Adaptive Unsupervised Learning with Enhanced Feature Representation for Intra-tumor Partitioning and Survival Prediction for Glioblastoma

Yifan Li[1], Chao Li[2], Yiran Wei[2], Stephen Price[2], Carola-Bibiane Schönlieb[3], and Xi Chen[1,4(✉)]

[1] Department of Computer Science, University of Bath, Bath, UK
{yl3548,xc841}@bath.ac.uk
[2] Division of Neurosurgery, Department of Clinical Neurosciences, University of Cambridge, Cambridge, UK
{cl647,yw500,sjp58}@cam.ac.uk
[3] Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK
cbs31@cam.ac.uk
[4] Department of Physics, University of Cambridge, Cambridge, UK
xc253@mrao.cam.ac.uk

Abstract. Glioblastoma is profoundly heterogeneous in regional microstructure and vasculature. Characterizing the spatial heterogeneity of glioblastoma could lead to more precise treatment. With unsupervised learning techniques, glioblastoma MRI-derived radiomic features have been widely utilized for tumor sub-region segmentation and survival prediction. However, the reliability of algorithm outcomes is often challenged by both ambiguous intermediate process and instability introduced by the randomness of clustering algorithms, especially for data from heterogeneous patients.

In this paper, we propose an adaptive unsupervised learning approach for efficient MRI intra-tumor partitioning and glioblastoma survival prediction. A novel and problem-specific Feature-enhanced Auto-Encoder (FAE) is developed to enhance the representation of pairwise clinical modalities and therefore improve clustering stability of unsupervised learning algorithms such as K-means. Moreover, the entire process is modelled by the Bayesian optimization (BO) technique with a custom loss function that the hyper-parameters can be adaptively optimized in a reasonably few steps. The results demonstrate that the proposed approach can produce robust and clinically relevant MRI sub-regions and statistically significant survival predictions.

Keywords: Glioblastoma · MRI · Auto-encoder · K-means clustering · Bayesian optimization · Survival prediction

C. Li—Equal contribution.

A. Crimi and S. Bakas (Eds.): BrainLes 2021, LNCS 12962, pp. 124–139, 2022.
https://doi.org/10.1007/978-3-031-08999-2_10

1 Introduction

Glioblastoma is one of the most aggressive adult brain tumors characterized by heterogeneous tissue microstructure and vasculature. Previous research has shown that multiple sub-regions (also known as tumor habitats) co-exist within the tumor, which gives rise to the disparities in tumor composition among patients and may lead to different patient treatment response [9,10]. Regional differences within the tumour are often seen on imaging and may have a prognostic significance [30]. The intra-tumor heterogeneity is near ubiquitous in malignant tumors and likely to reflects cancer evolutionary dynamics [12,25]. Therefore, this intra-tumoral heterogeneity has significantly challenged the precise treatment of patients. Clinicians desire a more accurate identification of intra-tumoral invasive sub-regions for targeted therapy.

Magnetic resonance imaging (MRI) is a non-invasive technique for tumor diagnosis and monitoring. MRI radiomic features [22] provide quantitative information for both tumor partition and survival prediction [7,8]. Mounting evidence supports the usefulness of the radiomic approach in tumor characterization, evidenced by the Brain Tumor Image Segmentation (BraTS) challenge, which provides a large dataset of structural MRI sequences, i.e., T1-weighted, T2-weighted, post-contrast T1-weighted (T1C), and fluid attenuation inversion recovery (FLAIR). Although providing high tissue contrast, these weighted MRI sequences are limited by their non-specificity in reflecting tumor biology, where physiological MRIs, e.g., perfusion MRI (pMRI) and diffusion MRI (dMRI), could complement. Specifically, pMRI measures vascularity within the tumor, while dMRI estimates the brain tissue microstructure. Incorporating these complementary multi-modal MRI has emerged as a promising approach for more accurate tumor characterization and sub-region segmentation for clinical decision support.

Unsupervised learning methods have been widely leveraged to identify the intra-tumoral sub-regions based on multi-modal MRI [4,17,19,26,29,31]. Standard unsupervised learning methods, e.g., K-means, require a pre-defined class number, which lacks concrete determination criteria, affecting the robustness of sub-region identification. For instance, some researchers used pre-defined class numbers according to empirical experience before clustering [4,17]. Some other work [14,31] introduced clustering metrics, e.g., the Calinski-Harabasz (CH) index, which quantifies the quality of clustering outcomes to estimate the ideal class number. However, the CH index is sensitive to data scale [14,31], limiting its generalization ability across datasets. Some other clustering techniques, e.g., agglomerative clustering, do not require a pre-defined class number and instead require manual classification. A sensitivity hyper-parameter, however, is often needed *a priori*. The clustering results can be unstable during iterations and across datasets. Due to the above limitations, the generalization ability of clustering methods has been a significant challenge in clinical applications, particularly when dealing with heterogeneous clinical data.

Further, the relevance of clustering results is often assessed using patient survival in clinical studies [2,6,11,17]. However, existing research seldom addressed

the potential influence of instability posed by the unsupervised clustering algorithms. Joint hyper-parameter optimization considering both clustering stability and survival relevance is desirable in tumor sub-region partitioning.

In this paper, we propose a variant of auto-encoder (AE), termed Feature-enhanced Auto-Encoder (FAE), to identify robust latent feature space constituted by the multiple input MRI modalities and thus alleviate the impact brought by the heterogeneous clinical data. Additionally, we present a Bayesian optimization (BO) framework [24] to undertake the joint optimization task in conjunction with a tailored loss function, which ensures clinical relevance while boosting clustering stability. As a non-parametric optimization technique based on Bayes' Theorem and Gaussian Processes (GP) [21], BO learns the representation of the underlying data distribution that the most probable candidate of the hyper-parameters is generated for evaluation in each step. Here, BO is leveraged to identify the (sub)optimal hyper-parameter set with the potential to effectively identify robust and clinically relevant tumor sub-regions. The primary contributions of this work include:

- Developing a novel loss function that balances the stability of sub-region segmentation and the performance of survival prediction.
- Developing an FAE architecture in the context of glioblastoma studies to further enhance individual clinical relevance between input clinical features and improve the robustness of clustering algorithms.
- Integrating a BO framework that enables automatic hyper-parameter search, which significantly reduces the computational cost and provides robust and clinically relevant results.

