Imbalance-aware loss functions improve medical image classification

Daniel Scholz[1,2] DANIEL.SCHOLZ@MRI.TUM.DE
[1] *Department of Neuroradiology, Technical University of Munich*
[2] *Institute for Artificial Intelligence and Informatics in Medicine, Technical University of Munich*

Ayhan Can Erdur[2,3] CAN.ERDUR@TUM.DE
[3] *Department of Radiation Oncology, Technical University of Munich*

Josef Buchner[3] J.BUCHNER@TUM.DE
Jan C. Peeken[3] JAN.PEEKEN@TUM.DE
Daniel Rueckert[*2] DANIEL.RUECKERT@TUM.DE
Benedikt Wiestler[*1] B.WIESTLER@TUM.DE

Editors: Accepted for publication at MIDL 2024

Abstract

Deep learning models offer unprecedented opportunities for diagnosis, prognosis, and treatment planning. However, conventional deep learning pipelines often encounter challenges in learning unbiased classifiers within imbalanced data settings, frequently exhibiting bias towards minority classes. In this study, we aim to improve medical image classification by effectively addressing class imbalance. To this end, we employ differentiable loss functions derived from classification metrics commonly used in imbalanced data settings: Matthews correlation coefficient (MCC) and the F1 score. We explore the efficacy of these loss functions both independently and in combination with cross-entropy loss and various batch sampling strategies on diverse medical datasets of 2D fundoscopy and 3D magnetic resonance images. Our findings demonstrate that, compared to conventional loss functions, we achieve notable improvements in overall classification performance, with increases of up to +12% in balanced accuracy and up to +51% in class-wise F1 score for minority classes when utilizing cross-entropy coupled with metrics-derived loss. Additionally, we conduct feature visualization to gain insights into the behavior of these features during training with imbalance-aware loss functions. Our visualization reveals a more pronounced clustering of minority classes in the feature space, consistent with our classification results. Our results underscore the effectiveness of combining cross-entropy loss with class-imbalance-aware loss functions in training more accurate classifiers, particularly for minority classes.

Keywords: Class imbalance, Deep learning, Loss Function, Unbiased Classifier

1. Introduction

Deep learning techniques have revolutionized medical image analysis by providing powerful tools for tasks such as diagnosis, prognosis, and treatment planning (Litjens et al., 2017; Oren et al., 2020; Pinto-Coelho, 2023). However, one significant challenge that persists in training deep learning models for medical image analysis is the presence of imbalanced data (Mazurowski et al., 2008; Buda et al., 2018; Johnson and Khoshgoftaar, 2019). In many

[*] Contributed equally as senior authors

medical datasets, the distribution of classes is often skewed, with certain classes representing the minority while others dominate. This class imbalance poses a substantial hurdle for conventional deep learning pipelines, as they tend to prioritize learning the majority classes at the expense of the minority ones. Consequently, models trained on imbalanced data may exhibit biased predictions, leading to suboptimal performance, particularly for the underrepresented classes critical for accurate diagnosis or prognosis. (Cluceru et al., 2022; Foltyn-Dumitru et al., 2023). The resulting bias against minority classes can have potentially grave consequences for patients when utilizing detection and diagnosis systems based on such biased deep-learning classifiers. This problem is further aggravated since commonly used metrics, such as unbalanced accuracy, convey overly optimistic results in imbalanced classification settings (Haixiang et al., 2017). More appropriate metrics for evaluating classification performance in imbalanced data settings, such as the Matthews correlation coefficient (MCC)(Matthews, 1975; Gorodkin, 2004) and F1 score, have been proposed.

Addressing class imbalance is crucial for developing robust and reliable deep learning models that can generalize well across diverse medical imaging datasets and yield clinically relevant insights. In this study, we focus on mitigating the effects of class imbalance in medical image classification tasks using novel approaches derived from imbalance-aware loss functions, aiming to improve the overall performance and equity across all classes. To this end, we comprehensively compare and analyze class-imbalance-aware loss functions in combination with and against established loss functions in two challenging datasets with imbalanced class distributions. In summary, our contributions are as follows:

1. We investigate different loss formulations and introduce **new combinations of class-imbalance-aware losses** by integrating the MCC and F1 score with a cross-entropy loss.

2. To this end, we **comprehensively compare** different strategies to address class imbalance, namely, class-imbalance-aware loss functions in combination with over-sampling and per-sample weighting, and established loss functions.

3. We experimentally show that **class-imbalance-aware loss functions increase the performance** of challenging classification tasks on diverse medical imaging datasets, as well as the **discernability of class representations**.

2. Related Work

2.1. Overcoming class imbalance

Considering how relevant a challenge class imbalance is for training clinically applicable deep learning models, several strategies have been developed to overcome this challenge. These can be broadly grouped into (i) sampling-based, (ii) loss-based, and (iii) synthesis-based approaches.

Sampling-based approaches typically aim to oversample the minority class(es) or adjust loss weights. While easy to implement and computationally cheap, this can lead to the model severely overfitting the few samples available in the minority classes (Zheng et al., 2015).

Alternatively, specific *loss functions* such as focal loss (Lin et al., 2017) may be used. Focal loss over-proportionally decreases the loss of an easy sample compared to difficult ones. Since the samples in the minority class are potentially more challenging to classify due to their low prevalence, the model is penalized strongly for misclassifying them. However, the focal formulation does not directly address the class imbalance but rather its "side effect" that minority samples are typically more challenging to classify.

