IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

A Mixed Focal Loss Function for Handling Class Imbalanced Medical Image Segmentation

MICHAEL YEUNG[1,2], EVIS SALA[1,3], CAROLA-BIBIANE SCHÖNLIEB[4], LEONARDO RUNDO[1,3]
[1]Department of Radiology, University of Cambridge, Cambridge CB2 0QQ, United Kingdom
[2]School of Clinical Medicine, University of Cambridge, Cambridge CB2 0SP, United Kingdom
[3]Cancer Research UK Cambridge Institute, Cambridge CB2 0RE, United Kingdom
[4]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge CB3 0WA, United Kingdom

Corresponding authors: Michael Yeung (e-mail: mjyy2@cam.ac.uk), Leonardo Rundo (e-mail: lr495@cam.ac.uk).

ABSTRACT Automatic segmentation methods are an important advancement in medical imaging analysis. Machine learning techniques, and deep neural networks in particular, are the state-of-the-art for most automated medical image segmentation tasks, ranging from the subcellular to the level of organ systems. Issues with class imbalance pose a significant challenge irrespective of scale, with organs, and especially with tumours, often occupying a considerably smaller volume relative to the background. Loss functions used in the training of segmentation algorithms differ in their robustness to class imbalance, with cross entropy-based losses being more affected than Dice-based losses. In this work, we first experiment with seven different Dice-based and cross entropy-based loss functions on the publicly available Kidney Tumour Segmentation 2019 (KiTS19) Computed Tomography dataset, and then further evaluate the top three performing loss functions on the Brain Tumour Segmentation 2020 (BraTS20) Magnetic Resonance Imaging dataset. Motivated by the results of our study, we propose a Mixed Focal loss function, a new compound loss function derived from modified variants of the Focal loss and Focal Dice loss functions. We demonstrate that our proposed loss function is associated with a better recall-precision balance, significantly outperforming the other loss functions in both binary and multi-class image segmentation. Importantly, the proposed Mixed Focal loss function is robust to significant class imbalance. Furthermore, we showed the benefit of using compound losses over their component losses, and the improvement provided by the focal variants over other variants.

INDEX TERMS Class imbalance, Loss function, Machine learning, Medical image segmentation, Computed Tomography, Magnetic Resonance Imaging

I. INTRODUCTION

Image segmentation involves partitioning an image into meaningful regions, based on the regional pixel characteristics, thus aiming at identifying objects of interest [1]. This task is fundamental in computer vision and has been applied widely in face recognition, autonomous driving, as well as medical image processing. In particular, automatic segmentation methods are an important advancement in medical image analysis, capable of demarcating structures across a range of imaging modalities including computed tomography (CT), magnetic resonance imaging (MRI) and positron emission tomography (PET).

Classical approaches for image segmentation include direct region detection methods such as the split-and-merge and region growing algorithms [2], graph-based methods [3], active contour and level set models [4]. Alongside these developments, later approaches have focused on applying and adapting traditional machine learning techniques [5], such as support vector machines (SVMs) [6], unsupervised clustering [7] and atlas-based segmentation [8]. In recent years, however, significant progress has been achieved using deep learning [9], [10].

The most well-known architecture in image segmentation, the U-Net architecture [11], is a modification of the convolutional neural network (CNN) architecture into an encoder-decoder network, similar to SegNet [12], which enables end-to-end feature extraction and pixel classification. Since its inception, many variants based on the U-Net architecture have been proposed [13], [14], including the 3D U-Net [15], Attention U-Net [16] and V-Net [17].

Once a model architecture is selected, optimisation of model parameters is based on minimisation of the loss function during training. The cross entropy loss is perhaps the most widely used loss function in classification problems [18] and is applied in U-Net [11], 3D U-Net [15] and SegNet [12]. In contrast, Attention U-Net [16] and V-Net [17] leverage the Dice loss function, which is based on the most commonly used metric for evaluating segmentation

performance, and therefore represents a form of direct loss minimisation. Broadly, loss functions used in image segmentation may be classified into distribution-based losses (such as the cross entropy loss), region-based losses (such as Dice loss), boundary-based losses (such as the boundary loss) [19], and more recently compound losses. Compound losses refer to the simultaneous minimisation of multiple, independent loss functions, such as the Combo loss, which minimises the sum of Dice and cross entropy loss [20].

A dominant issue in medical image segmentation is handling class imbalance, which refers to an unequal distribution of foreground and background elements. For example, automatic organ segmentation often involves organ sizes an order of magnitude smaller than the scan itself, resulting in a skewed distribution favouring background elements [21]. This issue is even more prevalent in oncology, where tumour sizes are themselves often significantly smaller than their organ of origin. In these class imbalanced circumstances, careful selection of the loss function is crucial, with the Dice loss generally better suited than the cross entropy loss function. Taghanaki *et al.* [20] distinguishes between input and output imbalance, the former as aforementioned, and the latter referring to classification biases arising during inference. These include false positives and false negatives, which respectively describe background pixels incorrectly classified as foreground objects, and foreground objects incorrectly classified as background. Both are particularly important in the context of medical image segmentation; in the case of image-guided interventions, false positives may result in a larger radiation field or excessive surgical margins, and conversely false negatives may lead to inadequate radiation delivery or incomplete surgical resection. Therefore, it is important to design a loss function that can be optimised to handle both input and output imbalances.

Due to the impracticality of experimenting and testing numerous loss functions, it is often the case that only a handful of loss functions are tested, from which the best performing model is selected. Despite its significance, few studies have focused on comparing large numbers of loss functions. Mun *et al.* [22] compared the performance of six loss functions on the Prostate MR Image Segmentation 2012 (PROMISE12) dataset [23], with the cosine similarity outperforming Dice-based and cross entropy-based losses amongst others. More recently, a comparison of fifteen loss functions using the NBFS Skull-stripping dataset [24] (brain CT segmentation), which also introduces the log-cosh Dice loss, concluded that Focal Tversky loss and Tversky loss are generally optimal [25].

Whilst these studies are based on organ segmentation, datasets involving tumour segmentation are associated with even greater degrees of class imbalance. Manual tumour delineation is both time-consuming and operator-dependent. Automatic methods of tumour delineation aim to address these issues, and public datasets, such as the Kidney Tumour Segmentation 19 (KiTS19) dataset for kidney tumour CT [26] and Brain Tumour Segmentation 2020 (BraTS20) for brain tumour MRI [27], have accelerated progress towards this goal. In fact, there has been recent developments for translating the BraTS20 dataset into clinical and scientific practice [28].

For the KiTS19 dataset, the current state-of-the-art is the "no-new-Net" (nnU-Net) [29], [30], an automatically configurable deep learning-based segmentation method involving the ensemble of 2D, 3D and cascaded 3D U-Nets. This framework was optimised using the Dice and cross entropy loss. Recently, an ensemble-based method obtained comparable results to nnU-Net, and involved initial independent processing of kidney organ and kidney tumour segmentation by 2D U-Nets trained using the Dice loss, followed by suppression of false positive predictions of the kidney tumour segmentation using the network trained for kidney organ segmentation [31]. When the dataset size is small, results from an active learning-based method using CNN-corrected labeling, also trained using the Dice loss, showed a higher segmentation accuracy over nnU-Net [32].

For the BraTS20 dataset, a popular approach is to use a multi-scale architecture where different receptive field sizes allow the independent processing of both local and global contextual information [33], [34]. Kamnitsas *et al.* used a two-phase training process involving initial upsampling of under-represented classes, followed by a second-stage where the output layer is retrained on a more representative sample [33]. Similarly, Havaei *et al.* used a sampling rule to impose equal probability of foreground or background pixels at the centre of a patch, and used the cross entropy loss for optimisation [34].

It is apparent that for both the KiTS19 and BraTS20 datasets, class imbalance is largely handled by altering either the training or input data sampling process, and rarely with adapting the loss function. Even state-of-the-art solutions typically use either Dice loss, cross entropy loss or a combination of the two. However, popular methods—such as upsampling the underrepresented class—are inherently associated with an increase in false positive predictions, and more complicated, often multi-stage training processes require more computational resources. In contrast, adapting the loss function provides a simpler, ubiquitous solution at no additional cost in terms of computation.

