A Dempster-Shafer approach to trustworthy AI with application to fetal brain MRI segmentation

Lucas Fidon, Michael Aertsen, Florian Kofler, Andrea Bink, Anna L. David, Thomas Deprest, Doaa Emam, Frédéric Guffens, András Jakab, Gregor Kasprian, Patric Kienast, Andrew Melbourne, Bjoern Menze, Nada Mufti, Ivana Pogledic, Daniela Prayer, Marlene Stuempflen, Esther Van Elslander, Sébastien Ourselin, Jan Deprest, Tom Vercauteren

Abstract—Deep learning models for medical image segmentation can fail unexpectedly and spectacularly for pathological cases and images acquired at different centers than training images, with labeling errors that violate expert knowledge. Such errors undermine the trustworthiness of deep learning models for medical image segmentation. Mechanisms for detecting and correcting such failures are essential for safely translating this technology into clinics and are likely to be a requirement of future regulations on artificial intelligence (AI). In this work, we propose a trustworthy AI theoretical framework and a practical system that can augment any backbone AI system using a fallback method and a fail-safe mechanism based on Dempster-Shafer theory. Our approach relies on an actionable definition of trustworthy AI. Our method automatically discards the voxel-level labeling predicted by the backbone AI that violate expert knowledge and relies on a fallback for those voxels. We demonstrate the effectiveness of the proposed trustworthy AI approach on the largest reported annotated dataset of fetal MRI consisting of 540 manually annotated fetal brain 3D T2w MRIs from 13 centers. Our trustworthy AI method improves the robustness of four backbone AI models for fetal brain MRIs acquired across various centers and for fetuses with various brain abnormalities. Our code is publicly available here.

1 Introduction

Automatic segmentation of medical images is needed for personalized medicine and to study anatomical development in healthy populations as well as populations with a pathology. Artificial Intelligence (AI) algorithms for medical image segmentation can reach super-human accuracy on average [1] and yet most radiologists do not trust them [2], [3]. This is partly because, for some cases, AI algorithms fail spectacularly with errors that violate expert knowledge about the segmentation task when the AI was applied across imaging protocol and anatomical pathologies [2], [4], [5] (Fig.1b). This sense of distrust is exacerbated by the current lack of clear fit-for-purpose regulatory requirements for AI-based medical image software [6].

The legal framework for the deployment in clinics of AI tools for medical segmentation is likely to soon become more stringent once the European Union has proposed its Artificial Intelligence Act to regulate AI and AI trust is at the core of this proposal [7], [8]. Guidelines for trustworthy AI claim that AI trustworthiness must precede trust in the deployment of AI systems to avoid miscalibration of the human trust with respect to the trustworthiness of an AI system [7]. In Psychology, trust of humans in AI can be defined as the belief of the human user that the AI system will satisfy the criteria of a set of contracts of trust. This contract-based definition of human-AI trust reflects the plurality and the context-dependency of human-AI trust. In particular, the user may trust an AI system for one population or one type of scanner but not trust it for others. An AI system is trustworthy to a contract of trust if it can maintain this contract in all situations within the contract scope. The EU ethics guidelines for trustworthy AI, that upheld the AI Act, advocate that "AI systems should have safeguards that enable a *fallback* plan in case of problems" [7]. We argue that those safeguards should implement a *fail-safe* mechanism in relation with a collection of contracts of trust so as to improve the trustworthiness of the overall system.

In this article, we propose the first trustworthy AI framework with a fail-safe and a fallback for medical image segmentation. The proposed framework consists of three main components: first, a backbone AI algorithm, that can be any AI algorithm for the task at hand, second, a fallback segmentation algorithm, that is more robust than the backbone AI algorithm to out-of-distribution data but potentially less precise, and third, a fail-safe method that automatically detects local conflicts between the backbone AI algorithm prediction and the contracts of trust and switches to the fallback algorithm in case of conflicts. This is illustrated for fetal brain 3D MRI segmentation in Fig. 1. The proposed principled fail-safe method is based on Dempster-Shafer (DS) theory. DS theory allows to model partial information about the task, which is typically the case for expert knowledge.

• *L. Fidon, A. Melbourne, N. Mufti, S. Ourselin, and T. Vercauteren were with the School of Biomedical Engineering & Imaging Sciences, King's College London, London, UK. M. Aertsen, T. Deprest, D.Emam, E. van Elslander, and J. Deprest were with the Department of Radiology, University Hospitals Leuven, Leuven, Belgium. F. Kofler and B. Menze were with the Department of Informatics, Technical University Munich, Munich, Germany. A. Bink was with the Department of Neuroradiology and Clinical Neuroscience Center, University Hospital Zurich and University of Zurich, Zurich, Switzerland. A. L. David was with the Institute for Women's Health, University College London, London, UK. D. Emam was with the Department of Gynecology and Obstetrics, University Hospitals Tanta, Tanta, Egypt. A. Jakab was with the Center for MR Research, University Children's Hospital Zurich, University of Zurich, Zurich, Switzerland. G. Kasprian, P. Kienast, I. Pogledic, D. Prayer, and M. Stuempflen were with the Department of Biomedical Imaging and Image-guided Therapy, Medical University of Vienna, Vienna, Austria.*
E-mail: lucas.fidon@kcl.ac.uk

For example, in human brain anatomy, the cerebellum is known to be located in the lower back part of the brain. This gives us information only about the segmentation of the cerebellum while a segmentation algorithms will typically compute segmentation for many other tissue types in addition to the cerebellum. The Dempster's rule of combination of DS theory is an efficient mathematical tool to combine independent sources of information that discards contradictions among the sources. In our framework, the AI-based segmentation and each expert knowledge are treated as independent sources of information and the Dempster's rule of combination is employed to act as the fail-safe.

To demonstrate the applicability of the developed trustworthy AI framework, we propose one implementation for fetal brain segmentation in MRI. For the backbone AI model, we used the state-of-the-art deep learning-based segmentation pipeline nnU-Net [1]. For the fallback model we used a registration-based segmentation method inspired by the state-of-the-art multi-atlas method GIF [9]. We also show that our fail-safe formulation is flexible enough to model both spatial location-based and intensity-based contracts of trust about the regions of interest to be segmented. Spatial location-based fail-safes are implemented using the masks of the regions of interest computed by the registration-based fallback algorithm. However, in the fail-safe, the masks are interpreted differently. In this case, the mask of a region R is used only to exclude labeling voxels outside of the mask as belonging to R. This is illustrated in Fig. 1c. Inspired by the margins used for safety in radiation therapy planning to account for spatial registration errors [10], we first add spatial margins to the fallback masks before excluding the labels seen as anatomically unlikely according to the dilated fallback masks. While allowing the masks to overlap, this helps preventing miscoverage of the regions of interest that is the only source of error in this formulation of the fail-safe.

We evaluated the proposed trustworthy AI method on fetal brain segmentation into eight tissue types using 3D T2w MRI. The segmentation of fetal brain MRI is essential to study normal and abnormal fetal brain development. In the future, reliable analysis and evaluation of fetal brain structures could also support diagnosis of central nervous system pathology, patient selection for fetal surgery, evaluation and prediction of outcome, hence also parental counselling. In particular, fetal brain 3D T2w MRI segmentation presents multiple challenges for trustworthy AI [4]. There are variations in T2w MRI protocols used across clinical centers and there is a spectacular variation of the fetal brain anatomy across gestational ages and across normal and abnormal fetal brain anatomy.

2 Related Works

2.1 Information fusion for medical image segmentation

Information fusion methods based on probability theory have been proposed to combine different segmentations [11], [12]. The Simulataneous Truth And Performance Level Estimation (STAPLE) algorithm weighs each segmentation by estimating the sensitivity and specificity of each segmentations [11]. In particular, these methods define only image-wise weights to combine the segmentations. Fusion methods with weights varying spatially have been proposed for the special case of atlas-based algorithms [9], but not in general as in our method. In the context of deep learning-based segmentation methods, simple averaging is used in state-of-the-art pipelines [1]. Perhaps more importantly, fusion methods based on probability theory only cannot model imprecise or partial prior expert-knowledge [11]. In contrast, the use of Dempster-Shafer (DS) theory in our method allows us a larger diversity of prior knowledge that is typically robust but imprecise. We show that our approach based on DS can model prior given by either atlases or voxel intensity prior distributions in the case of fetal brain segmentation and more priors could be modeled as well in other segmentation tasks.

2.2 Dempster-Shafer for medical image segmentation

Only a few papers have proposed to use DS theory in the context of information fusion for medical image segmentation [13], [14], [15], [16], [17]. DS has been used to combine different image modalities [13], neighbouring voxels [17], or both [14], [15] for brain MRI segmentation.

In contrast, in this work DS is used to combine two arbitrary probabilistic segmentation algorithms with prior information about the segmentation tasks. In this case, we show that Dempster's rule of combination allows to detect segmentation failures of the first segmentation algorithms and switch to the second locally at the voxel level.

2.3 Domain generalization

Domain generalization (DG) aims at improving the out-of-distribution (OOD) generalization of AI algorithms without using OOD images during training [18]. DG methods include the use of various data augmentations [1], [19], self-supervised learning pre-training [20], and new training methods [21]. Closer to this work are DG methods based on the integration of anatomical prior [22], [23]. In those methods, atlas-based probabilities are fused using concatenation inside the deep neural network.

DG can be used to improve generalizability but it does not offer any guarantees in terms of trustworthiness. As such, it is a complementary approach to the one we propose. In our experiments, our approach is compared and combined with three DG methods as backbone AI leading to improved robustness.

3 Methods

3.1 Background on Human-AI trust

Artificial intelligence is defined as any automation perceived by the individual using it as having an intent [24].