Another strategy to tackle class imbalance is creating *synthetic examples*. Earlier examples of such techniques include the Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002). In SMOTE, the minority class is over-sampled by taking each sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors. Linear interpolation in image space, however, rarely gives sensible synthetic samples. Hence, generative models, such as generative adversarial networks (GANs) (Goodfellow et al., 2014; Qasim et al., 2020; Li et al., 2023) or diffusion models (Qin et al., 2023; Dhariwal and Nichol, 2021) have been developed to create realistic examples to supplement the minority classes. While these are powerful approaches, they are also computationally expensive and require considerable effort to develop generative models that produce realistic, helpful minority class examples, which is again challenging. Extending them to different tasks and classes usually requires extensive re-training.

3. Method

3.1. Class-imbalance-aware loss functions

To mitigate the imbalanced data issue in medical imaging datasets, imbalance-aware loss functions emerge to enhance the performance of minority classes. These loss functions generally assign larger loss values to misclassified instances of these less prevalent classes. This adaptation serves to rectify the disparity in their impact on the overall loss calculation. We compare the focal loss with two loss functions derived from the MCC and the F1 score.

3.1.1. Focal Loss

Focal Loss (Lin et al., 2017) is an often-used loss function for imbalanced deep-learning classification problems. Difficult-to-classify examples often stem from minority classes. These examples are often predicted with low confidence, yielding higher loss values. Hence, the deep learning model is incentivized to optimize for all classes equally. The loss function is given as

$$\mathcal{L}_{\text{focal}}(p_t) = -(1 - p_t)^{\gamma} \log(p_t) \tag{1}$$

where the exponent γ determines the strength of penalization for samples of class t with predicted probability p_t.

3.1.2. Soft F1 Loss

The F1 score is a valuable metric for assessing classification performance since it summarizes precision and recall into a single number through the harmonic mean. By macro-averaging the F1 score for all classes, we obtain a balanced assessment of the classifier. We leverage

this property by deriving a negative differentiable F1 score as a loss function. To this end, we use differentiable true positives (TP), false positives (FP), and false negatives (FN):

$$TP = \sum_{i \in I} y_i \cdot \hat{y}_i; \quad FP = \sum_{i \in I} (1 - y_i) \cdot \hat{y}_i; \quad FN = \sum_{i \in I} y_i \cdot (1 - \hat{y}_i) \tag{2}$$

where y_i is the label and $\hat{y}_i$ is the prediction for index i. The precision, recall, and F1 score are defined as:

$$\begin{aligned} \text{precision} = \frac{TP}{TP + FP}; \quad \text{recall} = \frac{TP}{TP + FN} \\ F1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \end{aligned} \tag{3}$$

We define the corresponding loss function as $\mathcal{L}_{F1} = 1 - F1_{\text{soft}}$, where $F1_{\text{soft}}$ is the macro average of the F1 score for each class using differentiable (*soft*) TP, FP, and FN.

3.1.3. Soft MCC Loss

Matthew's correlation coefficient (MCC) (Matthews, 1975) is a metric that encompasses all four entries of the confusion matrix, namely TP, TN, FP, and FN, into a single value in the binary classification case. It has been argued that the MCC is superior to many other metrics, such as accuracy, F1 score, and the receiver operating characteristic (ROC) area under the curve (AUC) (Chicco and Jurman, 2020; Chicco et al., 2021b,a; Chicco and Jurman, 2023), because of the normalization term accounting for class imbalance. We define (*soft*) TP, FP, and FN as above and additionally calculate true negative (TN):

$$TN = \sum_{i \in I} (1 - y_i) \cdot \hat{y}_i \tag{4}$$

From these definitions, MCC_{soft} is defined as:

$$MCC_{soft} = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{5}$$

The loss formulation is given as: $\mathcal{L}_{MCC} = 1 - MCC_{soft}$ (Abhishek and Hamarneh, 2021).

3.1.4. Combined loss functions

The established cross-entropy loss has desirable theoretical properties. The imbalance-aware losses presented might focus too heavily on the minority class, leading to an unwanted decrease in performance for the majority class. Hence, we also evaluate weighted sums of the F1 and the MCC loss with the cross-entropy loss with equal weights for each loss term.

3.2. Addressing class imbalance with sampling and weighting

In balanced data scenarios, the samples in a batch are drawn uniformly, i.e., with the same probability, to obtain an equal uniform distribution of each class in a batch.

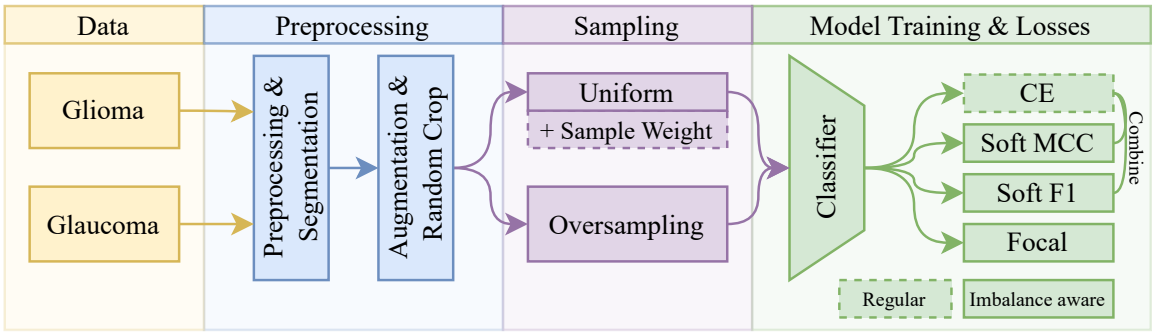

Figure 1: Overview of our study design. We systematically investigate combinations of different sampling strategies, sample weightings, and loss functions (including class-imbalance-aware loss functions).