In this paper, we propose the following contributions:

(a) We summarise and extend the knowledge provided by previous studies that compare loss functions using 2D U-Nets for binary classification problems, and evaluate multiple loss functions using 3D U-Nets for both binary and multi-class, highly class imbalanced classification problems.
(b) We introduce a new compound loss function, the Mixed Focal loss, which enables tuning to optimise for both input and output imbalances.
(c) Our proposed loss function improves segmentation quality over six other related loss functions across multiple classes and datasets, is associated with a better

recall-precision balance, and is robust to class imbalance.

(d) We provide evidence demonstrating the benefit of using compound losses over their component loss functions, and using focal variants over other variants of Dice or cross entropy-based losses in dealing with class imbalanced problems.

The manuscript is organised as follows. Section II provides a summary of the loss functions used. Section III describes the chosen medical imaging datasets, introduces the proposed Mixed Focal loss function, and defines the segmentation evaluation metrics used. Section IV presents and discusses the experimental results. Finally, Section V provides conclusive remarks and future directions.

II. BACKGROUND

Minimisation of the loss function represents the optimisation problem that occurs during training to generate optimal model parameters. This paper focuses on semantic segmentation, a sub-field of image segmentation where pixel-level classification is performed directly, in contrast to instance segmentation where an additional object detection stage is required. We describe seven loss functions that belong to either distribution-based, region-based or compound losses. A graphical overview of loss functions in these categories is provided in Fig. 1. First, the distribution-based functions are introduced, followed by region-based loss functions, and finally concluding with compound loss functions.

A. CROSS ENTROPY LOSS

The cross entropy loss is one of the most widely used loss functions in deep learning. With origins in information theory, cross entropy measures the difference between two probability distributions for a given random variable or set of events. As a loss function, it is superficially equivalent to the negative log likelihood loss and, for binary classification, the binary cross entropy loss ($\mathcal{L}_{\text{BCE}}$) is defined as the following:

$$\mathcal{L}_{\text{BCE}}(\mathbf{y}, \hat{\mathbf{y}}) = -(\mathbf{y} \log(\hat{\mathbf{y}}) + (\mathbf{1} - \mathbf{y}) \log(\mathbf{1} - \hat{\mathbf{y}})). \quad (1)$$

Here, $\mathbf{y}, \hat{\mathbf{y}} \in \{0, 1\}^N$, where $\hat{\mathbf{y}}$ refers to the predicted value and $\mathbf{y}$ refers to the ground truth label. This can be extended to multi-class problems, and the categorical cross entropy loss ($\mathcal{L}_{\text{CCE}}$) is computed as:

$$\mathcal{L}_{\text{CCE}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \cdot \log\left(p_{i,c}\right), \quad (2)$$

where $y_{i,c}$ uses a one-hot encoding scheme of ground truth labels, $p_{i,c}$ is a matrix of predicted values for each class, and where indices c and i iterate over all classes and pixels, respectively. Cross entropy loss is based on minimising pixel-wise error, leading to over-representation of larger objects in the loss, and consequently resulting in poorer quality segmentation of smaller objects.

B. FOCAL LOSS

The Focal loss is a variant of the binary cross entropy loss that addresses the issue of class imbalance faced by the standard cross entropy loss by down-weighting the contribution of easy examples enabling learning of harder examples [35]. To derive the Focal loss function, we first simplify the loss (1) as:

$$\text{CE}(p, y) = \begin{cases} -\log(p), & \text{if } y = 1 \\ -\log(1-p), & \text{if } y = 0 \end{cases}. \quad (3)$$

Next, we define the probability of predicting the ground truth class, p_t, as:

$$p_t = \begin{cases} p, & \text{if } y = 1 \\ 1-p, & \text{if } y = 0 \end{cases}. \quad (4)$$

The binary cross entropy loss ($\mathcal{L}_{\text{BCE}}$) can therefore be rewritten as:

$$\mathcal{L}_{\text{BCE}(p,y)} = \text{CE}\left(p_t\right) = -\log\left(p_t\right). \quad (5)$$

The Focal loss ($\mathcal{L}_{\text{F}}$) adds a modulating factor to the binary cross entropy loss:

$$\mathcal{L}_{F(p_t)} = \alpha \left(1 - p_t\right)^{\gamma} \cdot \mathcal{L}_{\text{BCE}(p,y)}, \quad (6)$$

The Focal loss is parameterised by α and γ, which control the class weights and degree of down-weighting of easy examples, respectively (Fig. 2a). When $\gamma = 0$, the Focal loss simplifies to the binary cross entropy loss.

To use the Focal loss for multi-class classification, we define the categorical Focal loss ($\mathcal{L}_{\text{CF}}$):

$$\mathcal{L}_{\text{CF}} = \boldsymbol{\alpha} \left(1 - \left(p_{t,c}\right)\right)^{\gamma} \cdot \mathcal{L}_{\text{CCE}}, \quad (7)$$

where $\boldsymbol{\alpha}$ is now a vector of class weights, $p_{t,c}$ is a matrix of ground truth probabilities for each class, and $\mathcal{L}_{\text{CCE}}$ is the categorical cross entropy loss as defined in Eq. (2).

C. DICE LOSS

The Sørensen–Dice index, known as the Dice similarity coefficient (DSC) when applied to Boolean data, is the most commonly used metric for evaluating segmentation accuracy. We can define DSC in terms of the per voxel classification of true positives (TP), false positives (FP) and false negatives (FN):

$$\text{DSC} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}. \quad (8)$$

For notational convenience and to highlight its similarity to the Tversky index (TI), from now on, we define a modified Dice similarity coefficient (mDSC) according to Eq. (9):

$$\text{mDSC} = \frac{\sum_{i=1}^{N} p_{0i} g_{0i}}{\sum_{i=1}^{N} p_{0i} g_{0i} + \delta \sum_{i=1}^{N} p_{0i} g_{1i} + (1-\delta) \sum_{i=1}^{N} p_{1i} g_{0i}}, \quad (9)$$

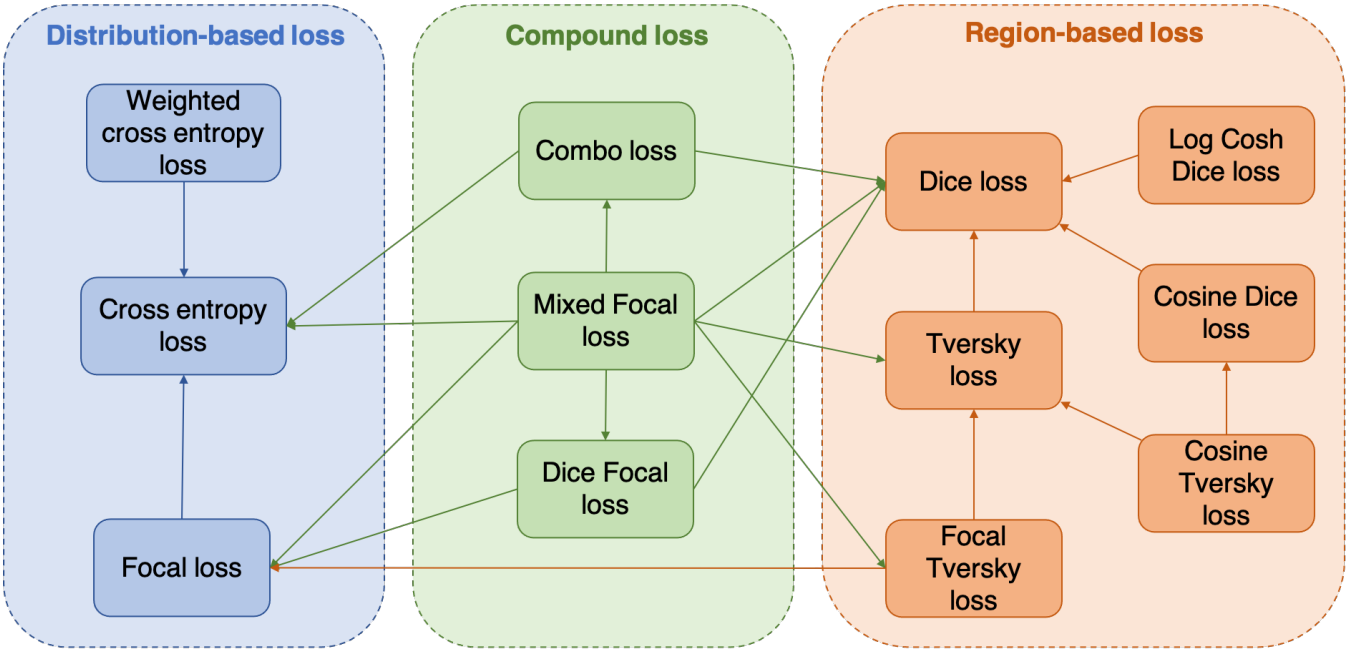

FIGURE 1: Overview of the various distribution-based, region-based and compound loss functions. The arrows connect related loss functions, with the direction of the arrows indicating the inheritance relationship.