Human-AI trust is multi-dimensional. For example, the user can trust an AI medical image segmentation algorithm for a given tissue type, for images coming from a given type of scanner, or for a given population and not another. This observation that trust has several facets and is context-dependent [25] has motivated the introduction of *contractual trust* [24]. A **contract of trust** is an attribute of the AI algorithm which, if not fulfilled, causes a risk in using the AI algorithm. A contract of trust is not necessarily related to the accuracy of the AI algorithm for the task at hand.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2023.3346330

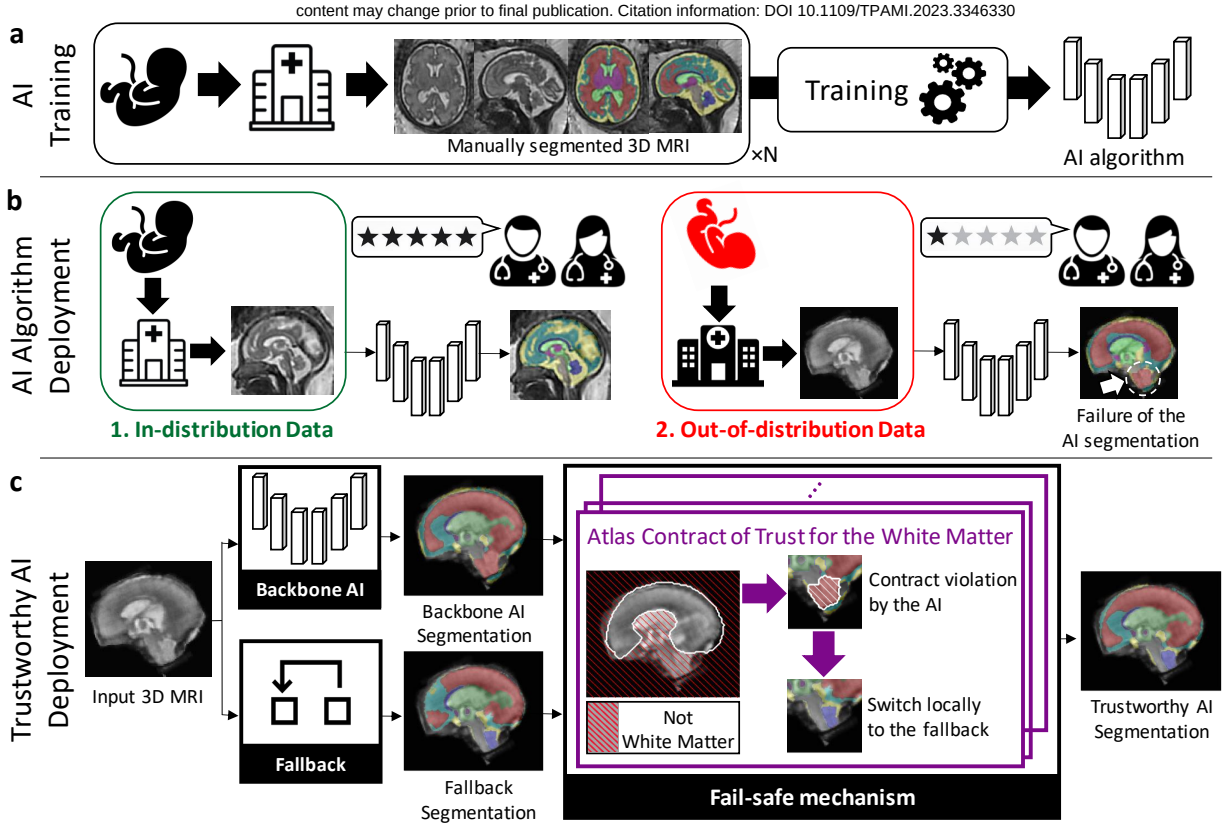


Fig. 1. **Schematics of our principled method for trustworthy AI applied to medical image segmentation. a.** Deep neural networks for medical image segmentation (AI algorithm) are typically trained on images from a limited number of acquisition centers. This is usually not sufficient to cover all the anatomical variability. **b.** When such a trained AI algorithm is deployed, it will typically give satisfactory accuracy for images acquired with the same protocol as training images and with a health condition represented in the training dataset (left). However, an AI algorithm might fail with errors that are not anatomically plausible, for images acquired with a slightly different protocol as training images and/or representing anatomy underrepresented in the training dataset (right). **c.** Schematic of the proposed trustworthy AI algorithm. A backbone AI segmentation algorithm is coupled with a *fallback* segmentation algorithm. Experts knowledge about the anatomy is modeled in the *fail-safe mechanism* using Dempster-Shafer theory. A rich variety of experts knowledge can be modeled as *contracts of trust*, such as, but not only, atlas-based prior and intensity-based prior (not shown here). When part of the AI segmentation is found to contradict one of the contracts of trust for a voxel, our trustworthy AI algorithm automatically switches continuously to the fallback segmentation for this voxel.

For the automatic segmentation of the heart on chest CT images, one contract of trust could be: "The heart labels are always on the left side of the body." The AI algorithm can fulfil this contract and yet compute an inaccurate segmentation of the heart. This contract can also be restricted to CT images of sufficient quality. Context can be added to the contract and several contracts can be derived from the contract above for different contexts. In the previous example, context is implied in that this contract does not apply to individuals with dextrocardia.

An AI algorithm is defined as **trustworthy** with respect to a contract of trust if it provides guarantees that it will abide by the obligations of said contract [24].

The requirements proposed in the EU guidelines for trustworthy AI [7] are examples of contracts of trust. One important requirement of trustworthy AI is the technical robustness and safety [7] which is the focus of our work. The EU guidelines propose to achieve trustworthiness in practice using a *fallback plan*. However, no technical means of implementing such plan have been provided or published. Problems with the backbone AI algorithm should be detected using a *fail-safe* algorithm and a *fallback* algorithm should be available. In the remaining of this section we will present a theoretical framework for the implementation of a trustworthy AI system leveraging Dempster-Shafer theory to implement a failsafe and a fallback plan. And we will show how our framework can be used to maintain several concrete contracts of trust for fetal brain MRI segmentation.

3.2 Background on Dempster-Shafer theory

In Dempster-Shafer theory [26], *basic probability assignments* are a generalization of probabilities that allow to model partial and imprecise information and to combine different sources of information using Dempster's rule.

Let $\mathbf{C}$ be the set of all classes and $2^{\mathbf{C}}$ the set of all subsets of $\mathbf{C}$. A **basic probability assignment (BPA)** on $\mathbf{C}$ is a function $m : 2^{\mathbf{C}} \mapsto [0, 1]$ that satisfies

$$m(\emptyset) = 0 \quad \text{and} \quad \sum_{A \subset \mathbf{C}} m(A) = 1 \tag{1}$$

Probabilities on $\mathbf{C}$ are functions $p : \mathbf{C} \mapsto [0, 1]$ that satisfy $\sum_{c \in \mathbf{C}} p(c) = 1$. Probabilities are equivalent to the BPAs that assign non-zeros weights only to the singletons, i.e. the sets $A = \{c\}$ for $c \in \mathbf{C}$. Given a probability p the BPA $m^{(p)}$ associated with p is defined as: $\forall c \in \mathbf{C}$, $m^{(p)}(\{c\}) = p(c)$ and $\forall A \subset \mathbf{C}$ with $|A| \neq 1$, $m^{(p)}(A) = 0$. Basic probability assignments are therefore more general than probabilities.

For $A \subset \mathbf{C}$, $m(A)$ is the probability that our knowledge about the true label is exactly and only: "the true class is one of the classes in A". In particular, it does not imply that $m(B) > 0$ for any set B such that $B \subsetneq A$ or $A \subsetneq B$. This

is in contrast to probabilities that can weight only the individual classes. BPAs allow to represent more precisely than probabilities what we know (and don't know) about the true class of a voxel. For example, the extreme case where we know nothing about the true class can be represented by the BPA m such that $m(\mathbf{C}) = 1$. The best one can do to try representing this case with probabilities is to define a probability p such that $\forall c \in \mathbf{C}, p(c) = \frac{1}{|\mathbf{C}|}$. However, this choice of p corresponds to the knowledge that the class distribution is uniformly random which is different from knowing nothing about the true class.

Finally, two BPAs on $\mathbf{C}$, m_1 and m_2, are said to be **completely contradictory** if

$$\sum_{E,F\subset\mathbf{C}|E\cap F=\emptyset} m_1(E)m_2(F) = 1 \tag{2}$$

Using (1), m_1 and m_2 are completely contradictory if and only if one cannot form a pair of overlapping sets of classes (A, B) such that m_1 commits some belief to A, i.e. $m_1(A) > 0$, and m_2 commits some belief to B, i.e. $m_2(B) > 0$.

3.2.1 Dempster's rule of combination

Dempster's rule of combination allows to combine any pair (m_1, m_2) of BPAs on $\mathbf{C}$ that are not completely contradictory using the formula, $\forall A \subset \mathbf{C}$,

$$(m_1 \oplus m_2)(A) = \begin{cases} \frac{\sum_{E,F\subset\mathbf{C}|E\cap F=A} m_1(E)m_2(F)}{1-\sum_{E,F\subset\mathbf{C}|E\cap F=\emptyset} m_1(E)m_2(F)} & \text{if } A \neq \emptyset \\ 0 & \text{if } A = \emptyset \end{cases} \tag{3}$$

It is worth noting that $m_1 \oplus m_2$ is also a BPA on $\mathbf{C}$. In addition, the relation $\oplus$ is symmetrical and associative.

One particular case that will be useful for our method is the combination of a probability p on $\mathbf{C}$ with a generic BPA m on $\mathbf{C}$ using Dempster's rule of combination.

Since p is a probability, for $A \subset \mathbf{C}$, $p(A)$ can be nonzeros only if A is a singleton, i.e. if it exists a class $c \in \mathbf{C}$ such that $A = \{c\}$. For simplicity, we will therefore use the abusive notation when considering p as a BPA: $p(c) := p(A) = p(\{c\})$. The relation of complete contradiction (2) between p and m can be simplified

$$\sum_{E,F\subset\mathbf{C}|E\cap F=\emptyset} p(E)m(F) = \sum_{c\in\mathbf{C}} \sum_{F\subset(\mathbf{C}\setminus\{c\})} p(c)m(F) = 1 \tag{4}$$

Similarly, if p and m are not completely contradictory, the Dempster's rule between p and m can be simplified. Let $A \subset \mathbf{C}$, $A \neq \emptyset$, using (3) we have

$$(p \oplus m)(A) = \frac{\sum_{c\in\mathbf{C}, F\subset\mathbf{C}|\{c\}\cap F=A} p(c)m(F)}{1-\sum_{c\in\mathbf{C}}\sum_{F\subset(\mathbf{C}\setminus\{c\})} p(c)m(F)} \tag{5}$$

We remark that $p \oplus m$ is also a probability on $\mathbf{C}$. Indeed, let $c \in \mathbf{C}$, it can exist $F \subset \mathbf{C}$ such that $\{c\} \cap F = A$ only if A is the singleton $A = \{c\}$ (we have assumed $A \neq \emptyset$). Therefore if A is not a singleton, i.e. $|A| > 1$, the sum on the numerator is empty and equal to 0. As a result, $\forall c \in \mathbf{C}$,

$$(p \oplus m)(c) = \frac{p(c)\sum_{F\subset\mathbf{C}|c\in F} m(F)}{1-\sum_{c'\in\mathbf{C}}\sum_{F\subset(\mathbf{C}\setminus\{c'\})} p(c')m(F)} \tag{6}$$

3.3 A Dempster-Shafer approach to Trustworthy AI

Our trustworthy AI segmentation method consists of three main components: 1) a backbone AI segmentation algorithm; 2) a fallback segmentation algorithm; and 3) and a fail-safe method that detects area of conflict between the AI algorithm segmentation and the contracts of trust and switches to the fallback algorithm for those regions. An illustration is given in Fig. 1.

