Artificial Intelligence In Medicine 167 (2025) 103154

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Research paper

Symbolic and hybrid AI for brain tissue segmentation using spatial model checking☆

Gina Belmonte [a], Vincenzo Ciancia [b], Mieke Massink [b,*]

[a] S. C. Fisica Sanitaria Nord, Azienda Toscana Nord Ovest, Lucca, Italy
[b] Istituto di Scienza e Tecnologie dell'Informazione 'A. Faedo', Consiglio Nazionale delle Ricerche, Pisa, Italy

ARTICLE INFO

Keywords:
Closure spaces
Spatial model checking
Formal methods
Machine learning
Brain segmentation
Glioblastoma

ABSTRACT

Segmentation of 3D medical images, and brain segmentation in particular, is an important topic in neuroimaging and in radiotherapy. Overcoming the current, time consuming, practise of manual delineation of brain tumours and providing an accurate, explainable, and replicable method of segmentation of the tumour area and related tissues is therefore an open research challenge.

In this paper, we first propose a novel symbolic approach to brain segmentation and delineation of brain lesions based on *spatial model checking*. This method has its foundations in the theory of closure spaces, a generalisation of topological spaces, and spatial logics. At its core is a high-level declarative logic language for image analysis, `ImgQL`, and an efficient spatial model checker, `VoxLogicA`, exploiting state-of-the-art image analysis libraries in its model checking algorithm. We then illustrate how this technique can be combined with Machine Learning techniques leading to a hybrid AI approach that provides accurate and explainable segmentation results.

We show the results of the application of the symbolic approach on several public datasets with 3D magnetic resonance (MR) images. Three datasets are provided by the 2017, 2019 and 2020 international MICCAI BraTS Challenges with 210, 259 and 293 MR images, respectively, and the fourth is the BrainWeb dataset with 20 (synthetic) 3D patient images of the normal brain. We then apply the hybrid AI method to the BraTS 2020 training set. Our segmentation results are shown to be in line with the state-of-the-art with respect to other recent approaches, both from the accuracy point of view as well as from the view of computational efficiency, but with the advantage of them being explainable.

1. Introduction

This paper presents a state-of-the-art, fully automated, *symbolic* paradigm for medical image analysis, along with its related software tool `VoxLogicA`. The proposed approach is based on *model checking*, a technique borrowed from the realm of formal methods. `VoxLogicA` can be used to describe imaging-related domain knowledge in a concise, unambiguous, executable specification language, rooted in the theory of topological spaces, exploiting *spatial logics* in a foundational way. Model checking is then used to check which parts of an image satisfy a given spatial-logical specification.

Although the work we present is aimed at building a novel foundational approach, where model checking plays a pivotal role (as opposed to, e.g., logical deduction), in the last decade, a number of case studies, of which two are presented in detail in this paper, have contributed to prove applicability in different (medical) imaging and video-analysis related scenarios.

This paper also studies a novel *hybrid symbolic-subsymbolic procedure*, based on `VoxLogicA`, that combines the strength of the state-of-the-art deep learning system nnU-Net [1] with our symbolic, logic-based method, obtaining excellent performance in determining the *Gross Tumour Volume* and Clinical Target Volume on a public brain tumour

☆ This research is partially supported by the Italian national MUR project PRIN 20228KXFN2 "STENDHAL", the bilateral project between CNR (Italy) and SRNSFG (Georgia) "Model Checking for Polyhedral Logic" (#CNR-22-010), and the European Union - Next Generation EU project in the context of The National Recovery and Resilience Plan, Investment 1.5 Ecosystems of Innovation PRI ECS00000017, Project "Tuscany Health Ecosystem" (THE), CUP: B83C22003920001. The names of the authors of the present paper are listed in the front-page in alphabetical order. All co-authors have contributed equally to the work described in the present paper and to the development of the paper. We thank Diego Latella for having contributed actively to this paper when he was still Senior Researcher with CNR. Since Sep. 1, 2024 he has retired.

* Corresponding author.
E-mail addresses: vincenzo.ciancia@isti.cnr.it (V. Ciancia), mieke.massink@isti.cnr.it (M. Massink).

https://doi.org/10.1016/j.artmed.2025.103154
Received 4 October 2024; Received in revised form 2 April 2025; Accepted 1 May 2025
Available online 24 May 2025

segmentation dataset, while retaining a symbolic, human-intelligible, accountable, fast and simple *white-box* analysis procedure.

Below we provide an introduction to the relevant concepts.

Image segmentation. Segmentation of medical images is an important and active topic of research in many applications in neuroimaging. Many automatic and semi-automatic methods have been proposed (see for example [2–7]). In this paper we focus on image segmentation of the brain. In particular we address segmentation of glioblastomas, which are the most common malignant intracranial tumours, and the segmentation of tissues in the healthy brain, such as white and grey matter. Neuroimaging protocols are commonly used before and after the treatment of glioblastoma to decide the best treatment and to evaluate the effect of treatment strategies. In the current practise, this evaluation is usually performed manually, which is a mostly accurate but time consuming procedure. To the best of our knowledge, the main reason for the hesitation to insert fully automated methods in the clinical setting is their lack of transparency which hampers their accountability. Therefore, much research is dedicated to finding sufficiently accurate, but at the same time *explainable and replicable* methods that can assist domain experts in the task of segmentation of the tumour area and related tissues to facilitate the evaluation and accountability necessary for showing adherence to existing treatment protocols and contouring guidelines.

Neuro-symbolic artificial intelligence. In this paper we first introduce the symbolic *spatial model checking approach* and related *spatial model checker* with which we will perform brain segmentation and delineation of brain lesions. We then illustrate how this technique can be combined with Machine Learning leading to a *hybrid* method. The field of hybrid – that is, *Neuro-symbolic* – Artificial Intelligence has seen a recent rise in research [8]. A neuro-symbolic system combines neural networks with symbolic reasoning or knowledge representation. The choice of the symbolic part may radically change the methodology and expected results. An effort into building a hybrid extension of `VoxLogicA` is ongoing (see [9], and the discussion in Section 7). Employing model checking for the symbolic part is, to the best of our knowledge, a novel contribution of this paper. Our proposed method – even if it has unique characteristics, among which being fully automatic, and computationally efficient – can be considered of Type 3 (that is, "Neuro|Symbolic"), according to Kautz's taxonomy[1] (see [8,10]). Our study demonstrates that this approach may provide accurate and, at the same time, explainable segmentation results. This approach enhances the accountability of the proposed contouring methods as it enables the domain expert to show that the contouring adheres to the contouring guidelines (see for example [11,12]).

Mathematical foundation. The origins of spatial logic can be traced back to the forties of the previous century when McKinsey and Tarski recognised the possibility of reasoning on space using topology as a mathematical framework for the interpretation of modal logic (see [13] for a thorough introduction). Since then, qualitative spatial reasoning (QSR), in the sense of reasoning about spatial entities without resorting to traditional quantitative techniques – such as those commonly used in the fields of computer graphics and computer vision [14] – has been a very active area of research. As we will see, in our work, as in QSR, topology, or better the theory of *closure spaces* [15–18], a generalisation of topological spaces, plays a central role and is used as the underlying mathematical foundation for reasoning about spatial models. However, we will approach the problem of spatial reasoning from a different perspective, namely that of *model checking*. The problem of model checking was originally formulated by Church (see e.g. [19]) in what he called the 'decision problem': given a model (of a circuit in his case) and a logic formula ϕ, does ϕ hold in the model? Model checking has made enormous progress since its conception by its pioneers [20] as a high-level, automatic method for the verification of, mostly, temporal logic properties of systems modelled as labelled transition systems, with main applications in the field of concurrent and distributed systems [19,21,22].

Spatial model checking. In recent work, we developed a novel variant of the model checking problem moving from checking temporal logic properties to *spatial* logic properties (and, in fact, also to the combination of reasoning on time and space in spatio-temporal model checking [23–26]). In the present article we focus on *spatial* model checking [27,28] and, in particular, its efficient application to a new domain, namely the analysis of 3D medical images. We show how spatial model checking provides a bridge between artificial intelligence (in particular the area of models of space – specifically closure spaces – and related logics) and automatic means to reason at a qualitative, and, to a certain extent, quantitative, level about properties of medical images. The proposed spatial logic for image analysis, Image Query Language (`ImgQL`), encompasses several domain oriented operators to accommodate the level of abstract spatial reasoning by domain experts [29,30]. The combination of basic logical operators, spatial operators and domain oriented operators provides powerful building blocks to develop concise, human readable and explainable image segmentation methods. In other words, spatial model checking builds on the mathematical foundation of closure spaces because the latter provides suitably abstract spatial models and a suitable, extendable logic to express a large variety of spatial queries on medical images. The query language `ImgQL` provides a flexible and compositional way to specify a range of different contouring methods as opposed to providing single, specialised methods. Furthermore, the building blocks that can be provided through the definition of derived operators are close to the level of abstraction at which domain experts themselves reason about segmentation procedures, facilitating explainability, confidence in, and potential adoption of, the method.

Case studies. In the present article, we show the feasibility of the spatial model checking approach applying it to two image segmentation case studies. The first study concerns the automatic segmentation of glioblastoma (GBM) and its associated oedema, one of the most common malignant intracranial tumours composed of infiltrating necrotic masses. Automatic contouring of GBM is an open challenging topic, since GBM is an intrinsically heterogeneous – in appearance, shape, and histology – brain tumour [4]. Since 2012 a yearly challenge is organised by the Medical Image Computing and Computer Assisted Intervention Society (MICCAI) Conference, namely the Brain Tumour Image Segmentation Benchmark (BraTS) [31]. Although the actual trend is the use of Machine Learning to solve this problem (see for example [32] for a recent review), legal aspects about the accountability and the explainability of decisions may arise, especially in radiotherapy (RT). For the treatment of glioblastomas, neuroimaging protocols are used before and after treatment to evaluate the effect of treatment strategies and to monitor the evolution of the disease. In clinical studies and routine treatment, magnetic resonance images (MRI) are evaluated based mostly on qualitative criteria such as the presence of hyperintense tissue appearing in the images [31]. The study and development of automatic and semi-automatic segmentation algorithms is aiming at overcoming the current time consuming practise of manual delineation of such tumours and at providing an accurate, reliable and reproducible method of segmentation of the tumour area and related tissues. We validated one of our segmentation methods, specified in `ImgQL`, using

[1] In a Neuro|Symbolic system, a neural network converts non-symbolic input, such as the pixels of an image, into a symbolic data structure, which is then processed by a symbolic reasoning system. Our method actually turns non-symbolic input into Boolean regions, which are then used to inform a logical language about the best possible choice of some parameters. Rather than being pure symbols, the Boolean regions are used to bridge the symbolic and non-symbolic worlds, by their dual nature of a non-symbolic encoding (a Boolean-valued image) of symbolic information (an atomic proposition denoting a specific image feature).

the 2017, 2019 and 2020 BraTS training datasets containing multi-institutional pre-operative MRI scans of 210, 259 and 293 patients, respectively, affected by high grade gliomas. All the imaging datasets provided by BraTS have been segmented manually and were approved by experienced neuro-radiologists. These images provide a 'ground truth' with which we compare our own results. This work builds on our previous results in Belmonte et al. [29].

The second case study concerns the segmentation of tissues of a normal (healthy) brain, such as white and grey matter. We present a further specification in ImgQL for the segmentation of these tissues and validate it on the BrainWeb [33] dataset,[2] which is composed of twenty synthetic brain MRI's. Also this dataset provides ground truth images for various types of brain tissues. We use this ground truth to assess the quality of our method. This work extends the preliminary results we presented in [34].

Validation. For the validation of our segmentation methods we use several commonly accepted similarity indexes, among which the Dice similarity index, the sensitivity (i.e. fraction of voxels that the segmentation and the ground truth have in common) and the specificity (i.e. the fraction of voxels that are not identified by the segmentation and are also not part of the ground truth). These indexes can be used to compare the quality of our method with that of other state-of-the-art approaches in the literature.

Directly related work. The current work builds on some of our previous work, in particular on [29,30,34]. In [30] a first feasibility study of the application of spatial model checking on the segmentation of brain lesions is described. That study has been performed with a predecessor of VoxLogicA for ImgQL, namely the general purpose spatio-temporal model checker topochecker. The analysis concerned several 2D and 3D MR images of patients with high-grade glioma. In [29] the work on spatial model-checking has been taken further with the development of the domain specific, and much more efficient, *spatial* model checker VoxLogicA and its application on the BraTS 2017 dataset. In [34] a feasibility study of the segmentation of tissues present in the *normal* brain was undertaken, restricted to the analysis of the first 2 of the 20 patients provided by the BrainWeb dataset. In the present article the various preliminary ideas are presented in a common framework and applied to *full* public datasets in such a way that the results are amenable to comparison with other methods proposed in the literature.

Original contribution. In summary, the original contributions of this work are:

- Introduction of a symbolic paradigm for image analysis consisting of an extendable spatial logic, ImgQL, and related model checking technique and software tool, VoxLogicA, for spatial models based on closure spaces, which we believe to be the first in its kind;
- Specification of segmentation procedures for brain tissues in terms of ImgQL;
- Validation of these symbolic procedures on two public medical imaging benchmarks of 3D brain MRI, namely the Brats 2017, 2019 and 2020 datasets and the BrainWeb dataset, and comparison with the state-of-the-art.
- Presentation and validation of a novel, symbolic-subsymbolic, explainable, hybrid AI segmentation method that combines spatial logic and deep learning methods and its validation on the BraTS 2020 training set. The validation shows that a considerably higher similarity score can be obtained compared to the pure symbolic approach (average Dice of 0.87 instead of 0.85), while still retaining a good level of explainability.

[2] https://brainweb.bic.mni.mcgill.ca/brainweb/anatomic_normal_20.html.

Synopsis. This article is organised as follows. In Section 2 we briefly recall the topological origins of the spatial model checking approach. We also recall ImgQL and the Spatial Logic for Closure Spaces (SLCS) on which it is based, and show the detailed algorithm for efficient model-checking for SLCS. In Section 3 we present the spatial model checker VoxLogicA for ImgQL, that provides an efficient implementation of the algorithm and further features to facilitate the analysis of MR images. In Section 4 the BraTS glioblastoma and the BrainWeb case studies are presented. Section 5 links our spatial logic approach to Machine Learning and presents the hybrid AI method for glioblastoma segmentation. Section 6 discusses further related work and in Section 7 we draw conclusions and provide an outlook on further research.

2. Topological origins of spatial model checking

We briefly recall some of the main topological notions on which our spatial model checking approach is based, providing the definition of the logic kernel of ImgQL, the *Image Query Language* [29], and introduce the relevant notation. ImgQL is a spatial logic language developed for the analysis of medical images using a spatial model checking algorithm implemented in the spatial model checker VoxLogicA, that will be briefly described in Section 3. ImgQL is based on SLCS (*Spatial Logic for Closure Spaces*) [27,28]. Closure spaces are a generalisation of topological spaces. We first present the basic notions underlying the approach. We refer to [15–18] and to our previous work [27–30] for further details on the relevant theoretical aspects.

2.1. Closure spaces

SLCS is interpreted over models based on *closure spaces*. A closure space – CS for short – is a pair (X, C) where X is a set (of points) and $C : 2^X \to 2^X$ is a function satisfying the following three axioms:

(i) $C(\emptyset) = \emptyset$;
(ii) $Y \subseteq C(Y)$ for all $Y \subseteq X$;
(iii) $C(Y_1 \cup Y_2) = C(Y_1) \cup C(Y_2)$ for all $Y_1, Y_2 \subseteq X$.

Closure spaces are a generalisation of topological spaces in the sense that the closure operator is characterised by only three of the usual four axioms from topology, omitting the fourth one that requires that the closure operator is idempotent. The *interior* of a set $Y \subseteq X$ is obtained by duality, i.e.

$$I(Y) = \overline{C(\overline{Y})}$$

where $\overline{Y} = X \setminus Y$ is the complement of Y. Given any relation $R \subseteq X \times X$, (X, C_R), with $C_R(Y) = Y \cup \{x \mid \exists y \in Y . y\, R\, x\}$, is a CS, specifically, a quasi-discrete CS (QdCS). In particular, a digital image can be modelled as a finite CS (every finite CS is also a QdCS) where X is the set of voxels and R their *adjacency relation*.[3] In the present paper, we will consider only QdCSs and use the *ortho-diagonal* relation in 3D, i.e. the reflexive and symmetric relation where two voxels are *adjacent* if and only if they share a face, an edge or a vertex.

We also introduce the notion of *paths* for QdCSs. Let $(\mathbb{N}, C_{succ})$ be the CS of the natural numbers $\mathbb{N}$ with the binary successor relation $succ = \{(m, n) \in \mathbb{N}^2 \mid n = m + 1\}$. A (discrete) *path* over QdCS (X, C) is a continuous function[4] from $(\mathbb{N}, C_{succ})$ to (X, C).

