Deep learning-based optical coherence tomography image analysis of human brain cancer

Nathan Wang,1 Cheng-Yu Lee,2 and Hyeon-Cheol Park,1 David W. Nauen,3 Kaisorn L. Chaichana,4 Alfredo Quinones-Hinojosa,4 Chetan Bettegowda,5 Xingde Li1

1Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21205, USA

2Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD 21218, USA

3Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA

4Department of Neurologic Surgery, Mayo Clinic, Jacksonville, FL 32224, USA

5Department of Neurosurgery, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA

\*xingde@jhu.edu

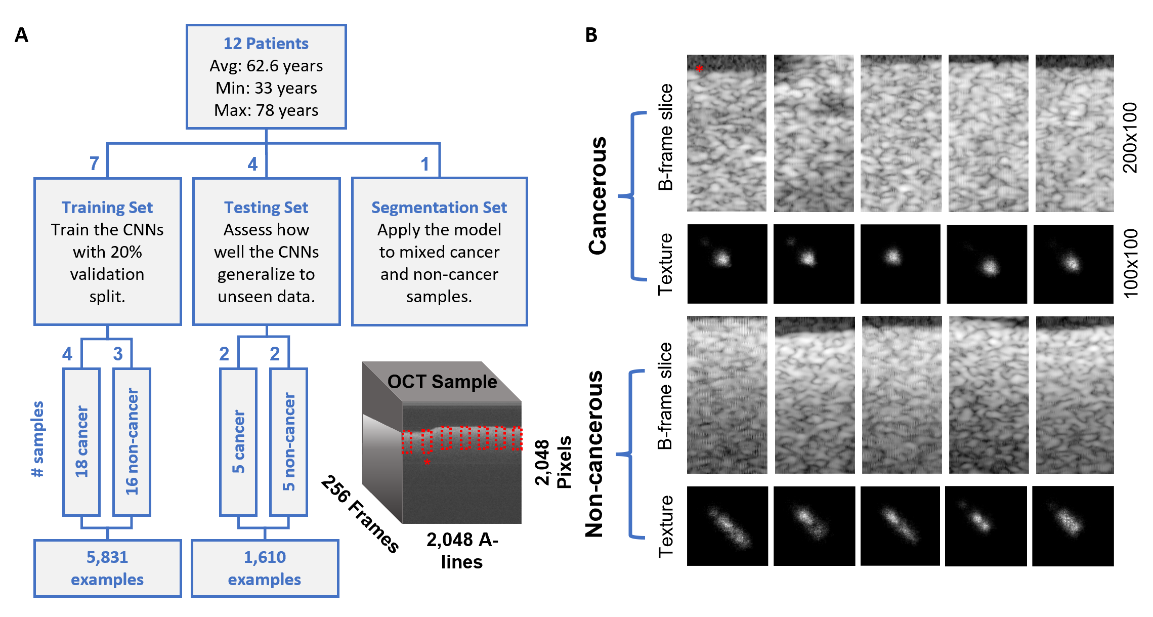
**Abstract:** Real-time intraoperative delineation of cancer and non-cancer brain tissues, especially in the eloquent cortex, is critical for thorough cancer resection, lengthening survival, and improving quality of life. Prior studies have established that thresholding optical attenuation values reveals cancer regions with high sensitivity and specificity. However, threshold of a single value disregards local information important to making more robust predictions. Hence, we propose deep convolutional neural networks (CNNs) trained on labeled OCT images and co-occurrence matrix features extracted from these images to synergize attenuation characteristics and texture features, allowing for accurate and reliable decision-making. Specifically, we adapt a deep ensemble model trained on 5,831 examples in a training dataset of 7 patients and obtain 93.31% sensitivity and 97.04% specificity on a holdout set of 4 patients with high confidence. The segmentation maps produced by parsing the OCT volume and tiling the outputs of our model are in excellent agreement with attenuation mapping-based methods. Our new approach for this important application has considerable implications for clinical translation.