The AI segmentation algorithm is a high-accuracy segmentor that can be, for example, a state-of-the-art convolutional neural network. The fallback segmentation algorithm is a segmentor that might achieve lower accuracy than the AI, but is superior to the AI for other desirable properties such as robustness. It is worth noting that AI and fallback segmentors are interchangeable in theory, and that either of them could consist of manual or semi-automatic segmentation methods. The AI and fallback segmentation algorithms take as input an image to be segmented and compute for each voxel of the image a probabilities vector with one probability for each class to be segmented.

The fail-safe mechanism aims at detecting erroneous predictions of the AI segmentation algorithm that contradict one of the contracts of trust. The contracts of trust embed domain knowledge such as "there can't be white matter in this part of the brain" or "hyperintense voxels on T2 fetal brain MRI are always cerebrospinal fluid". Most contract will not enforce a specific segmentation but rather impose that the automatic segmentation meets certain constraints. In the context of image segmentation, contract of trusts can only reduces the set of possible classes and reweights the class probabilities of the segmentation of a pixel or voxel. To implement the fail-safe mechanism, we propose to use a basic probability assignment (BPA) that acts on the backbone AI and the fallback class probabilities using Dempster's rule of combination (6). In addition, we assume that the fallback class probabilities never completely contradict the BPA representing the contracts of trust. As a result, Dempster's rule of combination can be used to switch automatically between the backbone AI algorithm and the fallback algorithm when the AI class probabilities completely contradict the BPA. Formally, the trustworthy segmentation prediction is defined for an input image I and for all voxel position $\mathbf{x}$ as

$$p_{I,\mathbf{x}}^{\text{TWAI}} = \left((1-\epsilon)p_{I,\mathbf{x}}^{\text{AI}} + \epsilon p_{I,\mathbf{x}}^{\text{fallback}}\right) \oplus m_{I,\mathbf{x}}^{\text{fail-safe}} \tag{7}$$

where $\oplus$ is the Dempster's combination rule (3), $p_{I,\mathbf{x}}^{\text{AI}}$ is the class probability prediction of the AI segmentation algorithm for voxel $\mathbf{x}$ of image I, $p_{I,\mathbf{x}}^{\text{fallback}}$ is the class probability prediction of the fallback segmentation for voxel $\mathbf{x}$ of image I, and $m_{I,\mathbf{x}}^{\text{fail-safe}}$ is the BPA of the fail-safe mechanism for voxel $\mathbf{x}$ of image I. The parameter ϵ is a constant in $]0, 1]$. A toy example is given in Appendix A.4

3.3.1 Fail-safe mechanism

In our framework, we assume that the fallback segmentation algorithm always produces segmentation probabilities that do not contradict entirely the BPA of the contracts of trust. A trivial example of such fallback, is the uniform segmentation algorithm that assigns an equal probability to all the classes to be segmented and for all voxels. In contrast, we do not

make such compatibility assumption for the AI segmentation algorithm. Not only does this make our approach applicable with any AI segmentation algorithm, but our method also relies on the incompatibility between the AI segmentation algorithm prediction and the contracts of trust to detect failure of the AI segmentation algorithm and to switch to the fallback segmentation algorithm. Formally, however small the weight ϵ given to the fallback is, as long as $\epsilon > 0$, when $p^{\text{AI}}_{I,\mathbf{x}}$ is completely contradictory with $m^{\text{fail-safe}}_{I,\mathbf{x}}$, we obtain that $p^{\text{TWAI}}_{I,\mathbf{x}}$ depends only on $p^{\text{fallback}}_{I,\mathbf{x}}$ and not on $p^{\text{AI}}_{I,\mathbf{x}}$. On the contrary, when $p^{\text{AI}}_{I,\mathbf{x}}$ is not completely contradictory with $m^{\text{fail-safe}}_{I,\mathbf{x}}$, we obtain that $p^{\text{TWAI}}_{I,\mathbf{x}}$ depends mainly on $p^{\text{AI}}_{I,\mathbf{x}}$ and the contribution of $p^{\text{fallback}}_{I,\mathbf{x}}$ is negligible for ϵ small enough. Here, we consider the case in which the AI algorithm predicted probability $p^{\text{AI}}_{I,\mathbf{x}}$ is completely contradictory with $m^{\text{fail-safe}}_{I,\mathbf{x}}$ for a voxel $\mathbf{x}$. Using (4)

$$\begin{cases} \sum\limits_{c' \in \mathbf{C}} \left(\sum\limits_{\mathbf{C}' \subset (\mathbf{C} \setminus \{c'\})} p^{\text{AI}}_{I,\mathbf{x}}(c')\, m^{\text{fail-safe}}_{I,\mathbf{x}}(\mathbf{C}') \right) = 1 \\ \forall c' \in \mathbf{C},\, \forall \mathbf{C}' \subset \mathbf{C} \,|\, c' \in \mathbf{C}', \quad p^{\text{AI}}_{I,\mathbf{x}}(c')\, m^{\text{fail-safe}}_{I,\mathbf{x}}(\mathbf{C}') = 0 \end{cases}$$

Using Dempster's rule of combination (6) we obtain,

$$\forall c \in \mathbf{C}, \quad p^{\text{TWAI}}_{I,\mathbf{x}}(c) = \left(p^{\text{fallback}}_{I,\mathbf{x}} \oplus m^{\text{fail-safe}}_{I,\mathbf{x}} \right)(c) \tag{8}$$

However small $\epsilon > 0$ can be, the trustworthy AI prediction for voxel $\mathbf{x}$ does not depend anymore on the AI algorithm probability but only on the fallback algorithm probability. In other words, we have switched totally from the backbone AI algorithm to the fallback algorithm.

3.3.2 General case with multiple contracts of trust

In general, $m^{\text{fail-safe}}_{I,\mathbf{x}}$ is a sum of contracts of trust BPAs that are not completely contradictory and can be written as

$$m^{\text{fail-safe}}_{I,\mathbf{x}} = \bigoplus_{k=1}^{K} m^{(k)}_{I,\mathbf{x}} \tag{9}$$

where each $m^{(k)}_{I,\mathbf{x}}$ is a basic probability assignment (BPA), K is the number of BPAs, and $\bigoplus_{k=1}^{K}$ is the Dempster's rule of combination (3) of K BPAs computed in any order. The $m^{(k)}_{I,\mathbf{x}}$ represent the contracts of trust in our framework.

Specifically, for medical image segmentation we propose the following trustworthy AI model:

$$p^{\text{TWAI}}_{I,\mathbf{x}} = \left((1-\epsilon) p^{\text{AI}}_{I,\mathbf{x}} + \epsilon p^{\text{fallback}}_{I,\mathbf{x}} \right) \oplus m^{\text{anatomy}}_{I,\mathbf{x}} \oplus m^{\text{intensity}}_{I,\mathbf{x}} \tag{10}$$

where $m^{\text{anatomy}}_{I,\mathbf{x}}$ is the anatomical contract of trust BPA for voxel $\mathbf{x}$ of image I, and $m^{\text{intensity}}_{I,\mathbf{x}}$ is the intensity contract of trust BPA for voxel $\mathbf{x}$ of image I. The definitions of $m^{\text{anatomy}}_{I,\mathbf{x}}$ and $m^{\text{intensity}}_{I,\mathbf{x}}$ will be derived in sections 3.3.3 and 3.3.4.

3.3.3 Dempster-Shafer anatomical contracts of trust

In this section, we describe our proposed anatomical prior basic probability assignment (BPA) m^{anatomy} that is used in our trustworthy AI method (10).

Our anatomical prior is computed using the segmentations computed using a multi-atlas segmentation algorithm [9]. Atlas-based segmentation algorithms are anatomically-constrained due to the spatial smoothness that is imposed to the spatial transformation used to compute the segmentation. In practice, this is achieved thanks to the parameterization of the spatial transformation and the regularization loss in the registration optimization problem [9], [27]. Therefore, if implemented correctly, atlas-based automatic segmentations can inherit from the anatomical prior represented by segmentation atlases.

In terms of contract of trust, every binary segmentation mask corresponding to a ROI in a fetal brain atlas is associated with an anatomical contract of trust. Each of those binary segmentation masks represents the anatomy of a given tissue type, for a given gestational age and a given population of fetuses. The anatomical contracts derived from atlas-based segmentation are therefore specific to a class, to a gestational age, and to the population of fetuses that was used to compute the atlas. Since only neurotypical fetal brain atlases [28], [29] and a spina bifida fetal brain atlas [30] are available in our work, our anatomical contract of trust will hold only for those two populations.

Due to the spatial smoothness imposed to the spatial transformation, atlas-based automatic segmentations will usually be correct up to a spatial margin. Therefore, we propose to to compute the BPAs of our anatomical contract of trust by adding spatial margins to the atlas-based segmentation. This approach is inspired by the safety margins used in radiotherapy to account for spatial registration errors [10]. Formally, let M^c a 3D (binary) mask from an atlas-based algorithm for class $c \in \mathbf{C}$. We propose to define the BPA map $m^{(c)} = \left(m^{(c)}_{\mathbf{x}} \right)_{\mathbf{x} \in \Omega}$ associated with M^c as

$$\forall \mathbf{x}, \quad \begin{cases} m^{(c)}_{\mathbf{x}}(\mathbf{C} \setminus \{c\}) = 1 - \phi(d(\mathbf{x}, M^c)) \\ m^{(c)}_{\mathbf{x}}(\mathbf{C}) = \phi(d(\mathbf{x}, M^c)) \end{cases} \tag{11}$$

where $d(\mathbf{x}, M^c)$ is the Euclidean distance from $\mathbf{x}$ to M^c, and $\phi : \mathbb{R}_+ \to [0, 1]$ with $\phi(0) = 1$ and ϕ non-increasing. We note that while $m^{(c)}$ is a BPA, it cannot be considered as a probability as it does not sum to one. In the following, we use the function

$$\forall d \geq 0, \quad \phi(d) = \begin{cases} 1 & \texttt{if } d \leq \eta \\ 0 & \texttt{otherwise} \end{cases} \tag{12}$$

where $\eta > 0$ is a hyper-parameter homogeneous to a distance and can be interpreted as a *safety margin* for the anatomical prior. We describe a method to tune the margins at training time for each class in appendix A.8. The BPA for this function ϕ can be implemented efficiently without computing explicitly the distance between every voxel $\mathbf{x}$ and the mask M^c. With the definition of the BPA $m^{(c)}_{\mathbf{x}}$ in (11), we formalize the following belief: far enough from the mask M^c we know for sure that the true class is not c, i.e. $m^{(c)}_{\mathbf{x}}(\mathbf{C} \setminus \{c\}) = 1$, otherwise we do not know anything for sure regarding class c, i.e. $m^{(c)}_{\mathbf{x}}(\mathbf{C}) > 0$.