[3] All the theory and related model checkers work both for 2D and 3D even though we use only 3D in the present paper. In the current work we use the word 'voxel' for 3D 'pixels'.

[4] A *continuous* function from CS (X_1, C_1) to CS (X_2, C_2) is a function $f : X_1 \to X_2$ such that $f(C_1(Y)) \subseteq C_2(f(Y))$ for all $Y \subseteq X_1$.

2.2. SLCS with distance and similarity: Syntax and Semantics

For given set AP of *atomic predicates* p the syntax of SLCS is the following:

$$\Phi ::= p \mid \neg\Phi \mid \Phi_1 \wedge \Phi_2 \mid \vec{\rho}\,\Phi_1[\Phi_2] \mid \overleftarrow{\rho}\,\Phi_1[\Phi_2].$$

Defined predicates are elements p of AP for which a *defining equation* $p := \alpha$ is given, where α is a boolean expression. The operators $\neg$ and $\wedge$ are the usual Boolean operators of negation and conjunction, respectively. The forward and backward reachability operators $\vec{\rho}\,\Phi_1[\Phi_2]$, resp. $\overleftarrow{\rho}\,\Phi_1[\Phi_2]$, express that a point can reach, resp. can be reached from, a point that satisfies Φ_1 along a path consisting of points satisfying Φ_2.

We extend SLCS with a *distance* operator D^I, for interval $I \subseteq \mathbb{R}_{\geq 0}$ of non-negative real numbers. A point x satisfies $D^I\Phi$ if the distance between x and the points satisfying Φ is within interval I. The *distance function*[5] $d : X \times X \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ satisfies $d(x, y) = 0$ if and only if $x = y$; d is lifted to sets in the usual way: $d(x, \emptyset) = \infty$ and for $\emptyset \subset Y \subseteq X$ we have $d(x, Y) = \inf\{d(x, y) \mid y \in Y\}$.

SLCS is interpreted over *attributed distance closure models*, i.e. structures $((X, C), d, \mathcal{A}, \mathcal{V})$ where (X, C) is a CS, d and $\mathcal{A}$ are distance and attribute evaluation functions, respectively, and $\mathcal{V} : \mathtt{AP} \to 2^X$ is a valuation which maps the *atomic predicates* to the points satisfying them. The value $\mathcal{A}(a, x) \in V$ of attribute a of point x is given by the attribute evaluation function $\mathcal{A} : A \times X \to V$, for suitable set V of *values*.

The satisfaction relation for model $\mathcal{M} = ((X, C), d, \mathcal{A}, \mathcal{V})$, $x \in X$, and formulas Φ is defined recursively on the structure of SLCS formulas Φ as follows, where $[\![\Phi]\!]^{\mathcal{M}} = \{x \in X \mid \mathcal{M}, x \vDash \Phi\}$ is the set of points in $\mathcal{M}$ satisfying Φ:

$\mathcal{M}, x \vDash p \Leftrightarrow x \in \mathcal{V}(p)$;
$\mathcal{M}, x \vDash \neg\Phi \Leftrightarrow \mathcal{M}, x \vDash \Phi$ does not hold;
$\mathcal{M}, x \vDash \Phi_1 \wedge \Phi_2 \Leftrightarrow \mathcal{M}, x \vDash \Phi_1$ and $\mathcal{M}, x \vDash \Phi_2$;
$\mathcal{M}, x \vDash \vec{\rho}\,\Phi_1[\Phi_2] \Leftrightarrow$ there exist path π, index ℓ s.t.
  $\pi(0) = x$ and
  $\pi(\ell) \vDash \Phi_1$ and
  for all j such that $0 < j < \ell$:
    $\pi(j) \vDash \Phi_2$;

$\mathcal{M}, x \vDash \overleftarrow{\rho}\,\Phi_1[\Phi_2] \Leftrightarrow$ there exist path π and index ℓ s.t.
  $\pi(\ell) = x$ and
  $\pi(0) \vDash \Phi_1$ and
  for all j s.t. $0 < j < \ell$:
    $\pi(j) \vDash \Phi_2$;

$\mathcal{M}, x \vDash D^I\Phi \Leftrightarrow d(x, [\![\Phi]\!]^{\mathcal{M}}) \in I$.

Whenever p is a *defined* predicate with defining equation $p := \alpha$, we extend the satisfaction relation by letting $x \in \mathcal{V}(p)$ if and only if $\mathcal{A}(\alpha, x)$ is true, where $\mathcal{A}(\alpha, x)$ is the lifting of the attribute evaluation function to attribute expressions and is defined in the obvious way.

We point out here that, whenever the underlying relation is symmetric, like in the case of the adjacency relation for the voxels of a digital image, it is easy to see that $\vec{\rho}\,\Phi_1[\Phi_2]$ is logically equivalent to $\overleftarrow{\rho}\,\Phi_1[\Phi_2]$. Consequently, we use only one of the two, namely $\vec{\rho}\,\Phi_1[\Phi_2]$, and denote it simply by $\rho\Phi_1[\Phi_2]$.

[5] Several distance functions are defined in the literature; the specific distance to be used depends on the application. The interested reader is referred to [30]. In this work we use the Manhattan distance where 1 voxel is the unit distance.

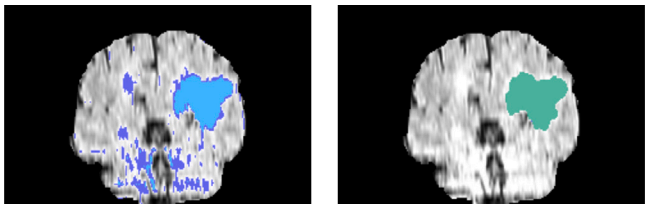


Fig. 1. Illustration of the grow operator. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Derived operators. The reachability operator is a basic, but rather expressive operator of the logic. In the following we show how a number of more abstract operators can be defined that are closer to the level of abstraction at which domain experts assess medical images. These abstract operators are defined as derived operators, i.e. their definition is given in terms of the reachability operator. Below we introduce some derived operators that are used in the case studies:

$\mathcal{N}\Phi \equiv \rho\Phi[\mathtt{false}]$
$\Phi_1\, S\, \Phi_2 \equiv \Phi_1 \wedge \neg\rho\neg(\Phi_1 \vee \Phi_2)[\neg\Phi_2]$
$\mathtt{touch}(\Phi_1, \Phi_2) \equiv \Phi_1 \wedge \rho\Phi_2[\Phi_1]$
$\mathtt{grow}(\Phi_1, \Phi_2) \equiv \Phi_1 \vee \mathtt{touch}(\Phi_2, \Phi_1)$
$\mathtt{smoothen}(r, \Phi) \equiv D^{\leq r}(D^{\geq r}\neg\Phi)$.

Point x satisfies $\mathcal{N}\Phi$ if it satisfies Φ or it is adjacent to a point satisfying Φ. Operator S denotes the surround operator. Point x satisfies $\Phi_1\, S\, \Phi_2$ if it satisfies Φ_1 and from it one cannot reach a point, by means of a sequence of adjacent voxels, *not* satisfying Φ_1 unless passing by a point satisfying Φ_2. Point x satisfies $\mathtt{touch}(\Phi_1, \Phi_2)$ if it satisfies Φ_1 and from it one can reach a point satisfying Φ_2 only passing by a sequence of adjacent points satisfying Φ_1. The grow operator $\mathtt{grow}(\Phi_1, \Phi_2)$ extends the set of points satisfying Φ_1 with those points satisfying Φ_2 that can reach points satisfying Φ_1 through paths the points of which satisfy Φ_2. In other words, the area satisfying Φ_1 is extended with an area of points satisfying Φ_2 under the condition that these areas touch each other. An example of the effect of the grow operator in the context of the segmentation of glioblastoma is shown in Fig. 1. Fig. 1(left) shows pixels with a very high intensity in blue and hyper intense pixels in cyan. It is known that hyper intense pixels are present in tumour tissue, whereas very intense ones are part of the oedema. However, not all very intense pixels are part of an oedema. The *grow* operator can be used to find only those very intense pixels that are in an area that is connected to an area of hyper intense pixels. The result of pixels satisfying the property $\mathtt{grow}(\mathtt{hyperIntense}, \mathtt{veryIntense})$ is shown in green in Fig. 1(right).

Finally, the smoothen operator first erodes and then dilates the area of points satisfying Φ by an amount of r, obtaining a smoothening effect.

Additional operators. ImgQL kernel can easily be extended with further logic operators that are of interest in a particular domain of application. One example of such an operator is the statistical similarity operator $\triangleq$. Further examples will follow in the course of this section. Their introduction as logical operators has the advantage that they can be easily combined with the more basic spatial logic operators, providing an expressive and flexible query-like language that can be adapted to the needs of a particular domain.

Let us first introduce a few basic notions required for the definition of the statistical similarity operator $\triangleq$. The first notion is that of histogram of an image, with respect to a given numerical attribute. It essentially provides information on the distribution, in the image, of the values taken by the attribute. More precisely, with reference to model $\mathcal{M} = ((X, C), \mathcal{A}, \mathcal{V})$, the *histogram* $\mathcal{H}(a, Y, m, M, k)$ of the distribution of the values of attribute a of the points in $Y \subseteq X$, in the interval $[m, M]$ with step size Δ and k bins, is the function $\mathcal{H} : A \times 2^X \times$

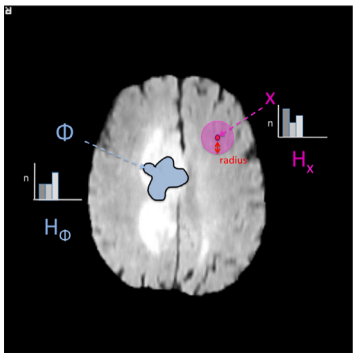


Fig. 2. Illustration of the ImgQL statistical similarity operator $\triangleq$

$\mathbb{R} \times \mathbb{R} \times \mathbb{N} \to (\mathbb{N} \to \mathbb{N})$ such that, for all values $m, M \in \mathbb{R}$, with $m < M$, and $k \in \mathbb{N} \setminus \{0\}$, and $i \in \{1, \ldots, k\}$, $\mathcal{H}(a, Y, m, M, k)(i) = |\{y \in Y \mid (i-1)\Delta \leq \mathcal{A}(y, a) - m < i\Delta\}|$, where $\Delta = \frac{M-m}{k}$. The *mean* $\overline{h}$ of histogram h is $\frac{1}{k}\sum_{i=1}^{k} h(i)$. Given histograms $h_1, h_2 : \{1, \ldots, k\} \to \mathbb{N}$, their *cross correlation* $\mathbf{r}(h_1, h_2)$ is given by

$$\mathbf{r}(h_1, h_2) = \frac{\sum_{i=1}^{k}\left(h_1(i) - \overline{h_1}\right)\left(h_2(i) - \overline{h_2}\right)}{\sqrt{\sum_{i=1}^{k}\left(h_1(i) - \overline{h_1}\right)^2}\sqrt{\sum_{i=1}^{k}\left(h_2(i) - \overline{h_2}\right)^2}}.$$

The value of $\mathbf{r}$ is *normalised* so that $-1 \leq \mathbf{r}(h_1, h_2) \leq 1$; $\mathbf{r}(h_1, h_2) = 1$ indicates that h_1 and h_2 are *perfectly correlated* (that is, $h_1 = \alpha h_2 + \beta$, with $\alpha > 0$); $\mathbf{r}(h_1, h_2) = -1$ indicates *perfect anti-correlation* (that is, $h_1 = \alpha h_2 + \beta$, with $\alpha < 0$). On the other hand, $\mathbf{r}(h_1, h_2) = 0$ indicates no correlation.[6]

We now have all the ingredients for completing the definition of the logical kernel of ImgQL by extending the syntax given above with

$$\triangleq_{\bowtie c}\begin{bmatrix} m & M & k \\ r & a & b \end{bmatrix},$$

and adding the following clause to the definition of the satisfaction relation, where $B(x, r) = \{y \in X \mid d(x, y) \leq r\}$ is the 'ball' of radius r centred in x, $h_a(i) = \mathcal{H}(a, B(x, r), m, M, k)(i)$, $h_b(i) = \mathcal{H}(b, [\![\Phi]\!]^{\mathcal{M}}, m, M, k)(i)$, and $\bowtie \in \{=, <, >, \leq, \geq\}$:

$$\mathcal{M}, x \vDash \triangleq_{\bowtie c}\begin{bmatrix} m & M & k \\ r & a & b \end{bmatrix}\Phi \Leftrightarrow \mathbf{r}(h_a, h_b) \bowtie c.$$

So $\triangleq_{\bowtie c}\begin{bmatrix} m & M & k \\ r & a & b \end{bmatrix}\Phi$ compares the region of the space constituted by the ball of radius r centred in x against the region characterised by Φ as illustrated in Fig. 2. The comparison is based on the cross correlation of the histograms of the values of the chosen attributes of (the points of) the two regions, namely attribute a for the points around x and attribute b for the points that satisfy Φ. Of course, these attributes may also be chosen to be the same. Both histograms share the same domain ($[m, M]$) and the same (number of) bins ($\{1, \ldots, k\}$). For further detailed examples of the application of the introduced operators we refer to Section 4, where they are used to define a method for the segmentation of glioblastoma in brain MRI scans and for the segmentation of white and grey matter tissue in synthetic MRI scans of the healthy brain.

[6] Note that normalisation makes the value of $\mathbf{r}$ undefined for constant histograms, having therefore standard deviation of 0; in terms of statistics, a variable with such standard deviation is only (perfectly) correlated to itself. This special case is handled by letting $\mathbf{r}(h_1, h_2) = 1$ when both histograms are constant, and $\mathbf{r}(h_1, h_2) = 0$ when only one of the h_1 or h_2 is constant.

2.3. Main spatial model checking algorithm

Spatial model checking consists of an automatic procedure that, given a *finite* spatial model $\mathcal{M} = ((X, C), d, \mathcal{A}, \mathcal{V})$ and an ImgQL formula Φ, it returns the *satisfaction set* $\mathtt{Sat}(\mathcal{M}, \phi) = [\![\Phi]\!]^{\mathcal{M}}$, i.e. the set of points in X that satisfy the formula Φ. This procedure is also known as a *global* model checking procedure, in the sense that the procedure checks the formula Φ on *all* points $x \in X$ at the same time. A part of the algorithm for the main operators is illustrated in Algorithm 1, where $\mathrm{CC}(\Phi)$ produces the set of all *connected components* of the subgraph of points in X that satisfy property Φ.

Algorithm 1: Main Spatial Model Checking Algorithm

```
1  Function Sat(M, φ)
       Input: Finite closure model M = ((X, ⊂_R), d, A, V), formula
              Φ
       Output: Set of points {x ∈ X | M, x ⊨ Φ}
2      Match φ
3          case ⊤ : return X;
4          case p : return V(p);
5          case ¬Φ_1 :
6              let P_1 = Sat(M, Φ_1);
7              return X \ P_1
8
9          case Φ_1 ∧ Φ_2 :
10             let P_1 = Sat(M, Φ_1);
11             let P_2 = Sat(M, Φ_2);
12             return P_1 ∩ P_2
13
14         case ρ Φ_1[Φ_2] :
15             let P_1 = Sat(M, Φ_1);
16             let P_2 = Sat(M, Φ_2);
17             return {x | x ∈⊂_R (P_1) ∨ x ∈⊂_R (Z) where Z ∈
                CC(P_2) ∧ (Z ∩ ⊂_R (P_1)) ≠ ∅}
18
```

The computation of the last case, $\rho\, \Phi_1[\Phi_2]$, is more involved. It is easy to see that a point x satisfies $\rho\, \Phi_1[\Phi_2]$ if and only if x satisfies $\mathcal{N}\Phi_1$ or there exist a connected component $C \subseteq [\![\Phi_2]\!]^{\mathcal{M}}$ and a point $y \in C$ such that $x \in C_R(C)$ and y satisfies $\mathcal{N}\Phi_1$. However, since $\mathcal{N}$ was defined as an operator derived from ρ in Section 2, this would lead to a circular definition. Note, though, that $\mathcal{N}$ can also very easily be defined in a direct way in terms of the closure operator as follows: $[\![\mathcal{N}\Phi]\!]^{\mathcal{M}} = C_R([\![\Phi]\!]^{\mathcal{M}})$. This leads to the formulation shown in **case** $\rho\, \Phi_1[\Phi_2]$ in Algorithm 1. The effective computation of this set can be performed in various ways. One way is the classical *flood-fill* approach that we have used for spatial model checking of directed graphs in our earlier work [28]. Another way is to exploit the operator CC as shown in Algorithm 1. Both solutions are available as primitives in various software libraries. We will discuss this in more detail in Section 3 addressing the implementation of VoxLogicA.