The remainder of this paper is organized as follows. Section 2 describes the overall study design, the proposed framework, and techniques. Section 3 reports numerical results, and Sect. 4 is the concluding remarks.

2 Problem Formulation and Methodology

Consider an N patients multi-modal MRI dataset $\boldsymbol{\Omega}$ with M modalities defined as $\{\mathbf{X}_m\}_{m=1}^{M}$. $\mathbf{X}_m$ denotes the mth (pixel-wise) modality values over a collection of N patients. $\mathbf{X}_m = \{\mathbf{x}_{m,n}\}_{n=1}^{N}$, where $\mathbf{x}_{m,n} \in \mathbb{R}^{I_{m,n} \times 1}$ and $I_{m,n}$ denotes total pixel number of an individual MRI image for the mth modality of the nth patient.

Our goal is to conduct sub-region segmentation on MRI images and perform clinically explainable survival analysis. Instead of running unsupervised learning algorithms directly on $\mathbf{X}_m$, we introduce an extra latent feature enhancement scheme (termed FAE) prior to the unsupervised learning step to further improve the efficiency and robustness of clustering algorithms.

As shown in Fig. 1(A), FAE aims to produce a set of latent features $\{\mathbf{Z}_{m'}\}_{m'=1}^{M}$ that represent the original data $\{\mathbf{X}_m\}_{m=1}^{M}$. Unlike a standard AE that takes all modalities as input, FAE 'highlights' pairwise common features and produces $\mathbf{Z}$ through a set of encoders (denoted as E) and decoders (denoted

as D). The latent features are then used in unsupervised clustering to classify tumor sub-region $\{\mathbf{P}_n\}_{n=1}^{N}$ for all patients. As an intermediate step, we can now produce spatial features $\{\mathbf{F}_n\}_{n=1}^{N}$ from the segmented figures through radiomic spatial feature extraction methods such as gray level co-occurrence matrix (GLCM) and Gray Level Run Length Matrix (GLRLM) [15].

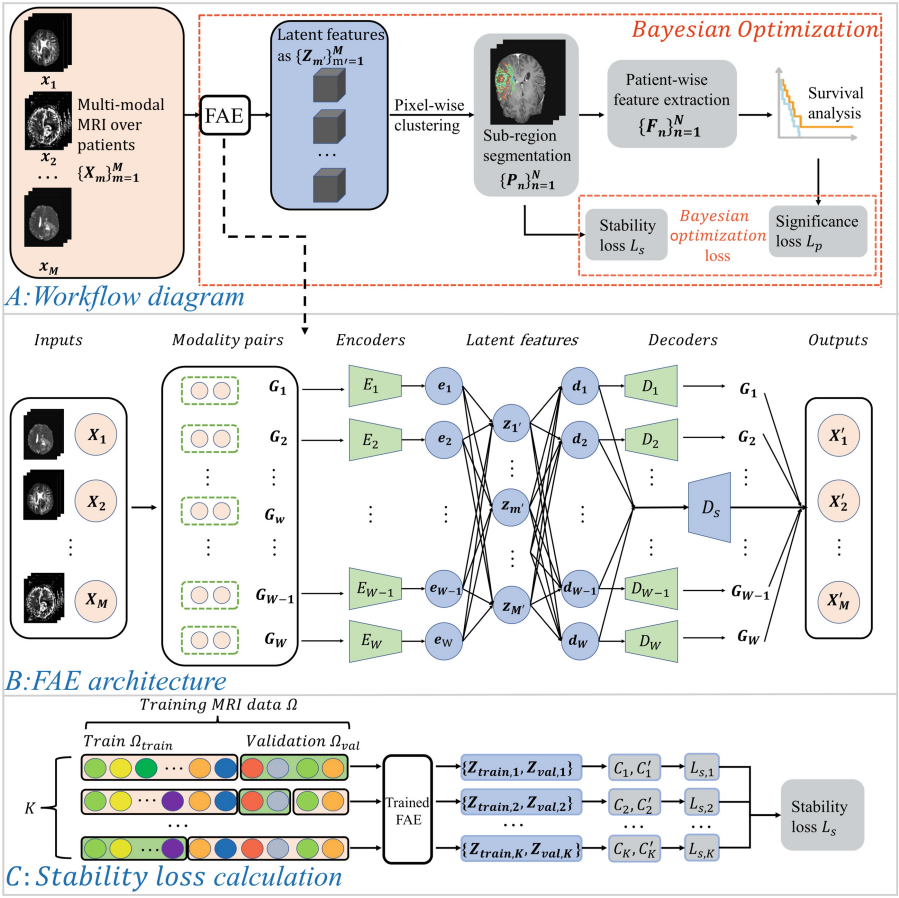


Fig. 1. A: Workflow of the proposed approach. The entire process is modelled under a Bayesian optimization framework. B: Architecture of FAE. The light orange circle represents modality $\mathbf{X}_m$ overall patients and the blue circle is the latent feature $\mathbf{Z}_{m'}$. The green dotted frame denotes the modality pair, and the green trapezoid represents feature-enhanced encoder E and decoder D. The blue trapezoid indicates the fully connected decoders D_s. C: Illustration of stability loss calculation. Circles in different colours represent individual patient MRI data, which are then randomly shuffled for K times to split into train/validation sets. (Color figure online)

2.1 Feature-Enhanced Auto-Encoder

FAE is developed on Auto-encoder (AE), a type of artificial neural network used for dimensionality reduction. A standard AE is a 3-layer symmetric network that has the same inputs and outputs. As illustrated in Fig. 1(B), FAE contains W feature-enhanced encoder layers $\{E_w\}_{w=1}^{W}$ to deal with $\{\mathbf{G}_w\}_{w=1}^{W}$ pairs of modalities, where $W = \binom{M}{2}$ pairs of modalities (from combination) given M inputs. The wth encoder takes a pair of modalities from $\{\mathbf{X}_m\}_{m=1}^{M}$ and encodes to a representation $\mathbf{e}_w$. The central hidden layer of FAE contains $\{\mathbf{Z}_{m'}\}_{m'=1}^{M}$ nodes that represents M learnt abstract features. FAE also possesses a 'mirrored' architecture similar to AE, where W feature-enhanced decoder layers $\{D_w\}_{w=1}^{W}$ are connected to the decoded representations $\{d_w\}_{w=1}^{W}$.