© 2022 Optica Publishing Group under the terms of the [Optica Publishing Group Open Access Publishing Agreement](https://doi.org/10.1364/OA_License_v2)

1. Introduction

Optimal brain tumor resection, defined as maximal removal of cancerous tissues while sparing surrounding noncancerous tissues, is the key to delaying tumor recurrence and minimizing surgery-inflicted neurological damage, thus increasing patient quality of life and survival time [1,2]. This is a challenging problem, especially in the eloquent cortical and subcortical regions where high resection precision is required [3]. Hence, it is critical for a modern intraoperative brain cancer imaging system to provide an accurate segmentation map with high spatial resolution and in real time. In this regard, optical coherence tomography (OCT) is advantageous for its rapid, non-contact, label-free, and micron-resolution imaging capability. In contrast, other modalities such as CT and MRI, which have made significant impact on diagnosis and pre-surgery planning, are less suitable for intra-operative, continuous, real-time guidance due to the drawbacks or tradeoffs in speed, resolution, safety, and cost.

Previous studies identified tissue optical attenuation as a single parameter that can be used to distinguish between cancerous and non-cancerous brain tissue. In general, proliferation of cancer in white matter leads to myelin degradation, resulting in a decreased optical attenuation [4,5]. For a 3D OCT data set consisting of multiple B-frames, an *enface* attenuation map can be generated with each pixel color-coded for cancerous and non-cancerous tissues depending on whether its value is above or below a threshold. However, the performance of thresholding is strictly limited by the overlap in the distribution of attenuation values for cancer and non-cancer tissues [6,7]. To increase robustness while achieving high performance, we propose a novel cancer classifier that is capable of integrating local and global features in an OCT image via both network-based multi-resolution analysis and classic texture analysis. For this purpose, deep convolutional neural networks (CNNs) are selected to perform end-to-end classification on OCT images and their textures. These two information domains are then linked by concatenating their feature embeddings via ensemble learning.



*Figure 1. Dataset description and visualization. (A) The dataset of 12 patients was divided for training (including validation), testing, and segmentation purposes. The blue numbers indicate the number of patients associated with the adjacent subset. ~175 200x100 B-frame slices were sparsely selected from each sample. (B) Differences between cancer and non-cancer classes are visible in representative training examples. All B-frame slices have a corresponding 100x100 grey level co-occurrence matrix (GLCM) texture. Separate networks are trained on slices and textures and then connected via ensemble learning.*

In the next section, the full data processing pipeline is described, including the dataset characteristics, sparse-sampling and texture extraction methods, and ensemble learning architecture. In the third section, the results obtained using our ensemble model is reported in comparison to its two subnetworks standalone and the network behavior is explained with gradient class activation maps (grad-CAM). In the last section, issues are discussed, and promising directions of related research are identified.

1. Methods and Materials

The dataset in this study comprises 23 cancerous and 21 non-cancerous brain cortex samples from 12 patients averaging 62.6 years of age (min: 33 years, max: 78 years). Tissue samples were obtained at the Johns Hopkins Hospital under an approved Institutional Review Board protocol. All the samples used in the study were evaluated by a pathologist to be mostly homogeneous and were imaged with a home-built SS-OCT over a 2 mm x 2 mm area along a depth of 1.8 mm in tissue. Each 3D OCT dataset consists of 256 B-frames where one B-frame is a collection of 2,048 A-lines measuring 2,048 pixels per line. The dataset was allocated according to Figure 1A: seven patients were used for training the deep networks, four different patients were reserved for testing, and a segmentation map was generated on the last patient’s sample containing mixed cancer and non-cancer regions.