3. The Spatial Model Checker VoxLogicA

The *spatial model checker* VoxLogicA (*Voxel-based Logical Analyser*)[7] provides a novel, *rapid-development*, declarative, logic-based approach to image analysis and segmentation and is a free and open source tool. As we shall see in the case studies, the approach is particularly suitable to reason at the "macro-level", by exploiting the *relative* spatial relations between tissues or organs at risk in medical images. VoxLogicA is specifically designed for the analysis of (possibly

[7] VoxLogicA: https://github.com/vincenzoml/VoxLogicA.

multi-dimensional, e.g. 3D) *digital images*. It combines efficient spatial model checking algorithms with state-of-the-art open source libraries, borrowed from computational image processing.

Image types. VoxLogicA has a static, simple, strong typing mechanism. The type system distinguishes between *boolean-valued* images, that can be arguments or results of the application of ImgQL spatial logic operators, and *number-valued* or *grey-scale* images. In the former case, each point in the image has a Boolean value representing the truth-value of the ImgQL formula that is applied to each voxel of the image. In the latter case each point has a *numeric* value resulting from imaging primitives such as the intensity of the voxel or the distance of the voxel from a selected point (or set of points) in the image.

For instance, the formula `intensity(flair) <. 0.01` denotes the voxels that, in the image denoted by `flair`, have intensity lower than 0.01, by producing a boolean image which is true on these voxels and false elsewhere. The type of `flair` is that of a possibly multi-channel image. The type of `intensity(flair)` is that of a greyscale image. The dot after < indicates that at the right of this operator a scalar constant value is expected, whereas on the left of <, without dot, a grey-scale image is expected. Operator .<. expects (expressions evaluating to) scalar values on both sides. In a similar way, the statistical similarity operator can be used without supplying a threshold, resulting in a grey-scale image providing the cross-correlation score for each voxel of the image under consideration. Further examples are provided in the case studies in Section 4.

ITK library. VoxLogicA is implemented in the programming language **FSharp** and uses the state-of-the-art imaging library **ITK**, via the **SimpleITK** glue library.[8] Most of the operators of VoxLogicA are implemented directly by a library call. Notably, this includes the *Maurer distance transform* used to efficiently implement the distance operators of ImgQL.

Novel algorithms. The two most relevant operators that do not have a direct implementation in **ITK** are **mayReach** and **crossCorrelation**, implementing, respectively, the logical reachability operator ρ, and the statistical similarity operator $\triangleq$ described in Section 2. As already briefly mentioned in Section 2, the computation of the voxels satisfying $\rho\ \Phi_1[\Phi_2]$ can be implemented either using the (classical, in computer graphics) *flood-fill* primitive, or by exploiting the *connected components* of Φ_2 as a reachability primitive; both primitives are available in **SimpleITK**. The flooding primitive in **ITK** is optimised to start from a *single* voxel, rather than with respect to a *set* of voxels. In our experiments that use this library from **FSharp**, the variant exploiting connected components performs better than the flood-fill approach for large images, therefore, this solution is adopted in VoxLogicA. Moreover, several important logical derived operators (e.g. *surrounded* and *touch*), are defined in terms of **mayReach**. Therefore, an optimised algorithm for **mayReach** is also a key performance improvement with respect to the algorithm in our previous work [28].

The **crossCorrelation** operation is resource-intensive, as it uses the histogram of a multi-dimensional hyperrectangle at each voxel. Pre-computation methods such as the *integral histogram* [35], would not yield the expected benefits, because cross-correlation is usually called only a few times on the same image. Therefore we designed a novel parallel algorithm exploiting *additivity* of histograms. Given two sets of values P_1, P_2, let h_1, h_2 be their respective histograms, and let h'_1, h'_2 be the histograms of $P_1 \setminus P_2$ and $P_2 \setminus P_1$. For i a bin, we have $h_2(i) = h_1(i) - h'_1(i) + h'_2(i)$. An example is shown in Fig. 3.

This property leads to a particularly efficient algorithm when P_1 and P_2 are two hyperrectangles each centred over two adjacent voxels, respectively, as $P_1 \setminus P_2$ and $P_2 \setminus P_1$ are *hyperfaces*, having one dimension less than the hyperrectangles themselves. Our algorithm equally divides the image into as many parts as the number of available processors. It then computes a *Hamiltonian path*[9] for each part, passing by each of its voxels exactly once.

All parts are visited in parallel, in the order imposed by such Hamiltonian paths; the histogram of the hyperrectangle for each voxel is computed in an incremental manner as described above; finally the cross-correlation score for the voxel is computed and stored as an attribute of that voxel in the image.

Memoizing execution semantics. Sub-formulas in VoxLogicA are *by construction* identified up-to syntactic equality and assigned a number, representing a unique identifier (UID). UIDs start from 0 and are contiguous, therefore admitting the use of a simple array of all existing unique sub-formulas, mapping them to pre-computed valuations of expressions without further use of hashing. So in VoxLogicA no unique sub-formula is computed more than once.

Algorithmic complexity. The asymptotic algorithmic complexity of the implementation of ImgQL primitives in VoxLogicA is linear in the product $k \times n$ of the number of *tasks* (k) to be executed, and the number of voxels (n). The number of tasks is equal to the number of (syntactically) *distinct* sub-formulas of the given formula.[10] The **cross-Correlation** operator has complexity $O(r \cdot n)$, where n is the number of voxels, and r is the size of the largest hyperface of the considered hypercube, which is the same as the radius of the 'area of interest' around voxel x.

3.1. Functionality of VoxLogicA

A VoxLogicA specification consists of a text file containing a sequence of **commands**; the following commands are available in the current implementation of VoxLogicA:

load `x` = "`s`"
: loads an image from file "`s`" and binds it to `x` for subsequent usage;

save "`s`" `e`
: stores the image resulting from evaluation of expression `e` to file "`s`";

print "`s`" `e`
: prints to the log-file the string **s** followed by the numeric, or boolean, result of computing **e** (the value of **e** must be a single boolean or numerical value);

let `f(x1, ...,xN) = e`
: is used for *function declaration*, also in the form **let** `f = e` *(constant declaration)*, and with special syntactic provisions to define *infix* operators. After execution of the command, name `f` is bound to a function or constant that evaluates to `e` with the appropriate substitutions of parameters; the current implementation of VoxLogicA does not support recursive function definitions;

import "`s`"
: imports a library of declarations from file "`s`"; subsequent import declarations for the same file are not processed; furthermore, such imported files can only contain **let** or **import** commands.

We conclude this section by showing in Table 1 the syntax of the basic operators of SLCS in ImgQL for VoxLogicA. In the table, `f` stands for the VoxLogicA representation of formula Φ. Note that, since VoxLogicA implements a global model-checking algorithm, the result of `crossCorrelation(r,a,b,f,m,M,k)` is a grey-scale image in

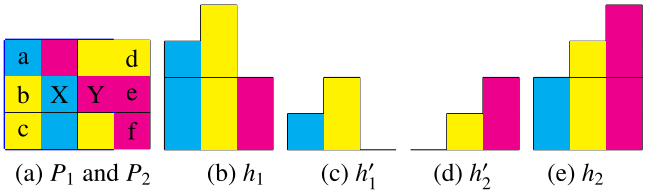
[8] See https://itk.org and http://www.simpleitk.org.

[9] A Hamiltonian path is a path through a graph in which each vertex is visited exactly once.

[10] Note furthermore that for images the cardinality of the adjacency relation is linear in the number of voxels, therefore it does not affect the calculation of the asymptotic complexity, as it does on general graphs.

Table 1
Syntax of basic SLCS operators in ImgQL.

SLCS :	$\neg$	$\wedge$	$\vee$	ρ	$D^{\leq r}\Phi$	$D^{\geq r}\Phi$	$\mathcal{N}\Phi$	$\triangleq_{\bowtie c}\left[\begin{smallmatrix} m & M & k \\ r & a & b \end{smallmatrix}\right]\Phi$
ImgQL in VoxLogicA :	!	&	|	~>	distleq(r,f)	distgeq(r,f)	N f	crossCorrelation(r,a,b,f,m,M,k) $\bowtie c$



(a) P_1 and P_2 (b) h_1 (c) h'_1 (d) h'_2 (e) h_2

Fig. 3. Example: (a) Small image with a hyperrectangle (P_1) centred on voxel X and a hyperrectangle (P_2) centred on voxel Y. Hyperface $P_1 \setminus P_2$ consists of the voxels a, b and c, whereas hyperface $P_2 \setminus P1$ consists of the voxels d, e and f; (b) Histogram h_1 of P_1; (c) Histogram h'_1 of $P_1 \setminus P_2$; (d) Histogram h'_2 of $P_2 \setminus P1$; (e) Histogram h_2 of P_2, with $h_2(i) = h_1(i) - h'_1(i) + h'_2(i)$, for each bin i.

which each voxel stores the value of the cross correlation computed for the corresponding voxel of the input image and, similarly, the value of crossCorrelation(r, a, b, f, m, M, k) $\bowtie c$ is a Boolean image in which each voxel is true if and only if the cross-correlation value of the corresponding voxel in the input image satisfies the bound $\bowtie$ with respect to the scalar c.

Examples of their use will be provided in the next section where the case studies are presented.

4. Symbolic methods for brain segmentation

In this section we use spatial model checking to address both the delineation of brain tumours, in particular glioblastomas (GBM), and that of tissues of the normal brain, in particular white and grey matter. In clinical studies and routine treatment, magnetic resonance images are evaluated based mostly on qualitative criteria such as the presence of hyperintense tissue appearing in the images [31].

In the sequel we illustrate the VoxLogicA approach to the automatic segmentation of brain lesions and brain tissues. We validate the approach on two different case studies for which public benchmark datasets are available. One concerns the segmentation of brain tumours (high grade glioblastoma) and the other the segmentation of normal brain tissues, in particular white and grey matter. We first show the definition of some derived operators common to both case studies, that were introduced in Section 2, and define a set of commonly used similarity indexes. This is followed by a detailed description of the two case studies.

4.1. Additional Operators in ImgQL

We make use of a number of ImgQL operators that are common to both case studies. Specification 2 shows the ImgQL version of the derived operators introduced in Section 2 and a shorthand for the cross correlation operator. For readability, in these definitions, as a convention, we use f and g to stand for formulas, r for a radius value, img for a grey-scale image and k for the number of bins in a histogram. The operators near (N) and surrounded are pre-defined and part of the stdlib.imgql file.

For the case studies on brain segmentation two further operators turn out to be very useful, the *maxvol* operator and the *percentiles*

ImgQL Specification 2: Derived Operators in ImgQL

```
import "stdlib.imgql"
let grow(f,g) = (f | touch(g,f))
let smoothen(r,f) = distleq(r,distgeq(r,!f))
let similarTo(r,f,img,k) = crossCorrela-
  tion(r,img,img,f,min(img),max(img),k)
```

operator. A point satisfies $maxvol\,\Phi$ if it belongs to a *largest* connected component of (the subspace induced by the) points that satisfy Φ. If there are more than one of such largest components, then the points of all such largest components satisfy the property $maxvol\,\Phi$.

The *percentiles* operator is a *quantitative* operator (see Section 3). It has three arguments $img, mask$, and c. It considers the set of points S identified by the Boolean-valued mask $mask$ in the grey-scale image img and returns a grey-scale image in which the centile of the attribute value $\mathcal{A}(intensity, x)$ of voxel x is the fraction of points in S that have an attribute value *below* that of x in img; more precisely, the centile value $\mathcal{A}(intensity, x)$ is defined by

$$\mathcal{A}(intensity, x) = \frac{l(x) + (c \cdot e(x))}{N}$$

where $l(x)$ is the number of voxels in S having an attribute value *below* that of x, $e(x)$ is the number of voxels in S that have an attribute value *equal to* that of x and N is the total number of voxels in S. The constant c is used to explicitly specify the fraction of the voxels to consider that have an attribute value *equal* to that of x.[11] Histograms of the intensity levels of MR images of the brain may differ from each other due to inter-patient or inter-scanner differences or depending on the actual acquisition volume or the file format used to store the image.[12] Various normalisation procedures have been proposed in the literature to overcome this problem. A common method is the *equalisation of histograms*, frequently used for texture analysis [36]. We do not use this method as it changes the relationship between intensity levels of different structures in the image, which we use rather prominently for differentiating different tissues. For our purposes normalisation of image intensity is sufficient. In our previous work [30] we divided the intensity of each pixel by the average of the intensity levels of all the *significant* pixels in the image. A pixel is considered significant when it does not belong to the background. In the present work we use a more robust form of normalisation using the percentiles operator to find areas of voxels with a (relative) level of intensity of interest and refine such areas subsequently, as illustrated in detail in the specifications in the case studies.

[11] The parameter c can take any of the three values 0, 0.5 or 1, characterising the specific variants of interest of the percentiles operator. For $c = 0.5$ one obtains the standard percentile rank.

[12] For instance, *jpeg* images, as downloaded from *Radiopaedia.org*, typically use 8-bit precision (typical range 0–255) whereas *dicom* images saved by scanners typically use 12 or 16-bit (for MR images, the typical range is 0–4096 or 0–65536, respectively).

4.2. Similarity indexes

Commonly used similarity indexes to compare the segmentation results with the ground truth are the *Dice–Sørensen similarity index*, the *sensitivity* (i.e. fraction of voxels that the segmentation and the ground truth have in common) and the *specificity* (i.e. the fraction of voxels that are not identified by the segmentation and are also not part of the ground truth) [37].

The definitions of the similarity indexes are given in ImgQL in Specification 3. In these definitions, x and y are two boolean images of the same size, one produced by the VoxLogicA segmentation method and one being the ground truth provided by the data set. They are used as shown in Specification 7 to collect the results in a convenient way in a spreadsheet format. All coefficients give a result between 0 (no similarity) and 1 (perfect similarity). The Dice index is defined as the fraction of twice the volume of corresponding voxels in the image in which both the segmentation, i.e. voxels satisfying formula f, and the ground truth, i.e. voxels satisfying formula g have value 1, and the sum of the volumes of the segmentation and the ground truth, respectively. The sensitivity and specificity are defined in a similar way.

ImgQL Specification 3: Similarity indexes

```
1 let dice(f,g) = (2 .*. volume(f & g)) ./.
    (volume(f) .+. volume(g))
2 let sensitivity(f,g) = volume(f & g) ./.
    (volume(f & g) .+. volume((!f) & (g)))
3 let specificity(f,g) = volume((!f) & (!g)) ./.
    (volume((!f) & (!g)) .+. volume((f) & (!g)))
```

4.3. Segmentation of glioblastoma in 3D medical images

4.3.1. Datasets and methodology

For this case study we evaluate VoxLogicA on the *Brain Tumour Image Segmentation Benchmarks* (BraTS) of 2017, 2019 and 2020 [31, 38]. These datasets, originally intended for training purposes of learning based algorithms, contains 210 (respectively, 259 and 293) multi-contrast MRI scans (in NIfTI format) of high grade glioma patients that were obtained from multiple institutions and were acquired with different clinical protocols and various scanners. In the sequel we will refer to this dataset as the *BraTS training datasets*. All the training datasets provided as part of the BraTS Challenge include a manual segmentation, where voxels that are part of tumour regions have been labelled manually by domain experts. These manual segmentations (also in NIfTI format) were approved by experienced neuro-radiologists. We use this segmentation as the *ground truth*. In this case study we used the T2 Fluid Attenuated Inversion Recovery (FLAIR) type of scans, which is one of the four modalities provided in the benchmark. The numeric thresholds of the VoxLogicA specification, that we present in the sequel, were manually calibrated against a subset of the first 20 cases of the BraTS 2017 training data set.

The BraTS 2017, 2019 and 2020 training datasets are partially overlapping, in the sense that they share a subset of images. This does not affect our setting as the purpose of our validation is to evaluate the VoxLogicA procedure on the datasets and show that the results are in line with the state-of-the-art of segmentation results produced by other methods in the literature for each year of the BraTS Challenge, and related BraTS dataset, as shown in the leader board scores provided by the BraTS Challenge. We refer to Section 4.3.3 for a detailed discussion of the results and comparison.

For what concerns the *pre-processing* of the images in the BraTS data sets, we refer to [31] for details. We have not applied any further pre-processing to the images as provided by the BraTS datasets. The only step that could be considered as a form of pre-processing is that we used the percentiles of the intensity of pixels instead of their absolute intensities. However, this step is an explicit and integral part of the tumour segmentation method that we propose (see Specification 6 line 1).