Unlike the standard symmetric AE, FAE has a 'dual decoding' architecture that an extra fully-connected decoder layer D_s is added to the decoding half of the networks to connect $\{d_w\}_{w=1}^{W}$ directly to the outputs $\{\mathbf{X'}_m\}_{m=1}^{M}$. Decoder D_s aims to pass all outputs information (and correlations) rather than the pairwise information from $\mathbf{G}_w$ in the back-propagation process. As a result, node weights $\{\mathbf{Z}_{m'}\}_{m'=1}^{M}$ are updated by gradients from both $\{D_w\}_{w=1}^{W}$ and D_s. In practice, $\mathbf{Z}$ and the encoders are iteratively amended by $\{D_w\}_{w=1}^{W}$ (i.e., reconstruction loss from pairwise AEs) and D_s (i.e., global reconstruction loss) in turns.

FAE enhances the latent features in every pair of input modalities before reducing the dimensionality from W to M. For instance, $\mathbf{e}_w$ is a unique representation that only depends on (and thus enhances the information of) the given input pair $\mathbf{G}_w$. Under this dual decoding architecture, FAE takes advantage of highlighting the pairwise information in $\{\mathbf{Z}_{m'}\}_{m'=1}^{M}$ while retaining the global correlation information from D_s. Another advantage of FAE lies in its flexibility to the dimensionality of input features. The FAE presented in this paper always produces the same number of latent features as the input dimension. The latent dimension might be further reduced manually depending on computational/clinical needs.

2.2 Patient-Wise Feature Extraction and Survival Analysis

We implement Kaplan-Meier (KM) survival analysis [2,17] on spatial features and sub-region counts $\{\mathbf{F}_n\}_{n=1}^{N}$ to verify the relevance of clustering sub-regions. To characterize the intratumoral co-existing sub-regions, we employed the commonly used texture features from the GLCM and GLRLM families, i.e., Long Run Emphasis (LRE), Relative mutual information (RMI), Joint Energy, Run Variance (RV) and Non-Uniformity. These features are formulated to reflect the spatial heterogeneity of tumor sub-regions. For example, LRE indicates the prevalence of a large population of tumor sub-regions. The formulas and interpretations of all these features are detailed in [27]. We next use the k-medoids technique to classify N patients into high- and low-risk subgroups based on $\{\mathbf{F}_n\}_{n=1}^{N}$ and then perform KM analysis to analyze the survival significance of the subgroups to determine the L_p, as described in Sect. 2.4 and Eq. 2.

2.3 Constructing Problem-Specific Losses

Stability Loss. We first introduce a stability quantification scheme to evaluate clustering stability using pairwise cluster distance [13,28], which will serve as part of the loss function in hyper-parameter optimization. Specifically, we employ a Hamming distance method (see [28] for details) to quantify the gap between clustering models. We first split the MRI training dataset $\boldsymbol{\Omega}$ into train and validation sets, denoted as $\boldsymbol{\Omega}_{train}$ and $\boldsymbol{\Omega}_{val}$ respectively. We then train two clustering models C (based on $\boldsymbol{\Omega}_{train}$) and C' (based on $\boldsymbol{\Omega}_{val}$). The stability loss aims to measure the performance of model C on the unseen validation set $\boldsymbol{\Omega}_{val}$. The distance $d(\cdot)$ (also termed as L_s) is defined as:

$$L_s = d(C, C') = \min_{\pi} \frac{1}{I_{val}} \sum_{\boldsymbol{\Omega}_{val}} \mathbb{1}_{\{\pi(C(\boldsymbol{\Omega}_{val}) \neq C'(\boldsymbol{\Omega}_{val}))\}}, \tag{1}$$

where I_{val} denotes the total number of pixels over all MRI images in the validation set $\boldsymbol{\Omega}_{val}$. $\mathbb{1}$ represents the Dirac delta function [32] that returns 1 when the inequality condition is satisfied and 0 otherwise, and function $\pi(\cdot)$ denotes the repeated permutations of dataset $\boldsymbol{\Omega}$ to guarantee the generalization of the stability measure [28].

Figure 1 (C) shows the diagram for L_s calculation, where N patients are randomly shuffled for K times to mitigate the effect of randomness. K pairs of intermediate latent features $\{\mathbf{Z}_{train,k}, \mathbf{Z}_{val,k}\}_{k=1}^{K}$ are generated through FAE for training the clustering models C and C'. We then compute L_s over K repeated trials. L_s is normalized to range $[0, 1]$, and smaller values indicates more stable clusterings.

Significance Loss. We integrate prior knowledge from clinical survival analysis and develop a significance loss L_p to quantify clinical relevance between the clustering outcomes and patient survival, as demonstrated in the below equation:

$$L_p = \log(\frac{\tau}{p}) \tag{2}$$

where p represents p-value (i.e., statistical significance measure) of the log-rank test in the survival analysis and τ is a predefined threshold.

This follows the clinical practice that a lower p-value implies that the segmented tumor sub-regions can provide sensible differentiation for patient survival. In particular, given threshold τ, for p less than the threshold, the loss equation returns a increasing positive reward. Otherwise, for p greater than or equal to τ, the segmented tumor sub-regions are considered undesirable and the penalty increases with p.