*2.1 Dataset sparse sampling*

The transverse resolution of our OCT system is 25 µm, meaning that B-frames in this range experience no significant changes. Hence, we adopt sparse sampling methods to ensure that the examples included in the dataset are sufficiently representative of the volumetric dataset. For each OCT scan of a cortex sample, a step size of 80 µm is used to select ~25 representative B-frames. A small proportion of B-frames with significant fragmenting at the tissue surface are manually removed. Of the remaining frames, a 5x5 median filter is applied to suppress noise followed by conventional OCT image processing. Each B-frame is evenly divided into seven slices of 100 A-lines separated by 290 µm. Finally, within each slice, the A-lines are truncated to cover the single scattering region, determined to be approximately 200 pixels (418 µm in total) below the sample surface. In this way, 5,831 total examples are generated (34 samples \* ~25 B-frames/sample \* 7 slices/B-frame). As shown in Figure 1B, non-cancerous B-frame slices exhibit a stronger white to black gradient from top to bottom in comparison to cancerous B-frame slices. This is a consequence of healthy white matter attenuating the intensity of the incident light more rapidly with depth. Prior studies utilizing statistical and machine learning methods have demonstrated that image texture is also discriminative for this task [8].

*2.2 Texture embedding*

To capture the global image texture, the grey level co-occurrence matrix (GLCM) is used to measure each B-frame slice. First, the global maximum and minimum of the training set are used to linearly transform the values in each B-frame slice to a new range of 0 to 99 inclusive. Then, a 100x100 GLCM texture map is computed for each slice using an adjacency condition of 15 pixels at an angle of 0 degrees selected by trial and error. The row and column indices of the GLCM represent pixel intensities and the matrix elements show how frequently the two given intensities are adjacent [9]. For example, if intermediate intensities (say 49 ± 10 on a 0-99 scale) are adjacent to other intermediate intensities, then the resulting GLCM will exhibit a bright white spot near its center. Figure 1B shows typical textural features in the training set. Qualitatively, concentrated circles are characteristic of cancer slices whereas textures of non-cancer slices are more oblong and smeared.

*2.3 Ensemble learning architecture*

In CNNs for classification, the dimension of the first layer is defined by the input dimensions which are 200x100 for B-frame slices and 100x100 for textures. The size of the final layer is two nodes, corresponding to the probability of cancer and non-cancer in the given input. Compared to the texture CNN, the B-frame CNN is wider in order to preserve visual information. Specifically, the number of feature maps increases as the size of each feature map decreases. By training separate CNNs on the textures and B-frame slices, their embeddings (i.e., the activation of the penultimate layer) can be extracted independently, concatenated together, and connected to a multilayer perceptron which synergistically makes the final prediction. Figure 2 shows the layer-by-layer details of the networks. Particularly, note that the penultimate fully connected layer is chosen as the embedding and the third to last convolutional layer is selected for grad-CAM visualization. grad-CAM weights the feature maps in the given layer by the average of the derivative of each pixel in the feature map with respect to the output. Overlaying the grad-CAM output on the input image heuristically indicates which regions are more or less important.