4.3.2. ImgQL Specification of GBM

The specification of the segmentation method for GBM in ImgQL is surprisingly simple and concise. We present the various parts of the specification below. It is divided into four parts: (1) loading of the image and the ground truth; (2) identification of the voxels that are part of the brain and those that are part of the background; (3) segmentation of the tumour and related oedema; (4) saving the results.

Part (1) is shown in Specification 4. In lines 1 and 3 the 3D MRI scan and the manually segmented ground truth images are loaded, respectively, in NIfTI format. In line 2 the 3D grey scale image flair is defined that stores, at each point, the intensity of the corresponding voxel in the 3D MRI scan. In a similar way, in line 4 the ground truth for the Gross Tumour Volume (GTV) is defined as a 3D Boolean image where true is denoted by 1, corresponding to those voxels with a positive intensity, and zero elsewhere.

ImgQL Specification 4: Obtain image and ground truth

```
1 load imgFLAIR = "Brats17_2013_2_1_flair.nii.gz"
2 let flair = intensity(imgFLAIR)
3 load imgGrndTruth = "Brats17_2013_2_1_seg.nii.gz"
4 let grndTruthGTV = intensity(imgGrndTruth) >. 0
```

Part (2), shown in Specification 5, identifies the background and the brain. It makes use of a built-in operator border that identifies the voxels situated at the border of the image. We make use of the knowledge that the background of MRI brain scans is very dark with voxels having an intensity below 0.1, and that the area composed of all the voxels in the background touches the border. This way we avoid to include also dark voxels within the area of the brain such as those corresponding to cerebrospinal fluid (CSF). The brain itself (line 2) is identified as the complement of the background .

ImgQL Specification 5: Identification of background and brain

```
1 let background = touch(flair <. 0.1,border)
2 let brain = !background
```

Part (3) concerns the actual tumour segmentation and is shown in Specification 6. The procedure consists of three main steps that are also illustrated in the form of images in Fig. 4:

1. *Initial identification of hyperintense regions in the FLAIR image.* Hyperintense regions are likely to be part of the tumour tissue. However, the average intensity of voxels that are part of the brain in the images may differ somewhat between one acquisition and another for various reasons. In line 1 the percentiles operator is used as a way to normalise such differences in intensity. This way the same specification can be applied to segment all images in the dataset without the need to recalibrate or explicitly normalise the intensities of each individual image. Hyperintense regions are areas of voxels with intensity above the 0.95 centile (line 2). Very intense regions have voxels with intensity above the 0.88 centile (line 3). Both areas are smoothened (lines 4–5) such that isolated voxels occupying too small areas or having thin protrusions of 5.0, resp. 2.0, millimetres are not considered, as clinical experience shows that these are unlikely to be part of the tumour or oedema tissue. The hyperintense and very intense voxels are shown in row (b) in Fig. 4 for one patient of the BraTS 2017 data set as the cyan, resp. blue, coloured partially transparent overlays on the original image.

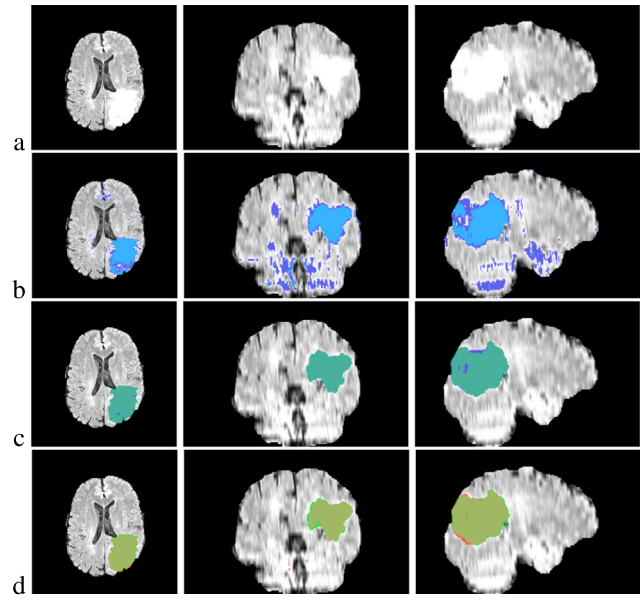


Fig. 4. (a) Cross section of Brats17_TCIA_335_1 MRI (fLTR: axial, coronal, sagittal view); (b) hyperIntense (cyan) and veryIntense (blue); (c) growTum (green) and gtv (blue); (d) gtv (green) and GTV ground truth (red). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

2. *Extending the hyperintense areas and searching for voxels surrounded by similar tissue.* Not all very intense areas in the brain are tumour or oedema tissue. Only very intense areas that are close to and 'touching' the hyperintense areas are very likely to be part of the areas of interest. We extend the hyperintense area 'growing' it with very intense voxels (line 6). The result is shown in Fig. 4 in row (c). Note that in row (b) there are substantially more very intense areas than in the images in row (c). We then search for voxels that have a neighbourhood with a texture that is sufficiently similar to the area of voxels satisfying the formula `growTum` (lines 7–8).
3. *Identification of Gross Tumour Volume.* To find the GTV we extend the tumour area `growTum` with the voxels found in the previous step (line 9). Row (d) in Fig. 4 shows the segmentation result for gtv in green and the GTV ground truth in red as overlays on the original image.

The *Clinical Tumour Volume* (`ctv`) is an extension of the `gtv`. For glioblastomas this margin is a 2–2.5 cm isotropic expansion of the GTV volume within the brain (line 10). This concludes the specification of the whole tumour segmentation method.

ImgQL Specification 6: Tumour segmentation method

```
1  let pflair = percentiles(flair,brain,0)
2  let hI = pflair > . 0.95
3  let vI = pflair > . 0.88
4  let hyperIntense = smoothen(5.0,hI)
5  let veryIntense = smoothen(2.0,vI)
6  let growTum = grow(hyperIntense,veryIntense)
7  let tumSim = similarTo(5,growTum,flair,100)
8  let tumStatCC = smoothen(2.0,(tumSim > . 0.6))
9  let gtv= grow(growTum,tumStatCC)
10 let ctv = distleq(25,gtv) & brain
```

Specification 7 shows an excerpt of the way the relevant information can be saved in ImgQL. In particular, in line 1 the ground truth for the CTV is defined, and in line 2 the image of the tumour segmentation `gtv` is saved in the NIfTI (.nii.gz) format. In fact, any of the intermediate results of the method shown in Specification 6 can be saved this way and manually inspected using any MRI viewer. In lines 3–8 the values of the similarity indexes are saved.

ImgQL Specification 7: Save results

```
1 let grndTruthCTV = distleq(25,grndTruthGTV) & brain
2 save "complete-FLAIR_FL-seg.nii" gtv
3 print "SensGTV" sensitivity(gtv,grndTruthSegGTV)
4 print "SpecGTV" specificity(gtv,grndTruthSegGTV)
5 print "DiceGTV" dice(gtv,grndTruthSegGTV)
6 print "SensCTV" sensitivity(ctv,grndTruthSegCTV)
7 print "SpecCTV" specificity(ctv,grndTruthSegCTV)
8 print "DiceCTV" dice(ctv,grndTruthSegCTV)
```

4.3.3. Validation and results of the GBM segmentation method

Validation[13] of the method was conducted as follows. A priori, 17 of the 210 cases from the BraTS 2017 data set have been excluded[14] contain some form of artefact in the FLAIR image or we judged our current procedure unsuitable for the exhibited pathology. This is due to the presence of multi-focal tumours (different tumours in different areas of the brain), or due to clearly distinguishable artefacts in the FLAIR acquisition, or because the hyperintense area is too large and clearly not significant (possibly by incorrect acquisition). The judgement was made based on expert visual inspection of the cases. However, for completeness, we present the results both for the full dataset (210 cases), and for the subset without these problematic cases (193 cases) to allow for a fair comparison with other methods.

Table 2 shows the mean, the range and the median values of the similarity indexes (Dice, sensitivity and specificity) for both the GTV and CTV volumes for the specification in Section 4.3.2 applied to the BraTS 2017 training dataset. The top-scoring methods of the BraTS 2017 Challenge [39] can be considered a good sample of the state-of-the-art in this domain. Among those, in order to collect significant statistics, we selected the 18 techniques that have been applied to at least 100 cases of the BraTS 2017 training dataset (210 HGG and 76 LGG cases). For the latter techniques the *median* and *range of values* of the Dice score, sensitivity and specificity indexes for the GTV segmentation of the whole tumour were, respectively, 0.88 (ranging from 0.64 to 0.96), 0.88 (0.55 to 0.97) and 0.99 (0.98 to 0.999). Full data for these techniques are available in the BratS 2017 training phase leaderboard.[15] Compared to these numbers, the results in Table 2 show that our results are well in line with the state-of-the-art in this domain as we obtained an average Dice score for the GTV of 0.85 (StdDev 0.09) on 193 cases of the training set with a median of 0.87. This is very interesting, and to some extent also surprising, considering the extreme simplicity of the presented specification.

We have further validated the same ImgQL specification for the same threshold values on the datasets of BraTS 2019 (see results in Table 3) and BraTS 2020 (see results in Table 4).

In 2019, 74 teams submitted to the BraTS Challenge using mainly machine learning techniques and applying their method to at least 100 cases of the BraTS 2019 training dataset (consisting in total of 259 HGG and 76 LGG cases). The three best teams reached an average Dice score

[13] Full results are available from the authors and will be made available in a public repository when the article will be published.

[14] These concern the following: CBICA_AYI_1, CBICA_BHK_1, TCIA_473_1, 2013_11_1, 2013_3_1, 2013_4_1, CBICA_AAL_1, CBICA_AVG_1, CBICA_AWI_1, TCIA_198_1, TCIA_149_1, TCIA_247_1, TCIA_338_1, TCIA_411_1, TCIA_499_1, TCIA_314_1, CBICA_ABN_1.

[15] https://www.cbica.upenn.edu/BraTS17/lboardTraining.html of Nov. 5, 2018.

Table 2
VoxLogicA evaluation on the HGG cases of the BraTS 2017 training data set (vI = 88, hI = 95).

	Dice (193 cases)			Sensitivity (193 cases)			Specificity (193 cases)		
	Mean (Stdev)	Range	Median	Mean (Stdev)	Range	Median	Mean (Stdev)	Range	Median
GTV	0.85 (0.09)	0.57–0.97	0.87	0.86 (0.12)	0,45–1,0	0,91	1.0 (0.00)	0.99–1.0	1.0
CTV	0.91 (0.08)	0.53–0.99	0.94	0.94 (0.07)	0.58–1.0	0.97	0.99 (0.01)	0.93–1.0	1.0
	Dice (all 210 cases)			Sensitivity (all 210 cases)			Specificity (all 210 cases)		
GTV	0.81 (0.16)	0.0–0.97	0.87	0.84 (0.18)	0–1.0	0.9	0.998 (0.0)	0.99–1.0	1.0
CTV	0.88 (0.15)	0–0.99	0.93	0.92 (0.14)	0–1.0	0.96	0.99 (0.1)	0.91–1.0	1.0

Table 3
VoxLogicA evaluation on the HGG cases of the BraTS 2019 training data set (vI = 88, hI = 95).

	Dice (242 cases)			Sensitivity (242 cases)			Specificity (242 cases)		
	Mean (Stdev)	Range	Median	Mean (Stdev)	Range	Median	Mean (Stdev)	Range	Median
GTV	0.85 (0.09)	0.44–0.97	0.88	0.87 (0.13)	0.44–1.0	0.92	1.0 (0.00)	0.99–1.0	1.0
CTV	0.91 (0.08)	0.53–0.99	0.94	0.94 (0.09)	0.52–1.0	0.97	0.99 (0.01)	0.93–1.0	1.0
	Dice (all 259 cases)			Sensitivity (all 259 cases)			Specificity (all 259 cases)		
GTV	0.82 (0.15)	0–0.97	0.87	0.85 (0.17)	0–1,0	0.92	0.998 (0.0)	0.99–1.0	1.0
CTV	0.89 (0.14)	0–0.99	0.93	0.92 (0.14)	0–1.0	0.97	0.99 (0.01)	0.91–1.0	1.0

Table 4
VoxLogicA evaluation on the HGG cases of the BraTS 2020 training data set (vI = 88, hI = 95).

	Dice (276 cases)			Sensitivity (276 cases)			Specificity (276 cases)		
	Mean (Stdev)	Range	Median	Mean (Stdev)	Range	Median	Mean (Stdev)	Range	Median
GTV	0.85 (0.11)	0.0–0.97	0.88	0.859 (0.14)	0.0–1.0	0.91	0.999 (0.0)	0.99–1.0	1.0
CTV	0.91 (0.1)	0.0–0.99	0.94	0.937 (0.10)	0.0–1.0	0.97	0.993 (0.01)	0.93–1.0	1.0
	Dice (all 293 cases)			Sensitivity (all 293 cases)			Specificity (all 293 cases)		
GTV	0.82 (0.16)	0.0–0.97	0.87	0.843 (0.17)	0.0–1.0	0.91	0.998 (0.0)	0.99–1.0	1.0
CTV	0.889 (0.14)	0.0–0.99	0.94	0.923 (0.14)	0.0–1.0	0.97	0.992 (0.01)	0.91–1.0	1.0

of 0.88 to 0.89 for the segmentation of the whole tumour area on the final test set consisting of 166 cases (mixed HGG and LGG) [40–42]. With our method we reached an average Dice score of 0.85 (StdDev 0.09) but on the larger set of 242 HGG cases of the training set (see footnote[14] for details on excluded cases).

In the 2020 BraTS Challenge 84 teams submitted their techniques and that have been applied to at least 100 cases of the training dataset (consisting in total of 293 HGG and 76 LGG cases). The four best performing teams reached an average Dice score of between 0.8828 to 0.8895 on the final test set of 166 cases [1,43–45]. With our method we reached an average Dice score of 0.85 (StdDev 0.11), which is a bit lower, but we applied it on the larger set of 276 HGG cases of the training set (see footnote[14] for details on excluded cases).

Full results for these techniques, applied to the training sets, are available in the BraTS 2019 leaderboard.[16] and the BraTS 2020[17] leaderboard[18] Note that the final BraTS test sets[19] contain both HGG and LGG cases and that these cases and their ground truth are not publicly available. So, unfortunately, we could apply our method only on the cases of the larger BraTS training sets. Nevertheless, for the BraTS 2019 and 2020 training sets our results are close to the state-of-the-art.

An interesting aspect of our proposed technique is that it directly considers *all dimensions* of the 3D image at once, instead of performing segmentation layer-by-layer. This makes it easier to identify contiguous areas of interest in the three dimensions at once. Furthermore, a single specification gives good results on a large set of images. So no calibration for individual images is needed, the method does not require a training phase and is explainable. Of course, if higher accuracy is required, one can calibrate some thresholds for individual patients. We address this idea further in Section 5.1.

This simple specification shows the potential of the method, but has also some limitations. As we mentioned before, it has not been designed to segment multi-focal tumours. It can also not be used on images with clearly distinguishable artifacts in the acquisition. Finally, it seems to be slightly less accurate compared to the most recent machine learning techniques. In Section 5.2 we address a proposal for a hybrid AI approach that combines spatial logic with nnU-Net that has the advantages of both worlds, namely higher accuracy while retaining the explainability of the spatial logic based approach.

4.3.4. Computational performance

The 3D images used in our analysis have size 240 × 240 × 155 (about 9M voxels). The evaluation of each patient of the BraTS training data sets takes about 5 s on a desktop computer equipped with an Intel Core I9 9900K processor (with 8 cores) and 32 GB of RAM.

4.4. Segmentation of normal brain tissues

4.4.1. Datasets and methodology

For this second case study we evaluate VoxLogicA on the BrainWeb [33] data set.[20] consisting of 20 synthetic[21] MRI's of the normal

[16] https://www.cbica.upenn.edu/BraTS19/lboardTraining.html.
[17] Last edition for which publications are available at the time of writing.
[18] https://www.cbica.upenn.edu/BraTS20/lboardTraining.html of June 6, 2023.
[19] This is the dataset on which each technique was evaluated in the final BraTS Challenge.
[20] https://brainweb.bic.mni.mcgill.ca/brainweb/anatomic_normal_20.html.
[21] The MRI's of the brain were obtained with T1-weighted simulated data with the following specific parameters: SFLASH (spoiled FLASH) sequence with TR = 22 ms, TE = 9.2 ms, flip angle = 30 deg and 1 mm isotropic voxel size.

brain, i.e. a brain without lesions. Synthetic ground truth images for various types of brain tissues are provided as part of the dataset, such as for white and grey matter, but also for cerebrospinal fluid (CSF), fat, muscle, skull, blood vessels, dura matter and bone marrow. In this study we focus on the segmentation of white and grey matter.

We have used the randomly selected case Pat04 to develop and calibrate the segmentation method.