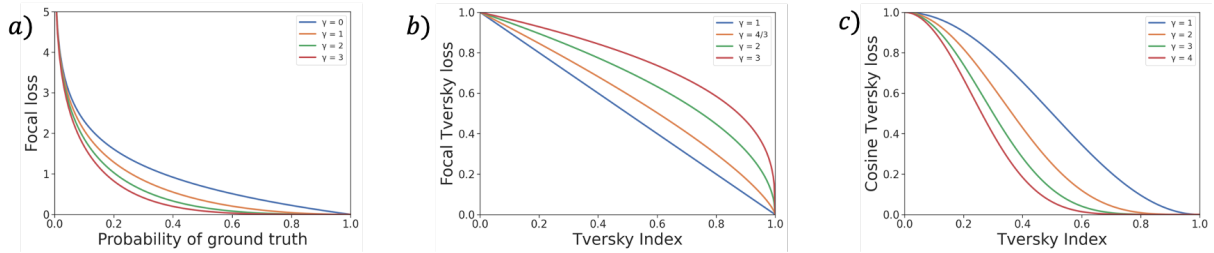

FIGURE 2: Effect of altering the parameter γ for the (a) Focal loss, (b) Focal Tversky loss, and (c) Cosine Tversky loss.

This is equivalent to Eq. (8) when $\delta = \frac{1}{2}$:

$$\mathrm{DSC} = \frac{\sum_{i=1}^{N} p_{0i} g_{0i}}{\sum_{i=1}^{N} p_{0i} g_{0i} + \frac{1}{2}\sum_{i=1}^{N} p_{0i} g_{1i} + \frac{1}{2}\sum_{i=1}^{N} p_{1i} g_{0i}}, \tag{10}$$

where p_{0i} is the probability of pixel i belonging to the foreground class and p_{1i} is the probability of pixel belonging to background class. p_{0i} is 1 for foreground and 0 for background and conversely $g_{1\mathrm{i}}$ takes values of 1 for background and 0 for foreground.

The Dice loss ($\mathcal{L}_{\mathrm{DSC}}$), for C classes, can therefore be defined as:

$$\mathcal{L}_{\mathrm{DSC}} = \sum_{c=1}^{C} (1 - \mathrm{DSC}). \tag{11}$$

Other variants of the Dice loss include the Generalised Dice loss [36], [37] where the class weights are corrected by the inverse of their volume, and the Generalised Wasserstein Dice loss [38], which combines the Wasserstein metric with the Dice loss and is adapted for dealing with hierarchical data, such as the BraTS20 dataset [27].

Even in its most simple formulation, the Dice loss is partially robust to class imbalance, with equal weighting provided to each class.

D. TVERSKY LOSS

The Tversky index [39] is closely related to the Dice score, but enables optimisation for output imbalance by altering the weights assigned to false positives and false negatives. In its most general form it is equivalent to the Eq. (9), but is most commonly used by setting $\delta = \frac{7}{10}$:

$$\mathrm{TI} = \frac{\sum_{i=1}^{N} p_{0i} g_{0i}}{\sum_{i=1}^{N} p_{0i} g_{0i} + \frac{7}{10}\sum_{i=1}^{N} p_{0i} g_{1i} + \frac{3}{10}\sum_{i=1}^{N} p_{1i} g_{0i}}, \tag{12}$$

To use TI as a loss function, we define the Tversky loss, $\mathcal{L}_{\mathrm{T}}$, for C classes as:

$$\mathcal{L}_{T} = \sum_{c=1}^{C} (1 - \mathrm{TI}). \tag{13}$$

When the Dice loss function is applied to highly class imbalanced problems, the resulting segmentation often exhibits high precision but low recall rate [39]. By assigning a greater weight to false negatives, recall rate is improved leading to a better balance of precision and recall.

E. FOCAL TVERSKY LOSS

Analogous to the way the Focal loss adapts the cross entropy loss to focus on harder examples, the Focal Tversky loss [40] adapts the Tversky loss by down-weighting easy to classify regions in favour of more difficult regions.

Using the definition of TI from Eq. (12), we can define the Focal Tversky loss ($\mathcal{L}_{\text{FT}}$) as:

$$\mathcal{L}_{\text{FT}} = \sum_{c=1}^{C} (1 - \text{TI})^{\frac{1}{\gamma}}, \tag{14}$$

where higher values of γ increases the degree of focusing on harder examples (Fig. 2b) and simplifies to the Tversky loss when $\gamma = 1$.

F. COSINE TVERSKY LOSS

Inspired by results from [22], we test another variant of the Tversky loss, closely related to the Cosine Dice loss proposed in [41]. Here, we define the Cosine Tversky loss ($\mathcal{L}_{\text{cosT}}$), again using the TI from Eq. (12):

$$\mathcal{L}_{\text{cosT}} = \sum_{c=1}^{C} \cos^{\gamma}\left(\frac{\pi}{2} \cdot \text{TI}\right), \tag{15}$$

where γ is analogous to the focal parameters in the Focal loss and Focal Tversky loss (Fig. 2c).

G. COMBO LOSS

The Combo loss [20] belongs to the class of compound losses, where multiple loss functions are minimised in unison. The Combo loss ($\mathcal{L}_{\text{combo}}$) is defined as a weighted sum of the Dice similarity coefficient in Eq. (10) and a modified form of the cross entropy loss ($\mathcal{L}_{\text{mCE}}$):

$$\mathcal{L}_{\text{combo}} = \alpha\left(\mathcal{L}_{\text{mCE}}\right) - (1 - \alpha) \cdot \text{DSC}, \tag{16}$$

where:

$$\mathcal{L}_{\text{mCE}} = -\frac{1}{N}\sum_{i=1}^{N} \beta\left(t_i - \log\left(p_i\right)\right) + (1-\beta)\left[(1 - t_i)\ln\left(1 - p_i\right)\right] \tag{17}$$

and α in the range of $[0, 1]$ controls the relative weighting of the Dice and cross entropy terms contribution to the loss, and β controls the relative weights assigned to false positives and negatives. A value of $\beta > \frac{1}{2}$ penalises false negative predictions more than false positives.

For our experiments, we use a simplified, multi-class variant of the Combo loss:

$$\mathcal{L}_{\text{combo}} = \mathcal{L}_{\text{CCE}} - \text{DSC}. \tag{18}$$

Firstly, we assign equal weights to the Dice and cross entropy loss, which is equivalent to the optimal value of $\alpha = \frac{1}{2}$ [20]. Secondly, we use the standard cross entropy loss, given that the optimal value of β is dependent on the dataset used.

Confusingly, the term "Dice and cross entropy loss" has been used to refer to both the sum of cross entropy loss and DSC [20], [29], as well as the sum of the cross entropy loss and Dice loss, such as in the Dice Focal loss [42], [43]. Here, we decide to use the former implementation, which is consistent with both Combo loss and the loss function used in the state-of-the-art for the KiTS19 dataset [29].