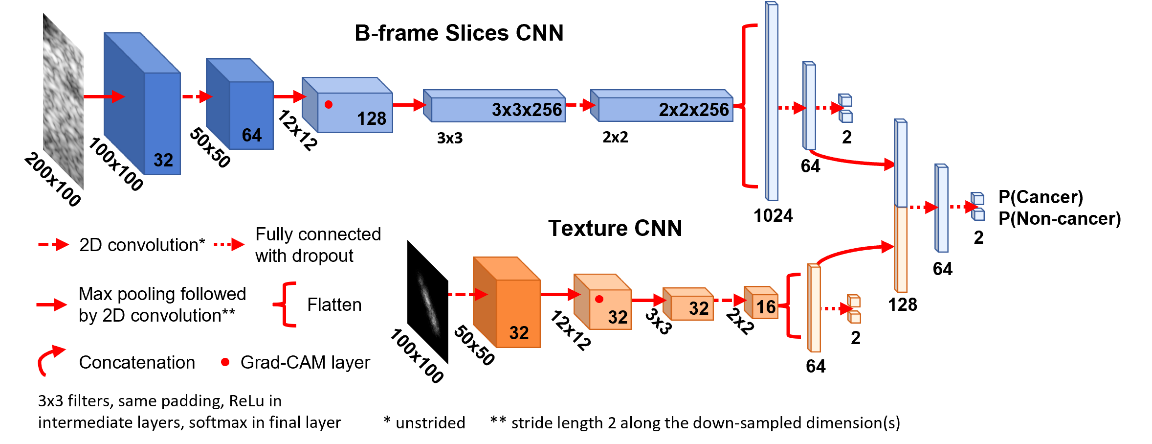


Figure 1. Ensemble learning architecture integrating B-frame slice pixels and texture features for OCT cancer tissue analysis. The dimension of each layer is indicated, where the two dimensions outside each layer is the feature map shape and the number inside each layer is the number of maps/filters. The B-frame Slices CNN and Texture CNN are trained separately. Once they have converged, they are used to create embeddings of their respective inputs, which are linked to a fully connected network that makes the final classification.

The final layer of the model is subject to the softmax activation. Binary cross entropy is selected as the loss function with balanced class weights regularization used to proportionally re-weight the loss in order to prevent bias towards the majority class [10]:

. (1)

In Eq. 1, the loss *L* is a function of the model parameters . *N*=8 is the batch size, *M=2* is the number of classes, and *y* and *p* are scalars corresponding to the label and prediction respectively. For example, *pij* denotes the model prediction for class *j* on training example *i*. Furthermore, all labels are one-hot encoded and smoothed according to a well-established protocol with , preventing network overconfidence [11]. This is especially important in this application since the cortex samples may contain inhomogeneities and mislabeled examples would impede training. Additionally, random up-down and left-right flipping was applied to the inputs to augment the data and the dropout rate was set to 0.5 for each dense layer to discourage overfitting [12]. Global min-max normalization was applied to all training examples. All network experiments were conducted on a GeForce RTX 3060 ti GPU and AMD Ryzen 3700x CPU.

1. Results and Discussion

The ensemble model outperformed its component subnetworks on the testing set, achieving (94.90% ± 0.11%) accuracy on the testing set after 50 epochs of training. The model is also compact, capable of processing 1 B-frame every 7 milliseconds (143 fps). The gradient class activation maps demonstrate that the model makes decisions based on relevant features and therefore is able to be interpreted.

*3.1 Network results*

The training graphs in Figure 3A exhibit smooth convergence and well-correlated training and validation curves. For many epochs, the validation accuracy even exceeded the training accuracy. This phenomenon is likely due to augmentation and dropout being inactivated during validation, making the predictions more straightforward during the validation phase. The texture CNN stabilized more quickly and at a higher accuracy than the B-frame slices CNN, indicating that the problem was more easily solved in the texture domain. Additionally, the ensemble model reached high accuracy after just one epoch and achieved a slightly higher validation accuracy than either subnetwork. This is reasonable because the embeddings are sufficiently high-level such that a good prediction can already be made from a simple linear transformation.