As in the previous case study, we use the Dice index, the sensitivity index and the specificity index for assessing the quality of the segmentation.

No further preprocessing of the images have been performed, apart from the fact that we used percentiles of the image intensity instead of the absolute values of pixel intensity. However, this is an explicit part of the segmentation procedure (see line 1 of Specification 9).

4.4.2. *ImgQL specification for normal brain*

In this section, we briefly describe the main parts of the VoxLogicA specification and illustrate the intermediate results of some steps in Figs. 5 and 6. The specification uses the same operators as previously defined for the glioblastoma case in Specification 2. Specification 8 shows the loading of the image (line 1) and the definition of the grey scale image of intensities of the original image (line 2). The image is a 3D MRI of type T1, short for T1-weighted-FLAIR, in NIfTI file format.

ImgQL Specification 8: Obtain image file of type T1

```
1 load imgT1 = "pat04_t1.nii.gz"
2 let t1 = intensity(imgT1)
```

The segmentation procedure consists of three main parts: the distinction between the head and the background, the segmentation of white matter and the segmentation of grey matter, followed by a minor step to refine the segmentation of the white matter after the grey matter has been identified.

Specification 9 shows the segmentation of head and background as a series of steps that get closer and closer to the goal. This procedure differs from that in the glioblastoma case study because the images in the BrainWeb data set are not skull-stripped. As before, the percentiles operator is used to deal in a normalised way for what concerns the voxel intensities (line 1). In line 2 the 'percentile' image (i.e. grey-level image with a percentile level as attribute for each voxel) is used to identify all voxels from which a border point can be reached passing only through relatively dark (low intensity) voxels that satisfy `bg <. 0.6`. This results in image `!bg1`; the relevant points are shown in cyan in row (b) of Fig. 5. Note that the formula is satisfied also by some points inside the skull, for example in the area of the ears. A first approximation of the (smoothened) area of the head (`head1`) is composed of voxels that are *not* part of the background (`!bg1`), of which the maximum such area is taken (`maxvol`). The latter operator makes sure that all smaller areas in the background, that are clearly not part of the head, are excluded from the head (line 3). We can also observe these in row (b) in Fig. 5 as small dark specks within the cyan area of the background. We apply a similar approach to the area of the head (line 4–5) by first including voxels that are at 3 mm from a voxel satisfying `head1` and then identifying the background as the maximal volume of the complement of the head. The result is shown in row (c) of Fig. 5. We then add again the voxels in the area of 3 mm of the head and obtain a good segmentation of the background (line 6). The final segmentation of the head is then obtained as the complement of the background (line 7), shown in violet in row (d) of Fig. 5.

The segmentation method for white matter is shown in Specification 10 and finalised in Specification 12 after the grey matter has been segmented. This method exploits the level of self-similarity of the

ImgQL Specification 9: Segmentation of head and background

```
1 let bg = percentiles(t1, t1 >. 0, 0.5)
2 let bg1 = touch(bg <. 0.6,border)
3 let head1 = maxvol(smoothen(2,!bg1))
4 let head2 = distleq(3,head1)
5 let bg2 = maxvol(!head2)
6 let background = distleq(3,bg2)
7 let head=!background
```

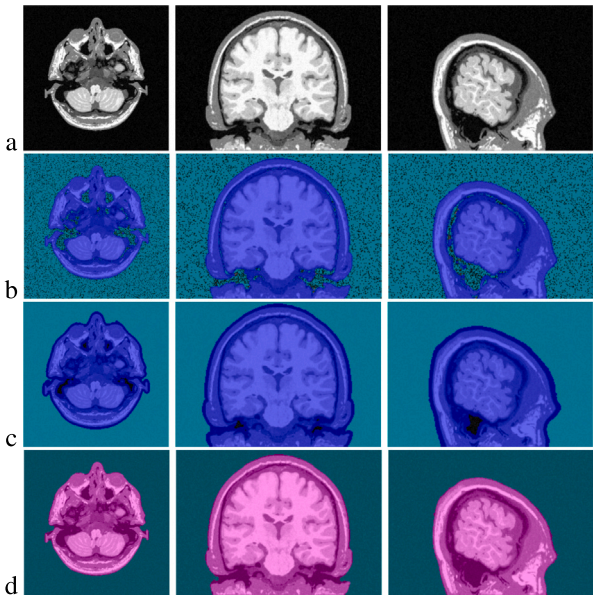


Fig. 5. Identification of head and background in BrainWeb pat04 MRI at slice (x,y,z) = (191,122,22), (fLTR: axial, coronal, sagittal view): (a) Original view; (b) In cyan voxels satisfying `bg1` and in blue those satisfying `head1`; (c) in cyan voxels satisfying `bg2` and in blue those satisfying `head2`; (d) in violet voxels satisfying `head` and in green those satisfying `background`. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

neighbourhood of a voxel belonging to the head to that of the *entire* head, and in a second phase, to that of the likely white matter found in the first approximation. Again, percentiles are used as substitute for a normalisation of the intensity of the image (line 1), but now only with respect to the area of the head that was defined in Specification 9. In line 2 the similarity score is obtained of voxels belonging to the head with respect to the whole head, considering their neighbourhood of 3 mm and histograms with 30 bins. This score is used in `white1` (line 4) to obtain a first approximation of the white matter, using the empirical knowledge that voxels in white matter have a similarity (cross correlation) score of between 0.2 and 0.6, and are situated in the internal part of the head, specified by `headInt` (line 3). However, as shown in row (b) of Fig. 6, this gives only a first rough approximation of the white matter. We apply the same approach of self-similarity a second time. This time we look for similarity with respect to the white matter defined by `white1`. We also consider a smaller neighbourhood around each voxel, for better precision (line 5–6), and require a higher level of similarity (cross correlation score > 0.6). Finally, in line 7 some further voxels are admitted to the white matter defined in `white2`. These are less similar, but are surrounded by white matter and are situated very close (less than or equal to 1.0 mm) to `white2`. Finally, in Specification 12 (line 1), some further white matter is added to `white3` that lays in between the grey matter and `white3`, leading to the final specification of white matter `white`. Row (c) of Fig. 6 shows `white3` (in green) and `white` (in blue). Row (d) of Fig. 6

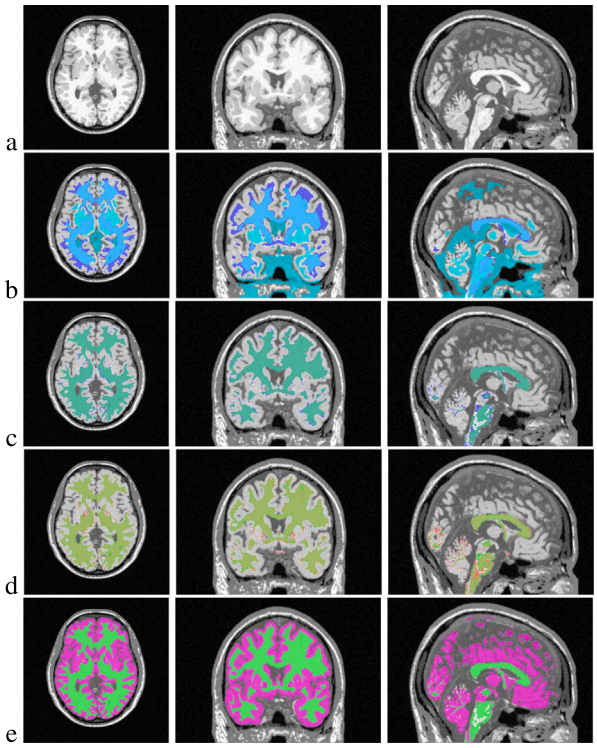


Fig. 6. (a) Cross section of BrainWeb pat04 MRI at slice (x,y,z) = (129,147,78), (fLTR: axial, coronal, sagittal view): (a) Original view; (b) `white1` (cyan) and `white2` (blue); (c) `white3` (green) and `white` (blue); (d) `white` (green) and white ground truth (red); (e) `white` (green) and `grey` (violet). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

shows the comparison of `white` (in green) with respect to the ground truth of white matter (shown in red). As can be observed, there is an excellent correspondence between `white` and the ground truth. Only in a few places the ground truth does not completely overlap with the segmentation. These voxels show up as clearly red (when they are only part of the ground truth) or as bright green (when they are only part of the segmentation).

ImgQL Specification 10: Segmentation of white matter

```
1 let pt1 = percentiles(t1,head,0.5)
2 let headSim = similarTo(3,head,t1,30)
3 let headInt = head & !(distleq(30,!head))
4 let white1 = maxvol((headSim >. 0.2) & (headSim
    <. 0.6) & headInt)
5 let whiteT1 = similarTo(1,white1,t1,30)
6 let white2 = maxvol(whiteT1 >. 0.6)
7 let white3 = white2 | ((headSim >. 0.3) &
    surrounded((headSim >. 0.3),white2) &
    (distleq(1,white2)))
```

The segmentation method for grey matter is defined in Specification 11. The method is again exploiting self-similarity scores, but also the knowledge that white and grey matter are touching each other. Grey matter can be found closer to the skull, so in this specification a larger interior of the head is considered (line 1) than in Specification 10. In line 2 a first approximation of voxels in grey matter is defined that consists of those voxels that are sufficiently similar to the head, but that do not have a too high intensity (centile < 0.8) and laying within the internal part of the head as defined by `headInt2`. In line 3 this set of voxels is further restricted to those areas that touch the white matter only via (paths passing by) other voxels that satisfy `grey1`. This gives the approximation `grey2`. In line 4–5 a second self-similarity score is exploited and some further restrictions on the intensity and similarity to white matter are applied. Finally, in line 6 some areas enclosed between the white and grey matter obtained so far are added leading to the final segmentation of grey matter, shown in violet in the bottom row of Fig. 6.

ImgQL Specification 11: Segmentation of grey matter

```
1 let headInt2 = head & !(distleq(10,!head))
2 let grey1 = (headSim >. 0.5) & (pt1 <. 0.8) &
    headInt2
3 let grey2 = touch(grey1,white3)
4 let greyT1 = similarTo(3,grey2,t1,30)
5 let grey4 = (greyT1 >. 0.3) & (whiteT1 <. 0.8) &
    (pt1 >. 0.4) & (pt1 <. 0.8)
6 let grey = touch(grey4,white3) &
    distleq(9,white3) & !white3
```

The segmentation of the grey matter also gives a further opportunity to refine the segmentation of the white matter, adding additional voxels between white and grey that had escaped so far. This is shown in Specification 12.

ImgQL Specification 12: Further white matter

```
1 let white = white3 | ((pt1 >. 0.7) &
    (distleq(5,white3)) & (distleq(3,grey)) &
    (!(grey | white3)))
2 let brain = white | grey
```

This concludes the segmentation of white and grey matter illustrated on a single case of the BrainWeb data set (pat04). We validate the method on the whole BrainWeb data set in the next section.

4.4.3. *Validation and results of the method for grey and white matter*

Table 5 shows the results of the application of the `VoxLogicA` specification in Section 4.4.2 to the whole BrainWeb dataset containing 20 cases.[22] This concerns the segmentation of white and grey matter for this dataset. The first interesting observation is that the Dice coefficient for both the white and grey matter is above 0.9 on average. This indicates an excellent quality of the, still rather simple, segmentation procedures. Also the standard deviation is relatively small, in particular for the white matter. In none of the twenty BrainWeb cases the Dice score is below 0.88.

Equally good results are obtained for the sensitivity and the specificity. The mean values for the sensitivity are 0.97 for the white matter, and 0.89 for the grey matter. In both cases the standard deviation is very small (0.02 and 0.03, respectively) which indicates a very high stability of the scores. This is interesting also because the BrainWeb images are synthetic images, which means that the ground truth can be established in a more precise way than would be possible with images from actual persons. The score for the specificity is even higher (0.99 for both white and grey matter). This indicates that the `VoxLogicA` segmentation procedure produces very few false positives.

Overall, from the qualitative point of view, the results so far for this novel segmentation approach for white and grey matter seem very promising. Future research is needed to see whether the specifications give equally good results on MRI images of actual persons.

[22] Full results are available from the authors and will be made available in a public repository when the article will be published.

Table 5
VoxLogicA evaluation on the BrainWeb data set of 20 synthetic normal brains.

	Dice (20 cases)			Sensitivity (20 cases)			Specificity (20 cases)		
	Mean (Stdev)	Range	Median	Mean (Stdev)	Range	Median	Mean (Stdev)	Range	Median
White	0.93 (0.02)	0.88–0.95	0.93	0.97 (0.02)	0.92–0.99	0,97	0.99 (0.00)	0.98–1.0	0.99
Grey	0.91 (0.16)	0.88–0.93	0.92	0.89 (0.03)	0.83–0.94	0.90	0.99 (0.00)	0.99–1.0	0.99

4.4.4. Computational performance

The average analysis time per patient in the BrainWeb dataset for the specification in Section 4.4.2, segmenting both white and grey matter, is 10,321.4 ms (stdev 242 ms), ranging from 10,074 ms to 11,112 ms on a desktop computer equipped with an Intel Core I9 9900K processor (with 8 cores) and 32 GB of RAM. Each image consists of circa 12 M voxels (i.e. 256 × 256 × 181). This time is larger than that for the segmentation of tumours in the BraTS datasets. This is partially due to the larger number of voxels of the images and to the fact that in the BrainWeb case two different tissues are segmented instead of one, involving more operations.

5. Hybrid method: Combining logic and machine learning

In Section 4 we discussed the potential of a symbolic spatial-logic-based AI approach for the automatic segmentation of brain tumours. In this section we return to brain tumour segmentation and investigate a *hybrid setup*, combining our symbolic method with a deep learning system.

The first question we address is what is the *best possible accuracy* of the logical specification presented in Section 4 in terms of average Dice score. To answer this question, we perform an exhaustive parameter search. For the sake of simplicity and computational efficiency, we employ just the "region growing" part of the specification, that is, lines 1 – 6 of Specification 6, leaving out the part on fine-tuning by cross-correlation. As we discussed in Section 4, the simplified specification depends on two threshold values: one for hyperIntense voxels (hI) and one for very intense (vI) voxels. In Section 4 we used a single pair of (hI, vI) thresholds for the segmentation of *all* the images in the BraTS datasets. These gave very acceptable results. However, in Section 5.1 we show that each *individual* image is amenable to much more accurate logic-based segmentation, by finding a specific optimal pair of thresholds for each image.

Another question is whether Machine Learning could play a pivotal role in identifying (approximations of) such individual optimal thresholds, after training on a limited subset of representative elements from the BraTS 2020 dataset. In Section 5.2 we provide a procedure to do so, paving the way to an interesting combined use of spatial model checking and deep learning for brain tumour segmentation.

In particular, we first obtain a segmentation by a trained deep learning algorithm. Such segmentation is used to find thresholds for which the logic-based segmentation method obtains the best match, by optimising the similarity index w.r.t. the trained deep learning segmentation algorithm. If the Dice score for such a match is sufficiently high, this means that we found two completely different methods to agree on the segmentation, and, as a side effect, we found a pair of thresholds that are close to the optimal ones w.r.t. the segmentation result obtained with the deep learning algorithm. A detailed flow diagram of the hybrid approach is provided in Fig. 7.

The advantage of a method with built-in redundancy is that if the methods show good agreement in the segmentation, one obtains for free a human understandable justification for how the results have been obtained from the logics based part of the method. This helps experts to have justified confidence in the quality of the results. If there is no satisfactory level of agreement, this triggers a warning that there might be something exceptional going on with the image which deserves a manual expert inspection.

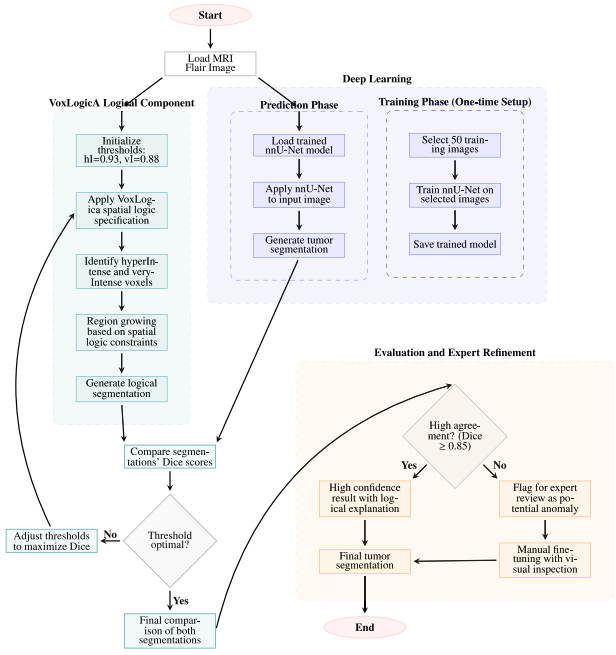


Fig. 7. Hybrid learning and model checking segmentation method.

