing, and M. Feng. A review of uncertainty quantification in medical image  
845 analysis: Probabilistic and non-probabilistic methods. *Medical Image Analysis*, 97:103223, 2024.
- 846 [40] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Koehler,  
847 T. Norajitra, S. Wirkert, et al. nnu-net: Self-adapting framework for u-net-based medical image  
848 segmentation. *arXiv preprint arXiv:1809.10486*, 2018.
- 849 [41] E. Kerfoot, J. Clough, I. Oksuz, J. Lee, A. P. King, and J. A. Schnabel. Left-ventricle quantification  
850 using residual u-net. In *International Workshop on Statistical Atlases and Computational Models of*  
851 *the Heart*, pages 371–380. Springer, 2018.
- 852 [42] S. U. R. Khan, M. Zhao, S. Asif, and X. Chen. Hybrid-net: A fusion of densenet169 and advanced  
853 machine learning classifiers for enhanced brain tumor diagnosis. *International Journal of Imaging*  
854 *Systems and Technology*, 34(1):e22975, 2024.
- 855 [43] M. Kim, D. K. Han, and H. Ko. Joint patch clustering-based dictionary learning for multimodal image  
856 fusion. *Information fusion*, 27:198–214, 2016.
- 857 [44] D. Lahat, T. Adali, and C. Jutten. Multimodal data fusion: An overview of methods, challenges, and  
858 prospects. *Proceedings of the IEEE*, 103(9):1449–1477, 2015.
- 859 [45] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estima-  
860 tion using deep ensembles. *Advances in neural information processing systems*, 30, 2017.
- 861 [46] H. Li and X.-J. Wu. Crossfuse: A novel cross attention mechanism based infrared and visible image  
862 fusion approach. *Information Fusion*, 103:102147, 2024.
- 863 [47] X. Li, F. Zhou, H. Tan, W. Zhang, and C. Zhao. Multimodal medical image fusion based on joint  
864 bilateral filter and local gradient energy. *Information Sciences*, 569:302–325, 2021.
- 865 [48] C. Lian, S. Ruan, T. Denoeux, H. Li, and P. Vera. Joint tumor segmentation in PET-CT images  
866 using co-clustering and fusion based on belief functions. *IEEE Transactions on Image Processing*,

- 867 28(2):755–766, 2019.
- 868 [49] X. Liang, P. Hu, L. Zhang, J. Sun, and G. Yin. Mcfnnet: Multi-layer concatenation fusion network for  
869 medical images fusion. *IEEE Sensors Journal*, 19(16):7107–7119, 2019.
- 870 [50] Y. Liu, X. Chen, J. Cheng, and H. Peng. A medical image fusion method based on convolutional neural  
871 networks. In *2017 20th international conference on information fusion (Fusion)*, pages 1–7. IEEE, 2017.
- 872 [51] Y. Liu, S. Liu, and Z. Wang. A general framework for image fusion based on multi-scale transform and  
873 sparse representation. *Information fusion*, 24:147–164, 2015.
- 874 [52] Y. Liu, F. Mu, Y. Shi, and X. Chen. Sf-net: A multi-task model for brain tumor segmentation in  
875 multimodal mri via image fusion. *IEEE Signal Processing Letters*, 29:1799–1803, 2022.
- 876 [53] B. C. Lowekamp, D. T. Chen, L. Ibáñez, and D. Blezek. The design of simpleitk. *Frontiers in neuroin-*  
877 *formatics*, 7:45, 2013.
- 878 [54] H. M. Luu and S.-H. Park. Extending nn-unet for brain tumor segmentation. In *Brainlesion: Glioma,*  
879 *Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 173–186, Cham, 2022. Springer Inter-  
880 national Publishing.
- 881 [55] L. Ma and T. Denœux. Partial classification in the belief function framework. *Knowledge-Based*  
882 *Systems*, 214:106742, 2021.
- 883 [56] D. J. MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*,  
884 4(3):448–472, 1992.
- 885 [57] X. Meng, Q. Wei, L. Meng, J. Liu, Y. Wu, and W. Liu. Feature fusion and detection in Alzheimers  
886 disease using a novel genetic multi-kernel svm based on mri imaging and gene data. *Genes*, 13(5):837,  
887 2022.