Oversampling One measure to combat class imbalance in deep learning is oversampling the minority class(es) to obtain a more uniform distribution than the original distribution in the dataset. We implement a stratified oversampling technique, i.e., we allocate equal portions of a batch to each class corresponding to equal clinical relevance of each type.

Sample weighting Instead of oversampling the minority classes, we can assign higher importance to minority class samples by scaling the influence of a sample according to the prevalence of its class $c \in C$ in the loss calculation:

$$L = \frac{1}{n} \sum_{i} w_c^{(i)} L^{(i)} \tag{6}$$

We choose a normalized inverse frequency (Sparck Jones, 1972) scaling for each sample.

4. Experiment Setup

4.1. Experiment pipeline

An overview of our experimental design is shown in Figure 1. We conduct our experiments on two diverse and imbalanced datasets. We apply standard preprocessing and augmentation techniques. Two sampling strategies are used to train a ResNet classifier (He et al., 2016). We compare six loss functions, cross-entropy (CE), focal loss, soft F1, soft MCC, and CE with MCC and F1. For all experiments using the cross-entropy-loss we also compare sample weighting while sampling uniformly.

We share our dataset configurations and the code used for our study at `https://github.com/daniel-scholz/address-class-imbalance`. For further details of our experiments implementation, refer to Appendix B.

4.2. Datasets

To allow the reproduction of our results, our study only uses publicly available datasets.

Glioma The first dataset comprises 3D MR images (T1w -/+ contrast, T2w, FLAIR) from large public datasets of adult patients with newly diagnosed gliomas, namely UCSF-PDGM (Calabrese et al., 2022), EGD (van der Voort et al., 2021), and TCGA (Bakas et al., 2017). Besides having all four imaging sequences outlined above available, we require biomarker testing for *IDH* mutation and 1p/19q status in order to classify samples according to the 2021 WHO classification of brain tumors into (a) *IDH* wildtype glioblastoma, (b) *IDH* mutant and 1p/19q intact astrocytoma and (c) *IDH* mutant and 1p/19q codeleted oligodendroglioma (Wen and Packer, 2021). In total, our dataset contains preoperative MRIs of 1174 patients. The prevalence of glioblastoma (∼80%) in comparison to oligodendroglioma (∼8%) and astrocytoma (∼12%) is striking and consistent throughout all the available datasets, mirroring the real-world distribution. A visualization of the class distributions is shown in the Appendix C, Figure 5. We hold the TCGA dataset out for testing and use the remaining data for training. For additional robustness analysis, we run each experiment with four different network initializations.

Glaucoma The second dataset consists of 1542 individual 2D RGB fundus photographs, of which 786 are healthy controls, 289 photographs show early glaucoma, and 467 are from advanced glaucoma patients (Ahn et al., 2018). We randomly split the dataset into training ($\frac{3}{4}$) and testing ($\frac{1}{4}$) data, stratified by class: {no, early, advanced} glaucoma.

5. Results

5.1. Global classification results

Our main results on the test sets are shown in Table 1 (and Appendix A, Table 3, 4, and 5). The baseline model (CE loss, uniform sampling) shows reasonable performance, which improves when adding sample weights or oversampling the minority classes for both datasets. The most considerable improvement over the baseline qua balanced accuracy is achieved when using the CE + soft F1 loss combination and oversampling, with a relative improvement of 12.7% on the glioma dataset. We also observe smaller standard deviations for the loss combinations compared to the single class-imbalance-aware losses, F1 and MCC, indicating better training stability for the combination. We explore this further in an ablation study (Appendix A, Table 6) for small batch size regimes. The performance on the glaucoma dataset also improves most with the CE + F1 loss combination and oversampling (+10.5%).

5.2. Per-class analysis

In addition, we perform a per-class analysis of our methods using the class-wise F1 score, which balances precision and recall (Table 2). The baseline training setup yields a classifier biased towards the majority class (glioblastoma / no glaucoma) while performing poorly on the minority classes. The overall improvements in classification performance can be directly traced to improved minority class performance, since majority class performance stays constant across almost all experiments. The CE + F1 loss tremendously improves classification performance on the astrocytoma minority class (+20.3%). The largest improvement when using CE + F1 with oversampling loss is observed on the least prevalent early glaucoma (+51.3%). However, we also observe that using only a class imbalance-aware loss sometimes yields classifiers entirely ignoring one class (e.g., F1 loss in the glaucoma).

Table 1: Multi-class classification results for six different loss functions, two sampling methods, and two different medical imaging datasets in terms of **balanced accuracy**(↑). Results for the glioma dataset are mean$_{\pm \text{std}}$.