5.1. Demonstrating Existence of Optimal Thresholds for hI and vI via Exhaustive Analysis

We first present the results of a pre-study indicating that there is indeed a clear association between single images and a specific threshold pair (hI, vI) that leads to an optimal individual segmentation with respect to the provided ground truth for the ImgQL specification (i.e. the version in Specification 6 without the cross correlation in lines 7–8). In this study we use the HGG cases of the BraTS 2020 dataset. The values considered for the threshold hI range from 0.92 to 0.94 and those considered for the threshold vI range from 0.83 to 0.91, in steps of 0.01, as for those combinations the average Dice score is sufficiently high to be of interest (i.e. having a value above 0.75).

Fig. 8 shows which combination of thresholds hI and vI lead to which average value for the Dice score calculated over *all* HGG cases of the BraTS 2020 dataset (293 MRI scans), if no individual optimisation takes place. The figure shows that the value of hI has much less impact on the Dice score, for the range of interest, than the value of vI. The values for which the best average Dice score is obtained is hI = 0.93 and vI = 0.88, leading to an average Dice score of 0.826 (StDev 0.141).

Next, we proceed to investigate existence of optimal values for each individual case, and the impact of optimisation on the average Dice score for the dataset. We analyse each HGG image of the dataset using the ImgQL specification, identifying, by means of tabulation (that is, exhaustive search), the combination of values for hI and vI that leads to the best possible Dice score with respect to the ground truth for that image. The result shows a clear increase in both the Dice score of the single image and the average Dice score of these individually optimised segmentations. Whereas the best score for a fixed pair of values was 0.826 (StDev 0.141), as we have seen above, the average of

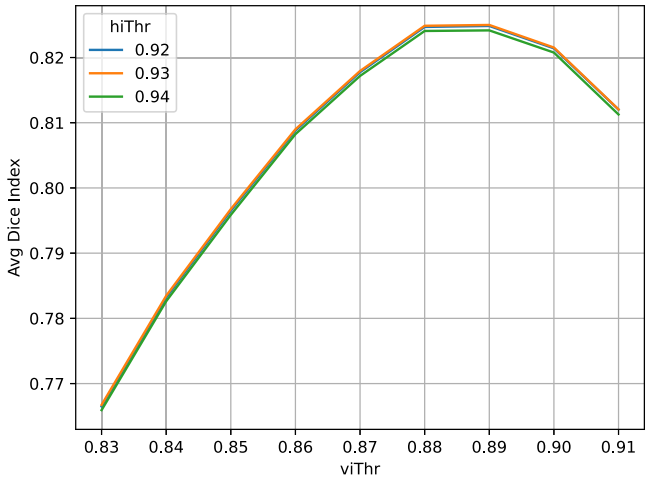


Fig. 8. Average Dice score for various combinations of thresholds for vI and hI for all the 293 HGG cases in the BraTS 2020 Dataset.

the individually best Dice scores for 234 cases is 0.884 (StDev 0.075) (see first row of Table 6). This means that fine-tuning of the hI and vI values for individual images may indeed increase the accuracy, in terms of Dice score, considerably. For the BraTS 2020 dataset we found improvements of the Dice score of up to 0.2 units, e.g. the image BraTS20_Training_135 improved the Dice score from 0.66 to 0.86 using the thresholds (0.92,0.83) instead of (0.93, 0.89) for (hI,vI).

5.2. Finding Thresholds for hI and vI via nnU-Net

Exhaustive search proves that selecting individual optimal values of hI and vI, maximising the Dice score, enhances the result considerably. However, this does not yet provide a practical method to find close approximations for the ideal optimal values of such parameters because in the real world setting no ground truth is available when practitioners have to segment a new tumour lesion. In this section we investigate the usefulness of the Convolution Neural Network (CNN) based algorithm nnU-Net for this purpose. The latter consists of a deep learning based segmentation framework that automatically configures itself, including preprocessing, network architecture, training and post processing [1]. We have chosen nnU-Net as our reference deep learning technique as it currently is one of the state-of-the-art deep learning techniques available and suitable for this feasibility study on brain tumour segmentation. In future work other suitable sub-symbolic techniques could be considered as this field is in rapid evolution (see for example [46,47]).

The assumption here is that an appropriately trained system can provide a good segmentation of a tumour. One could then use the ImgQL specification to find another segmentation, with related hI and vI values, in such a way that the similarity, in terms of Dice score, with the segmentation obtained via nnU-Net is as high as possible. If a sufficiently high Dice score is found, this implies a high confidence on the correctness of the segmentation since it has been obtained using two completely different segmentation methods that give very similar results. In case the two segmentations are not sufficiently aligned, this may be an indication that a human expert needs to look at it as there may be some form of anomaly present in the image. Furthermore, in case the segmentations nearly coincide, the ImgQL specification can provide a human understandable explanation and justification for the result. Finally, the obtained result could be amenable to further manual refinement and fine-tuning. For example a domain expert could slightly vary the vI value and use visual inspection to see whether an even more satisfactory segmentation could be obtained without much further effort but with a meaningful human-centric control over the final result. Note that such 'tuneability', which is not possible with plain machine-learning based approaches similar to nnU-Net, is a desiderata in the clinical setting, as different sites may decide to adopt different segmentation protocols, and the final decision is up to the practitioner.

Before describing our method in more detail and studying its performance, we note in passing that, in the classification proposed by [10], a system is called Neuro|Symbolic (also "of Type 3", see [8]), when a neural network converts non-symbolic input, such as the pixels of an image, into a symbolic data structure, which is then processed by a symbolic reasoning system. Our method, which is rather novel and peculiar, also uses a neural network to encode non-symbolic information and feed it into a symbolic system (the model checker), although it turns non-symbolic input into Boolean regions, which have a dual nature. In fact, a Boolean region is as a non-symbolic encoding (a Boolean-valued image) of symbolic information (an atomic proposition denoting a specific image feature). By this, we consider our method as Neuro|Symbolic in turn.

A new convolutional neural network (CNN)-based algorithm, nnU-Net, was recently developed [1], offering a fully automated deep learning-based segmentation framework. Unlike other methods available in the literature, nnU-Net is designed to self-configure across all stages of the segmentation pipeline, including preprocessing, network architecture, training, and postprocessing. This adaptability allows nnU-Net to achieve high performance without requiring manual intervention or expert knowledge, making it a versatile tool for a wide range of medical imaging tasks. Its automated nature makes it an ideal companion to our fully-automated model-checking based approach, which is, however, independent from the specific neural network used, which could be easily replaced by others for further experimentation.

Hybrid method. The method we follow (see Fig. 7) is to train nnU-Net (3D full resolution U-Net configuration) on 50 randomly selected percentiles-normalised Flair images (without background) from the HGG part of the BraTS 2020 dataset, using the provided ground truth, and then use the rest of the dataset as test set for validation.

We then select by tabulation the best values for the thresholds by maximising the Dice score with respect to the nnU-Net segmentation. Figs. 11 to 13 show the frequency distribution of the Dice score for the various GTV segmentations for 234 BraTS 2020 HGG images (obtained by excluding the 50 cases used for training and 17 discarded ones, see footnote[14]) w.r.t. the ground truth. Fig. 11 shows the frequency of Dice scores for the individually optimised ImgQL segmentations w.r.t. the ground truth. Fig. 12 shows the frequency of the Dice scores w.r.t. the ground truth for the ImgQL segmentation in which the vI and hI thresholds have been obtained by maximising the Dice score with respect to the nnU-Net segmentation instead of w.r.t. the original ground truth. Fig. 13 shows the frequency of the Dice scores for the segmentation obtained in the validation phase with the nnU-Net predictor that was trained on 50 cases w.r.t. the ground truth.

Results and comparison. Table 6 summarises the Dice, sensitivity and specificity scores for all four methods. The $*$-symbol denotes the Wilcoxon signed-rank statistical test with p-value < 0.01 with respect to the segmentation obtained with the purely symbolic method with a fixed pair of pre-established thresholds (vI = 88, hI = 93) used for all 234 cases.

Illustration of differences. Fig. 9 illustrates the segmentation results for the image BraTS20_Training_099 for the various methods. This case has been specifically selected because it shows maximum difference between segmentations, even if the Dice scores for this case are not that high.[23] We can clearly see that the segmentation with the method of Section 4 with (vI = 88, hI = 93), in red, identifies a too large area w.r.t. the ground truth (in green). This can be explained by

[23] Dice values for BraTS20_Training_099 case: GTV-nnU: 0.68, GTV-nnAsGT: 0.74, GTV-optimal: 0.78, GTV-symbolic: 0.50.

Table 6
GTV-optimal using tabulation to get optimal thresholds w.r.t. BraTS 2020 ground truth; GTV-nnAsGT using nn-Unet result as ground truth for finding optimal thresholds; GTV-nnU segmentation result from ML algorithm trained with 50 cases; GTV-symbolic segmentation with pure symbolic specification of Section 4.

	Dice (234 cases)			Sensitivity (234 cases)			Specificity (234 cases)		
	Mean (Stdev)	Range	Median (IQR)	Mean (Stdev)	Range	Median (IQR)	Mean (Stdev)	Range	Median (IQR)
GTV-optimal	0.884* (0.075)	0.58–0.97	0.913 (0.07)	0.87* (0.10)	0.43–0.98	0.91 (0.09)	1.00* (0.00)	0.99–1.00	1.00 (0.00)
GTV-nnAsGT	0.867* (0.087)	0.46–0.97	0.895 (0.10)	0.85* (0.122)	0.42–1.00	0.89 (0.13)	1.00* (0.001)	0.99–1.00	1.00 (0.00)
GTV-nnU	0.891* (0.082)	0.42–0.98	0.914 (0.07)	0.92* (0.841)	0.27–1.00	0.94 (0.08)	0.99* (0.00)	0.99–1.00	1.00 (0.00)
GTV-symbolic	0.848 (0.11)	0.0–0.97	0.884 (0.103)	0.88 (0.126)	0.00–1.00	0.92 (0.13)	1.00 (0.00)	0.98–1.00	1.00 (0.00)

* Denotes Wilcoxon signed-rank statistical test with p-value < 0.01 with respect to the GTV-symbolic segmentation with a fixed pair of pre-established thresholds (vI = 88, hI = 93) used for all cases.

the effect of using too generous thresholds for this specific case. By choosing the optimal thresholds for this case (vI = 95, hI = 97), the segmentation result, shown in yellow, is much closer to the ground truth. The segmentation result for the thresholds vI = 93 and hI = 97, obtained from approximating the nnU-net segmentation, is shown in blue and is just a little larger than the ground truth and the optimal method. It provides an explainable segmentation for the area that was also identified by the subsymbolic, and thus more difficult to justify, nnU-net method. When used together, the hybrid method provides accountability whereas the nnU-net segmentation, when showing close correspondence with the hybrid segmentation, may also be used to increase accuracy.

A 3D visualisation of the optimal segmentation for case BraTS20_Training_099 is shown in Fig. 10.

5.3. Discussion

Our study shows that although a single pair of vI and hI thresholds used in the ImgQL specification leads already to very good average Dice scores of the HGG segmentation for the BraTS datasets, this score can be improved considerably by tuning the thresholds for individual images.

Using a trained nnU-Net predictor as a substitute of the ground truth to find suitable thresholds for the ImgQL specification also seems to provide very acceptable results, with a mean of 0.867 (StDev 0.087) and median 0.895 for the 234 HGG cases of the BraTS 2020 training set that were not used for training of the nnU-Net predictor (see Table 6). This mean Dice score is considerably higher than that obtained with the symbolic specification of Section 4 when only a single fixed pair of threshold values is used for the whole set of 234 HGG cases, which was 0.848 (StDev 0.11).

Furthermore, the histogram in Fig. 11 shows that individual fine-tuning of the thresholds for each image can reach even better results with a mean of 0.884 (StDev 0.075) and median 0.913. This confirms that a further fine-tuning with visual inspection by an expert might be a viable way to combine nnU-Net and logic based automatic segmentation with a meaningful, but simple and time saving, human control and adjustment of the results. Moreover, the ImgQL specification provides the expert with a meaningful, high-level explanation and justification of the segmentation, even if the starting point of the method was an automatic Machine Learning based segmentation that served as an first approximation of ground truth.

In conclusion, the hybrid spatial logic and nnU-Net approach provides a quick and accurate segmentation method that is amenable to a meaningful human control and fine-tuning of the segmentation of HGG in patient's MRI scans. Furthermore, the symbolic segmentation with a predefined threshold can be used for comparison in order to check whether the hybrid method indeed improves the segmentation for the individual case at hand. The hybrid, intrinsically and intensionally redundant, approach increases the reliability of the method and facilitates

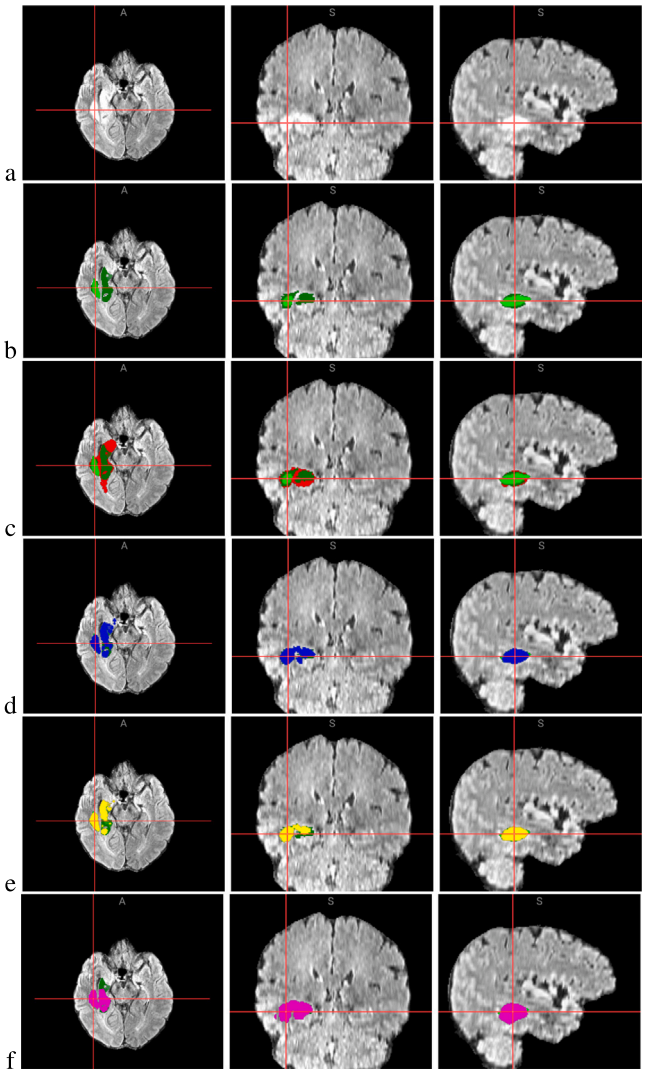


Fig. 9. Visual comparison of the proposed methods compared to manual segmentation on case BraTS20_Training_099. Row (a) FLAIR (fLTR: axial, coronal, sagittal view); Row (b) Ground truth (GT) in green; Row (c) Segmentation result of the specification of Section 4 in red, GT in green; Row (d) Segmentation obtained using the hybrid approach (GTV-nnAsGT), with thresholds $hI = 0.97$ and $vI = 0.93$ in blue, GT in green; Row (e) Best-performing segmentation using VoxLogicA (GTV-optimal), with thresholds $hI = 0.97$ and $vI = 0.95$ in yellow, GT in green; Row (f): nnUNet segmentation (GTV-nnU) in magenta, GT in green. The screenshots have been obtained using VoxLogicA-UI (see [48]). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

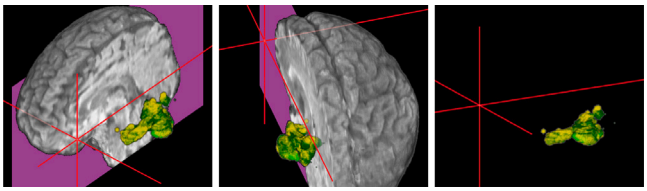


Fig. 10. 3D visualisation of case BraTS20_Training_099 of GTV-optimal in yellow and ground truth in green. The screenshots have been obtained using `VoxLogicA-UI` (see [48]). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

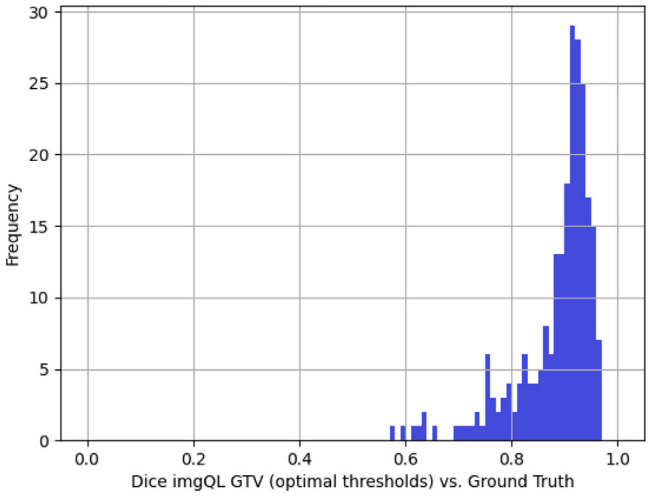


Fig. 11. Frequency distribution of Dice for individually optimised `ImgQL` segmentations w.r.t. ground truth. Mean 0.884 (StDev 0.075) and median 0.913.