III. MATERIALS AND METHODS

A. DATASET DESCRIPTIONS

1) KiTS19 dataset

Kidney tumour segmentation is a challenging task due to the widespread presence of hypodense tissue, as well as highly heterogeneous appearance of tumours on CT [44], [45]. To evaluate our loss functions, we select the Kidney Tumour Segmentation 2019 (KiTS19) dataset [26], a highly class imbalanced, multi-class problem. Briefly, this dataset consists of 300 arterial phase abdominal CT scans from patients who underwent partial removal of the tumour and surrounding kidney or complete removal of the kidney including the tumour at the University of Minnesota Medical Center, USA. Kidney and tumour boundaries were manually delineated by two students, with class labels of either kidney, tumour or background assigned to each voxel resulting in a semantic segmentation task [26]. 210 scans and their associated segmentations are provided for training, with the segmentation masks for the other 90 scans withheld from public access for testing. We therefore exclude the 90 scans without segmentation masks, and further exclude another 6 scans (case 15, 23, 37, 68, 125 and 133) due to concern over ground truth quality [46], leaving 204 scans for use.

2) BraTS20 dataset

To assess for generalisation, we further evaluate the top three performing loss functions on the Brain Tumour Segmentation 2020 (BraTS20) dataset [27], [47], [48]. This is currently the largest, publicly available and fully-annotated dataset for medical image segmentation, and comprises of 494 multi-modal scans of patients with either low-grade glioma or high-grade glioblastoma. Whilst kidney tumours are well visualised on CT scans, MRI is better suited for brain tumours. The BraTS20 dataset provides images for the following MRI sequences: T1-weighted (T1), T1-weighted contrast-enhanced using gadolinum contrast agents (T1-CE), T2-weighted (T2) and fluid attenuated inverse recovery (FLAIR) sequence. Images were manually annotated, with regions associated with the tumour labelled as: necrotic and non-enhancing tumour core, peritumoural oedema or gadolinum-enhancing tumour. From the 494 scans provided, 125 scans are used for validation with reference segmentation masks withheld from public access, and therefore are excluded.

We further exclude T1, T2 and FLAIR sequences to focus on gadolinum-enhancing tumour segmentation using the T1-CE sequence [49], [50], which not only appears to be the the most difficult class to segment [51], but is also the most clinically relevant for radiation therapy [52]. We further exclude another 27 scans without enhancing tumour regions, leaving 342 scans for use.

B. THE PROPOSED MIXED FOCAL LOSS

Combo loss [20] and Dice Focal loss [42] are two compound loss functions that inherit benefits from both Dice-based and cross entropy-based loss functions. The Combo loss is better adapted to handle output imbalance, with a modifiable β parameter in its cross entropy component loss. However, the Combo loss lacks an equivalent tunable parameter for its Dice component loss, and neither Dice nor cross entropy loss are adapted to handle highly class imbalanced inputs. In contrast, the Dice Focal loss is better adapted to handle input imbalance, with its Focal parameter in the Focal loss component. However, similar to the Combo loss, its Dice component is not adapted to handling highly class imbalanced data.

Here, we propose a novel compound loss function, namely the Mixed Focal loss function, which involves further modifications of Dice-based and cross entropy based loss functions, incorporating tunable parameters to handle output imbalance, as well as focal parameters to handle input imbalance, for both the Dice and cross entropy-based component losses.

Firstly, to provide the Dice component of the loss with a parameter to optimise the weighting of false positive and false negative predictions, we define a modified Dice loss using Eq. (9):

$$\mathcal{L}_{\text{mD}} = \sum_{c=1}^{C}(1 - \text{mDSC}), \tag{19}$$

where the parameter δ in Eq. (9) controls the relative contribution of false positive and false negative predictions to the loss.

Using this formulation, we can combine the modified Dice loss with the modified cross entropy loss function (17) to define a modified Combo loss ($\mathcal{L}_{\text{mCombo}}$):

$$\mathcal{L}_{\text{mCombo}} = \alpha\left(\mathcal{L}_{\text{mCE}}\right) - (1-\alpha)\cdot\left(\mathcal{L}_{\text{mD}}\right), \tag{20}$$

where the parameters β in Eq. (17) and δ in Eq. (9) control the weights of the false positive and false negatives for the modified cross entropy and modified Dice loss, respectively.

Whilst this enables tuning for output imbalance, the standard Dice and cross entropy losses are maladapted for handling highly class imbalanced inputs, whereas loss functions using the focal parameter γ appear more suitable. Therefore, next we add separate focal parameters to both the modified cross entropy loss and modified Dice loss, to produce the modified Focal loss ($\mathcal{L}_{\text{mF}}$) and modified Focal Dice loss ($\mathcal{L}_{\text{mFD}}$) respectively:

$$\mathcal{L}_{\text{mF}} = -\alpha\left(1-p_t\right)^{\gamma}\cdot\mathcal{L}_{\text{mCE}}, \tag{21}$$

$$\mathcal{L}_{\text{mFD}} = \sum_{c=1}^{C}(1 - \text{mD})^{\frac{1}{\gamma}}, \tag{22}$$

Using these equations, we define the Mixed Focal loss ($\mathcal{L}_{\text{MF}}$) as the weighted sum of the modified Focal loss and modified Focal Dice loss:

$$\mathcal{L}_{\text{MF}} = \lambda\mathcal{L}_{\text{mF}} + (1-\lambda)\mathcal{L}_{\text{mFD}}, \tag{23}$$

where $\lambda \in [0, 1]$ and determines the relative weighting of the two component loss functions.

To enable a fair comparison with the simplified Combo loss in Eq. (18), we implement a simplified, Categorical variant of the Mixed Focal loss ($\mathcal{L}_{\text{CMF}}$) where equal weights are assigned to the component losses, with parameters chosen to equate the modified Focal Dice loss to the Focal Tversky loss, and modified Focal loss to the categorical Focal loss Eq. (7):

$$\mathcal{L}_{\text{CMF}} = \mathcal{L}_{\text{FT}} + \mathcal{L}_{\text{CF}}. \tag{24}$$

C. EXPERIMENTAL SETUP

For our experiments, we make use of the Medical Image Segmentation with Convolutional Neural Networks (MIScnn) open-source Python library [43].

For both the KiTS19 and BraTS20 dataset, images and ground truth segmentation masks are provided in an anonymised NIfTI file format. For the KiTS19 dataset, the original image resolution is 512×512 in the axial plane, with an average of 216 slices in coronal plane. Pixel values are normalised to $[0, 1]$ using the z-score, Hounsfield units (HU) are clipped to $[-79, 304]$ HU and voxel spacing resampled to $3.22 \times 1.62 \times 1.62$. We perform patch-wise analysis using random patches of size of $80 \times 160 \times 160$ and patch-wise overlap of $40 \times 80 \times 80$, with a batch size of 2. For our model architecture, we use the standard 3D U-Net as described in [15] with a final softmax activation layer.

For the BraTS20 dataset, the original image resolution is $240 \times 240 \times 155$. The provided data is already pre-processed, with the skull stripped and images interpolated to the same resolution of 1mm^3. We further normalise the pixel values to $[0, 1]$ using the z-score. We perform patch-wise analysis using random patches of size of $96 \times 96 \times 96$ and patch-wise overlap of $48 \times 48 \times 48$, with a batch size of 2. We use the same model architecture as for the KiTS19 dataset.

D. IMPLEMENTATION DETAILS

For both the KiTS19 and BraTS20 datasets, we perform five-fold cross validation on remaining cases after exclusion. Since all scans belong to unique individuals, we perform a single random assignment of scans to each fold and use the resulting configuration to evaluate all loss functions.

We evaluate the following loss functions: Focal loss, Dice loss, Tversky loss, Focal Tversky loss, Cosine Tversky loss, Combo loss and Mixed Focal loss. We set $\alpha = 0.25$ and

$\gamma = 2$ for Focal loss as in [35]. In contrast, we use $\gamma = 4/3$ for Focal Tversky loss as in [40] and use $\gamma = 1$ for Cosine Tversky loss. For the Mixed Focal loss, we use the same parameters as for the individual Focal loss and Focal Tversky loss.

Model parameters are initialised randomly, and we again make use of use MIScnn that leverages the 'batchgenerators' library to perform the following data augmentations: rotation, mirroring, brightness, contrast, gamma, elastic deformation and Gaussian noise.