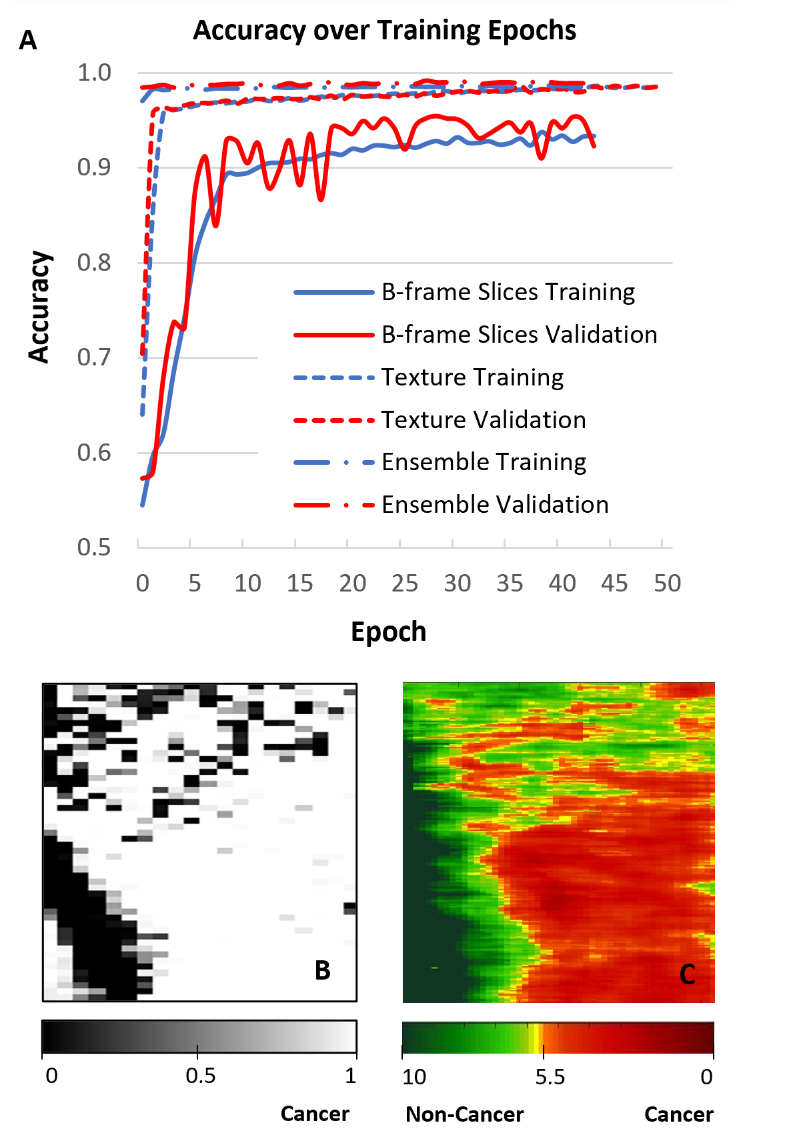


Figure 2. Network training curves and segmentation output. (A) Accuracy metric for all models during training. (B) Ensemble model generated segmentation map. (D). Color-coded attenuation map.

A segmentation map was generated from an additional patient sample including both cancer and non-cancer regions. The sample contained 256 B-frames with 20 slice and texture pairs extracted from each frame. Each pair of data were input sequentially into the ensemble model, generating a binary prediction which were then tiled in a 256x20 map. After resizing this map to 20x20 for better viewing, the width and height of each pixel are approximately 0.1 mm in resolution. Figure 3B, 3C, and 3D provide a side-by-side comparison of the network-generated segmentation map and the color-coded attenuation map generated using traditional method respectively all of which are in good agreement.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Mean Accuracy** | **Standard Deviation** | **B-frame Slices** | **Texture** | **Sub-optimal Texture** |
| *Ensemble* | 94.90% | 0.11% | t=132.21, F=23.44 | t=-0.39,  F=1.81 | -- |
| *B-frame Slices* | 85.01% | 0.52% | -- | t=-130.21, F=12.94 | -- |
| *Texture* | 94.91% | 0.14% | t=130.21, F=12.94 | -- | -- |
| *Sub-optimal Ensemble* | 90.03% | 0.17% | t=65.11,  F=9.38 | -- | t=26.40,  F=5.68 |
| *Sub-optimal Texture* | 88.39% | 0.41% | -- | t=-107.26, F=7.91 | -- |

Table 1. Summary of statistical testing. The t-statistic compares the mean accuracies between models, and the F-ratio compares variance in accuracies between models. Magnitudes of t greater than 1.661 or magnitudes of F greater than 1.607 indicate that the observe difference in the compared statistic is significant.

*3.2 Network ablation analysis*

To determine whether the performance gain is significant with respect to individual network components, we quantified model confidence by computing the standard deviation in accuracy over 50 evaluations on the testing set with dropout active. Because dropout randomly disconnects nodes, a different testing accuracy is generated over each iteration. Models with a low spread in testing accuracies and thus higher confidence indicate that the learned feature representations are more robust [12,13]. Hence, we calculate the t-statistic to compare model accuracies and F-ratio to compare model variances/confidences. At an alpha level of 0.05, the critical t-value is 1.661 and the critical F-ratio is 1.607.