- 888 [58] D. Mercier, B. Quost, and T. Denœux. Refined modeling of sensor reliability in the belief function  
889 framework using contextual discounting. *Information fusion*, 9(2):246–258, 2008.
- 890 [59] C. Pei, K. Fan, and W. Wang. Two-scale multimodal medical image fusion based on guided filtering  
891 and sparse representation. *IEEE Access*, 8:140216–140233, 2020.
- 892 [60] H. Peiris, M. Hayat, Z. Chen, G. Egan, and M. Harandi. A robust volumetric transformer for accurate  
893 3d tumor segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI*  
894 *2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V*, pages  
895 162–172. Springer, 2022.
- 896 [61] F. Pichon, D. Dubois, and T. Denœux. Quality of information sources in information fusion. In  
897 É. Bossé and G. L. Rogova, editors, *Information Quality in Information Fusion and Decision Making*,  
898 pages 31–49. Springer International Publishing, Cham, 2019.
- 899 [62] F. Pichon, D. Mercier, E. Lefèvre, and F. Delmotte. Proposition and learning of some belief function  
900 contextual correction mechanisms. *International Journal of Approximate Reasoning*, 72:4–42, 2016.
- 901 [63] Z. Qu, Y. Li, and P. Tiwari. Qnmf: A quantum neural network based multimodal fusion system for  
902 intelligent diagnosis. *Information Fusion*, 100:101913, 2023.
- 903 [64] B. Quost, M.-H. Masson, and T. Denœux. Classifier fusion in the Dempster-Shafer framework using  
904 optimized t-norm based combination rules. *International Journal of Approximate Reasoning*, 52(3):353–  
905 374, 2011.
- 906 [65] G. L. Rogova and V. Nimier. Reliability in information fusion: literature survey. In *Proceedings of the*  
907 *seventh international conference on information fusion*, volume 2, pages 1158–1165, 2004.
- 908 [66] A.-J. Rousseau, T. Becker, J. Bertels, M. B. Blaschko, and D. Valkenborg. Post training uncertainty cal-  
909 ibration of deep networks for medical image segmentation. In *2021 IEEE 18th International Symposium*  
910 *on Biomedical Imaging (ISBI)*, pages 1052–1056. IEEE, 2021.
- 911 [67] M. Safari, A. Fatemi, and L. Archambault. Medfusiongan: multimodal medical image fusion using an  
912 unsupervised deep generative adversarial network. *BMC Medical Imaging*, 23(1):203, 2023.
- 913 [68] M. Sensoy, L. Kaplan, and M. Kandemir. Evidential deep learning to quantify classification uncertainty.  
914 In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors,  
915 *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.
- 916 [69] G. Shafer. *A mathematical theory of evidence*, volume 42. Princeton University Press, 1976.
- 917 [70] H. R. Shahdoosti and Z. Tabatabaei. Mri and pet/spect image fusion at feature level using ant colony

- 918 based segmentation. *Biomedical Signal Processing and Control*, 47:63–74, 2019.
- 919 [71] Y. Shi, C. Zu, P. Yang, S. Tan, H. Ren, X. Wu, J. Zhou, and Y. Wang. Uncertainty-weighted and  
 920 relation-driven consistency training for semi-supervised head-and-neck tumor segmentation. *Knowledge-*  
 921 *Based Systems*, 272:110598, 2023.
- 922 [72] R. Singh, M. Vatsa, and A. Noore. Multimodal medical image fusion using redundant discrete wavelet  
 923 transform. In *2009 Seventh International Conference on Advances in Pattern Recognition*, pages 232–  
 924 235. IEEE, 2009.
- 925 [73] P. Smets. Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem.  
 926 *International Journal of Approximate Reasoning*, 9:1–35, 1993.
- 927 [74] P. Smets and R. Kennes. The Transferable Belief Model. *Artificial Intelligence*, 66:191–243, 1994.
- 928 [75] W. Tan, W. Thitø, P. Xiang, and H. Zhou. Multi-modal brain image fusion based on multi-level  
 929 edge-preserving filtering. *Biomedical Signal Processing and Control*, 64:102280, 2021.
- 930 [76] W. Tang, F. He, Y. Liu, and Y. Duan. Matr: Multimodal medical image fusion via multiscale adaptive  
 931 transformer. *IEEE Transactions on Image Processing*, 31:5134–5149, 2022.