For the KiTS19 dataset, we train each model for 500 epochs with 100 iterations per epoch, using an Adam optimiser [53] with an initial learning rate of 3.0×10^{-4} and minimum learning rate of 1.0×10^{-6}, and use batch shuffling after each epoch. To account for the larger dataset size of the BraTS20 dataset, here we train each model for 400 epochs and 150 iterations per epoch. Validation loss is evaluated after each epoch, and the model with the lowest validation loss is selected as the final model. All experiments are programmed using Keras with TensorFlow backend and trained using NVIDIA P100 GPUs. Source code is available at: https://github.com/mlyg/mixed-focal-loss.

E. EVALUATION METRICS

To assess segmentation accuracy, we use three commonly used metrics [54]: Dice similarity coefficient (DSC), recall and precision. DSC is defined as Eq. (8), and recall and precision are defined similarly per voxel and according to Eqs. (25) and (26), respectively:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{25}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{26}$$

F. STATISTICS

To provide a statistical comparison of loss function performance, we perform pair-wise Wilcoxon rank sum tests comparing kidney and tumour DSC validation scores for the KiTS19 dataset, and enhancing tumour DSC validation score for the BraTS20 dataset. To account for multiple comparisons, p-values are adjusted using the Holm-Bonferroni method [55]. Statistical tests were implemented using the SciPy package, and p-value adjustments with the 'statsmodels' package.

IV. EXPERIMENTAL RESULTS

In this section, we first describe the results for the KiTS19 dataset, and then for the BraTS20 dataset. The results for the KiTS19 dataset are shown in Table 1.

Our proposed loss function, the Mixed Focal loss, outperformed all other loss functions with a score of 0.945 ± 0.017 and 0.751 ± 0.017 for DSC kidney and DSC tumour, respectively. Furthermore, the Mixed Focal loss is associated with the highest recall score for both kidney and tumour, and with similarly strong performance for precision score. Despite a poor DSC tumour score, Focal loss was associated with the highest precision score for kidney segmentation. For tumour segmentation, the highest precision scores were seen with both Dice loss and Combo loss. Despite high precision scores for the Dice loss, Combo loss and Focal loss, these loss functions were associated with poorer recall scores, and subsequently lower DSC values. In contrast, higher recall scores were obtained by the Tversky loss and its variants across both kidney and tumour segmentations, although this was balanced by lower precision scores. Comparisons between the Tversky loss variants showed that the Focal Tversky loss performed the best across all metrics, whilst the Cosine Tversky loss was generally the worst, only outperforming the Tversky loss for tumour recall score. Comparing compound losses with their component losses, besides equivalent scores for tumour precision, the Combo loss outperformed the Dice loss across all other metrics. Similarly, the Mixed Focal loss outperformed both the Focal Tversky loss and Focal loss, except for the kidney precision score. Finally, comparisons between the two compound losses showed better recall-precision balance with the Mixed Focal loss, outperforming the Combo loss for both the DSC and recall metrics.

Results from statistical comparisons using the Wilcoxon rank sum test for the KiTS19 dataset are shown in Table 2. For kidney DSC values, the Mixed Focal loss is the only loss function which performed significantly better than Tversky loss ($p = 5.41 \times 10^{-3}$). The Cosine Tversky loss was associated with the lowest kidney DSC, and significantly better performance was seen with the Focal loss ($p = 4.68 \times 10^{-3}$), Focal Tversky loss ($p = 1.45 \times 10^{-3}$), Combo loss ($p = 2.49 \times 10^{-3}$) and Mixed Focal loss ($p = 2.52 \times 10^{-5}$). For tumour DSC, Focal Tversky loss ($p = 8.30 \times 10^{-4}$), Combo loss (2.09×10^{-2}) and Mixed Focal loss ($p = 1.30 \times 10^{-4}$) performed significantly better than the Focal loss.

Examples of image segmentations for the KiTS19 dataset are shown in Fig. 3. Whilst kidney segmentations are generally similar, the Focal loss kidney segmentation is noticeably different from the other loss functions, with an apparent over-prediction of the kidney class. This is an expected consequence of the over-representation of the larger kidney class in the loss. On the other hand, tumour segmentation quality is noticeably different amongst all loss functions. The tumour appears under-segmented with the Focal loss, again resulting from under-representation of the smaller class and reflects its higher precision but lower recall score. In contrast, the tumour appears over-segmented with Tversky variant loss functions, in agreement with the higher recall but lower precision scores observed. The segmentation resulting from training with compound loss functions produces the most accurate tumour shape, with the highest quality segmentation seen with Mixed Focal loss, followed by Combo loss.

Based on the results from the KiTS19 dataset, we select the Mixed Focal loss, Combo loss and Focal Tversky loss as the top three performing loss functions and evaluate these on the BraTS20 dataset. The results are shown in Table 3. In agreement with results from the KiTS19 dataset, the Mixed

TABLE 1: Performance on KiTS19 dataset. Values are in the form mean ± 95% confidence intervals. Numbers in boldface denote the highest values for each metric.

Loss function	DSC kidney	Precision kidney	Recall kidney	DSC tumour	Precision tumour	Recall tumour
Focal loss	0.940±0.020	**0.952±0.011**	0.935±0.021	0.576±0.079	0.718±0.036	0.549±0.086
Dice loss	0.938±0.018	0.948±0.012	0.934±0.019	0.695±0.066	**0.794±0.034**	0.684±0.084
Tversky loss	0.936±0.018	0.928±0.010	0.949±0.021	0.711±0.081	0.754±0.073	0.732±0.084
Cosine Tversky loss	0.933±0.013	0.924±0.013	0.947±0.017	0.722±0.046	0.742±0.026	0.767±0.066
Focal Tversky loss	0.943±0.015	0.934±0.011	0.955±0.016	0.740±0.061	0.772±0.038	0.764±0.068
Combo loss	0.943±0.014	0.951±0.013	0.938±0.018	0.723±0.047	**0.794±0.051**	0.722±0.051
Mixed Focal loss	**0.945±0.017**	0.936±0.012	**0.957±0.017**	**0.751±0.057**	0.791±0.026	**0.778±0.069**

TABLE 2: Matrix of adjusted p-values from pairwise Wilcoxon rank sum scores for KiTS19 dataset kidney Dice (top) and tumour Dice scores (bottom). Row variables are compared with column variables, and positive statistic values are shaded in green, whilst negative statistic values are shaded in red. * $p < 0.05$, ** $p < 0.01$, ‡ $p < 0.001$.

	Focal loss	Dice loss	Tversky loss	Cosine Tversky loss	Focal Tversky loss	Combo loss	Mixed Focal loss
Focal loss	-						
Dice loss	1.00	-					
Tversky loss	0.209	1.00	-				
Cosine Tversky loss	0.00468**	0.0962	1.00	-			
Focal Tversky loss	1.00	1.00	0.0950	0.00145**	-		
Combo loss	1.00	1.00	0.165	0.00249**	1.00	-	
Mixed Focal loss	1.00	0.476	0.00541**	2.52×10-5‡	1.00	1.00	-

	Focal loss	Dice loss	Tversky loss	Cosine Tversky loss	Focal Tversky loss	Combo loss	Mixed Focal loss
Focal loss	-						
Dice loss	0.286	-					
Tversky loss	0.0841	1.00	-				
Cosine Tversky loss	0.0841	1.00	1.00	-			
Focal Tversky loss	8.30×10-4‡	0.267	0.757	0.757	-		
Combo loss	0.0209*	1.00	1.00	1.00	1.00	-	
Mixed Focal loss	1.30×10-4‡	0.0841	0.331	0.308	1.00	1.00	-