2.4 Bayesian Optimization

Hyper-parameters tuning is computational expensive and often requires expert knowledge, both of which raise practical difficulties in clinical applications. In

this paper, we consider two undetermined hyper-parameters: a quantile threshold $\gamma \in [0, 1]$ that distinguishes outlier data points from the majority and cluster number η for the pixel-wise clustering algorithm. We treat the entire process of Fig. 1(A) as a *black-box system*, of which the input is the hyper-parameter set $\boldsymbol{\theta} = [\gamma, \eta]$ and the output is a joint loss $\mathcal{L}$ defined as:

$$\mathcal{L} = \alpha L_s + (1 - \alpha) L_p \tag{3}$$

where α is a coefficient that balances L_s and L_p and ranges between $[0, 1]$.

Algorithm 1: Bayesian optimization for hyper-parameter tuning

```
Initialization of GP surrogate f and the RBF kernel 𝒦(·)
while not converged do
    Fit GP surrogate model f with {θ_j, ℒ_j}_{j=1}^J
    Propose a most probable candidate θ_{j+1} through Equation (4)
    Run Algorithm 2 with θ_{j+1}, and compute loss ℒ_{j+1}
    Estimate current optimal θ_{j+2} of the constructed GP surrogate f'
    Run Algorithm 2 with θ_{j+2}, calculate the loss ℒ_{j+2}
    J = J + 2
end
Obtain (sub)optimal θ_* upon convergence
```

We address the hyper-parameter tuning issue by modelling the black-box system under BO, a sequential optimization technique that aims to approximate the search space contour of $\boldsymbol{\theta}$ by constructing a Gaussian Process (GP) surrogate function in light of data. BO adopts an *exploration-exploitation scheme* to search for the most probable $\boldsymbol{\theta}$ candidate and therefore minimize the surrogate function mapping $f : \Theta \rightarrow \mathcal{L}$ in J optimization steps, where Θ and $\mathcal{L}$ denote input and output spaces respectively. The GP surrogate is defined as: $f \sim \mathcal{GP}(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$; where $\boldsymbol{\mu}$ is the $J \times 1$ mean function vector and $\boldsymbol{\Sigma}$ is a $J \times J$ co-variance matrix composed by the pre-defined kernel function $\mathcal{K}(\cdot)$ over the inputs $\{\boldsymbol{\theta}_j\}_{j=1}^J$. In this paper, we adopt a standard radial basis function (RBF) kernel (see [3] for an overview of GP and the kernel functions).

Given training data $\boldsymbol{\Omega}_B = \{\boldsymbol{\theta}_j, \mathcal{L}_j\}_{j=1}^J$, BO introduces a so-called acquisition function $a(\cdot)$ to propose the most probable candidate to be evaluated at each step. Amongst various types of acquisition functions [24], we employ an EI strategy that seeks new candidates to maximize *expected improvement* over the current best sample. Specifically, suppose f' returns the best value so far, EI searches for a new $\boldsymbol{\theta}$ candidate that maximizes function $g(\boldsymbol{\theta}) = \max\{0, f' - f(\boldsymbol{\theta})\}$. The EI acquisition can thus be written as a function of $\boldsymbol{\theta}$:

$$a_{EI}(\boldsymbol{\theta}) = \mathbb{E}(g(\boldsymbol{\theta})|\boldsymbol{\Omega}_B) = (f' - \boldsymbol{\mu})\Phi(f'|\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \boldsymbol{\Sigma}\mathcal{N}(f'|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{4}$$

where $\Phi(\cdot)$ denotes CDF of the standard normal distribution. In practice, BO step J increases over time and the optimal θ_* can be obtained if the predefined convergence criteria is satisfied. Pseudo-code of the entire process is shown in both Algorithms 1 and Algorithm 2.

2.5 Experiment Details

Data from a total of $N = 117$ glioblastoma patients were collected and divided into training set $\boldsymbol{\Omega} = 82$ and test set $\boldsymbol{\Omega}_{test} = 35$, where the test set was separated for out-of-sample model evaluation. We collected both pMRI and dMRI data and co-registered them into T1C images (details in Appendix 5.1), containing approximately 11 million pixels per modality over all patients. $M = 3$ input modalities were calculated, including rCBV (denoted as $\mathbf{r}$) from pMRI, and isotropic/anisotropic components (denoted as $\mathbf{p}/\mathbf{q}$) of dMRI, thus $\mathbf{X} = \{\mathbf{p}, \mathbf{q}, \mathbf{r}\}$. Dataset $\boldsymbol{\Omega}$ was used for stability loss calculation with $\boldsymbol{\Omega}_{train} = 57$, $\boldsymbol{\Omega}_{val} = 25$. L_s was evaluated over $K = 10$ trials for all following experiments. The BO is initialized with $J = 10$ data points $\boldsymbol{\Omega}_B$, $\gamma \in [0, 1]$ and η is an integer ranges between 3 and 7. The models were developed on Pytorch platform [18] under Python 3.8. Both encoder E and decoder D employed a fully connected feed-forward NN with one hidden layer, where the hidden node number was set to 10. We adopted *hyperbolic tangent* as the activation function for all layers, *mean squared error (MSE)* as the loss function, and *Adam* as the optimiser.