The BPAs m_c defined as in (11) are nowhere completely contradictory with each other. A proof can be found in Appendix A.11. Therefore, we can define the anatomical prior BPA used in (10) for image I and voxel $\mathbf{x}$ as

$$m^{\text{anatomy}}_{I,\mathbf{x}} = \bigoplus_{c \in \mathbf{C}} m^{(c)}_{\mathbf{x}} \tag{13}$$

where m_c is the BPA associated to the mask M^c_I for class c of the segmentation obtained using the multi-atlas fallback

Algorithm 1 Dempster-Shafer anatomical contract of trust.

Require: $(p_{I,\mathbf{x}})_{\mathbf{x}\in\Omega}$: input class probability map.
Require: $(M^c)_{c\in\mathbf{C}}$: binary masks prior for all classes.
for $c \in \mathbf{C}$ **do**
  $M^c \leftarrow Dilate_{\eta_c}(M^c)$ ▷ Dilate mask by margin η_c
  $\forall \mathbf{x} \in \Omega, \quad p_{I,\mathbf{x}}(c) \leftarrow p_{I,\mathbf{x}}(c) \times M^c_{\mathbf{x}}$ ▷ Mask
$\forall \mathbf{x} \in \Omega, \forall c \in \mathbf{C} \quad p_{I,\mathbf{x}} \leftarrow \frac{p_{I,\mathbf{x}}(c)}{\sum_{c'} p_{I,\mathbf{x}}(c')}$ ▷ Normalize
Output: $(p_{I,\mathbf{x}})_{\mathbf{x}\in\Omega}$

segmentation algorithm (see Appendix A.7). We prove in Appendix A.11 that the proposed anatomical prior BPA is never completely contradictory with the fallback.

We prove that for all voxel $\mathbf{x}$ and for all subset of classes $\mathbf{C}' \subset \mathbf{C}$, the anatomical BPA mass that the true label of $\mathbf{x}$ is not in $\mathbf{C}'$ is equal to

$$m^{\text{anatomy}}_{I,\mathbf{x}}(\mathbf{C} \setminus \mathbf{C}') = \prod_{c\in\mathbf{C}} \Big(\delta_c(\mathbf{C}')m^{(c)}_{\mathbf{x}}(\mathbf{C} \setminus \{c\}) + (1-\delta_c(\mathbf{C}'))m^{(c)}_{\mathbf{x}}(\mathbf{C})\Big) \tag{14}$$

where for all $c \in \mathbf{C}$, δ_c is the Dirac measure defined as

$$\forall \mathbf{C}' \subset \mathbf{C}, \quad \delta_c(\mathbf{C}') = \begin{cases} 1 & \text{if } c \in \mathbf{C}' \\ 0 & \text{if } c \notin \mathbf{C}' \end{cases} \tag{15}$$

The proof of (14) can be found in the Appendix.

In practice, we are particularly interested in summing the anatomical prior BPA with probabilities using the particular case of Dempster's rule in (6). Let $\mathbf{x}$ a voxel and $p_{I,\mathbf{x}}$ a probability on $\mathbf{C}$ for voxel $\mathbf{x}$ of image I that is not completely contradictory with $m^{\text{anatomy}}_{I,\mathbf{x}}$. For all $c \in \mathbf{C}$, we can show that

$$\left(p_{I,\mathbf{x}} \oplus m^{\text{anatomy}}_{I,\mathbf{x}}\right)(c) = \frac{p_{I,\mathbf{x}}(c)m^{(c)}_{\mathbf{x}}(\mathbf{C})}{\sum_{c'\in\mathbf{C}} p_{I,\mathbf{x}}(c')m^{(c')}_{\mathbf{x}}(\mathbf{C})} \tag{16}$$

A proof of this equality can be found in the Appendix and the pseudo-code can be found in Algorithm 1 . It is worth noting that, due to the specific form of $m^{\text{anatomy}}_{I,\mathbf{x}}$ and because $p_{I,\mathbf{x}}$ is a probability, the computational cost of $p_{I,\mathbf{x}} \oplus m^{\text{anatomy}}_{I,\mathbf{x}}$ is $\mathcal{O}(\mathbf{C})$ even though there are $2^{|\mathbf{C}|}$ elements in $2^{\mathbf{C}}$. Another important remark is that when $p_{I,\mathbf{x}}$ is completely contradictory with $m^{\text{anatomy}}_{I,\mathbf{x}}$, we have $\sum_{c'\in\mathbf{C}} p_{I,\mathbf{x}}(c')m^{(c')}_{\mathbf{x}}(\mathbf{C}) = 0$.

Tuning the margins: The margins η were tuned for each class and each condition independently using the 3D MRIs of the fold 0 of the training dataset. More details can be found in the appendix A.8.

3.3.4 Dempster-Shafer intensity-based contracts of trust

In this section, we describe our proposed intensity prior BPA $m^{\text{intensity}}$ that is used in our trustworthy AI method for fetal brain 3D MRI segmentation (10).

In T2-weighted MRI, hyper-intense voxels inside the brain are highly likely to be part of the cerebrospinal fluid (CSF). Voxels outside the brain (*background* class) can also be hyper-intense but not the non-CSF tissue types. We therefore propose to model this intensity prior about high intensities as a contract of trust. Regarding hypo-intense voxels, it is unclear how to derive similar prior because even the CSF classes contain hypo-intense voxels, such as the choroid plexus for the intra-axial CSF class and the vein of Galena and straight sinus for the extra-axial CSF class [31].

Let $\mathbf{C}_{high} \subset \mathbf{C}$ be the subset of classes that contain all the classes that partition the entire CSF (intra-axial CSF and extra-axial CSF) and the background. Let $I = \{I_{\mathbf{x}}\}_{\mathbf{x}\in\Omega}$ be the volume and Ω the volume domain of a fetal brain 3D MRI. We propose to fit a Gaussian mixture model (GMM) with two components to the image intensity distribution of I. The two components of parameters $(\mu_{high}, \sigma_{high})$ and $(\mu_{low}, \sigma_{low})$ are associated to high and low intensities. We propose to define the intensity prior BPA for all voxels $\forall \mathbf{x}$, up to a normalization factor, as

$$\begin{cases} m^{\text{intensity}}_{I,\mathbf{x}}(\mathbf{C}_{high}) \propto \dfrac{1}{\sigma_{high}} \exp\left(\dfrac{1}{2}\left(\dfrac{I_{\mathbf{x}} - \mu_{high}}{\sigma_{high}}\right)^2\right) \\ m^{\text{intensity}}_{I,\mathbf{x}}(\mathbf{C}) \propto \dfrac{1}{\sigma_{low}} \exp\left(\dfrac{1}{2}\left(\dfrac{I_{\mathbf{x}} - \mu_{low}}{\sigma_{low}}\right)^2\right) \end{cases} \tag{17}$$

It is worth noting that $m^{\text{intensity}}_{I,\mathbf{x}}(\mathbf{C}) > 0$. Therefore, no probability will be set to 0 using the Dempster's rule of combination with $m^{\text{intensity}}$. In other words, $m^{\text{intensity}}$ does not forbid any assignment. This is in contrast with the anatomical BPAs defined in section 3.3.3.

Let $\mathbf{x}$ a voxel and $p_{I,\mathbf{x}}$ a probability on $\mathbf{C}$ for voxel $\mathbf{x}$ of image I. Since $m^{\text{intensity}}_{I,\mathbf{x}}(\mathbf{C}) > 0$, $p_{I,\mathbf{x}}$ is not completely contradictory with $m^{\text{intensity}}_{I,\mathbf{x}}$. Using Dempster's rule, we have, for all class $c \in \mathbf{C}$

$$\left(p_{I,\mathbf{x}} \oplus m^{\text{intensity}}_{I,\mathbf{x}}\right)(c) \propto \begin{cases} \left(1 + \dfrac{m^{\text{intensity}}_{I,\mathbf{x}}(\mathbf{C}_{high})}{m^{\text{intensity}}_{I,\mathbf{x}}(\mathbf{C})}\right) p_{I,\mathbf{x}}(c) \text{ if } c \in \mathbf{C}_{high} \\ p_{I,\mathbf{x}}(c) \text{ otherwise} \end{cases} \tag{18}$$

This can be interpreted as a soft-thresholding operation Thus, only the probabilities for the background and CSF classes in $\mathbf{C}_{high}$ are increased in the case of a voxel $\mathbf{x}$ with relatively high intensity. In particular, the probabilities remain approximately unchanged for a voxel $\mathbf{x}$ with relatively low or medium intensity. This reflects the fact that the background and CSF classes also contain hypo-intense voxels. The hyper-intense voxels must be in $\mathbf{C}_{high}$ while we ca not say anything about hypo-intense voxels in general. There are hypo-intense voxels in every class.

4 Experiments

4.1 Evaluation on a large multi-center dataset.

To effectively evaluate the performance of our trustworthy AI framework as a suitable method to improve the trustworthiness of a backbone AI model using a fallback model, we have selected the task of fetal brain segmentation in 3D MRI. This task is clinically relevant and is characterized by large image protocol variations and large anatomical variations.

Deep learning-based AI methods for fetal brain MRI segmentation have recently defined state-of-the-art segmentation performance [31], [32], [33], [34], [35], [36], [37], [38], gradually replacing image registration-based segmentation methods [39] in the literature. Most previous work on deep learning for fetal brain MRI segmentation trained

TABLE 1
Evaluation of our contracts of trust for different AI models. Best values for each AI model are in **bold** and best values overall are underlined. IN: in-scanner distribution, OUT: out-of-scanner distribution, int.: intensity contract of trust, anat.: anatomical contract of trust.