Table 1 reports that the CNN trained on texture is significantly more accurate than the CNN trained on B-frame slices. Due to this performance disparity, the accuracy of the ensemble is not significantly increased. However, the ensemble becomes more confident in its decision-making as expected when linking two complementary information domains. In future work such as distinguishing between cancer grades, it is conceivable that the texture differences will be more subtle leading to lower accuracy in the texture CNN. We simulated this scenario by saving the weights of a “sub-optimal texture” model at an earlier epoch. When the resulting “sub-optimal ensemble” is constructed in the same way, both its accuracy and confidence increased relative to its subnetwork counterparts. Hence, the utility of ensemble learning is further justified.

*3.3 Network interpretability*

Figure 4 displays representative grad-CAMS for both the B-frame slice and texture CNNs in each of the four possible prediction scenarios. Grad-CAM heatmaps offers a visual way to explain network behavior by highlighting pixels that contribute most greatly to the final decision [14]. In the true positive and true negative cases of the texture CNN, the superimposed heat maps outline the relevant areas of high intensity. In other words, the network learned that the shape of these textures is an important feature. On the other hand, the B-frame slice CNN identifies the very bottom part of a slice as a key feature. This is consistent with the observation that lower optical attenuation in cancerous white matter leads to more uniform intensity throughout the slice. In non-cancerous tissues, the signal becomes significantly attenuated towards the bottom of the single-scattering region. Hence, the behavior of the network correlates to the underlying principles of OCT.

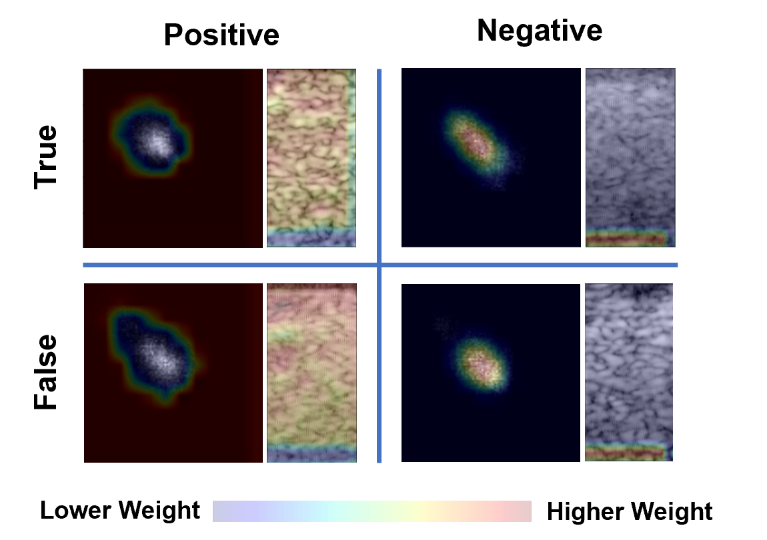


Figure 3. Grad-CAM visualizations. The positive case indicates the presence of cancer while the negative case indicates the absence of cancer. In the textures, the network has identified the shape as a key feature. In the B-frame slices, the bottom/deeper region of the frame is learned as a distinguishing feature.

1. Conclusion

A deep network cancer detector capable of producing accurate segmentation maps in real-time has been presented in this paper. Specifically, the ensemble model comprising a texture CNN and a B-frame CNN has been designed and validated to be sensitive, specific, fast, stable, interpretable, and superior to its individual components alone.

Although single parameter methods like attenuation thresholding remain attractive for their clear explainability, deep learning models are quickly approaching a point where they can be just as or even more robust, as we have demonstrated. Deep learning approaches also have the potential to differentiate between low and high-grade cancers as well as between cancers in white or grey matter based on analysis of structure in the speckle noise given that speckle noise may present in subtly different patterns in different tissues. Additionally, deep networks may also be created that are robust to challenges in the clinical setting such as motion artifacts, image defocusing, and presence of blood. These are all important direction of future work which needs to utilize datasets of in-vivo animal model studies or more patients.