Algorithm 2: Pseudo-code of the workflow as a component of BO

`// Initialization`
Prepare MRI data $\boldsymbol{\Omega}$ with N patients and M modalities, perform data filtering with quantile threshold γ
`// FAE training follows Figure 1(B)`
Compose W pairs of modalities $G_{w=1}^{W}$, where $W = \binom{M}{2}$
Train FAE on $\{\mathbf{X}_m\}_{m=1}^{M}$ to generate latent features $\{\mathbf{Z_m}'\}_{m'=1}^{W}$
`// Stability loss calculation follows Figure 1(C)`
for k =1,2,...,K **do**
    Randomly divide $\boldsymbol{\Omega}$ into train ($\boldsymbol{\Omega}_{train}$) and validation ($\boldsymbol{\Omega}_{val}$) sets
    Produce latent pairs $\{\mathbf{Z}_{train,k}, \mathbf{Z}_{val,k}\}_{k=1}^{K}$
    `// Pixel-wise clustering`
    Obtain C_k and C_k' through standard K-means with η clusters
    Compute kth stability loss $L_{s,k}$ by Eq (1)
end
Compute stability score L_s by averaging over $\{L_{s,k}\}_{k=1}^{K}$
`// Sub-region segmentation`
Obtain patient-wise sub-region segments $\{\mathbf{P}_n\}_{n=1}^{N}$
`// Patient-wise feature extraction`
Extract $\{\mathbf{F}_n\}_{n=1}^{N}$ for all N patients
`// Survival analysis`
Cluster patients into high/low risk subgroups based on $\{\mathbf{F}_n\}_{n=1}^{N}$ using a standard K-Medoids algorithm. Perform survival analysis and obtain p
`// BO loss calculation`
Compute clinical significance score L_p by Eq (2)
Compute joint loss L follows Eq (3)

3 Results and Discussions

We first present the clustering stability of the models incorporating FAE architecture, which contains 1 hidden layer with 10 hidden nodes. The hyper-parameter choice of FAE architecture, which is simple to be compared in numerical experiments, are determined by empirical experiences. Other AE variants against the baseline model and then compare the performance of the proposed methodology under different experimental settings. We finally demonstrate the results of survival analysis and independent test.

3.1 Evaluation of FAE Based Clustering

The results comparing the models are detailed in Table 1. One sees that all three AE variants show better stability performance than that of the baseline model in the varying cluster numbers. Of note, our proposed FAE architecture, which incorporates both standard AE and ensemble AE, outperforms other models in majority comparisons.

Table 1. Stability performance of cluster algorithms under different AE variants. Baseline represents the original model without AE. The standard AE represents a standard 3-layer (with 1 hidden layer) feed-forward network and the ensemble AE is the FAE without dual decoder D_s. The hidden layer contains 10 nodes for all AE variants.

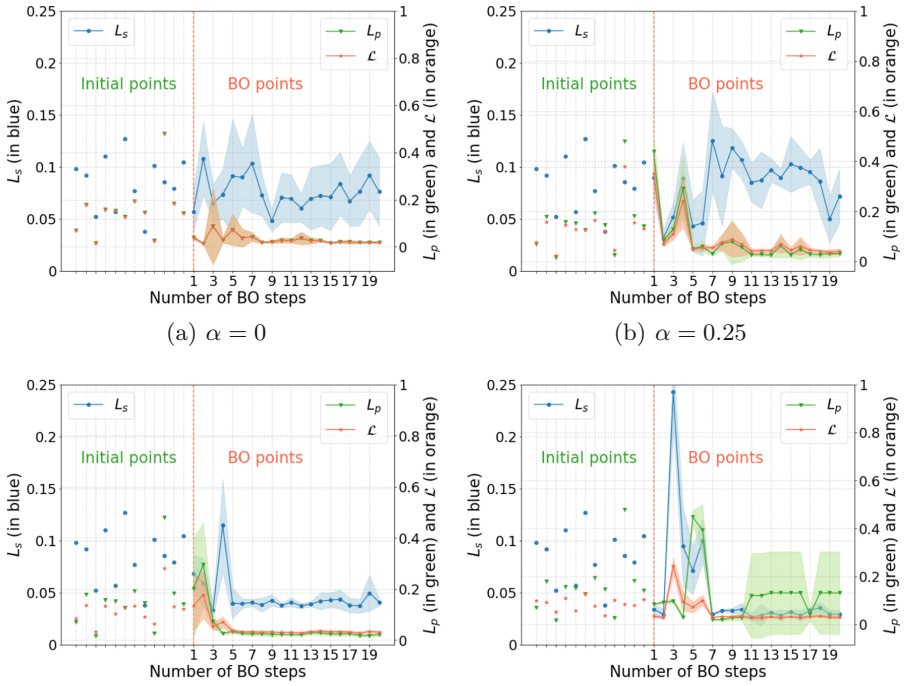
Clusters	3	4	5	6
Stability score				
Baseline	0.761±0.026	0.890±0.04	0.744±0.027	0.761±0.035
Standard AE	0.909±0.024	0.896±0.063	0.859±0.06	0.836±0.061
Ensemble AE	**0.972±0.013**	0.921±0.028	0.872±0.046	0.881±0.046
FAE	0.909±0.048	**0.923±0.029**	**0.911±0.038**	**0.891±0.048**
Calinski-Harabasz (CH) score				
Baseline (10^6)	4.12±0.00003	5.16±0.00013	4.82±0.00003	4.73±0.00009
Standard AE (10^6)	5.94±0.63	5.74±0.51	5.50±0.41	5.36±0.28
Ensemble AE (10^6)	10.43±0.67	10.99±0.52	10.98±0.89	11.09±1.00
FAE (10^6)	**13.85±4.45**	**14.85±4.49**	**15.09±4.19**	**15.34±4.14**

As expected, all AE variants enhance the clustering stability and quality, shown by the stability score and CH score. The latter is relatively sensitive to data scale but can provide reasonable evaluation for a fixed dataset. In our case, as the dimensions of the original input modalities and the latent features remain identical ($M = 3$), the considerably improved stability of the models

incorporating FAE architecture suggests the usefulness of the FAE in extracting robust features for the unsupervised clustering. Additionally, our experiments show that the FAE demonstrates remarkably stable performance in the clustering when the training data is randomly selected, which further supports the resilience of the FAE in extracting generalizable features for distance-based clustering algorithms.