Backbone AI model	Contract of trust		Mean-ROI Dice Score (in %)						Mean-ROI HD95 (in mm)					
			Neurotypical		**Spina Bifida**		**Other Path.**		**Neurotypical**		**Spina Bifida**		**Other Path.**	
	int.	**anat.**	IN	OUT	IN	OUT	IN	OUT	IN	OUT	IN	OUT	IN	OUT
Fallback	NA	NA	85.7 (2.2)	84.1 (3.6)	78.8 (6.2)	76.2 (10)	78.6 (9.2)	82.5 (6.0)	1.5 (0.3)	1.5 (0.4)	2.3 (0.7)	2.8 (1.6)	3.7 (2.6)	2.4 (1.6)
nnU-Net [1]	✗	✗	90.4 (1.8)	86.6 (3.8)	80.6 (6.9)	75.2 (14)	**83.6 (8.7)**	82.7 (5.6)	2.0 (0.4)	1.9 (0.4)	3.5 (1.7)	4.5 (2.9)	3.5 (2.3)	3.3 (2.5)
	✓	✗	90.4 (1.8)	86.7 (3.9)	80.8 (6.9)	75.3 (14)	**83.6 (8.6)**	83.0 (5.0)	2.0 (0.4)	1.8 (0.4)	3.4 (1.7)	4.6 (2.9)	3.5 (2.3)	3.3 (2.4)
	✗	✓	91.0 (1.7)	**87.4 (3.7)**	82.1 (5.8)	**78.1 (10)**	83.3 (9.7)	84.0 (5.7)	**1.2 (0.2)**	**1.3 (0.3)**	**2.3 (1.1)**	**2.8 (1.6)**	**3.2 (2.6)**	2.1 (1.3)
	✓	✓	**91.1 (1.6)**	**87.4 (3.7)**	**82.2 (5.8)**	78.0 (11)	83.3 (9.7)	**84.2 (5.4)**	**1.2 (0.2)**	**1.3 (0.3)**	**2.3 (1.1)**	**2.8 (1.6)**	**3.2 (2.6)**	**2.0 (1.3)**
SwinUNETR with SSL pre-training [20]	✗	✗	84.3 (5.3)	77.9 (3.6)	74.2 (11)	65.2 (14)	79.9 (10)	79.2 (5.7)	2.9 (1.0)	3.1 (0.8)	5.3 (3.1)	6.6 (3.0)	4.9 (3.0)	6.1 (2.4)
	✓	✗	84.8 (5.1)	78.2 (3.6)	75.1 (10)	66.5 (14)	80.5 (10)	79.6 (5.6)	2.8 (1.0)	3.0 (0.8)	5.2 (3.1)	6.4 (3.1)	4.7 (2.8)	6.1 (2.5)
	✗	✓	86.5 (4.3)	80.7 (4.0)	78.4 (7.5)	70.9 (11)	81.1 (10)	81.1 (6.0)	1.4 (0.3)	1.7 (0.4)	**2.4 (1.0)**	3.1 (1.5)	**3.3 (2.6)**	**2.2 (1.3)**
	✓	✓	**87.0 (3.9)**	**81.4 (3.9)**	**78.8 (7.3)**	**71.6 (11)**	**81.5 (10)**	**81.4 (6.0)**	**1.3 (0.2)**	**1.6 (0.3)**	**2.4 (1.0)**	**3.0 (1.5)**	**3.3 (2.6)**	**2.2 (1.2)**
nnU-Net [1] with atlas features fusion [22], [23]	✗	✗	90.5 (1.9)	86.6 (3.8)	80.4 (7.0)	76.8 (11)	83.4 (8.5)	82.5 (5.5)	1.9 (0.3)	1.8 (0.3)	3.5 (1.8)	3.9 (1.8)	3.5 (2.3)	2.7 (1.2)
	✓	✗	90.5 (1.9)	86.7 (3.8)	80.5 (7.0)	76.7 (12)	**83.5 (8.4)**	82.8 (4.9)	1.9 (0.3)	1.8 (0.3)	3.5 (1.7)	3.9 (1.8)	3.5 (2.3)	2.6 (1.1)
	✗	✓	91.1 (1.7)	87.3 (3.7)	82.0 (5.9)	**78.5 (10)**	82.9 (9.8)	83.4 (6.2)	**1.2 (0.2)**	**1.3 (0.3)**	**2.3 (1.1)**	**2.7 (1.5)**	**3.3 (2.6)**	**2.1 (1.3)**
	✓	✓	**91.2 (1.6)**	**87.4 (3.7)**	**82.1 (5.8)**	78.4 (10)	82.9 (9.8)	**83.6 (5.8)**	**1.2 (0.2)**	**1.3 (0.3)**	**2.3 (1.1)**	**2.7 (1.6)**	**3.3 (2.6)**	**2.1 (1.2)**
Ensemble nnU-Net [1] + atlas	✗	✗	90.9 (1.7)	87.7 (3.8)	81.6 (6.5)	77.9 (11)	83.5 (9.3)	84.6 (5.4)	1.6 (0.3)	1.5 (0.2)	3.1 (1.6)	3.4 (1.6)	3.4 (2.4)	2.5 (1.5)
	✓	✗	91.0 (1.7)	87.7 (3.9)	81.7 (6.6)	77.9 (12)	**83.9 (8.6)**	84.9 (4.8)	1.6 (0.2)	1.5 (0.2)	3.1 (1.6)	3.4 (1.6)	3.4 (2.5)	2.4 (1.5)
	✗	✓	**91.2 (1.7)**	**87.9 (3.8)**	**82.5 (5.7)**	**79.0 (10)**	83.2 (10)	84.8 (6.2)	**1.1 (0.2)**	**1.3 (0.3)**	**2.3 (1.1)**	**2.7 (1.6)**	**3.3 (2.6)**	**2.0 (1.3)**
	✓	✓	**91.2 (1.7)**	**87.9 (3.8)**	**82.5 (5.8)**	78.6 (11)	83.2 (9.9)	**85.0 (5.7)**	**1.1 (0.2)**	**1.3 (0.3)**	**2.3 (1.1)**	**2.7 (1.6)**	**3.3 (2.6)**	**2.0 (1.2)**

and evaluated their models using only MRIs of healthy fetuses or only MRIs acquired at one center. However, the segmentation performance of deep learning methods typically degrades when images from a different center or a different scanner vendor as the one used for training are used or when evaluating the segmentation performance on abnormal anatomy [40], [41], [42], [43], [44], [45]. One study has reported such issues for fetal brain MRI segmentation [4]. Thus, we have evaluated the proposed trustworthy AI approach with four different backbone AI algorithms based on deep learning [1], [20], [22], [23] and a fallback algorithm consisting of a registration-based segmentation method [9]. Details of the backbone AI and fallback methods can be found in the appendixA.5A.6A.7. We have used a large multi-centric fetal brain MRI dataset that consists of a total of 540 3D MRIs with neurotypical or abnormal brain development, with gestational ages ranging from 19 weeks to 40 weeks, and with MRIs acquired at 13 hospitals across six countries. The task consists of segmenting automatically a fetal brain 3D MRI into eight tissue types: the corpus callosum, the white matter, the cortical gray matter, the deep gray matter, the cerebellum, the brainstem, the intra-axial cerebrospinal fluid (CSF), and the extra-axial CSF.

4.2 Stratified evaluation across brain conditions and acquisition centers.

The evaluation of AI-based segmentation algorithms has shown that the performance of deep learning models can vary widely across clinically relevant populations and across data acquisition protocols [4] (Fig. 1b, Fig. 2).

Therefore, we performed a stratified comparison of the backbone AI algorithms, the fallback algorithm, and the trustworthy AI approach across two groups of acquisition centers and three groups of brain conditions. The composition of the dataset for each group is summarized in Fig. A.1 and detailed in section A.1. The acquisition centers were split into two groups, that we called *in-scanner distribution* and *out-of-scanner distribution*, depending if 3D MRIs acquired at a given center were present in the training dataset or not. Four out of thirteen data sources were used to train the backbone AI algorithms. In addition, the 3D MRIs were also separated based on the underlying brain condition of the fetus. The first group, *neurotypical*, contains the fetuses diagnosed by radiologists with a normal brain development using ultrasound and MRI. The second group, *spina bifida*, contains the fetuses with a condition called spina bifida aperta. We use the term *spina bifida* for short in this work. Cases of spina bifida aperta are typically accompanied by severe anatomical brain abnormalities [30], [46] with a type II Chiari malformation and an enlargement of the ventricles being most prevalent. The Chiari malformation type II is characterized by a small posterior fossa and hindbrain herniation in which the medulla, cerebellum, and fourth ventricule are displaced caudally into the direction of the spinal canal [47]. The third group, *other pathologies*, contains fetuses with various pathologies other than spina bifida and causing an abnormal brain development, such as corpus callosum agenesis and dysgenesis, intracranial hemorrhage and cyst, aqueductal stenosis, and Dandy-Walker malformation. Those other pathologies were not present in the training dataset of the backbone AI algorithms and spatio-temporal atlases are not available for the fallback and the fail-safe algorithm. Hence, testing 3D MRIs classified as other pathologies allow us to measure the segmentation performance of the trustworthy AI approach outside of the domain covered by the anatomical contracts of trust. Table 1 shows the results of the overall stratified evaluation in terms of Dice score and Hausdorff distances at 95% percentile for four different backbone AI models. The detailed results per ROI for nnU-Net as backbone AI can be found in the appendix (Fig. A.2,A.3). Statistical differences were evaluated using a Wilcoxon singed-rank test using the threshold 0.05

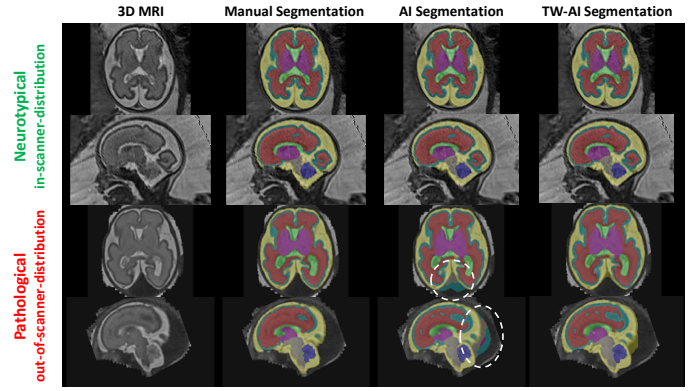


Fig. 2. Illustration of the improved robustness of the proposed trustworthy AI method (TW-AI) as compared to nnU-Net state-of-the-art backbone AI method. (Top) 3D MRI of a neurotypical fetus at 28 weeks of gestation acquired at the same center as the training data for the AI. (Bottom) 3D MRI of a fetus with a high-flow dural sinus malformation at 28 weeks of gestation acquired at a different center as the training data for the AI. Severe violations of the anatomy by the backbone AI are highlighted. The TW-AI does not make those errors.

for the p-values.

4.3 Comparison with other fusion methods.

In Table 1, we compared our Dempster-Shafer trustworthy AI with nnU-Net as backbone AI, *TW-nnU-Net* for short, to two other fusion methods based only on probabilities: the ensemble average of the predicted probabilities of nnU-Net [1] with the ones of the atlas-based fallback method [9], and *nnU-Net with atlas features fusion*, in which the fusion operation between the atlas features and nnU-Net features is learned during training. In this case, the atlas-based fallback probabilities are concatenated with the high-resolution deep features before the last level of the decoder [22], [23].