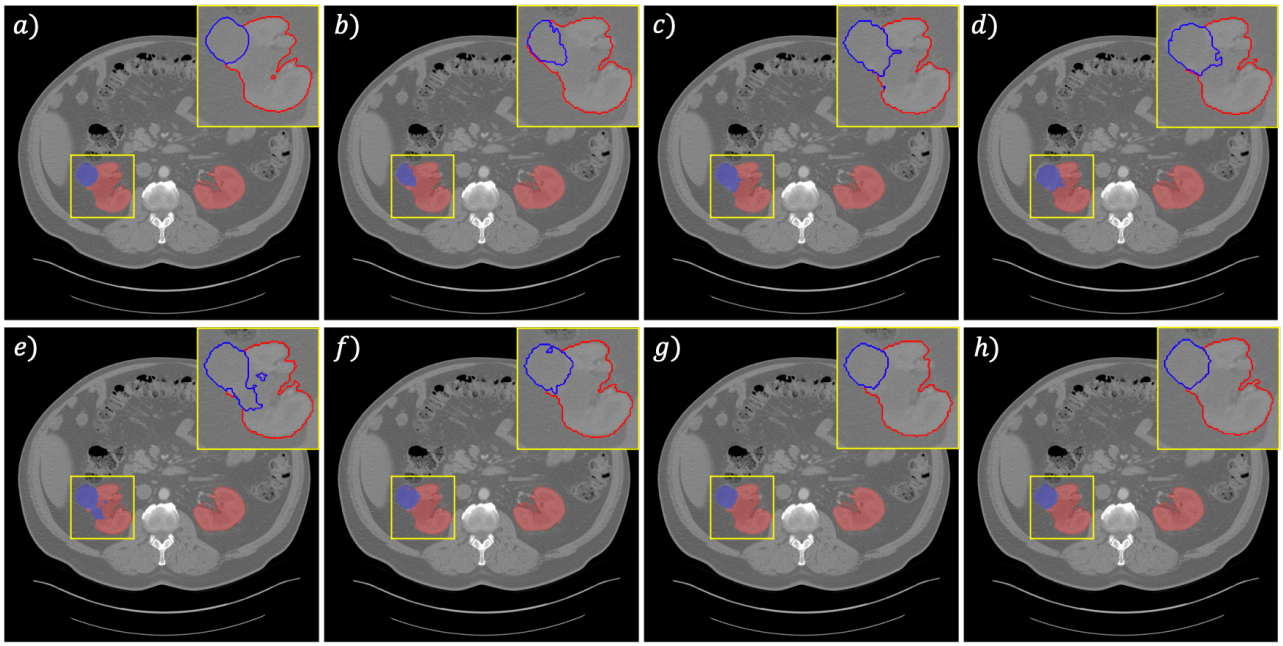

FIGURE 3: Axial CT slices of image segmentations generated from the KiTS19 dataset using (a) ground truth, (b) Focal loss, (c) Dice loss, (d) Tversky loss, (e) Cosine Tversky loss, (f) Focal Tversky loss, (g) Combo loss, (h) Mixed Focal loss. The kidney is highlighted in red and the tumour in blue. A magnified contour of the segmentation is provided in the top right-hand corner of each image.

Focal loss was associated with the best recall-precision balance, and outperformed the Focal Tversky loss and Combo loss for DSC tumour and recall tumour scores. The highest tumour precision score was seen with Combo loss, although

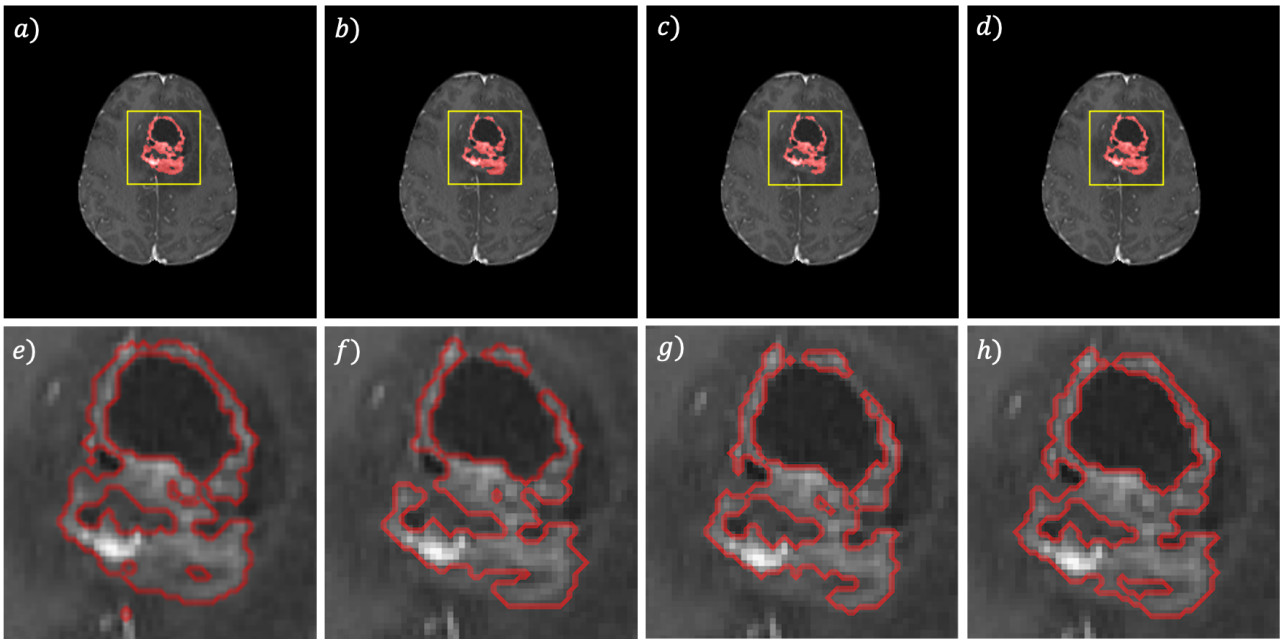

FIGURE 4: Axial MRI slices of image segmentations generated from the BraTS20 dataset using (a) ground truth, (b) Focal Tversky loss, (c) Combo loss and (d) Mixed Focal loss. Tumour is highlighted in red. A magnified contour of the segmentation is provided in (e-h) below each respective image.

TABLE 3: Performance on BraTS20 dataset. Values are in the form mean $\pm$ 95% confidence intervals. Numbers in boldface denote the highest values for each metric.

Loss function	DSC tumour	Precision tumour	Recall tumour
Focal Tversky loss	0.747±0.050	0.776±0.070	0.765±0.029
Combo loss	0.748±0.032	**0.833±0.039**	0.716±0.022
Mixed Focal loss	**0.775±0.033**	0.798±0.032	**0.795±0.027**

TABLE 4: Matrix of adjusted p-values from pairwise Wilcoxon rank sum scores for BraTS20 dataset tumour DSC. Row variables are compared with column variables, and positive statistic values are shaded in green, whilst negative statistic values are shaded in red. * $p < 0.05$, ** $p < 0.01$, ‡ $p < 0.001$.

	Focal Tversky loss	Combo loss	Mixed Focal loss
Focal Tversky loss	-		
Combo loss	0.312	-	
Mixed Focal loss	0.176	0.018*	-

this was also associated with the poorest tumour recall score.

Results from statistical comparisons using the Wilcoxon rank sum test for the BraTS20 dataset are shown in Table 4. Whilst the Combo loss is associated with a slightly better tumour DSC than the Focal Tversky loss, the Mixed Focal loss performed significantly better than Combo loss ($p = 0.018$) but not the Focal Tversky loss ($p = 0.0176$). This reflected generally better but less consistent segmentation quality with the Focal Tversky loss than the Combo loss.

Examples of image segmentations for the BraTS20 dataset are shown in Fig. 4. Image segmentations are similarly high quality for all three loss functions. Whilst the differences between the Focal Tversky loss and the Combo loss are more subtle, a higher segmentation quality with the Mixed Focal loss is apparent.

V. DISCUSSION AND CONCLUSIONS

In this study, we proposed a new compound loss function, the Mixed Focal loss, which is adapted to handle both input and output imbalances in semantic image segmentation tasks. The difference in model performance across the numerous loss functions compared highlights the importance of loss function choice in class imbalanced image segmentation tasks. Comparisons of compound losses with their respective component loss functions revealed consistent improvements across all metrics, with the Combo loss outperforming the Dice loss, and the Mixed Focal loss outperforming both Focal Tversky loss and Focal loss. Moreover, we showed that our proposed loss function outperformed the Combo loss, with higher DSC scores obtained across classes and datasets. These results were demonstrated for the KiTS19 dataset, a multi-class, class imbalanced dataset comprised of kidney tumour labelled CT scans, and further generalised to the BraTS20 dataset, which we adapted to solve a binary, highly class imbalanced brain tumour segmentation problem based on T1-CE MRI scans. Therefore, we evaluated our proposed loss function for both binary and multi-class classification, across two different modalities, sharing the common theme of class imbalance. The main metric we evaluate is the DSC, which is highest when both precision and recall scores are similarly high. Our results illustrated how loss functions tend to prioritise one of these component metrics over the other resulting in output imbalance, with higher precision scores observed with cross entropy-based losses and the Dice loss, and higher recall scores associated with Tversky vari-

ant losses. This is further complicated by input imbalance, where cross entropy based-losses such as the Focal loss are particularly susceptible, as shown by poorer tumour metrics, than Dice-based losses. The improved performance using the Mixed Focal loss reflects a better recall-precision balance, which is also robust to significant class imbalance.