3.2 Adaptive Hyper-parameter Tuning

Figure 2 shows the performance of the proposed approach in 4 different α values in terms of stability score (lower score value indicates better stability). 10 initial training steps and 20 follow-up BO steps are evaluated in the experiments, all the results are averaged over 10 repeated trials. One sees significant dispersion of initial points (dots in the left half of each figure) in all figures, indicating reasonable randomness of initial points in BO training. BO proposes a new



(a) $\alpha = 0$ (b) $\alpha = 0.25$

(c) $\alpha = 0.5$ (d) $\alpha = 1$

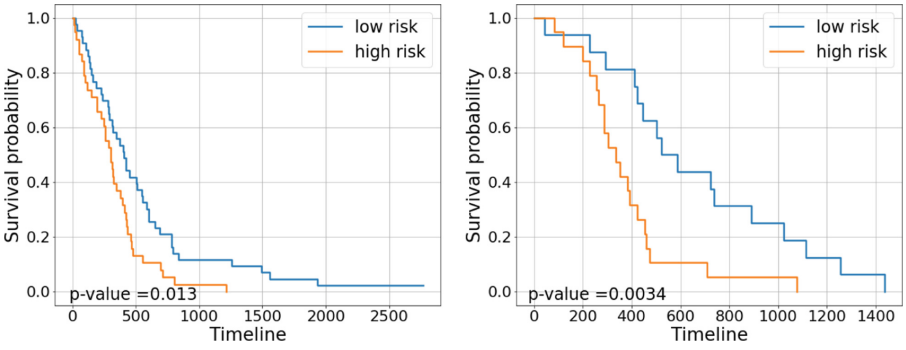
Fig. 2. Performance of the proposed approach with respect to BO step number (on x-axis). Each figure contains two y-axis: stability loss L_s (in blue) on the left y-axis, and both significant loss L_p (in green) and joint loss (in orange) on the right y-axis. All losses are normalized and the shadowed areas in different colors indicate error-bars of the corresponding curves. Figure (a)–(d) shows the performance with loss coefficient $\alpha = 0, 0.25, 0.5$ and 1, respectively. (Color figure online)

candidate $\boldsymbol{\theta}$ per step after the initial training. One observes that the joint loss $\mathcal{L}$ (orange curves) converges and the proposed approach successfully estimates (sub)optimal $\boldsymbol{\theta}_*$ in all α cases.

Figure 2(a) shows $\alpha = 0$ case, for which $\mathcal{L} = L_p$ according to Equation (3). In other words, the algorithm aims to optimize significance loss L_p (green curve) rather than stability loss L_s (blue curve). As a result, the orange and green curves overlap with each other, and the stability scores are clearly lower than that of L_s. A consistent trend can be observed across all four cases that the error-bar areas of L_s (blue shadowed areas) shrink as the weight of L_s increases in the joint loss. Similar observations can be seen in Fig. 2(d) where $\alpha = 1$ and $\mathcal{L} = L_s$, the error-bar area of L_p (green shadowed area) is considerably bigger than those in the rest α cases. Note that L_s and $\mathcal{L}$ also overlap with each other and the mismatch in the figure is caused by the differences of left and right y-axis scale. When $\alpha = 0.5$ (Fig. 2(c)), clustering stability can quickly converge in a few BO steps (around 6 steps in the orange curve), shows the advantage of the proposed BO integrated method in hyper-parameter optimization.

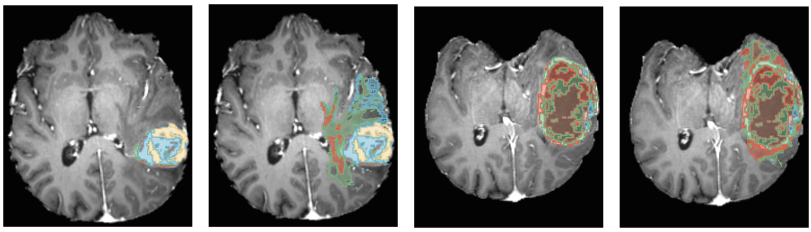
3.3 Statistical Analysis and Independent Test

Upon convergence of BO, we acquire well-trained FAE encoders to extract features from modalities, a well-trained clustering model for tumor sub-region segmentation and a population-level grouping model to divide patients into high-risk and low-risk subgroups. Eventually, we acquire 5 tumor sub-regions as $\{\mathbf{P}_n\}_{n=1}^N$ from features processed by the well-trained FAE, where $\mathbf{P}_n = \{\mathbf{p}_i\}_{i=1}^I$, $\mathbf{p}_i \in \{1, 2, 3, 4, 5\}$ denotes the sub-region labels for each pixel, and produce features $\{\mathbf{F}_n\}_{n=1}^N$, where $\mathbf{F}_n \in \mathbb{R}^{11\times 1}$ represents 9 spatial features and proportion of the 2 significant sub-regions, the details of clinical features could be found in Appendix 5.2. Subsequently, we apply these well-trained models to the test set with 35 patients. The results of KM analysis are shown in Fig. 3, illustrating that the spatial features extracted from tumor sub-regions could lead to patient-level clustering that successfully separates patients into distinct survival groups in both datasets (Train: p-value = 0.013 Test: p-value = 0.0034). Figure 4 shows two case examples from the high-risk and low-risk subgroups, respectively, where different colours indicate the partitioned sub-regions. Intuitively, these sub-regions are in line with the prior knowledge of proliferating, necrotic, and edema tumor areas, respectively.



(a) Train set $\boldsymbol{\Omega} = 82$ patients (b) Test set $\boldsymbol{\Omega}_{test} = 35$ patients

Fig. 3. KM survival curves for the train and test datasets.