We found that our TW-nnU-Net significantly outperformed nnU-Net with atlas features fusion for all groups and all metrics except in terms of Dice score for the in-scanner distribution other pathologies group for which they performed similarly. Our TW-nnU-Net also significantly outperformed the ensemble average for all groups in terms of Hausdorff distance. Our TW-nnU-Net and the ensemble average performed similarly in terms of Dice score.

4.4 Ablation study of the proposed contracts of trust.

The results in Table 1show the benefits of both intensity and anatomical contracts of trust. The anatomical contract of trust lead to similar or significantly better segmentation results for all backbone AI methods, all groups, and all metrics.We found that our intensity contract of trust applied to SwinUNETR [20] lead to significant improvement of the Dice score for the majority of the groups. For the other backbone AI models, all based on nnU-Net [1], the segmentation metrics were similar for all groups with or without the use of the intensity contract of trusts. We hypothesize that the various data augmentations used in nnU-Net, previously proposed in the domain generalization literature [19], allow those backbone AI models to learn robust intensity prior. Similarly, we found that combining anatomical and intensity contracts of trust lead to significant improvement of the mean-ROI Dice score for SwinUNETR for the majority of the groups but to similar segmentation performance for the other backbone AI models. Those results also show that our trustworthy AI approach can be successfully applied to various backbone AI models.

4.5 Scoring of trustworthiness by radiologists.

The Dice score and the Hausdorff distance are the two most standard metrics used for measuring the quality of automatic segmentations. However, those two metrics do not directly measure the trustworthiness of segmentation algorithms [48]. Therefore, we have also conducted an evaluation of the trustworthiness of the automatic segmentations as perceived by radiologists. We have asked a panel of eight experts from four different hospitals to score the trustworthiness of automatic segmentations from 0 (totally unacceptable) to 5 (perfect fit) for each region of interest and for the nnU-Net backbone AI, the fallback, and the corresponding trustworthy AI algorithms. Independent scoring were performed by raters at different hospitals. The scoring protocol and details about the panel of experts can be found in section A.2. The scoring was performed for the same 3D MRIs by all radiologists. We have used a subset of 50 3D MRIs from the out-of-distribution group of the testing dataset consisting of 20 neurotypical fetuses, 20 spina bifida fetuses, and 10 fetuses with other abnormalities. Those cases were selected per condition at random among the 3D MRIs of the publicly available FeTA dataset [31]. The *out-of-scanner distribution* group is the most relevant group for the evaluation of trustworthiness because this corresponds to the situation in which AI algorithms generalization is the most challenging and clinically relevant. The overall scoring results can be found in Fig. 3 and the detailed results per region of interest can be found in the appendix (Fig. A.6).

There were 7 volumes out of 50 for which the trustworthy AI approach achieved a lower average score than the backbone AI model. However, with a maximum decrease of -0.32 on our range of scores, we consider that there were no failure cases. In contrast, there were 33 volumes for which the trustworthy AI approach improved the backbone AI model average score by more than 0.32, including 7 volumes with an increase superior to 1 and a maximum of 1.8.

Expert raters noticed that the algorithms were dependent on the quality of the 3D MRIs they were based on for the *spina bifida* group. We found a positive correlation between the mean-class trustworthiness scores and the quality of the 3D MRI for the *spina bifida* group (Pearson $r = 0.43$). There was no correlation between scores and 3D MRI quality for the *neurotypical* group (Pearson $r = -0.1$) and the 3D MRIs of the *other pathologies* were all of high quality. In addition, the more structurally abnormal the brains were due to the pathologies, the more difficult it was to compare the algorithms. In the case of the Chiari malformations, this applies in particular to the cerebellum and brainstem.

4.6 Stratified evaluation across gestational ages.

The anatomy and the size of the fetal brain change significantly from 19 weeks of gestation until term for both neurotypical fetuses [28] and fetuses with spina bifida [30].

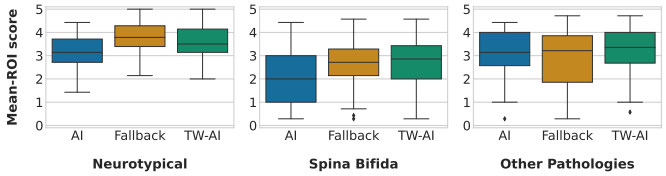


Fig. 3. **Mean-ROI Trustworthiness Scores for out-of-scanner distribution 3D MRIs.** We report four scoring by a panel of eight experts of the trustworthiness of the automatic segmentations for a subset of the out-of-scanner distribution testing 3D MRIs ($n = 50$). Each expert was asked to score from 0 (totally unacceptable) to 5 (perfect fit) the trustworthiness of each ROI. The scores displayed here are averaged across ROIs. AI corresponds to nnU-Net [1] here. Results per ROI can be found in the appendix (Fig. A.6).

This age-related variability is a challenge for segmentation algorithms for fetal brain MRI [4].

We analysed the performance of the proposed trustworthy AI algorithm for fetal brain segmentation as a function of the gestational age and compared it to the nnU-Net [1] backbone AI algorithm and the fallback algorithm based on image registration [9]. We grouped the fetuses with neurotypical or spina bifida condition with the same gestational age rounded to the closest week. The mean and the confidence intervals at 95% for the overall performance in terms of Dice sore (resp. Hausdorff distance) across regions of interest can be found in Fig. 4a (resp. Fig. 4b). The detailed results per region of interest can be found in the appendix (Fig. A.4,A.5). Overall, the nnU-Net backbone AI algorithm achieves higher Dice scores than the fallback algorithm, while the fallback achieved lower Hausdorff distances than the backbone AI method (Fig. 4). Our proposed trustworthy AI algorithm successfully combines backbone AI and fallback algorithms. It achieves higher or similar segmentation performance than those two algorithms in terms of established segmentation quality metrics such as the Dice score and the Hausdorff distance across all gestational ages for *neurotypical* and *spina bifida*.

5 Discussion and Conclusion

5.1 A principled and practical trustworthy AI method.

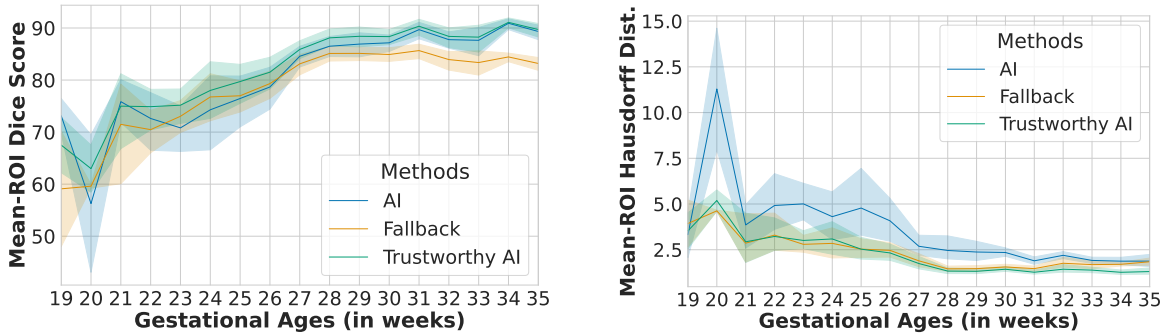
We have mathematically formalized a method for trustworthy AI with a fallback based on Dempster-Shafer theory. For application to fetal brain MRI segmentation, we have shown that our trustworthy AI method can be implemented using anatomy-based and intensity-based priors. We have proposed to interpret those priors as contracts of trust in Human-AI trust theory. Altogether, we showed that our principled trustworthy AI method improves the robustness and the trustworthiness of four state-of-the-art backbone AI algorithms and outperforms other fusion methods based on probabilities for fetal brain 3D MRI segmentation.

5.2 Complementarity of AI and atlas-based algorithms.

AI-based algorithms and registration-based algorithms have different error patterns. In several situations we have found that the registration-based method tends to achieve better segmentation performance in terms of Hausdorff distance as compared to the AI-based method while the AI-based method achieved better segmentation performance in terms of Dice score. We have found that the segmentation performance of the fallback algorithm decreases less than for the backbone AI algorithms, when comparing out-of-scanner distribution to in-scanner distribution for neurotypical and spina bifida fetal brain 3D MRIs. In our scoring of trustworthiness on out-of-scanner distribution data, we have also found that the fallback algorithm outperformed the nnU-Net backbone AI algorithm for neurotypical and spina bifida cases (Fig. 3). We think this is because the anatomical prior used by registration-based segmentation methods prevents mislabelling voxels far from the real anatomy. In contrast, AI-based method are unconstrained and such errors can occur. This is what we observe for the out-of-distribution cases displayed in Fig. 1c, 2. Our proposed fail-safe method uses the registration-based segmentation with added margins with the aim to automatically detect and discard such errors that were found to occur more often for AI-based approach than for registration-based approach.

5.3 The contracts of trust hold for sub-populations covered by brain atlases.

Our implementation of trustworthy AI for fetal brain segmentation depends on the availability of spatio-temporal segmentation atlases of the fetal brain in 3D MRI. While such atlases currently exist for neurotypical fetal brain [28], [29] and fetuses with spina bifida [30], it is not the case for other fetal brain pathologies. Therefore, our contracts of trust are not expected to hold for the group *other pathologies*. This illustrates how AI trustworthiness is context-dependent. We found that the *other pathologies* group is the only one for which radiologists associated the fallback method, based solely on the atlases, with a lower trustworthy scores than the backbone AI algorithm (Fig. 3). Surprisingly, we found that the trustworthy AI algorithm still performs better or on a par with the backbone AI algorithm for the *other pathologies* group. We associate this with the use of our margins and to the proposed voxel intensity prior for the cerebrospinal fluid that are specific to the trustworthy AI algorithm. For the *other pathologies* group, we used the margin values estimated for spina bifida. Our group *other pathologies* gathers diverse rare developmental diseases associated with different variations of the fetal brain anatomy. However, due to the low number of examinations available per pathology, grouping them was necessary for evaluation purposes. This introduces biases when comparing the segmentation performance for the *other pathologies* groups associated with in-scanner and out-of-scanner distribution. In particular, some of the fetuses with *other pathologies* in the in-scanner distribution had very severe brain anatomical abnormalities, such as acqueductal stenosis with large supratentorial ventricles and caudal discplacement of the cerebellum or intracranial hemorrhage with parenchymal destruction and ventriculomegaly. In contrast, the one in the out-of-scanner distribution have milder brain abnormalities, such as moderate ventriculomegaly, and there were no cases with parenchymal destruction. This explains why, for this condition, we observe more outliers with low Dice scores and high Hausdorff distances for the backbone AI algorithms for in-scanner-distribution as compared to out-of-distribution 3D MRIs (Table 1). This is also the only group



(a) **Mean and 95% CI of mean-ROI Dice Scores** (b) **Mean and 95% CI of mean-ROI HD95**

Fig. 4. **Comparison of the backbone AI, fallback, and trustworthy AI segmentation algorithms across gestational ages, for neurotypical and spina bifida cases.** AI corresponds to nnU-Net [1] here. Results per ROI can be found in the appendix (Fig. A.4, A.5).

for which the trustworthy AI method does not significantly outperform the backbone AI methods in terms of Dice score.