- 932 [77] A. Tannaz, S. Mousa, D. Sabalan, and P. Masoud. Fusion of multimodal medical images using non-  
 933 subsampled shearlet transform and particle swarm optimization. *Multidimensional Systems and Signal*  
 934 *Processing*, 31:269–287, 2020.
- 935 [78] M. Tanveer, T. Goel, R. Sharma, A. Malik, I. Beheshti, J. Del Ser, P. Suganthan, and C. Lin. Ensemble  
 936 deep learning for Alzheimers disease characterization and estimation. *Nature Mental Health*, pages 1–13,  
 937 2024.
- 938 [79] Z. Tong, P. Xu, and T. Denoëux. An evidential classifier based on Dempster-Shafer theory and deep  
 939 learning. *Neurocomputing*, 450:275–293, 2021.
- 940 [80] Z. Tong, P. Xu, and T. Denoëux. Evidential fully convolutional network for semantic segmentation.  
 941 *Applied Intelligence*, 51:6376–6399, 2021.
- 942 [81] F. G. Veshki and S. A. Vorobyov. Coupled feature learning via structured convolutional sparse coding  
 943 for multimodal image fusion. In *ICASSP 2022-2022 IEEE International Conference on Acoustics,*  
 944 *Speech and Signal Processing (ICASSP)*, pages 2500–2504. IEEE, 2022.
- 945 [82] Q. Wang, S. Li, H. Qin, and A. Hao. Robust multi-modal medical image fusion via anisotropic heat  
 946 diffusion guided low-rank structural analysis. *Information fusion*, 26:103–121, 2015.
- 947 [83] Z. Wang, Z. Cui, and Y. Zhu. Multi-modal medical image fusion by laplacian pyramid and adaptive  
 948 sparse representation. *Computers in Biology and Medicine*, 123:103823, 2020.
- 949 [84] Y. Weng, Y. Zhang, W. Wang, and T. Dening. Semi-supervised information fusion for medical image  
 950 analysis: Recent progress and future perspectives. *Information Fusion*, page 102263, 2024.
- 951 [85] P. Xu, X. Zhu, and D. A. Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions*  
 952 *on Pattern Analysis & Machine Intelligence*, 45(10):12113–12132, oct 2023.
- 953 [86] L. Yang, B. Guo, and W. Ni. Multimodality medical image fusion based on multiscale geometric  
 954 analysis of contourlet transform. *Neurocomputing*, 72(1-3):203–211, 2008.
- 955 [87] L. A. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1:3–28, 1978.
- 956 [88] J. Zhang, L. Jiao, W. Ma, F. Liu, X. Liu, L. Li, P. Chen, and S. Yang. Transformer based conditional  
 957 gan for multimodal image fusion. *IEEE Transactions on Multimedia*, 25:8988–9001, 2023.
- 958 [89] S. Zhang, J. Zhang, B. Tian, T. Lukasiewicz, and Z. Xu. Multi-modal contrastive mutual learning  
 959 and pseudo-label re-learning for semi-supervised medical image segmentation. *Medical Image Analysis*,  
 960 83:102656, 2023.
- 961 [90] Y.-D. Zhang, Z. Dong, S.-H. Wang, X. Yu, X. Yao, Q. Zhou, H. Hu, M. Li, C. Jiménez-Mesa, J. Ramirez,  
 962 et al. Advances in multimodal data fusion in neuroimaging: Overview, challenges, and novel orientation.  
 963 *Information Fusion*, 64:149–187, 2020.
- 964 [91] W. Zhao and H. Lu. Medical image fusion and denoising with alternating sequential filter and adaptive  
 965 fractional order total variation. *IEEE Transactions on Instrumentation and Measurement*, 66(9):2283–  
 966 2294, 2017.
- 967 [92] H.-Y. Zhou, J. Guo, Y. Zhang, X. Han, L. Yu, L. Wang, and Y. Yu. nnformer: Volumetric medical  
 968 image segmentation via a 3d transformer. *IEEE Transactions on Image Processing*, 2023.

- 969 [93] K. Zou, X. Yuan, X. Shen, M. Wang, and H. Fu. Tbrats: Trusted brain tumor segmentation. In  
970 *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages  
971 503–513. Springer, 2022.