(a) low-risk (CE) (b) low-risk (NE) (c) high-risk (CE) (d) high-risk (NE)

Fig. 4. Two case examples from the high-risk (a & b) and lower-risk (c & d) group, respectively. Different colours denote the partitioned sub-regions. The two patients have significantly different proportions of sub-regions with clinical relevance, which could provide clinical decision support. (Color figure online)

4 Conclusions

The paper is an interdisciplinary work that helps clinical research to acquire robust and effective sub-regions of glioblastoma for clinical decision support. The proposed FAE architectures significantly enhance the robustness of the clustering model and improve the quality of clustering results. Additionally, robust and reliable clustering solutions can be accomplished with minimal time investment by integrating the entire process inside a BO framework and presenting a unique loss function for problem-specific multi-task optimization. Finally, the independent validation of our methodology using a different dataset strengthens its viability in clinical applications.

Although we have conducted numerous repeating trials, it is inevitable to eliminate the randomness for clustering algorithm experiments. In future work, we could include more modalities and datasets to test the framework. To enhance the clinical relevance, more clinical variables could be included into the BO framework for multi-task optimization. To summarise, the BO framework combined with the suggested FAE and mixed loss represents a robust framework for obtaining clustering results that are clinically relevant and generalizable across datasets.

5 Appendix

5.1 Details of Dataset and Imagine Processing

Patients with surgical resection (July 2010–August 2015) were consecutively recruited, with data prospectively collected by the multidisciplinary team (MDT) central review. All glioblastoma patients underwent pre-operative 3D MPRAGE (pre-contrast T1 and T1C), T2-weighted FLAIR, pMRI and dMRI sequences. All patients have a radiogical diagnosis of de novo glioblastoma, aged 18 to 75, eligible for craniotomy and radiotherapy, and all images resolution were resampled to $1 \times 1 \times 1\,\mathrm{m}^3$.

Co-registration of the images was accomplished using the linear registration tool (FLIRT) included in the Oxford Centre for Functional MRI of the Brain Software Library (FSL) v5.0.0 (Oxford, UK) [5,23]. NordicICE was used to process dynamic susceptibility contrast (DSC), one of the most frequently utilised perfusion methods (NordicNeuroLab). The arterial input function was automatically defined. The diffusion toolbox in FSL was used to process the diffusion images (DTI) [1]. The isotropic (p) and anisotropic (q) components were computed after normalisation and eddy current correction [20].

5.2 Details for Clinical Features

In this study, through the BO, the tumor were divided into 5 sub-regions as $\{\mathbf{P}_n\}_{n=1}^{N}$ from $\{\mathbf{Z}_{m'}\}_{m'=1}^{M}$, the features processed by the well-trained FAE, where $\mathbf{P}_n = \{\mathbf{p}_i\}_{i=1}^{I}$, $\mathbf{p}_i \in \{1, 2, 3, 4, 5\}$ denotes the sub-region labels for each pixel. Rather than representing the numerical grey value of images, the value of each $\mathbf{p}_i$ represents sub-region labels, rendering the majority of features in the GLCM and GLRLM families invalid. Finally, the Table 2 summarises the selected features which remain meaningful for the label matrix. Eventually, the clinical features $\{\mathbf{F}_n\}_{n=1}^{N}$, where $\mathbf{F}_n \in \mathbb{R}^{11\times 1}$ include 9 spatial characteristics in Table 2 and the fraction of 2 significant sub-regions.

Table 2. Clinical features from GLCM matrix of size $N_g \times N_g$ and GLRLM matrix of size $N_g \times N_r$ family including Relative mutual information(RMI), Entropy, Joint Energy, Informational Measure of Correlation(IMC), Long Run Emphasis(LRE), Short Run Emphasis(SRE), Run Variance(RV) and Run Entropy(RE). $p(i,j|\theta)$ in the formula column describes the probability of the (i,j)th elements of matrices along angle θ, $\mu = \sum_{i=1}^{N_g} \sum_{j=1}^{N_r} p(i,j|\theta)i$ denotes the average run length of GLRLM matrix [15].

Feature name	Formula	Interpretation
RMI	$\frac{-(\sum_{j=1}^{N_g} p_y(j) \log_2 p_y(j)+\epsilon)+\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (p(i,j)) \log_2 p(i\|j)}{-\sum_{j=1}^{N_g} p_y(j) \log_2 p_y(j)+\epsilon}$	Uncertainty coefficient in landspace pattern [16]
Entropy	$-\sum_{i=1}^{N_g} p(i) \log_2(p(i)+\epsilon)$	The uncertainty/ randomness in the image values
Joint Energy	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (p(i,j))^2$	Energy is a measure of homogeneous patterns in the image
IMC	$\frac{HXY-HXY1}{\max\{HX,HY\}}$	Quantifying the complexity of the texture)
LRE	$\frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} (p(i,j))^2 j^2}{N_r(\theta)}$	LRE is a measure of the distribution of long run lengths
SRE	$\frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} \frac{(p(i,j))^2}{j^2}}{N_r(\theta)}$	SRE is a measure of the distribution of short run lengths
Non-uniformity	$\frac{\sum_{i=1}^{N_g} (\sum_{j=1}^{N_r} P(i,j\|\theta))}{N_r(\theta)^2}$	Measures the similarity of gray-level intensity values in the image
RV	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} p(i,j\|\theta)(j-\mu)^2$	Measure of the variance in runs for the run lengths
RE	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} p(i,j\|\theta) \log_2(p(i,j\|\theta)+\epsilon)$	Measures the uncertainty/randomness in the distribution of run lengths

References

1. Behrens, T.E., et al.: Characterization and propagation of uncertainty in diffusion-weighted MR imaging. Magn. Resonan. Med. Off. J. Int. Soc. Magn. Resonan. Med. **50**(5), 1077–1088 (2003)
2. Beig, N., et al.: Radiogenomic-based survival risk stratification of tumor habitat on Gd-T1w MRI is associated with biological processes in glioblastoma. Clin. Cancer Res. **26**(8), 1866–1876 (2020)
3. Brochu, E., Cora, V.M., De Freitas, N.: A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. arXiv preprint arXiv:1012.2599 (2010)
4. Dextraze, K., et al.: Spatial habitats from multiparametric MR imaging are associated with signaling pathway activities and survival in glioblastoma. Oncotarget **8**(68), 112992 (2017)
5. Jenkinson, M., Bannister, P., Brady, M., Smith, S.: Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage **17**(2), 825–841 (2002)