The two histograms of gestational ages for the training spina bifida 3D MRIs and the in-scanner-distribution testing spina bifida 3D MRIs are not uniform and have the same shape (see Fig. A.1). In contrast, the histogram of gestational ages for the out-of-scanner-distribution testing spina bifida is more uniform. This might partly explain the degradation of Dice scores and Hausdorff distances between in-scanner-distribution and out-of-scanner-distribution for the backbone AI algorithms (Table 1). Training and in-scanner-distribution testing spina bifida MRIs were mostly clinical data acquired at UHL. In this center, MRI of spina bifida are typically performed a few days before and after the surgery that is performed prior to 26 weeks of gestation. In addition, a follow-up MRI is sometimes performed one month after the surgery. This explains the two modes observed in the histograms for those two groups. In the training data, the use of the spina bifida atlas [30], that has a uniform gestational age distribution, makes the second mode less visible. Our results suggest the trustworthy AI algorithm is more robust than the AI algorithm to the gestational ages distributional shift between training and testing.

For gestational ages lower than 27 weeks, the Dice scores and Hausdorff distances degrade for all the algorithms (Fig. 4a,4b). For the nnU-Net backbone AI this is surprising given that more MRIs acquired at gestational ages lower than 27 weeks than higher were present in the training dataset (Fig. A.1). Poorer MRI quality, which is typical for younger fetuses, might explain this degradation. In addition, the ratio of spina bifida over neurotypical examinations is higher for gestational ages lower than 27 weeks in our dataset. The abnormal brain anatomy of spina bifida cases leads to more difficult segmentation compared to neurotypical cases. This is particularly the case for several classes: the cerebellum, the extra-axial cerebrospinal fluid (CSF), the cortical gray matter, and the brainstem (Fig. A.2,A.3,A.4,A.5,A.6). The cerebellum is more difficult to detect using MRI before surgery as compared to early or late after surgery [49], [50]. This has already been found to affect the segmentation performance of AI-based algorithms [4]. For neurotypical fetuses, the extra-axial CSF is present all around the cortex. However, for fetal brain MRI of spina bifida fetuses with gestational ages of 27 weeks or less this is often not the case and the extra-axial CSF might be reduced to several small connected components that do not embrace the entire cortex anymore. The spina bifida atlas does not cover well this variability of the extra-axial CSF [30]. Due to the explicit spatial regularization, medical image registration cannot tackle such differences of topology. Therefore, using the atlas currently available, the contract of trust for extra-axial CSF does not apply for this group of spina bifida cases. It can also influence nearby regions, such as the cortical gray matter in this case. For the fallback algorithm and the trustworthy AI algorithm, a further degradation of the segmentation performance for gestational ages lower than 21 weeks was expected because the fetal brain atlases used start at 21 weeks. For gestational ages of 21 weeks or higher, the trustworthy AI outperforms either the backbone AI-algorithm or the fallback algorithm and performs better or on a par with the best other algorithms for all regions of interest in terms of Dice score and Hausdorff distance (Fig. A.4,A.5). The confidence intervals are also similar or narrower for the trustworthy AI algorithm than for the other algorithms for gestational ages higher or equal to 21 weeks. This illustrates that our contracts of trust improve the robustness of the proposed trustworthy AI algorithm for spina bifida for the range of gestational ages covered by the atlas used [30].

5.4 Future work.

For this work we have created the largest manually segmented fetal brain MRI dataset to date that consists of 540 fetal brain 3D MRIs from 13 acquisition centers. A recent trend in medical image processing using AI is to gather even larger multi-institutional datasets using methods such as federated learning [51]. One can hypothesize that, with enough data, the AI algorithm would get more accurate even in the worst case until eventually reaching the same accuracy as the trustworthy AI algorithm in all cases. However, results of our stratified evaluation suggest that this will require manually annotated 3D MRIs for every scanner acquisition protocol, for every condition, and for every gestational age. To give an order of magnitude of the required dataset size, if we consider that 10 3D MRIs are required for each gestational age from 19 weeks to 38 weeks, for each of 10 conditions and each of 5 hospitals, we would

already need $10,000$ 3D MRIs for both training and testing. Given the low prevalence of some conditions [4] and the cost of obtaining fully-segmented data, classical supervised learning approaches might not be sufficient. This rough estimation does not even include important confounding factors such as ethnicity and gender. Altogether, this suggests that gathering more training data to improve the AI algorithm prior to deployment might not be sufficient to make the AI algorithm alone trustworthy.

The proposed trustworthy AI approach is not limited to fetal brain MRI and we expect it to be applicable to many medical image segmentation problems. The proposed fail-safe mechanism, that is part of our trustworthy AI method, could be used to help improving the backbone AI continuously after its deployment. An AI incident could be declared when a large part of the AI algorithm prediction was discarded by the *fail-safe mechanism*. This would allow automatic detection of images to correct and include in priority in the training set to update the backbone AI algorithm. In addition, reporting such incidents could help to further improve the trust of the user. In the context of trust, it is important to report such issues even when the incidents were handled correctly using the fallback segmentation algorithm. In addition, as part of the European Union Medical Device Regulations (EU MDR) Article 87 [52], it is a requirement for medical device manufacturers to report device-related incidents. Previous methods for global segmentation failures detection, i.e. at the image-level, were proposed [53], [54]. In contrast, our fail-safe mechanism approaches the problem locally, i.e. at the voxel-level, by using atlas-based and intensity-based priors.

The margins used in our trustworthy AI segmentation algorithm could also support interactive segmentation. Instead of providing voxel-level corrections or scribbles, the annotator could interact with the automatic segmentation by manually adapting the margins for its annotation. After manual adjustment, the voxels outside the margins are automatically marked as correctly labelled while for the voxel inside the margins will be assigned a set of possible labels. This yields partial annotations that can be exploited to improve the backbone AI method using partially-supervised learning methods [33]. This use of margins is similar, in terms of user interaction, to the safety margins that are used in clinics for radiation therapy planning [10].

The expert raters also emphasized that some frequent major violations in the cortex layer could be quickly removed manually and that they would have given higher scores to the segmentations if they could interact with them. This echoes previous work on computational-aided decision making that found that users are more satisfied with imperfect algorithms if they can interact with them [55]. Our findings suggest that allowing interactions would also increase the trust in AI algorithms for medical image segmentation.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement TRABIT No 765148. Tom Vercauteren is supported by a Medtronic / RAEng Research Chair [RCSRF1819\7\34].

References

[1] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[2] B. Allen, S. Agarwal, L. Coombs, C. Wald, and K. Dreyer, "2020 ACR data science institute artificial intelligence survey," *Journal of the American College of Radiology*, vol. 18, no. 8, pp. 1153–1159, 2021.

[3] F. Cabitza, "Biases affecting human decision making in AI-supported second opinion settings," in *International Conference on Modeling Decisions for Artificial Intelligence*. Springer, 2019, pp. 283–294.

[4] L. Fidon, M. Aertsen, N. Mufti, T. Deprest, D. Emam, F. Guffens, E. Schwartz, M. Ebner, D. Prayer, G. Kasprian *et al.*, "Distributionally robust segmentation of abnormal fetal brain 3D MRI," in *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis*. Springer, 2021, pp. 263–273.

[5] C. Gonzalez, K. Gotkowski, A. Bucher, R. Fischbach, I. Kaltenborn, and A. Mukhopadhyay, "Detecting when pre-trained nnU-Net models fail silently for Covid-19 lung lesion segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 304–314.

[6] K. G. van Leeuwen, S. Schalekamp, M. J. Rutten, B. van Ginneken, and M. de Rooij, "Artificial intelligence in radiology: 100 commercially available products and their scientific evidence," *European radiology*, vol. 31, no. 6, pp. 3797–3804, 2021.

[7] European Commission, "Ethics guidelines for trustworthy AI," European Commission, Report, 2019.

[8] ——, "Artificial intelligence act," European Commission, Report, 2021.

[9] M. J. Cardoso, M. Modat, R. Wolz, A. Melbourne, D. Cash, D. Rueckert, and S. Ourselin, "Geodesic information flows: spatially-variant graphs and their application to segmentation and fusion," *IEEE transactions on medical imaging*, vol. 34, no. 9, pp. 1976–1988, 2015.

[10] M. Niyazi, M. Brada, A. J. Chalmers, S. E. Combs, S. C. Erridge, A. Fiorentino, A. L. Grosu, F. J. Lagerwaard, G. Minniti, R.-O. Mirimanoff *et al.*, "Estro-acrop guideline "target delineation of glioblastomas"," *Radiotherapy and oncology*, vol. 118, no. 1, pp. 35–42, 2016.

[11] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation," *IEEE transactions on medical imaging*, vol. 23, no. 7, pp. 903–921, 2004.

[12] P. Welinder, S. Branson, P. Perona, and S. Belongie, "The multidimensional wisdom of crowds," *Advances in neural information processing systems*, vol. 23, 2010.

[13] I. Bloch, "Some aspects of dempster-shafer evidence theory for classification of multi-modality medical images taking partial volume effect into account," *Pattern Recognition Letters*, vol. 17, no. 8, pp. 905–919, 1996.

[14] A.-S. Capelle, O. Colot, and C. Fernandez-Maloigne, "Evidential segmentation scheme of multi-echo MR images for the detection of brain tumors using neighborhood information," *Information Fusion*, vol. 5, no. 3, pp. 203–216, 2004.

[15] J. Ghasemi, R. Ghaderi, M. K. Mollaei, and S. Hojjatoleslami, "A novel fuzzy dempster–shafer inference system for brain MRI segmentation," *Information Sciences*, vol. 223, pp. 205–220, 2013.

[16] B. Lelandais, I. Gardin, L. Mouchard, P. Vera, and S. Ruan, "Dealing with uncertainty and imprecision in image segmentation using belief function theory," *International Journal of Approximate Reasoning*, vol. 55, no. 1, pp. 376–387, 2014.

[17] J. Liu, X. Lu, Y. Li, X. Chen, and Y. Deng, "A new method based on dempster–shafer theory and fuzzy c-means for brain MRI segmentation," *Measurement Science and Technology*, vol. 26, no. 10, p. 105402, 2015.