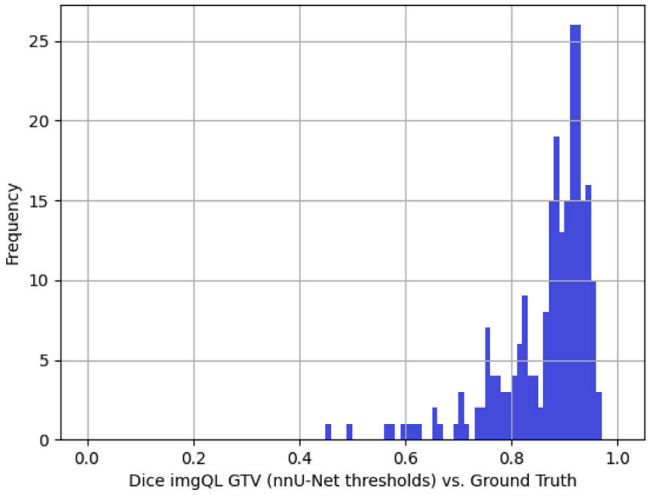


Fig. 12. Frequency distribution of Dice for individually optimised `ImgQL` segmentations w.r.t. nnU-Net segmentation. Mean 0.867 (StDev 0.087) and median 0.895.

the human justification of the results and the critical decisions that depend on those segmentations to propose the best possible treatment of the patient.

6. Related work

The idea of using model checking in medicine in general, and more specifically, spatial model checking for the analysis of medical images is relatively recent and there are only a few articles exploring this field so far. In work by Sundstrom et al. [49], spatio-temporal model checking techniques are used inspired by the work by Grosu et al. [50] – pursuing machine learning of the logical structure of image features – for the detection of tumours. In contrast, our approach is more focused on human-intelligible logical descriptions that can be reused and extended. Other interesting work is that by Pârvu and Gilbert [51], where spatio-temporal meta model checking is used for the analysis of biological processes, with a focus on *multi-scale* aspects. Two variants of spatial modalities have also been added to the Signal Temporal Logic [52,53] leading to the Signal Spatio-Temporal Logic (SSTL) [54]. In the first variant, a *bounded somewhere* operator is introduced that is reminiscent of the *somewhere* operator originally proposed by Reif and Sistla [55], while in the second one, a *bounded* surround operator is introduced that is inspired by the surround operator of the logic SLCS.

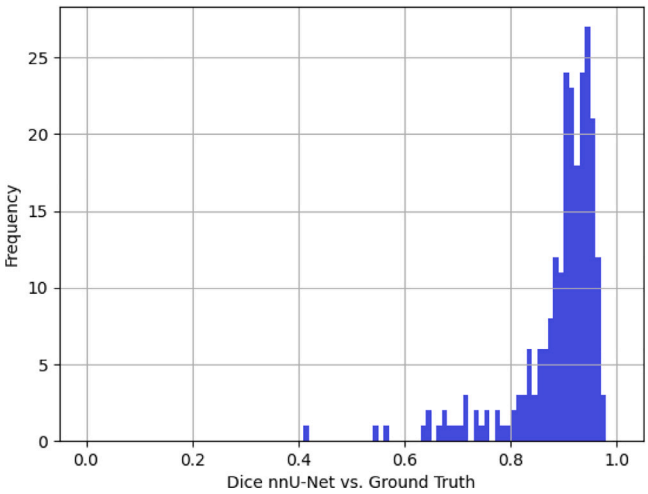


Fig. 13. Frequency distribution of Dice for nnU-Net segmentation w.r.t. ground truth. Mean 0.891 (StDev 0.082) and median 0.914.

In further work by Grosu et al. [56], a variant of spatial logic is proposed where spatial properties are expressed using quad trees. The authors show that very complex spatial structures can be identified with the support of model checking algorithms as well as machine learning procedures. However, the formulation of spatial properties becomes rather complex. The combination of this spatial logic with linear time signal temporal logic, defined with respect to continuous-valued signals, has recently led to the spatio-temporal logic SpaTeL [57].

There has been interesting work in the context of the verification of clinical guidelines in which temporal logic model checking (i.e. *not spatial* model checking) has been adopted (see for example the work by Groot at al. [58], and Bottrighi et al. [59]). Clinical guidelines play an important role in the medical area as a means to specify, in a precise way, the clinical procedures that are best suited to guarantee the quality of medical assistance or the optimal medical treatment path for a given case. In the work by Bottrighi et al. [59] the focus is on the integration of a computerised guideline management system and a model checker for a linear time logic to verify important properties of clinical guidelines. In particular, they show that such automatic verification approach is able to detect possible inconsistencies in the guidelines. In the work by Groot et al. [58], instead, the focus is on critiquing medical guidelines against patient data, i.e. to identify and analyse differences between the proposed treatments by a medical doctor and a set of 'ideal' actions as prescribed by (computerised) guidelines.

Many fully automated approaches for the segmentation of brain tissues and lesions that recently gained interest are based on machine learning and, in particular, *deep learning* (see for example a recent survey by Akkus et al. [32] and work by Xia et al. [46,47]). The latter work is based on self-supervised learning and semi-supervised

medical image segmentation of CT scans. The work in [47] exploits the enlarged discrepancies at the feature level to train differential decoders and then learn from these differential features in an iterative way. The method has been successfully applied to eight single organ/tumour segmentation sets and compared to eight state-of-the-art learning methods through a five-fold cross validation outperforming the other methods in most cases. The work in [46] concerns a self-supervised learner, CADS, based on cross-modal alignment and deep self-distillation. This approach enhances the ability of an encoder to characterise 3D CT scan volumes. The approach has shown to outperform some of the best state-of-the-art self-supervised learning methods and medical image segmentation methods such as nnUNet. Both works provide very promising results that show increasing progress in the accuracy of learning based segmentation. It would be very interesting to investigate whether a hybrid approach in this case could further improve accuracy in the sense that the logic procedure could delineate the rough borders within which the learning procedure is supposed to find a more accurate segmentation.

Although manual segmentation is still the standard for *in vivo* images, this method is expensive and time-consuming, difficult to reproduce and possibly inaccurate due to human error. In areas where large, reliable datasets are available, machine learning and deep learning approaches have shown promising results in pattern recognition. Deep learning is based on the use of artificial neural networks, consisting of several layers, that can extract a hierarchy of features from raw input data. These methods often depend heavily on the availability of large training datasets.

Furthermore, some machine learning approaches require the generation of manual *ground truth* labels, i.e. data sets in which segments of interest are indicated by experts manually in a standard way. This is a complicated task not only because it is very laborious, but also because of the relatively high intra-expert and inter-expert variability of $20 \pm 15\%$ and $28 \pm 12\%$, respectively, for manual segmentations of brain tumour images [60]. Nevertheless, much of the recent research in medical imaging and segmentation tasks in particular, has focussed on these (probabilistic) learning algorithms (see for example [31,39]). Therefore, interactive approaches based on spatial model checking may also be of help to improve the generation of manual ground truth labels in a more efficient, transparent and reproducible way. Furthermore, spatial model checking may be used as a method to monitor the quality of the results of deep learning approaches and signal cases in which one of the two approaches produce completely different results, which may indicate cases that need further, manual inspection.

7. Conclusions and future work

In this work we presented a novel symbolic and hybrid neuro-symbolic AI approach to the development of explainable, declarative segmentation procedures for glioblastoma and normal brain tissues in 3D brain MRI. The approach makes use of an (extendable) spatial logic, `ImgQL`, together with a spatial model checking procedure that is implemented in the model checker `VoxLogicA`. The spatial models are founded on the theory of closure spaces, a well-known extension of topological spaces that includes discrete spatial structures such as general graphs and images.

The symbolic approach has been validated on two public medical imaging benchmarks of 3D brain MRI scans, namely the Brats 2017, 2019 and 2020 datasets and the BrainWeb dataset, and compared with the state-of-the-art. This validation consisted in the development of three spatial logic specifications. One for the segmentation of high-grade glioblastoma in 3D MRI brain images based on a small subset of the Brats 2017 training dataset, one for the segmentation of white matter and one for the segmentation of grey matter in 3D synthetic brain images of the BrainWeb dataset. In all three cases the results have been compared with the ground truth segmentations that were supplied as part of the above mentioned datasets. The results are very promising, reaching average Dice similarity scores that are in line with those found using other state-of-the-art methods for what concerns the case study on brain tumour segmentation, whereas the results are excellent for the BrainWeb dataset case study. The additional strengths of our proposed spatial logic based symbolic contouring methods are that each `ImgQL` specification is very concise and amenable to human understanding and rapid, incremental prototyping. Furthermore, no learning phase is needed in the symbolic approach. All intermediate results and the final results of the segmentation procedures can be easily visualised as overlays on the medical images and readily inspected by domain experts. The asymptotic computational complexity of the approach is also interesting, being essentially linear in the number of voxels of the image, and taking only several seconds per 3D patient image.

A further innovative step to improve the accuracy of the spatial logic based symbolic method has been to combine it with sub-symbolic nnU-Net based segmentation. The latter essentially serves to produce a preliminary ground truth to find optimal thresholds used in the logic based segmentation procedure. The strength of this novel hybrid approach, that exploits an intrinsically redundant technique, has been shown to have a significantly better accuracy, in terms of Dice score, with respect to a purely symbolic approach, and, at the same time, it provides a human understandable explanation of the segmentation at a level of abstraction that is close to that used by domain experts in their contouring work. The latter is essential to enhance the accountability of, and the confidence in, the contouring methods. Moreover, it enables domain experts to show that the contouring adheres to established contouring guidelines. We remark that lack of accountability is likely the major reason why sub-symbolic methods alone are not yet widely adopted in the clinical setting.

Limitations. Although the proposed symbolic and hybrid methods are very promising, there are also a number of current limitations to the approach. The presented `ImgQL` specifications have been designed for the segmentation of MRI images showing single focus HGG. Images with multi-focal tumours and LGG have not been considered. Also images with clearly distinguishable artifacts in the acquisition have been excluded. Development of further `ImgQL` specifications to treat a wider range of images is ongoing. The `ImgQL` specifications presented in the current work have been validated on public datasets. Further validation on real-world clinical datasets is needed to provide further evidence of the robustness and generalisability of the approaches.

The presented `ImgQL` specifications depend on two thresholds, one for very intense and one for hyper intense pixels. Finding a good estimate of these values can be challenging. However, we have shown that a well-chosen fixed value based on expert human estimation gives acceptable results for a large set of images. Furthermore, we have shown that an estimate for these thresholds based on a nnU-net network trained on 50 HGG images also provides a good starting point for an accurate segmentation of individual images using this hybrid method. Furthermore, the thresholds could be further manually tuned by domain experts on a per-image base, which actually would give them more control over the segmentation. Further research is needed to investigate in more detail the potential impact on the results due to averse imaging conditions or scanner settings that may occur in a clinical setting, and establish minimal quality criteria for MRI images that are compatible with the proposed segmentation method.

Future work. Future work is planned in several directions. The spatial model checking approach lends itself particularly well for splitting the work in several parts that can be efficiently handled using GPU programming. Preliminary results can be found in Bussi et al. [61]. So far we have investigated a limited number of brain tissues. We are planning to apply the technique also on other brain tissues such as cerebral spine fluid or blood vessels. Such analysis might require the use of multiple co-registered MR images of the same patient contemporarily. This is already possible with the `VoxLogicA` model checker. In the longer term, clinical trials on real-world data are also planned. Limited steps

in that direction have been taken for the segmentation of glioblastoma, but real clinical data are often not publicly available and is therefore less suitable for the publication of larger scale comparative studies. We are also investigating the application of the approach in other medical domains, for example for the segmentation of nevi [62], and to develop a suitable graphical user interface for the VoxLogicA to facilitate its use by domain experts [63]. Finally, we are developing a novel model checking technique for – polyhedra models of – continuous space [64].

Spatial model technique could also be of use for checking clinical guidelines [58,59]. We envision that spatial model checking could enhance that work as it enables checking the delineations produced manually or by future automatic means such as Machine Learning, on which the diagnosis and treatment plans are based. Furthermore, the hybrid AI approach can be enriched by meaningful human control as further fine-tuning of the thresholds based on visual inspection of the image is a feasible and easy to perform option.

To further enhance the hybrid setup that we have introduced, and to apply similar approaches to further case studies, the model checker VoxLogicA will evolve into a multi-language microservice-based system that will encompass neural network *training* and *prediction* primitives of the language used in VoxLogicA, alongside logical operators [9], in order to guarantee not only reproducibility but also executable documentation of the *provenance* of the results. The combination of such ongoing development with a newly introduced user interface [48], inspired by principles of *cognitive load minimisation*, will lead to a first-of-its-kind human-centric approach to neuro-symbolic computation. The major advantages of this approach rely on the interaction between specific, example-based, black-box type networks with explainable, concise and unambiguous procedures that express the main part of the computation in a human-readable form. In particular the use of spatial logics for this part aims at encoding of domain expertise. Symbolic domain knowledge in such a system can be used both to augment the training dataset and to combine black-box predictions into higher level, symbolic procedures, leading to a human-centric simple, declarative approach to hybrid AI.

CRediT authorship contribution statement

Gina Belmonte: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Vincenzo Ciancia:** Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Mieke Massink:** Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Isensee F, Jäger PF, Full PM, Vollmuth P, Maier-Hein KH. nnU-Net for brain tumor segmentation. In: Crimi A, Bakas S, editors. Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries. Cham: Springer International Publishing; 2021, p. 118–32.

[2] Lemieux L, Hagemann G, Krakow K, Woermann F. Fast, accurate, and reproducible automatic segmentation of the brain in T1-weighted volume MRI data. Magn Reson Med 1999;42(1):127–35.

[3] Despotović I, Goossens B, Philips W. MRI segmentation of the human brain: Challenges, methods, and applications. Comput Math Methods Med 2015;2015:1–23. http://dx.doi.org/10.1155/2015/450341.

[4] Dupont C, Betrouni N, Reyns, Vermandel M. On image segmentation methods applied to glioblastoma: State of art and new trends. IRBM 2016;37(3):131–43. http://dx.doi.org/10.1016/j.irbm.2015.12.004.

[5] Fyllingen E, Stensjøen A, Berntsen E, Solheim O, Reinertsen I. Glioblastoma segmentation: Comparison of three different software packages. In: Pham D, editor. PLoS One 2016;11(10):e0164891. http://dx.doi.org/10.1371/journal.pone.0164891.

[6] Simi V, Joseph J. Segmentation of glioblastoma multiforme from MR images – A comprehensive review. Egypt J Radiol Nucl Med 2015;46(4):1105–10. http://dx.doi.org/10.1016/j.ejrnm.2015.08.001.

[7] Zhu Y, Young G, Xue Z, Huang R, You H, Setayesh K, Hatabu H, Cao F, Wong S. Semi-automatic segmentation software for quantitative clinical brain glioblastoma evaluation. Academic Radiol 2012;19(8):977–85. http://dx.doi.org/10.1016/j.acra.2012.03.026.

[8] Garcez Ad, Lamb LC. Neurosymbolic AI: the 3rd wave. Artif Intell Rev 2023;56(11):12387–406. http://dx.doi.org/10.1007/s10462-023-10448-w.

[9] Belmonte G, Bussi L, Ciancia V, Latella D, Massink M. Towards Hybrid-AI in imaging using VoxLogicA. In: Margaria T, Steffen B, editors. Leveraging applications of formal methods, verification and validation. software engineering methodologies - 12th international symposium, ISoLA 2024, Crete, Greece, October 27-31, 2024, proceedings, part IV. Lecture notes in computer science, vol. 15222, Springer; 2024, p. 205–21. http://dx.doi.org/10.1007/978-3-031-75387-9_13.