Dataset	Sampling	CE	CE + F1	CE + MCC	F1	Focal	MCC
Glioma	Uniform	$0.55_{\pm 0.03}$	$0.57_{\pm 0.02}$	$0.59_{\pm 0.03}$	$0.51_{\pm 0.12}$	$0.57_{\pm 0.04}$	$0.39_{\pm 0.10}$
	+ Weights	$0.59_{\pm 0.03}$	$0.60_{\pm 0.02}$	$\mathbf{0.62}_{\pm 0.03}$	-	-	-
	Oversampling	$0.61_{\pm 0.01}$	$\mathbf{0.62}_{\pm 0.01}$	$0.59_{\pm 0.04}$	$0.60_{\pm 0.03}$	$0.61_{\pm 0.04}$	$0.59_{\pm 0.03}$
Glaucoma	Uniform	0.69	0.70	0.72	0.59	0.70	0.59
	+ Weights	0.74	0.76	0.75	-	-	-
	Oversampling	0.75	**0.77**	0.76	0.58	0.74	0.50

Table 2: Comparison for different class performances in terms of **F1 score** (↑). The majority class performance has little variance while the minority class performance improves tremendously for class imbalance aware losses and cross-entropy with oversampling (Astro: astrocytoma, GB: glioblastoma, Oligo: oligodendroglioma).

Loss	Sampling	Glaucoma			Glioma		
		No	Early	Advanced	GB	Astro	Oligo
CE	Uniform	**0.85**	0.37	0.83	$0.87_{\pm 0.00}$	$0.54_{\pm 0.06}$	$0.30_{\pm 0.06}$
	+ Weights	0.80	0.54	0.81	$0.88_{\pm 0.01}$	$0.63_{\pm 0.02}$	$0.30_{\pm 0.06}$
	Oversampling	0.79	0.55	0.79	$0.88_{\pm 0.01}$	$0.63_{\pm 0.01}$	$0.34_{\pm 0.04}$
CE + F1	Uniform	**0.85**	0.42	0.81	$0.87_{\pm 0.01}$	$0.60_{\pm 0.04}$	$0.30_{\pm 0.05}$
	+ Weights	0.80	**0.56**	0.84	$0.88_{\pm 0.01}$	$0.60_{\pm 0.02}$	$0.34_{\pm 0.03}$
	Oversampling	0.81	0.54	**0.85**	$\mathbf{0.89}_{\pm 0.00}$	$0.65_{\pm 0.03}$	$0.34_{\pm 0.00}$
CE + MCC	Uniform	**0.85**	0.44	0.83	$0.87_{\pm 0.01}$	$0.60_{\pm 0.03}$	$\mathbf{0.38}_{\pm 0.11}$
	+ Weights	0.78	0.55	0.83	$\mathbf{0.89}_{\pm 0.01}$	$0.62_{\pm 0.06}$	$\mathbf{0.38}_{\pm 0.04}$
	Oversampling	0.82	0.53	0.83	$0.88_{\pm 0.01}$	$0.59_{\pm 0.06}$	$0.32_{\pm 0.07}$
F1	Uniform	0.81	0.00	0.74	$0.85_{\pm 0.04}$	$0.52_{\pm 0.35}$	$0.07_{\pm 0.15}$
	Oversampling	0.81	0.00	0.70	$0.89_{\pm 0.00}$	$0.66_{\pm 0.02}$	$0.25_{\pm 0.07}$
Focal	Uniform	0.84	0.44	0.83	$0.87_{\pm 0.02}$	$0.63_{\pm 0.03}$	$0.25_{\pm 0.11}$
	Oversampling	0.80	0.52	0.81	$0.87_{\pm 0.02}$	$0.62_{\pm 0.08}$	$0.35_{\pm 0.04}$
MCC	Uniform	0.82	0.05	0.73	$0.81_{\pm 0.04}$	$0.17_{\pm 0.33}$	$0.00_{\pm 0.00}$
	Oversampling	0.00	0.32	0.72	$0.88_{\pm 0.01}$	$\mathbf{0.70}_{\pm 0.01}$	$0.15_{\pm 0.14}$

5.3. Visual feature space analysis

To investigate the learned representations, we plot the features of the last ResNet layer for the glioma dataset. We use the popular t-distributed stochastic neighbor embedding (tSNE) (Van der Maaten and Hinton, 2008) to project the 256-dimensional feature vectors to 2D for visualization purposes (see Figure 2). We observe that representations of the oligodendroglioma are often poorly clustered in this feature space, corresponding to the inferior performance in this class observed in Table 2. CE + MCC loss with uniform sampling

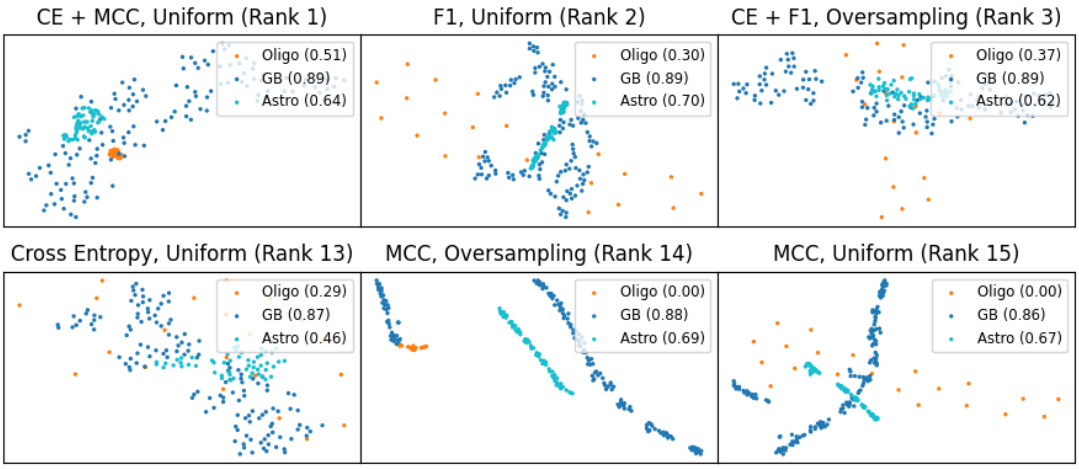

Figure 2: tSNE visualization of the feature representations before the last layer of the best and worst models (1st run) according to the F1 score (↑) with class-wise F1 scores.

shows a better clustering of the oligodendroglioma features compared to the baseline or the MCC-only loss.