6. Leone, J., Zwenger, A.O., Leone, B.A., Vallejo, C.T., Leone, J.P.: Overall survival of men and women with breast cancer according to tumor subtype. Am. J. Clin. Oncol. **42**(2), 215–220 (2019)
7. Li, C., et al.: Decoding the interdependence of multiparametric magnetic resonance imaging to reveal patient subgroups correlated with survivals. Neoplasia **21**(5), 442–449 (2019)
8. Li, C., et al.: Multi-parametric and multi-regional histogram analysis of MRI: modality integration reveals imaging phenotypes of glioblastoma. Eur. Radiol. **29**(9), 4718–4729 (2019)
9. Li, C., et al.: Intratumoral heterogeneity of glioblastoma infiltration revealed by joint histogram analysis of diffusion tensor imaging. Neurosurgery **85**(4), 524–534 (2019)
10. Li, C., et al.: Low perfusion compartments in glioblastoma quantified by advanced magnetic resonance imaging and correlated with patient survival. Radiother. Oncol. **134**, 17–24 (2019)
11. Mangla, R., et al.: Correlation between progression free survival and dynamic susceptibility contrast MRI perfusion in WHO grade III glioma subtypes. J. Neurooncol. **116**(2), 325–331 (2013). https://doi.org/10.1007/s11060-013-1298-9
12. Meacham, C.E., Morrison, S.J.: Tumour heterogeneity and cancer cell plasticity. Nature **501**(7467), 328–337 (2013)
13. Meilă, M.: Comparing clusterings by the variation of information. In: Schölkopf, B., Warmuth, M.K. (eds.) COLT-Kernel 2003. LNCS (LNAI), vol. 2777, pp. 173–187. Springer, Heidelberg (2003). https://doi.org/10.1007/978-3-540-45167-9_14
14. Meyer-Bäse, A., Saalbach, A., Lange, O., Wismüller, A.: Unsupervised clustering of fMRI and MRI time series. Biomed. Sig. Process. Control **2**(4), 295–310 (2007)
15. Mohanty, A.K., Beberta, S., Lenka, S.K.: Classifying benign and malignant mass using GLCM and GLRLM based texture features from mammogram. Int. J. Eng. Res. Appl. **1**(3), 687–693 (2011)
16. Nowosad, J., Stepinski, T.F.: Information theory as a consistent framework for quantification and classification of landscape patterns. Landscape Ecol. **34**(9), 2091–2101 (2019). https://doi.org/10.1007/s10980-019-00830-x
17. Park, J.E., Kim, H.S., Kim, N., Park, S.Y., Kim, Y.H., Kim, J.H.: Spatiotemporal heterogeneity in multiparametric physiologic MRI is associated with patient outcomes in IDH-wildtype glioblastoma. Clin. Cancer Res. **27**(1), 237–245 (2021)
18. Paszke, A., et al.: Pytorch: an imperative style, high-performance deep learning library. Adv. Neural. Inf. Process. Syst. **32**, 8026–8037 (2019)
19. Patel, E., Kushwaha, D.S.: Clustering cloud workloads: K-means vs gaussian mixture model. Procedia Comput. Sci. **171**, 158–167 (2020)
20. Pena, A., Green, H., Carpenter, T., Price, S., Pickard, J., Gillard, J.: Enhanced visualization and quantification of magnetic resonance diffusion tensor imaging using the p: q tensor decomposition. Br. J. Radiol. **79**(938), 101–109 (2006)
21. Rasmussen, C.E.: Gaussian processes in machine learning. In: Bousquet, O., von Luxburg, U., Rätsch, G. (eds.) ML -2003. LNCS (LNAI), vol. 3176, pp. 63–71. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-28650-9_4
22. Sala, E., et al.: Unravelling tumour heterogeneity using next-generation imaging: radiomics, radiogenomics, and habitat imaging. Clin. Radiol. **72**(1), 3–10 (2017)
23. Smith, S.M., et al.: Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage **23**, S208–S219 (2004)
24. Snoek, J., Larochelle, H., Adams, R.P.: Practical Bayesian optimization of machine learning algorithms. arXiv preprint arXiv:1206.2944 (2012)

25. Sottoriva, A., et al.: Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. Proc. Natl. Acad. Sci. **110**(10), 4009–4014 (2013)
26. Syed, A.K., Whisenant, J.G., Barnes, S.L., Sorace, A.G., Yankeelov, T.E.: Multiparametric analysis of longitudinal quantitative MRI data to identify distinct tumor habitats in preclinical models of breast cancer. Cancers **12**(6), 1682 (2020)
27. Van Griethuysen, J.J., et al.: Computational radiomics system to decode the radiographic phenotype. Can. Res. **77**(21), e104–e107 (2017)
28. Von Luxburg, U.: Clustering stability: an overview. Found. Trends Mach. Learn. **2**(3), 235–274 (2010)
29. Wu, J., et al.: Unsupervised clustering of quantitative image phenotypes reveals breast cancer subtypes with distinct prognoses and molecular pathways. Clin. Cancer Res. **23**(13), 3334–3342 (2017)
30. Wu, J., Gong, G., Cui, Y., Li, R.: Intra-tumor partitioning and texture analysis of DCE-MRI identifies relevant tumor subregions to predict pathological response of breast cancer to neoadjuvant chemotherapy. J. Magn. Resonan. Imaging (JMRI) **44**(5), 1107 (2016)
31. Xia, W., et al.: Radiogenomics of hepatocellular carcinoma: multiregion analysis-based identification of prognostic imaging biomarkers by integrating gene data-a preliminary study. Phys. Med. Biol. **63**(3), 035044 (2018)
32. Zhang, L.: Dirac delta function of matrix argument. Int. J. Theor. Phys. 1–28 (2020)