There are several limitations associated with our study. Firstly, we focused our experiments on seven loss functions, only a small proportion of all the currently available loss functions. In particular, we did not include any boundary-based loss functions [19], [56], another class of loss functions that instead use distance-based metrics to optimise contours rather than distributions or regions used by cross entropy and Dice-based losses, respectively. In favour of simplicity and fairness, we also did not optimise for hyperparameters, instead either simplifying or where possible relying on prior experiments to select hyperparameter values.

We conclude by highlighting several areas for future research. In this paper, we focused on simplified variants of the Combo loss and Mixed Focal loss, and there is scope for further improvement with a more careful hyperparameter selection. The additional hyperparameter introduced from combining loss functions provides another layer of complexity, which controls the contributions of each component loss function to the total loss. Furthermore, combining other classes of loss functions, such as the boundary-based losses, may provide complementary benefit to optimisation using distribution and region-based loss functions. Finally, it will be useful to experiment using the Mixed Focal loss with more complex network architectures, to assess whether the performance gains generalise to state-of-the-art deep learning methods, and whether this is able to complement or even replace alternatives, such as training or sampling-based methods for handling class imbalance.

ACKNOWLEDGMENTS

This work was partially supported by The Mark Foundation for Cancer Research and Cancer Research UK Cambridge Centre [C9685/A25177] and the CRUK National Cancer Imaging Translational Accelerator (NCITA) [C42780/A27066]. Additional support was also provided by the National Institute of Health Research (NIHR) Cambridge Biomedical Research Centre. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health and Social Care.

CBS acknowledges support from the Leverhulme Trust project on 'Breaking the non-convexity barrier', the Philip Leverhulme Prize, the Royal Society Wolfson Fellowship, the EPSRC grants EP/S026045/1, EP/T003553/1, EP/N014588/1, EP/T017961/1, the Wellcome Innovator Award RG98755, European Union Horizon 2020 research and innovation programmes under the Marie Skodowska-Curie grant agreement No. 777826 NoMADS and No. 691070 CHiPS, the Cantab Capital Institute for the Mathematics of Information and the Alan Turing Institute.

REFERENCES

[1] Pal, N.R., Pal, S.K.: A review on image segmentation techniques. Pattern recognition **26**(9) (1993) 1277–1294
[2] Rundo, L., Militello, C., Vitabile, S., Casarino, C., Russo, G., Midiri, M., Gilardi, M.C.: Combining split-and-merge and multi-seed region growing algorithms for uterine fibroid segmentation in MRgFUS treatments. Med. Biol. Eng. Comput. **54**(7) (2016) 1071–1084
[3] Chen, X., Pan, L.: A survey of graph cuts/graph search based medical image segmentation. IEEE Rev. Biomed. Eng. **11** (2018) 112–124
[4] Khadidos, A., Sanchez, V., Li, C.T.: Weighted level set evolution based on local edge features for medical image segmentation. IEEE Trans. Image Process. **26**(4) (2017) 1979–1991
[5] Rundo, L., Militello, C., Vitabile, S., Russo, G., Sala, E., Gilardi, M.C.: A survey on nature-inspired medical image analysis: a step further in biomedical data integration. Fundam. Inform. **171**(1-4) (2020) 345–365
[6] Wang, S., Summers, R.M.: Machine learning and radiology. Med. Image Anal. **16**(5) (2012) 933–951
[7] Ren, T., Wang, H., Feng, H., Xu, C., Liu, G., Ding, P.: Study on the improved fuzzy clustering algorithm and its application in brain image segmentation. Appl. Soft Comput. **81** (2019) 105503
[8] Wachinger, C., Golland, P.: Atlas-based under-segmentation. In: Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer (2014) 315–322
[9] Ker, J., Wang, L., Rao, J., Lim, T.: Deep learning applications in medical image analysis. IEEE Access **6** (2018) 9375–9389
[10] Rueckert, D., Schnabel, J.A.: Model-based and data-driven strategies in medical image computing. Proc. IEEE **108**(1) (2019) 110–124
[11] Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer (2015) 234–241
[12] Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(12) (2017) 2481–2495
[13] Liu, L., Cheng, J., Quan, Q., Wu, F.X., Wang, Y.P., Wang, J.: A survey on U-shaped networks in medical image segmentations. Neurocomputing **409** (2020) 244–258
[14] Rundo, L., Han, C., Nagano, Y., et al.: USE-Net: incorporating squeeze-and-excitation blocks into U-Net for prostate zonal segmentation of multi-institutional MRI datasets. Neurocomputing **365** (2019) 31–43
[15] Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer (2016) 424–432
[16] Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., Rueckert, D.: Attention gated networks: learning to leverage salient regions in medical images. Med. Image Anal. **53** (2019) 197–207
[17] Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: Proc. Fourth International Conference on 3D Vision (3DV), IEEE (2016) 565–571
[18] Liu, Y., Yang, G., Hosseiny, M., Azadikhah, A., Mirak, S.A., Miao, Q., Raman, S.S., Sung, K.: Exploring uncertainty measures in bayesian deep attentive neural networks for prostate zonal segmentation. IEEE Access **8** (2020) 151817–151828
[19] Kervadec, H., Bouchtiba, J., Desrosiers, C., Granger, E., Dolz, J., Ayed, I.B.: Boundary loss for highly unbalanced segmentation. In: Proc. International Conference on Medical Imaging with Deep Learning (MIDL), PMLR (2019) 285–296
[20] Taghanaki, S.A., Zheng, Y., Zhou, S.K., Georgescu, B., Sharma, P., Xu, D., Comaniciu, D., Hamarneh, G.: Combo loss: Handling input and output imbalance in multi-organ segmentation. Comput. Med.Imaging Graph. **75** (2019) 24–33
[21] Roth, H.R., Lu, L., Farag, A., Shin, H.C., Liu, J., Turkbey, E.B., Summers, R.M.: Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In: International conference on medical image computing and computer-assisted intervention, Springer (2015) 556–564
[22] Mun, J., Jang, W., Sung, D.J., Kim, C.: Comparison of objective functions in cnn-based prostate magnetic resonance image segmentation. In: Proc. International Conference on Image Processing (ICIP), IEEE (2017) 3859–3863
[23] Litjens, G., Toth, R., van de Ven, W., Hoeks, C., Kerkstra, S., van Ginneken, B., et al.: Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge. Med. Image Anal. **18**(2) (2014) 359–373

[24] Eskildsen, S.F., Coupé, P., Fonov, V., Manjón, J.V., Leung, K.K., Guizard, N., Wassef, S.N., Østergaard, L.R., Collins, D.L., Initiative, A.D.N., et al.: BEaST: brain extraction based on nonlocal segmentation technique. NeuroImage **59**(3) (2012) 2362–2373

[25] Jadon, S.: A survey of loss functions for semantic segmentation. In: Proc. Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), IEEE (2020) 1–7

[26] Heller, N., Sathianathen, N., Kalapara, A., Walczak, E., Moore, K., Kaluzniak, H., Rosenberg, J., Blake, P., Rengel, Z., Oestreich, M., et al.: The KiTS19 challenge data: 300 kidney tumor cases with clinical context. arXiv preprint arXiv:1904.00445 (2019)

[27] Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imaging **34**(10) (2014) 1993–2024

[28] Kofler, F., Berger, C., Waldmannstetter, D., Lipkova, J., Ezhov, I., Tetteh, G., Kirschke, J., Zimmer, C., Wiestler, B., Menze, B.H.: BraTS toolkit: Translating BraTS brain tumor segmentation algorithms into clinical and scientific practice. Front. Neurosci. **14** (2020)