6. Discussion and Conclusion

Class imbalance-aware loss functions relevantly enhance classification performance, primarily by improving recognition of minority classes. Notably, the most substantial enhancements are observed when combining F1 or MCC loss with standard cross-entropy loss. In particular, this combination seems to stabilize training. This points to similarities between classification and segmentation methodologies, exemplified in frameworks like nnUNet (Isensee et al., 2021), where the combination of Dice (which essentially is the F1 score) and cross-entropy is found to perform best, and underscores the effectiveness of such hybrid approaches. Further, it's worth noting that when employing only class imbalance-aware loss functions, there can be instances where certain classes may be somewhat neglected, a scenario not encountered in the combined approach. In summary, the demonstrated efficacy of class imbalance-aware loss functions, alongside their ease of implementation, computational efficiency, and adaptability across various medical imaging tasks, highlights their potential impact on advancing clinical applications and enhancing the accuracy and reliability of deep learning-based diagnostics in real-world healthcare settings. These properties call for future studies exploring more scenarios in medical image classification with class-imbalance-aware loss functions such as different network architectures or time-series data. Ultimately, we show that integrating loss functions derived from popular metrics such as the F1 score and the MCC with standard cross-entropy loss results in more robust classifiers, with particular benefits for the minority class, thus underscoring its clinical relevance.

Acknowledgments

This study was supported by the DFG, grant #504320104.

References

Kumar Abhishek and Ghassan Hamarneh. Matthews correlation coefficient loss for deep convolutional networks: Application to skin lesion segmentation. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 225–229. IEEE, 2021.

Jin Mo Ahn, Sangsoo Kim, Kwang-Sung Ahn, Sung-Hoon Cho, Kwan Bok Lee, and Ungsoo Samuel Kim. A deep learning model for the detection of both advanced and early glaucoma using fundus photography. *PLOS ONE*, 13(11):e0207982, November 2018. ISSN 1932-6203. doi: 10.1371/journal.pone.0207982.

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S. Kirby, John B. Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific Data*, 4(1):170117, September 2017. ISSN 2052-4463. doi: 10.1038/sdata.2017.117.

Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, October 2018. ISSN 0893-6080. doi: 10.1016/j.neunet.2018.07.011.

Evan Calabrese, Javier E. Villanueva-Meyer, Jeffrey D. Rudie, Andreas M. Rauschecker, Ujjwal Baid, Spyridon Bakas, Soonmee Cha, John T. Mongan, and Christopher P. Hess. The University of California San Francisco Preoperative Diffuse Glioma MRI Dataset. *Radiology: Artificial Intelligence*, 4(6):e220058, November 2022. doi: 10.1148/ryai.220058.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357, June 2002. ISSN 1076-9757. doi: 10.1613/jair.953.

Davide Chicco and Giuseppe Jurman. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1):6, January 2020. ISSN 1471-2164. doi: 10.1186/s12864-019-6413-7.

Davide Chicco and Giuseppe Jurman. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Mining*, 16(1):4, February 2023. ISSN 1756-0381. doi: 10.1186/s13040-023-00322-4.

Davide Chicco, Niklas Tötsch, and Giuseppe Jurman. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Mining*, 14(1):13, February 2021a. ISSN 1756-0381. doi: 10.1186/s13040-021-00244-z.

Davide Chicco, Matthijs J. Warrens, and Giuseppe Jurman. The Matthews Correlation Coefficient (MCC) is More Informative Than Cohen's Kappa and Brier Score in Binary

Classification Assessment. *IEEE Access*, 9:78368–78381, 2021b. ISSN 2169-3536. doi: 10.1109/ACCESS.2021.3084050.

Julia Cluceru, Yannet Interian, Joanna J. Phillips, Annette M. Molinaro, Tracy L. Luks, Paula Alcaide-Leon, Marram P. Olson, Devika Nair, Marisa LaFontaine, Anny Shai, Pranathi Chunduru, Valentina Pedoia, Javier E. Villanueva-Meyer, Susan M. Chang, and Janine M. Lupo. Improving the noninvasive classification of glioma genetic subtype with deep learning and diffusion-weighted imaging. *Neuro-Oncology*, 24(4):639–652, April 2022. ISSN 1523-5866. doi: 10.1093/neuonc/noab238.

Prafulla Dhariwal and Alexander Nichol. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794. Curran Associates, Inc., 2021.

Martha Foltyn-Dumitru, Marianne Schell, Aditya Rastogi, Felix Sahm, Tobias Kessler, Wolfgang Wick, Martin Bendszus, Gianluca Brugnara, and Philipp Vollmuth. Impact of signal intensity normalization of MRI on the generalizability of radiomic-based prediction of molecular glioma subtypes. *European Radiology*, September 2023. ISSN 1432-1084. doi: 10.1007/s00330-023-10034-2.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

J. Gorodkin. Comparing two K-category assignments by a K-category correlation coefficient. *Computational Biology and Chemistry*, 28(5):367–374, December 2004. ISSN 1476-9271. doi: 10.1016/j.compbiolchem.2004.09.006.

Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73:220–239, May 2017. ISSN 0957-4174. doi: 10.1016/j.eswa.2016.12.035.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

Justin M. Johnson and Taghi M. Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):27, December 2019. ISSN 2196-1115. doi: 10.1186/s40537-019-0192-5.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Florian Kofler, Christoph Berger, Diana Waldmannstetter, Jana Lipkova, Ivan Ezhov, Giles Tetteh, Jan Kirschke, Claus Zimmer, Benedikt Wiestler, and Bjoern H. Menze. BraTS Toolkit: Translating BraTS Brain Tumor Segmentation Algorithms Into Clinical and Scientific Practice. *Frontiers in Neuroscience*, 14:125, 2020. ISSN 1662-4548. doi: 10.3389/fnins.2020.00125.

Wei Li, Jinlin Chen, Jiannong Cao, Chao Ma, Jia Wang, Xiaohui Cui, and Ping Chen. EID-GAN: Generative Adversarial Nets for Extremely Imbalanced Data Augmentation. *IEEE Transactions on Industrial Informatics*, 19(3):3208–3218, March 2023. ISSN 1941-0050. doi: 10.1109/TII.2022.3182781.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, December 2017. ISSN 1361-8423. doi: 10.1016/j.media.2017.07.005.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic Gradient Descent with Warm Restarts. In *International Conference on Learning Representations*, 2016.

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, 2015.

B. W. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451, October 1975. ISSN 0005-2795. doi: 10.1016/0005-2795(75)90109-9.

Maciej A. Mazurowski, Piotr A. Habas, Jacek M. Zurada, Joseph Y. Lo, Jay A. Baker, and Georgia D. Tourassi. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural Networks*, 21(2):427–436, March 2008. ISSN 0893-6080. doi: 10.1016/j.neunet.2007.12.031.

Ohad Oren, Bernard J. Gersh, and Deepak L. Bhatt. Artificial intelligence in medical imaging: Switching from radiographic pathological data to clinically meaningful endpoints. *The Lancet Digital Health*, 2(9):e486–e488, September 2020. ISSN 2589-7500. doi: 10.1016/S2589-7500(20)30160-6.

Luís Pinto-Coelho. How Artificial Intelligence Is Shaping Medical Imaging Technology: A Survey of Innovations and Applications. *Bioengineering*, 10(12):1435, December 2023. ISSN 2306-5354. doi: 10.3390/bioengineering10121435.

Ahmad B. Qasim, Ivan Ezhov, Suprosanna Shit, Oliver Schoppe, Johannes C. Paetzold, Anjany Sekuboyina, Florian Kofler, Jana Lipkova, Hongwei Li, and Bjoern Menze. Red-GAN: Attacking class imbalance via conditioned generation. Yet another medical imaging perspective. In *Proceedings of the Third Conference on Medical Imaging with Deep Learning*, pages 655–668. PMLR, September 2020.

Yiming Qin, Huangjie Zheng, Jiangchao Yao, Mingyuan Zhou, and Ya Zhang. Class-Balancing Diffusion Models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18434–18443, Vancouver, BC, Canada, June 2023. IEEE. ISBN 9798350301298. doi: 10.1109/CVPR52729.2023.01768.

Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11), 2008.

Sebastian R. van der Voort, Fatih Incekara, Maarten M. J. Wijnenga, Georgios Kapsas, Renske Gahrmann, Joost W. Schouten, Hendrikus J. Dubbink, Arnaud J. P. E. Vincent, Martin J. van den Bent, Pim J. French, Stefan Klein, and Marion Smits. The Erasmus Glioma Database (EGD): Structural MRI scans, WHO 2016 subtypes, and segmentations of 774 patients with glioma. *Data in Brief*, 37:107191, August 2021. ISSN 2352-3409. doi: 10.1016/j.dib.2021.107191.

Stéfan van der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart, Tony Yu, and the scikit-image contributors. Scikit-image: Image processing in Python. *PeerJ*, 2:e453, June 2014. ISSN 2167-8359. doi: 10.7717/peerj.453.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Patrick Y. Wen and Roger J. Packer. The 2021 WHO Classification of Tumors of the Central Nervous System: Clinical implications. *Neuro-Oncology*, 23(8):1215, August 2021. doi: 10.1093/neuonc/noab120.

Zhuoyuan Zheng, Yunpeng Cai, and Ye Li. Oversampling method for imbalanced classification. *Computing and Informatics*, 34(5):1017–1037, 2015.

Appendix A. Further results

A.1. Additional metrics

Table 3: Multi-class classification results for two different medical imaging datasets in terms of **F1 score**(↑). Results for the glioma dataset are $\text{mean}_{\pm\text{std}}$.

Dataset	Sampling	CE	CE + F1	CE + MCC	F1	Focal	MCC
Glioma	Uniform	$0.57_{\pm 0.03}$	$0.59_{\pm 0.02}$	$0.62_{\pm 0.04}$	$0.48_{\pm 0.16}$	$0.59_{\pm 0.05}$	$0.32_{\pm 0.12}$
	+ Weights	$0.60_{\pm 0.03}$	$0.61_{\pm 0.02}$	$\mathbf{0.63}_{\pm 0.02}$	-	-	-
	Oversampling	$0.62_{\pm 0.01}$	$0.62_{\pm 0.01}$	$0.59_{\pm 0.04}$	$0.60_{\pm 0.03}$	$0.61_{\pm 0.04}$	$0.58_{\pm 0.05}$
Glaucoma	Uniform	0.71	**0.73**	**0.73**	0.52	0.70	0.53
	+ Weights	0.68	0.69	0.71	-	-	-
	Oversampling	0.72	**0.73**	0.72	0.50	0.71	0.35

Table 4: Multi-class classification results in terms of **macro-averaged Area under the ROC Curve**(↑). Results for the glioma dataset are $\text{mean}_{\pm\text{std}}$.