[18] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[19] L. Zhang, X. Wang, D. Yang, T. Sanford, S. Harmon, B. Turkbey, B. J. Wood, H. Roth, A. Myronenko, D. Xu *et al.*, "Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation," *IEEE transactions on medical imaging*, vol. 39, no. 7, pp. 2531–2540, 2020.

[20] Y. Tang, D. Yang, W. Li, H. R. Roth, B. Landman, D. Xu, V. Nath, and A. Hatamizadeh, "Self-supervised pre-training of swin transformers for 3d medical image analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 730–20 740.

[21] L. Fidon, M. Aertsen, T. Deprest, D. Emam, F. Guffens, N. Mufti, E. Van Elslander, E. Schwartz, M. Ebner, D. Prayer, G. Kasprian, A. L. David, A. Melbourne, S. Ourselin, J. Deprest, G. Langs, and T. Vercauteren, "Distributionally robust deep learning using hardness weighted sampling," *Machine Learning for Biomedical Imaging*, vol. 1, 2022.

[22] K. Kushibar, S. Valverde, S. Gonzalez-Villa, J. Bernal, M. Cabezas, A. Oliver, and X. Lladó, "Automated sub-cortical brain structure segmentation combining spatial and deep convolutional features," *Medical image analysis*, vol. 48, pp. 177–186, 2018.

[23] Q. Liu, C. Chen, Q. Dou, and P.-A. Heng, "Single-domain generalization in medical image segmentation via test-time adaptation from shape dictionary," 2022.

[24] A. Jacovi, A. Marasović, T. Miller, and Y. Goldberg, "Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 624–635.

[25] R. R. Hoffman, "A taxonomy of emergent trusting in the human–machine relationship," *Cognitive Systems Engineering*, pp. 137–164, 2017.

[26] G. Shafer, *A mathematical theory of evidence*. Princeton university press, 1976.

[27] M. Modat, G. R. Ridgway, Z. A. Taylor, M. Lehmann, J. Barnes, D. J. Hawkes, N. C. Fox, and S. Ourselin, "Fast free-form deformation using graphics processing units," *Computer methods and programs in biomedicine*, vol. 98, no. 3, pp. 278–284, 2010.

[28] A. Gholipour, C. K. Rollins, C. Velasco-Annis, A. Ouaalam, A. Akhondi-Asl, O. Afacan, C. M. Ortinau, S. Clancy, C. Limperopoulos, E. Yang *et al.*, "A normative spatiotemporal MRI atlas of the fetal brain for automatic segmentation and analysis of early brain growth," *Scientific reports*, vol. 7, no. 1, pp. 1–13, 2017.

[29] J. Wu, T. Sun, B. Yu, Z. Li, Q. Wu, Y. Wang, Z. Qian, Y. Zhang, L. Jiang, and H. Wei, "Age-specific structural fetal brain atlases construction and cortical development quantification for chinese population," *NeuroImage*, p. 118412, 2021.

[30] L. Fidon, E. Viola, N. Mufti, A. David, A. Melbourne, P. Demaerel, S. Ourselin, T. Vercauteren, J. Deprest, and M. Aertsen, "A spatio-temporal atlas of the developing fetal brain with spina bifida aperta," *Open Research Europe*, vol. 1, no. 123, 2021.

[31] K. Payette, P. de Dumast, H. Kebiri, I. Ezhov, J. C. Paetzold, S. Shit, A. Iqbal, R. Khan, R. Kottke, P. Grehten *et al.*, "An automatic multi-tissue human fetal brain segmentation benchmark using the fetal tissue annotation dataset," *Scientific Data*, vol. 8, no. 1, pp. 1–14, 2021.

[32] A. E. Fetit, A. Alansary, L. Cordero-Grande, J. Cupitt, A. B. Davidson, A. D. Edwards, J. V. Hajnal, E. Hughes, K. Kamnitsas, V. Kyriakopoulou *et al.*, "A deep learning approach to segmentation of the developing cortex in fetal brain MRI with minimal manual labeling," in *Medical Imaging with Deep Learning*. PMLR, 2020, pp. 241–261.

[33] L. Fidon, M. Aertsen, D. Emam, N. Mufti, F. Guffens, T. Deprest, P. Demaerel, A. L. David, A. Melbourne, S. Ourselin *et al.*, "Label-set loss functions for partial supervision: application to fetal brain 3D MRI parcellation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 647–657.

[34] L. Fidon, M. Aertsen, S. Shit, P. Demaerel, S. Ourselin, J. Deprest, and T. Vercauteren, "Partial supervision for the feta challenge 2021," *arXiv preprint arXiv:2111.02408*, 2021.

[35] J. Hong, H. J. Yun, G. Park, S. Kim, C. T. Laurentys, L. C. Siqueira, T. Tarui, C. K. Rollins, C. M. Ortinau, P. E. Grant *et al.*, "Fetal cortical plate segmentation using fully convolutional networks with multiple plane aggregation," *Frontiers in neuroscience*, p. 1226, 2020.

[36] N. Khalili, N. Lessmann, E. Turk, N. Claessens, R. de Heus, T. Kolk, M. Viergever, M. Benders, and I. Išgum, "Automatic brain tissue segmentation in fetal MRI using convolutional neural networks," *Magnetic resonance imaging*, vol. 64, pp. 77–89, 2019.

[37] L. Li, M. Sinclair, A. Makropoulos, J. V. Hajnal, A. David Edwards, B. Kainz, D. Rueckert, and A. Alansary, "CAS-Net: Conditional atlas generation and brain segmentation for fetal MRI," in *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis*. Springer, 2021, pp. 221–230.

[38] L. Zhao, J. Asis-Cruz, X. Feng, Y. Wu, K. Kapse, A. Largent, J. Quistorff, C. Lopez, D. Wu, K. Qing *et al.*, "Automated 3D fetal brain segmentation using an optimized deep learning approach," *American Journal of Neuroradiology*, 2022.

[39] A. Makropoulos, S. J. Counsell, and D. Rueckert, "A review on automatic fetal and neonatal brain MRI segmentation," *NeuroImage*, vol. 170, pp. 231–248, 2018.

[40] D. Alis, M. Yergin, C. Alis, C. Topel, O. Asmakutlu, O. Bagcilar, Y. D. Senli, A. Ustundag, V. Salt, S. N. Dogan *et al.*, "Inter-vendor performance of deep learning in segmenting acute ischemic lesions on diffusion-weighted imaging: a multicenter study," *Scientific reports*, vol. 11, no. 1, pp. 1–10, 2021.

[41] R. A. Kamraoui, V.-T. Ta, T. Tourdias, B. Mansencal, J. V. Manjon, and P. Coupé, "Deeplesionbrain: Towards a broader deep-learning generalization for multiple sclerosis lesion segmentation," *Medical Image Analysis*, vol. 76, p. 102312, 2022.

[42] G. Mårtensson, D. Ferreira, T. Granberg, L. Cavallin, K. Oppedal, A. Padovani, I. Rektorova, L. Bonanni, M. Pardini, M. G. Kramberger *et al.*, "The reliability of a deep learning model in clinical out-of-distribution MRI data: a multicohort study," *Medical Image Analysis*, vol. 66, p. 101714, 2020.

[43] L. Oakden-Rayner, J. Dunnmon, G. Carneiro, and C. Ré, "Hidden stratification causes clinically meaningful failures in machine learning for medical imaging," in *Proceedings of the ACM conference on health, inference, and learning*, 2020, pp. 151–159.

[44] C. S. Perone, P. Ballester, R. C. Barros, and J. Cohen-Adad, "Unsupervised domain adaptation for medical imaging segmentation with self-ensembling," *NeuroImage*, vol. 194, pp. 1–11, 2019.

[45] I. Redko, E. Morvant, A. Habrard, M. Sebban, and Y. Bennani, *Advances in domain adaptation theory*. Elsevier, 2019.

[46] J. Pollenus, L. Lagae, M. Aertsen, and K. Jansen, "The impact of cerebral anomalies on cognitive outcome in patients with spina bifida: A systematic review," *European Journal of Paediatric Neurology*, 2020.

[47] T. P. Naidich, R. M. Pudlowski, J. Naidich, M. Gornish, and F. Rodriguez, "Computed tomographic signs of the chiari II malformation. part I: Skull and dural partitions." *Radiology*, vol. 134, no. 1, pp. 65–71, 1980.

[48] F. Kofler, I. Ezhov, F. Isensee, F. Balsiger, C. Berger, M. Koerner, J. Paetzold, H. Li, S. Shit, R. McKinley *et al.*, "Are we using appropriate segmentation metrics? identifying correlates of human expert perception for CNN training beyond rolling the DICE coefficient," *arXiv preprint arXiv:2103.06205*, 2021.

[49] M. Aertsen, J. Verduyckt, F. De Keyzer, T. Vercauteren, F. Van Calenbergh, L. De Catte, S. Dymarkowski, P. Demaerel, and J. Deprest, "Reliability of MR imaging–based posterior fossa and brain stem measurements in open spinal dysraphism in the era of fetal surgery," *American Journal of Neuroradiology*, vol. 40, no. 1, pp. 191–198, 2019.

[50] E. Danzer, M. P. Johnson, M. Bebbington, E. M. Simon, R. D. Wilson, L. T. Bilaniuk, L. N. Sutton, and N. S. Adzick, "Fetal head biometry assessed by fetal magnetic resonance imaging following in utero myelomeningocele repair," *Fetal diagnosis and therapy*, vol. 22, no. 1, pp. 1–6, 2007.

[51] N. Rieke, J. Hancox, W. Li, F. Milletari, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein *et al.*, "The future of digital health with federated learning," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–7, 2020.

[52] European Union, "The European Union Medical Device Regulation – Regulation (EU) 2017/745 (EU MDR)," European Union, Regulation, 2017.

[53] F. Kofler, I. Ezhov, L. Fidon, C. M. Pirkl, J. C. Paetzold, E. Burian, S. Pati, M. El Husseini, F. Navarro, S. Shit *et al.*, "Robust, primitive, and unsupervised quality estimation for segmentation ensembles," *Frontiers in Neuroscience*, vol. 15, 2021.

[54] R. Robinson, V. V. Valindria, W. Bai, O. Oktay, B. Kainz, H. Suzuki, M. M. Sanghvi, N. Aung, J. M. Paiva, F. Zemrak *et al.*, "Automated quality control in image segmentation: application to the UK Biobank cardiovascular magnetic resonance imaging study," *Journal of Cardiovascular Magnetic Resonance*, vol. 21, no. 1, pp. 1–14, 2019.

[55] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them," *Management Science*, vol. 64, no. 3, pp. 1155–1170, 2018.