[29] Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P.F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S., et al.: nnU-net: Self-adapting framework for U-net-based medical image segmentation. arXiv preprint arXiv:1809.10486 (2018)

[30] Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods (2020)

[31] Fatemeh, Z., Nicola, S., Satheesh, K., Eranga, U.: Ensemble U-net-based method for fully automated detection and segmentation of renal masses on computed tomography images. Med. Phys. **47**(9) (2020) 4032–4044

[32] Kim, T., Lee, K., Ham, S., Park, B., Lee, S., Hong, D., Kim, G.B., Kyung, Y.S., Kim, C.S., Kim, N.: Active learning for accuracy enhancement of semantic segmentation with CNN-corrected label curations: Evaluation on kidney segmentation in abdominal CT. Sci. Rep. **10**(1) (2020) 1–7

[33] Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B.: Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med. Image Anal. **36** (2017) 61–78

[34] Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.M., Larochelle, H.: Brain tumor segmentation with deep neural networks. Med. Image Anal. **35** (2017) 18–31

[35] Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollar, P.: Focal loss for dense object detection. In: Proc. International Conference on Computer Vision (ICCV), IEEE (Oct 2017)

[36] Crum, W.R., Camara, O., Hill, D.L.G.: Generalized overlap measures for evaluation and validation in medical image analysis. IEEE Trans. Med. Imaging **25**(11) (2006) 1451–1461

[37] Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Cardoso, M.J.: Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer (2017) 240–248

[38] Fidon, L., Li, W., Garcia-Peraza-Herrera, L.C., Ekanayake, J., Kitchen, N., Ourselin, S., Vercauteren, T.: Generalised wasserstein dice score for imbalanced multi-class segmentation using holistic convolutional networks. In: Proc. International MICCAI Brainlesion Workshop, Springer (2017) 64–76

[39] Salehi, S.S.M., Erdogmus, D., Gholipour, A.: Tversky loss function for image segmentation using 3D fully convolutional deep networks. In: Proc. International Workshop on Machine Learning in Medical Imaging, Springer (2017) 379–387

[40] Abraham, N., Khan, N.M.: A novel focal Tversky loss function with improved attention U-Net for lesion segmentation. In: Proc. 16th International Symposium on Biomedical Imaging (ISBI), IEEE (2019) 683–687

[41] Chen, W., Zhang, Y., He, J., Qiao, Y., Chen, Y., Shi, H., Wu, E.X., Tang, X.: Prostate segmentation using 2D bridged U-net. In: Proc. International Joint Conference on Neural Networks (IJCNN), IEEE (2019) 1–7

[42] Zhu, W., Huang, Y., Zeng, L., Chen, X., Liu, Y., Qian, Z., Du, N., Fan, W., Xie, X.: AnatomyNet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy. Med. Phys. **46**(2) (2019) 576–589

[43] Müller, D., Kramer, F.: MIScnn: A framework for medical image segmentation with convolutional neural networks and deep learning. arXiv preprint arXiv:1910.09308 (2019)

[44] Linguraru, M.G., Yao, J., Gautam, R., Peterson, J., Li, Z., Linehan, W.M., Summers, R.M.: Renal tumor quantification and classification in contrast-enhanced abdominal CT. Pattern Recognit. **42**(6) (2009) 1149–1161

[45] Rundo, L., Beer, L., Ursprung, S., Martin-Gonzalez, P., Markowetz, F., Brenton, J.D., Crispin-Ortuzar, M., Sala, E., Woitek, R.: Tissue-specific and interpretable sub-segmentation of whole tumour burden on CT images by unsupervised fuzzy clustering. Comput. Biol. Med. **120** (2020) 103751

[46] Heller, N., Isensee, F., Maier-Hein, K.H., Hou, X., Xie, C., Li, F., Nan, Y., Mu, G., Lin, Z., Han, M., et al.: The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the KiTS19 challenge. Med. Image Anal. **67** (2021) 101821

[47] Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci. Data **4** (2017) 170117

[48] Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R.T., Berger, C., Ha, S.M., Rozycki, M., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. arXiv preprint arXiv:1811.02629 (2018)

[49] Rundo, L., Tangherloni, A., Cazzaniga, P., Nobile, M.S., Russo, G., Gilardi, M.C., et al.: A novel framework for MR image segmentation and quantification by using MedGA. Comput. Methods Programs Biomed. **176** (2019) 159–172

[50] Han, C., Rundo, L., Araki, R., Nagano, Y., Furukawa, Y., et al.: Combining noise-to-image and image-to-image GANs: brain MR image augmentation for tumor detection. IEEE Access **7**(1) (2019) 156966–156977

[51] Henry, T., Carre, A., Lerousseau, M., Estienne, T., Robert, C., Paragios, N., Deutsch, E.: Top 10 BraTS 2020 challenge solution: Brain tumor segmentation with self-ensembled, deeply-supervised 3D-Unet like neural networks. arXiv preprint arXiv:2011.01045 (2020)

[52] Rundo, L., Stefano, A., Militello, C., Russo, G., Sabini, M.G., D'Arrigo, C., Marletta, F., Ippolito, M., Mauri, G., Vitabile, S., Gilardi, M.C.: A fully automatic approach for multimodal PET and MR image segmentation in Gamma Knife treatment planning. Comput. Methods Programs Biomed. **144** (2017) 77–96

[53] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint `arXiv:1412.6980` (2014)

[54] Wang, Z., Wang, E., Zhu, Y.: Image segmentation evaluation: a survey of methods. Artif. Intell. Rev. **53**(8) (2020) 5637–5674

[55] Holm, S.: A simple sequentially rejective multiple test procedure. Scand. J. Statist. **6**(2) (1979) 65–70

[56] Zhu, Q., Du, B., Yan, P.: Boundary-weighted domain adaptive neural network for prostate mr image segmentation. IEEE Trans. Med. Imaging **39**(3) (2019) 753–763

MICHAEL YEUNG received his Bachelor's degree in Neuroscience in 2019 from the University of Cambridge, United Kingdom. He is currently a medical student at the School of Clinical Medicine, University of Cambridge, United Kingdom. He is a Senior Whitby Scholar at Downing College. His research interests include Machine learning, Radiology and Computer Vision.

EVIS SALA received her medical degree from University of Tirana, Albania, in 1991 and her PhD Degree in Epidemiology and Biostatistics from the Cambridge University, UK, in 2000. Currently, she is the Professor of Oncological Imaging at Cambridge University, UK, and co-leads the Advanced Cancer Imaging and the Integrated Cancer Medicine Programmes for the CRUK Cambridge Centre. Her research in the new field of radiogenomics has focused on understanding the molecular basis of cancer by demonstrating the phenotypic patterns that occur because of multiple genetic alterations that interact with the tumour microenvironment to drive the disease in several tumour types. She is also leading multiple research projects focusing on the applications of artificial intelligence methods for image reconstruction, segmentation, and data integration.

CAROLA-BIBIANE SCHÖNLIEB graduated from the Institute for Mathematics, University of Salzburg, Austria, in 2004. She received her PhD degree from the University of Cambridge in 2009. Currently, she is Professor of Applied Mathematics at the Department of Applied Mathematics and Theoretical Physics (DAMTP), University of Cambridge, United Kingdom. There, she is head of the Cambridge Image Analysis group, Director of the Cantab Capital Institute for Mathematics of Information, and co-Director of the EPSRC Centre for Mathematics of Information in Healthcare. Since 2011 she is a fellow of Jesus College Cambridge and since 2016 a fellow of the Alan Turing Institute, London. Her current research interests focus on variational methods, partial differential equations and machine learning for image analysis, image processing and inverse imaging problems.

LEONARDO RUNDO received the Bachelor's and Master's Degrees in Computer Science Engineering from the University of Palermo, Italy, in 2010 and 2013, respectively. In 2013, he was a Research Fellow at the Institute of Molecular Bioimaging and Physiology, National Research Council of Italy (IBFM-CNR). He obtained his PhD in Computer Science at the University of Milano-Bicocca, Italy, in 2019. Since November 2018, he is a Research Associate at the Department of Radiology, University of Cambridge, United Kingdom, tightly collaborating with Cancer Research UK. His main scientific interests include Biomedical Image Analysis, Machine Learning, Computational Intelligence, and High-Performance Computing.