Dataset	Sampling	CE	CE + F1	CE + MCC	F1	Focal	MCC
Glioma	Uniform	$0.80_{\pm 0.02}$	$0.81_{\pm 0.01}$	$0.81_{\pm 0.02}$	$0.74_{\pm 0.12}$	$0.79_{\pm 0.03}$	$0.57_{\pm 0.13}$
	+ Weights	$0.80_{\pm 0.01}$	$0.81_{\pm 0.02}$	$\mathbf{0.82}_{\pm 0.01}$	-	-	-
	Oversampling	$0.81_{\pm 0.01}$	$\mathbf{0.82}_{\pm 0.02}$	$0.81_{\pm 0.02}$	$0.77_{\pm 0.03}$	$0.79_{\pm 0.02}$	$0.76_{\pm 0.01}$
Glaucoma	Uniform	0.90	0.89	0.90	0.75	0.89	0.78
	+ Weights	0.90	0.90	0.90	-	-	-
	Oversampling	0.90	0.90	0.89	0.79	0.88	0.57

Table 5: Multi-class classification results for different medical imaging datasets in terms of **Matthews correlation coefficient (MCC)**(↑). Results for the glioma dataset are $\text{mean}_{\pm\text{std}}$.

Dataset	Sampling	CE	CE + F1	CE + MCC	F1	Focal	MCC
Glioma	Uniform	$0.47_{\pm 0.03}$	$0.51_{\pm 0.03}$	$0.52_{\pm 0.02}$	$0.41_{\pm 0.27}$	$0.51_{\pm 0.05}$	$0.13_{\pm 0.25}$
	+ Weights	$0.52_{\pm 0.03}$	$0.52_{\pm 0.02}$	$0.54_{\pm 0.05}$	-	-	-
	Oversampling	$0.53_{\pm 0.02}$	$\mathbf{0.55}_{\pm 0.01}$	$0.50_{\pm 0.04}$	$0.54_{\pm 0.02}$	$0.52_{\pm 0.05}$	$0.56_{\pm 0.03}$
Glaucoma	Uniform	0.63	0.64	**0.65**	0.53	0.63	0.53
	+ Weights	0.62	0.64	0.63	-	-	-
	Oversampling	0.61	0.64	0.64	0.52	0.61	0.27

A.2. Ablation: Small batch size regimes

To better understand the benefits of combining a class-imbalance aware loss function with cross-entropy loss, we performed an experiment in a small batch size regime, which is commonly found in 3D medical image analysis. The results from three independent runs are shown in Table 6. Noteworthy, we observe clearly lower standard deviations when combining soft MCC with cross-entropy loss, indicating a stabilizing effect of combining both losses.

Table 6: Comparison of soft MCC loss with and without additional cross-entropy loss in small batch size regimes (batch size = 10).

Loss	Sampling	Balanced Accuracy	MCC	AUC
MCC	Uniform	$0.33_{\pm 0.00}$	$0.00_{\pm 0.00}$	$0.50_{\pm 0.00}$
	Oversampling	$0.51_{\pm 0.12}$	$0.39_{\pm 0.28}$	$0.66_{\pm 0.12}$
CE + MCC	Uniform	$\mathbf{0.60}_{\pm 0.01}$	$\mathbf{0.53}_{\pm 0.01}$	$\mathbf{0.81}_{\pm 0.01}$
	Oversampling	$\mathbf{0.60}_{\pm 0.03}$	$0.51_{\pm 0.03}$	$0.81_{\pm 0.01}$

Appendix B. Implementation details

B.1. Glioma dataset

B.1.1. Image preprocessing and segmentation

All images were preprocessed and segmented using the publicly available BraTS Toolkit (Kofler et al., 2020). After tumor segmentation, images are [0;1] normalized within the brainmask. A 96^3 patch, centered around the center of mass of the tumor mask, is cropped from the image.

B.1.2. Data augmentation

We incorporate a range of randomized image intensity and geometry augmentations with a probability of 0.5. The set of intensity-changing augmentations consists of randomly adjusting gamma values within the range of 0.5 to 1.5 and Gaussian blur, with a standard deviation varying randomly between 0 and 1.5. In our geometric augmentations, we randomly flip along the sagittal, coronal, or axial planes and randomly crop with a randomized center, selecting a 64^3 cube within an already cropped tumor region to introduce more variability in the tumor's positioning.

B.2. Glaucoma dataset

B.2.1. Image preprocessing

All images were resized to 240×240 px and [0;1] normalized.