Funding National Institutes of Health (R01CA200399).

Disclosures The authors declare no conflicts of interest.

Data Availability The processing codes developed for this paper are available upon request.

References

1. N. Sanai, M.-Y. Polley, M. W. McDermott, A. T. Parsa, and M. S. Berger, "An extent of resection threshold for newly diagnosed glioblastomas," J. Neurosurg. **115**(1), 3–8 (2011).

2. M. J. McGirt, K. L. Chaichana, M. Gathinji, F. J. Attenello, K. Than, A. Olivi, J. D. Weingart, H. Brem, and A. R. Quiñones-Hinojosa, "Independent association of extent of resection with survival in patients with malignant brain astrocytoma," J. Neurosurg. **110**(1), 156–162 (2009).

3. G. E. Keles, D. A. Lundin, K. R. Lamborn, E. F. Chang, G. Ojemann, and M. S. Berger, "Intraoperative subcortical stimulation mapping for hemispherical perirolandic gliomas located within or adjacent to the descending motor pathways: evaluation of morbidity and assessment of functional outcome in 294 patients," J. Neurosurg. **100**(3), 369–375 (2004).

4. M. Alaminos, V. Dávalos, S. Ropero, F. Setién, M. F. Paz, M. Herranz, M. F. Fraga, J. Mora, N.-K. V. Cheung, W. L. Gerald, and M. Esteller, "EMP3, a myelin-related gene located in the critical 19q13.3 region, is epigenetically silenced and exhibits features of a candidate tumor suppressor in glioma and neuroblastoma," Cancer Res. **65**(7), 2565–2571 (2005).

5. F. J. van der Meer, D. J. Faber, M. C. G. Aalders, A. A. Poot, I. Vermes, and T. G. van Leeuwen, "Apoptosis- and necrosis-induced changes in light attenuation measured by optical coherence tomography," Lasers Med. Sci. **25**(2), 259–267 (2010).

6. W. Yuan, C. Kut, W. Liang, and X. Li, "Robust and fast characterization of OCT-based optical attenuation using a novel frequency-domain algorithm for brain cancer detection," Sci. Rep. **7**(1), 44909 (2017).

7. C. Kut, K. L. Chaichana, J. Xi, S. M. Raza, X. Ye, E. R. McVeigh, F. J. Rodriguez, A. Quinones-Hinojosa, and X. Li, "Detection of Human Brain Cancer Infiltration ex vivo and in vivo Using Quantitative Optical Coherence Tomography," Sci. Transl. Med. **7**(292), 292ra100 (2015).

8. J. Möller, A. Bartsch, M. Lenz, I. Tischoff, R. Krug, H. Welp, M. R. Hofmann, K. Schmieder, and D. Miller, "Applying machine learning to optical coherence tomography images for automated tissue classification in brain metastases," Int. J. Comput. Assist. Radiol. Surg. **16**(9), 1517–1526 (2021).

9. B. S. V, "Grey Level Co-Occurrence Matrices: Generalisation and Some New Features," Int. J. Comput. Sci. Eng. Inf. Technol. **2**(2), 151–157 (2012).

10. Z. Xu, C. Dan, J. Khim, and P. Ravikumar, "Class-Weighted Classification: Trade-offs and Robust Approaches," in *Proceedings of the 37th International Conference on Machine Learning* (PMLR, 2020), pp. 10544–10554.

11. R. Müller, S. Kornblith, and G. Hinton, "When Does Label Smoothing Help?," arXiv (2019).

12. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," J. Mach. Learn. Res. **15**(56), 1929–1958 (2014).

13. M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential Deep Learning to Quantify Classification Uncertainty," in (2018).

14. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," Int. J. Comput. Vis. **128**(2), 336–359 (2020).