[10] Kautz HA. The third AI summer: AAAI Robert S. Engelmore memorial lecture. AI Mag 2022;43(1):105–25. http://dx.doi.org/10.1002/aaai.12036, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/aaai.12036. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/aaai.12036.

[11] Niyazi M, Andratschke N, Bendszus M, Chalmers AJ, Erridge SC, Galldiks N, Lagerwaard FJ, Navarria P, Munck af Rosenschöld P, Ricardi U, van den Bent MJ, Weller M, Belka C, Minniti G. ESTRO-EANO guideline on target delineation and radiotherapy details for glioblastoma. Radiother Oncol 2023;184. http://dx.doi.org/10.1016/j.radonc.2023.109663.

[12] Kruser T, Bosch W, Badiyan S, Bovi J, Ghia A, Kim M, Solanki A, Sachdev S, Tsien C, Wang T, Mehta M, McMullen K. NRG brain tumor specialists consensus guidelines for glioblastoma contouring. J Neurooncol 2019;143(1):157–66. http://dx.doi.org/10.1007/s11060-019-03152-9, URL https://pubmed.ncbi.nlm.nih.gov/30888558/.

[13] Aiello M, Pratt-Hartmann I, van Benthem J. Handbook of spatial logics. Springer; 2007, http://dx.doi.org/10.1007/978-1-4020-5587-4.

[14] Cohn AG, Renz J. Qualitative spatial representation and reasoning. In: van Harmelen F, Lifschitz V, Porter BW, editors. Handbook of knowledge representation. Foundations of artificial intelligence, vol. 3, Elsevier; 2008, p. 551–96. http://dx.doi.org/10.1016/S1574-6526(07)03013-1.

[15] Galton A. The mereotopology of discrete space. In: Freksa C, David M, editors. Spatial information theory. cognitive and computational foundations of geographic information science. Lecture notes in computer science, vol. 1661, Springer Berlin Heidelberg; 1999, p. 251–66. http://dx.doi.org/10.1007/3-540-48384-5_17.

[16] Galton A. A generalized topological view of motion in discrete space. Theoret Comput Sci 2003;305(1–3):111–34. http://dx.doi.org/10.1016/S0304-3975(02)00701-6.

[17] Randell DA, Landini G, Galton A. Discrete mereotopology for spatial reasoning in automated histological image analysis. IEEE Trans Pattern Anal Mach Intell 2013;35(3):568–81. http://dx.doi.org/10.1109/TPAMI.2012.128.

[18] Galton A. Discrete mereotopology. In: Calosi C, Graziani P, editors. Mereology and the sciences: parts and wholes in the contemporary scientific context. Springer International Publishing; 2014, p. 293–321. http://dx.doi.org/10.1007/978-3-319-05356-1_11.

[19] Vardi MY. From church and prior to PSL. In: Grumberg O, Veith H, editors. 25 years of model checking - history, achievements, perspectives. Lecture notes in computer science, vol. 5000, Springer; 2008, p. 150–71. http://dx.doi.org/10.1007/978-3-540-69850-0_10.

[20] Clarke EM, Emerson EA, Sifakis J. Model checking: algorithmic verification and debugging. Commun ACM 2009;52(11):74–84. http://dx.doi.org/10.1145/1592761.1592781.

[21] Baier C, Katoen J. Principles of model checking. MIT Press; 2008.

[22] Clarke EM, Grumberg O, Peled D. Model checking. MIT Press; 2001, URL http://books.google.de/books?id=Nmc4wEaLXFEC.

[23] Ciancia V, Gilmore S, Latella D, Loreti M, Massink M. Data verification for collective adaptive systems: Spatial model-checking of vehicle location data. In: Eighth IEEE international conference on self-adaptive and self-organizing systems workshops. IEEE Computer Society; 2014, p. 32–7. http://dx.doi.org/10.1109/SASOW.2014.16.

[24] Ciancia V, Latella D, Massink M, Paškauskas R. Exploring spatio-temporal properties of bike-sharing systems. In: 2015 IEEE international conference on self-adaptive and self-organizing systems workshops, SASO workshops 2015, Cambridge, MA, USA, September 21-25, 2015. IEEE Computer Society; 2015, p. 74–9. http://dx.doi.org/10.1109/SASOW.2015.17.

[25] Ciancia V, Grilletti G, Latella D, Loreti M, Massink M. An experimental spatio-temporal model checker. In: Software engineering and formal methods - SEFM 2015 collocated workshops. Lecture notes in computer science, vol. 9509, Springer; 2015, p. 297–311. http://dx.doi.org/10.1007/978-3-662-49224-6_24.

[26] Ciancia V, Latella D, Massink M, Paškauskas R, Vandin A. A tool-chain for statistical spatio-temporal model checking of bike sharing systems. In: Margaria T, Steffen B, editors. Leveraging applications of formal methods, verification and validation: foundational techniques - 7th international symposium, ISoLA 2016, Imperial, Corfu, Greece, October 10-14, 2016, proceedings, part I. Lecture notes in computer science, vol. 9952, 2016, p. 657–73. http://dx.doi.org/10.1007/978-3-319-47166-2_46.

[27] Ciancia V, Latella D, Loreti M, Massink M. Specifying and verifying properties of space. In: Theoretical computer science - 8th IFIP TC 1/WG 2.2 international conference, TCS 2014, Rome, Italy, September 1-3, 2014. proceedings. Lecture notes in computer science, vol. 8705, Springer; 2014, p. 222–35. http://dx.doi.org/10.1007/978-3-662-44602-7_18.

[28] Ciancia V, Latella D, Loreti M, Massink M. Model checking spatial logics for closure spaces. Log Methods Comput Sci 2016;12(4). http://dx.doi.org/10.2168/LMCS-12(4:2)2016, URL http://lmcs.episciences.org/2067.

[29] Belmonte G, Ciancia V, Latella D, Massink M. VoxLogicA: A spatial model checker for declarative image analysis. In: Vojnar T, Zhang L, editors. Tools and algorithms for the construction and analysis of systems - 25th international conference, TACAS 2019, held as part of the European joint conferences on theory and practice of software, ETAPS 2019, Prague, Czech Republic, April 6-11, 2019, proceedings, part I. Lecture notes in computer science, vol. 11427, Springer; 2019, p. 281–98. http://dx.doi.org/10.1007/978-3-030-17462-0_16.

[30] Banci Buonamici F, Belmonte G, Ciancia V, Latella D, Massink M. Spatial logics and model checking for medical imaging. Int J Softw Tools Technol Transf 2020;22(2):195–217. http://dx.doi.org/10.1007/s10009-019-00511-9.

[31] Menze BH, et al. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans Med Imaging 2015;34(10):1993–2024. http://dx.doi.org/10.1109/TMI.2014.2377694.

[32] Akkus Z, Galimzianova A, Hoogi A, Rubin DL, Erickson BJ. Deep learning for brain MRI segmentation: State of the art and future directions. J Digit Imaging 2017;30:449–59. http://dx.doi.org/10.1007/s10278-017-9983-4.

[33] Aubert-Broche B, Griffin M, Pike G, Evans A, Collins D. Twenty new digital brain phantoms for creation of validation image data bases. IEEE Trans Med Imaging 2006;25(11):1410–6. http://dx.doi.org/10.1109/TMI.2006.883453.

[34] Belmonte G, Ciancia V, Latella D, Massink M. Innovating medical image analysis via spatial logics. In: ter Beek MH, Fantechi A, Semini L, editors. From software engineering to formal methods and tools, and back - essays dedicated to stefania gnesi on the occasion of her 65th birthday. Lecture notes in computer science, vol. 11865, Springer; 2019, p. 85–109. http://dx.doi.org/10.1007/978-3-030-30985-5_7.

[35] Porikli FM. Integral histogram: a fast way to extract histograms in Cartesian spaces. In: 2005 IEEE computer society conference on computer vi- sion and pattern recognition. CVPR'05, vol. 1, 2005, p. 829–36, vol. 1.

[36] Haralick R, et al. Textural features for image classification. IEEE Trans Syst Man Cybern 1973;3(6):610–21.

[37] Bahadure NB, Ray AK, Thethi HP. Comparative approach of MRI-based brain tumor segmentation and classification using genetic algorithm. J Digit Imaging 2018;31(4):477–89. http://dx.doi.org/10.1007/s10278-018-0050-6, PMID: 29344753; PMCID: PMC6113145.

[38] Bakas S, Akbari H, Sotiras A, Bilello M, Rozycki M, Kirby JS, Freymann JB, Farahani K, Davatzikos C. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci Data 2017;4. http://dx.doi.org/10.1038/sdata.2017.117, Online publication date: 2017/09/05.

[39] Spyridon (Spyros) Bakas, et al., editors. 2017 international MICCAI BraTS challenge: pre-conference proceedings. 2017, URL https://www.cbica.upenn.edu/sbia/Spyridon.Bakas/MICCAI_BraTS/MICCAI_BraTS_2017_proceedings_shortPapers.pdf.

[40] Jiang Z, Ding C, Liu M, Tao D. Two-stage cascaded U-Net: 1st place solution to BraTS Challenge 2019 segmentation task. In: Crimi A, Bakas S, editors. Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries. Cham: Springer International Publishing; 2020, p. 231–41.

[41] Zhao Y-X, Zhang Y-M, Liu C-L. Bag of tricks for 3D MRI brain tumor segmentation. In: Crimi A, Bakas S, editors. Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries. Cham: Springer International Publishing; 2020, p. 210–20.

[42] McKinley R, Rebsamen M, Meier R, Wiest R. Triplanar ensemble of 3D-to-2D CNNs with label-uncertainty for brain tumor segmentation. In: Crimi A, Bakas S, editors. Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries. Cham: Springer International Publishing; 2020, p. 379–87.

[43] Jia H, Cai W, Huang H, Xia Y. H2NF-Net for brain tumor segmentation using multimodal MR imaging: 2nd place solution to BraTS Challenge 2020 segmentation task. In: Crimi A, Bakas S, editors. Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries. Cham: Springer International Publishing; 2021, p. 58–68.

[44] Wang Y, Zhang Y, Hou F, Liu Y, Tian J, Zhong C, Zhang Y, He Z. Modality-pairing learning for brain tumor segmentation. In: Crimi A, Bakas S, editors. Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries. Cham: Springer International Publishing; 2021, p. 230–40.

[45] Yuan Y. Automatic brain tumor segmentation with scale attention network. In: Crimi A, Bakas S, editors. Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries. Cham: Springer International Publishing; 2021, p. 285–94.

[46] Ye Y, Zhang J, Chen Z, Xia Y. CADS: A self-supervised learner via cross-modal alignment and deep self-distillation for CT volume segmentation. IEEE Trans Med Imaging 2025;44(1):118–29. http://dx.doi.org/10.1109/TMI.2024.3431916.

[47] Zeng Q, Xie Y, Lu Z, Lu M, Zhang J, Xia Y. Consistency-guided differential decoding for enhancing semi-supervised medical image segmentation. IEEE Trans Med Imaging 2025;44(1):44–56. http://dx.doi.org/10.1109/TMI.2024.3429340.

[48] Strippoli A. VoxLogicA UI: Supporting declarative medical image analysis [Master's thesis], University of Pisa; 2025, URL https://etd.adm.unipi.it/theses/available/etd-01052025-215748/.

[49] Sundstrom A, Grabocka E, Bar-Sagi D, Mishra B. Histological image processing features induce a quantitative characterization of chronic tumor hypoxia. PLoS One 2016;11(4):1–30. http://dx.doi.org/10.1371/journal.pone.0153623.

[50] Grosu R, Smolka S, Corradini F, Wasilewska A, Entcheva E, Bartocci E. Learning and detecting emergent behavior in networks of cardiac myocytes. Commun ACM 2009;52(3):97–105.

[51] Pârvu O, Gilbert D. A novel method to verify multilevel computational models of biological systems using multiscale spatio-temporal meta model checking. PLoS One 2016;11(5):1–43. http://dx.doi.org/10.1371/journal.pone.0154847.

[52] Donzé A, Ferrère T, Maler O. Efficient robust monitoring for STL. In: Sharygina N, Veith H, editors. Computer aided verification - 25th international conference, CAV 2013, Saint Petersburg, Russia, July 13-19, 2013. Proceedings. Lecture notes in computer science, vol. 8044, Springer; 2013, p. 264–79. http://dx.doi.org/10.1007/978-3-642-39799-8_19.

[53] Maler O, Nickovic D. Monitoring temporal properties of continuous signals. In: Lakhnech Y, Yovine S, editors. Formal techniques, modelling and analysis of timed and fault-tolerant systems, joint international conferences on formal modelling and analysis of timed systems, FORMATS 2004 and formal techniques in real-time and fault-tolerant systems, FTRTFT 2004, Grenoble, France, September 22-24, 2004, proceedings. Lecture notes in computer science, vol. 3253, Springer; 2004, p. 152–66. http://dx.doi.org/10.1007/978-3-540-30206-3_12.

[54] Nenzi L, Bortolussi L, Ciancia V, Loreti M, Massink M. Qualitative and quantitative monitoring of spatio-temporal properties with SSTL. Log Methods Comput Sci 2018;14(4):1–38. http://dx.doi.org/10.23638/LMCS-14(4:2)2018, Published on line: 23 Oct. 2018. ISSN: 1860-5974.

[55] Reif J, Sistla A. A multiprocess network logic with temporal and spatial modalities. J Comput System Sci 1985;30(1):41–53. http://dx.doi.org/10.1016/0022-0000(85)90003-0.

[56] Grosu R, Bartocci E, Corradini F, Entcheva E, Smolka S, Wasilewska A. Learning and detecting emergent behavior in networks of cardiac myocytes. In: Egerstedt M, Mishra B, editors. Hybrid systems: computation and control. HSCC 2008, Lecture notes in computer science, vol. 4981, Springer; 2008, p. 229–43. http://dx.doi.org/10.1007/978-3-540-78929-1_17.

[57] Haghighi I, Jones A, Kong Z, Bartocci E, Grosu R, Belta C. SpaTeL: A novel spatial-temporal logic and its applications to networked systems. In: Proceedings of the 18th international conference on hybrid systems: computation and control. HSCC '15, New York, NY, USA: ACM; 2015, p. 189–98. http://dx.doi.org/10.1145/2728606.2728633.

[58] Groot P, Hommersom A, Lucas PJF, Merk R, ten Teije A, van Harmelen F, Serban R. Using model checking for critiquing based on clinical guidelines. Artif Intell Med 2009;46(1):19–36. http://dx.doi.org/10.1016/j.artmed.2008.07.007.

[59] Bottrighi A, Giordano L, Molino G, Montani S, Terenziani P, Torchio M. Adopting model checking techniques for clinical guidelines verification. Artif Intell Med 2010;48(1):1–19. http://dx.doi.org/10.1016/j.artmed.2009.09.003.

[60] Mazzara G, Velthuizen R, Pearlman J, Greenberg H, Wagner H. Brain tumor target volume determination for radiation treatment planning through automated MRI segmentation. Int J Radiat Oncol Biol Phys 2004;59(1):300–12.

[61] Bussi L, Ciancia V, Gadducci F. Towards a spatial model checker on GPU. In: Peters K, Willemse TAC, editors. Formal techniques for distributed objects, components, and systems - 41st IFIP WG 6.1 international conference, FORTE 2021, held as part of the 16th international federated conference on distributed computing techniques, DisCoTec 2021, Valletta, Malta, June 14-18, 2021, proceedings. Lecture notes in computer science, vol. 12719, Springer; 2021, p. 188–96. http://dx.doi.org/10.1007/978-3-030-78089-0_12.

[62] Belmonte G, Broccia G, Ciancia V, Latella D, Massink M. Feasibility of spatial model checking for nevus segmentation. In: 9th IEEE/ACM international conference on formal methods in software engineering, FormaliSE@iCSE 2021, Madrid, Spain, May 17-21, 2021. IEEE; 2021, p. 1–12. http://dx.doi.org/10.1109/FormaliSE52586.2021.00007.

[63] Ciancia V, Belmonte G, Latella D, Massink M. A hands-on introduction to spatial model checking using VoxLogicA - - invited contribution. In: Laarman A, Sokolova A, editors. Model checking software - 27th international symposium, SPIN 2021, Virtual Event, July 12, 2021, proceedings. Lecture notes in computer science, vol. 12864, Springer; 2021, p. 22–41. http://dx.doi.org/10.1007/978-3-030-84629-9_2.

[64] Bezhanishvili N, Ciancia V, Gabelaia D, Grilletti G, Latella D, Massink M. Geometric model checking of continuous space. Log Methods Comput Sci 2022;18(4):7:1–38. http://dx.doi.org/10.46298/LMCS-18(4:7)2022, URL https://lmcs.episciences.org/10348. Published on line: Nov 22, 2022. ISSN: 1860-5974.