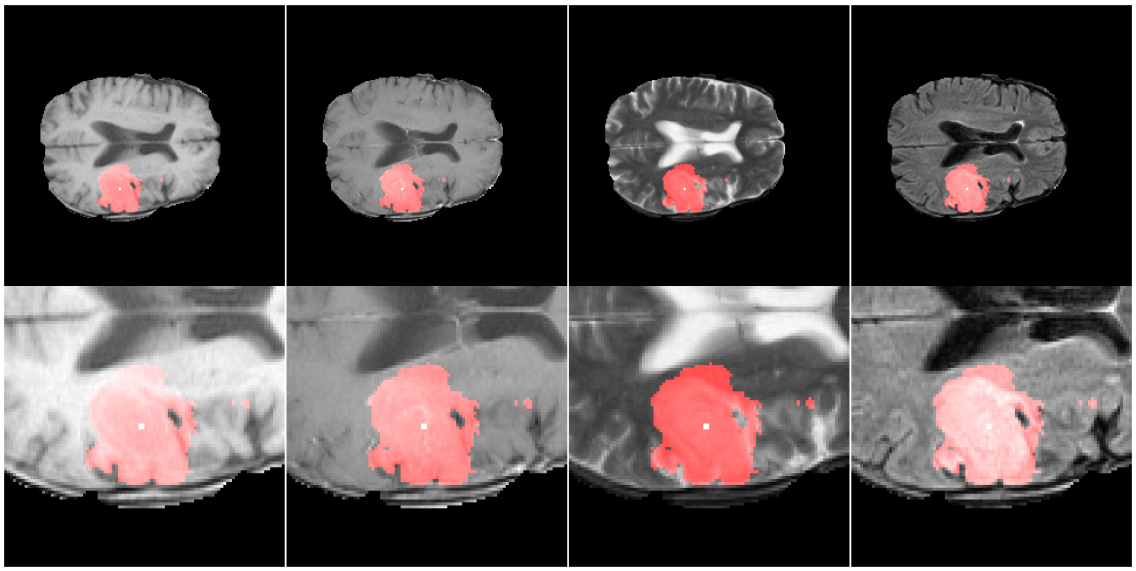

Figure 3: Visualization of the four input sequences available in our dataset. The top row shows entire slices (with the tumor segmentation overlaid in red), and the bottom row shows the crops used for model training.

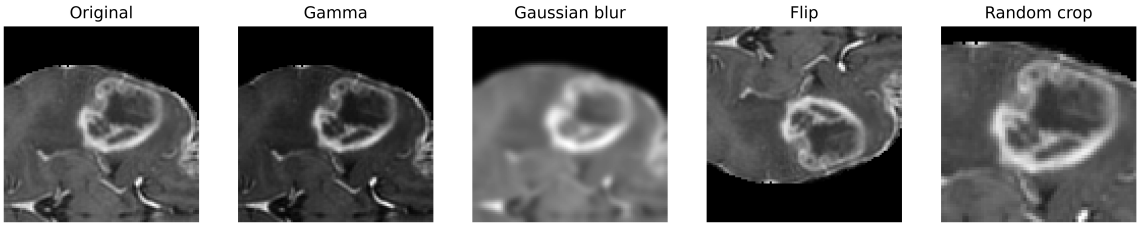

Figure 4: Visualization of the data augmentations used to train our classifier.

B.2.2. Data augmentation

For the glaucoma dataset, we also include a range of randomized image intensity and geometry augmentations with a probability of 0.5 each: The set of intensity-changing augmentations consists of randomly adjusting gamma values within the range of 0.5 to 1.5 and contrast adjustment with a gain randomly selected between [5.,10.]. In our geometric augmentations, we randomly flip along the horizontal or vertical axis.

B.3. Model training

Our classifier is a ResNet34 (He et al., 2016) architecture composed of [3;4;6;3] residual blocks, adapted to 3D. We implement the neural network and training using TensorFlow 2.14 (Martín Abadi et al., 2015), the gamma augmentations with Scikit-image 0.22.0 (van der Walt et al., 2014), and the Gaussian filtering with Scipy 1.11.3 (Virtanen et al., 2020). We use Adam optimizer (Kingma and Ba, 2015), with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, a

learning rate of 1e-3, and a batch size of 50. We also employ a cosine annealing learning rate scheduler (Loshchilov and Hutter, 2016), with a maximum of 250 epochs without warm-up.

Appendix C. Dataset Distributions

For improved visualization of the class imbalance present in the datasets used, we show the class distributions of all datasets in Figure 5. The distribution over the whole Glioma dataset is very similar to the individual sub-datasets, pointing to a skewed real-world distribution of Gliomas.

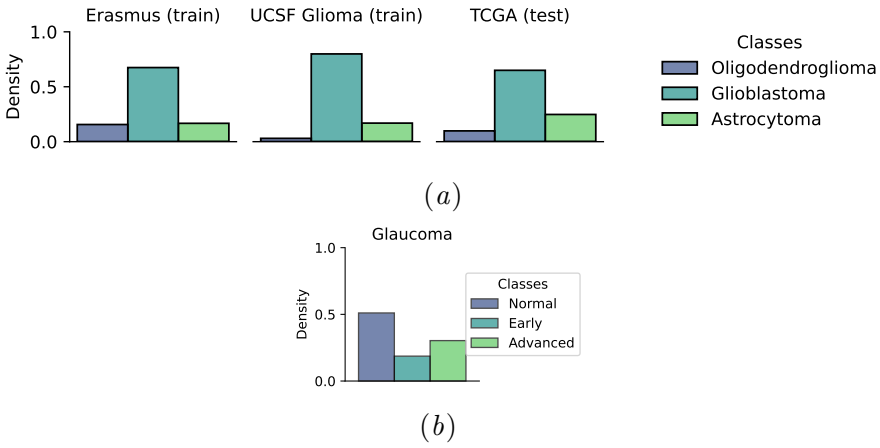

Figure 5: Class distributions for the Glioma and Glaucoma dataset.