Explaining Tumor Classifiers

Nikolaos Athanasopoulos

Department of Advanced Computing Sciences, Maastricht University

December 21, 2023

Preface

This report is part of a project for the 2023-2024 Period 2 course Deep Learning for Image and Video Processing. The codebook and the presentation can be found on Google Drive here. The dataset is acquired from Kaggle, and can be found here.

Abstract

Brain tumors are a type of neoplasm that can manifest either directly in the brain or by metastasis from another organ in the human body. How exactly these tumors present themselves is highly variable, often depending on the tumor's size, location, and growth rate, and can include symptoms such as headaches, seizures, and cognitive or motor deficits. This study aims to not only detect and classify such tumors by using Deep Learning models, but also take a deeper look on what features these models focus on and the reasons behind potential mistakes that they make, with the help of Explainable AI.

1 Introduction

According to the Surveillance, Epidemiology and End Results program of the United States of America National Cancer Institute, an estimated 25,000 new cases of brain tumors are discovered each year, accounting for 1.3% of all new cancer cases worldwide. The 5-year relative survival rate is calculated at 33%, and the lifetime risk of developing such a cancer is 0.6%. Magnetic Resonance Imaging (MRI) has greatly helped the detection and characterization of brain tumors. Unlike other imaging techniques MRI provides soft tissue contrast without the use of ionizing radiation, making it particularly suitable for brain imaging. This has allowed the introduction of various datasets of brain tumor classification, thanks to annotations from field experts, which can be used to train Deep Learning models such as Convolutional Neural Networks and Visual Transformers. Some other work in this field includes (Shanthi et al., 2022), (Shaimaa et al., 2023) and (Mercaldo et al., 2023). However, black box models cannot be used on their own on such scenarios due to their lack of inherent explainability. This gap, however, can be filled with the help of Explainable AI methods, which is what this report focuses on. With the help of saliency maps, Grad-CAM and Visual Attention Heatmaps, this report will show that the quantitative results by themselves can often be somewhat lacking when it comes to choosing a final model for such a scenario. The rest of this report is structured as follows: Section 2 will focus on the dataset itself and some insights about it. Section 3 will focus on the approach that was followed and the quantitative results from the Deep Learning models. Section 4 will discuss the Explainability methods applied to the models, and will show how such methods should be the backbone of such scenarios. Section 5 will discuss conclusions, and finally Section 6 will discuss potential future work.

As for previous experience with Deep Learning and Computer Vision, I have undertaken the Computer Vision class from Period 5 and during it, have tested many different CNN architectures of my own in order to see firsthand how different architecture designs lead to different results, for an emotion analysis scenario. Image stitching was the second assignment of that course and I have implemented a complete image stitching algorithm from scratch with Harris corners (Harris and Stephens, 1988), SIFT (Lowe, 2004) descriptors and RANSAC (Fischler and Bolles, 1981). Furthermore, a lot of experience has been gained from the Period 5 course Explainable AI, where almost all explainability methods discussed in this report have been thoroughly analyzed, and that was the main inspiration for the topic of this experiment.

2 Dataset

The dataset itself is procured from Kaggle, and consists of 5740 images from Magnetic Resonance Imaging. Each

image is accompanied by a label, which are 4 in total: meningioma, glioma, pituitary and no tumor. Each label (except for no tumor) describes the location of the tumor. Meningiomas form on the outermost layer of the brain, gliomas form on the temporal and frontal lobes and pituitary tumors form in the pituitary gland. The label split of the dataset is shown in Figure 1.

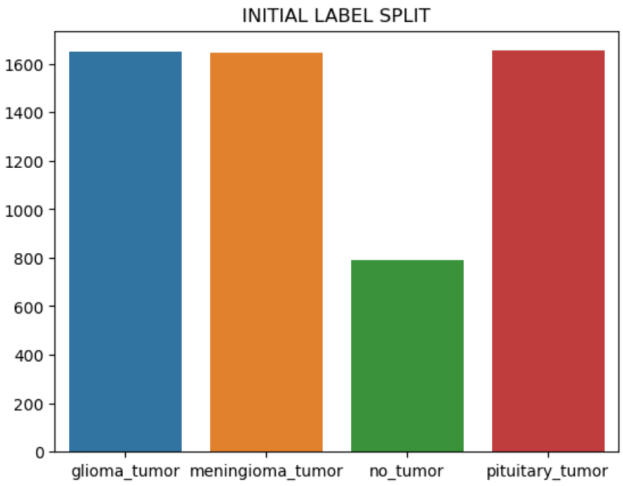

Figure 1: Original Label Split

It is immediately obvious that the no tumor class is under-represented, which, as will be seen in the results section, has a significant effect on the performance of the models. This is why augmentation and resampling was performed on the training set, and the results were compared to the the raw dataset in terms of both performance and explainability of the models. Of course, the validation and the test set remain original. Some example images from each label can be seen in Figure 2. The resampling strategy followed was random oversampling on the minority class only (no tumor), since there was no intention of producing synthetic datapoints. For the augmentation strategy, the following transforms were applied randomly to images: Affine transform, Gaussian Blur, Gaussian Noise and Salt and Pepper noise. These transforms ensure that the dataset is significantly harder for the Deep Learning models, which in turn is expected to make the models focus more on the important features of each tumor. In some cases, the augmented images are indeed very hard to analyze, which will lead to some interesting results afterwards. It should be noted that the original dataset comes pre-split into training and testing sets, but these were not the ones used in the actual models afterwards. These 2 sets were merged, shuffled and then re-split into training, validation and test sets, with a split of 60-20-20 respectively. The reason behind the large size of the test set is to ensure that there are enough images on which to test the various explainability methods afterwards. This, however, means that the actual non-augmented train set is somewhat small which may be the cause for some performance issues on the smaller models afterwards. Some images from the training set, after augmentation and resampling, can be seen in Figure 3.

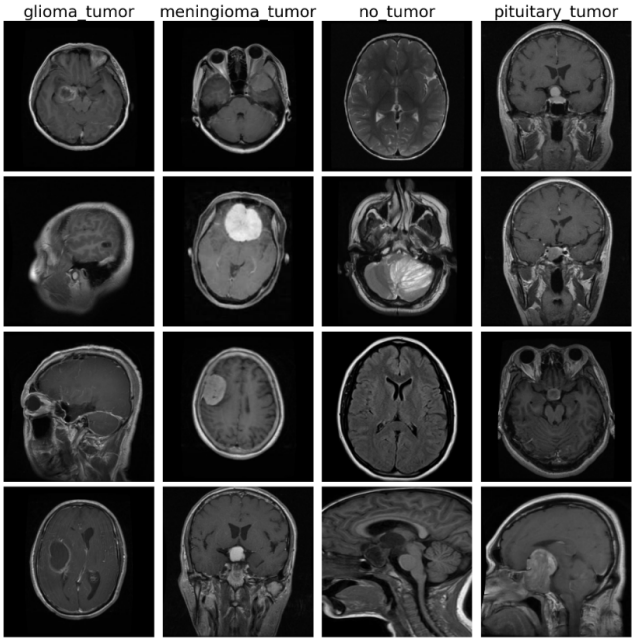

Figure 2: Example Images from the Dataset

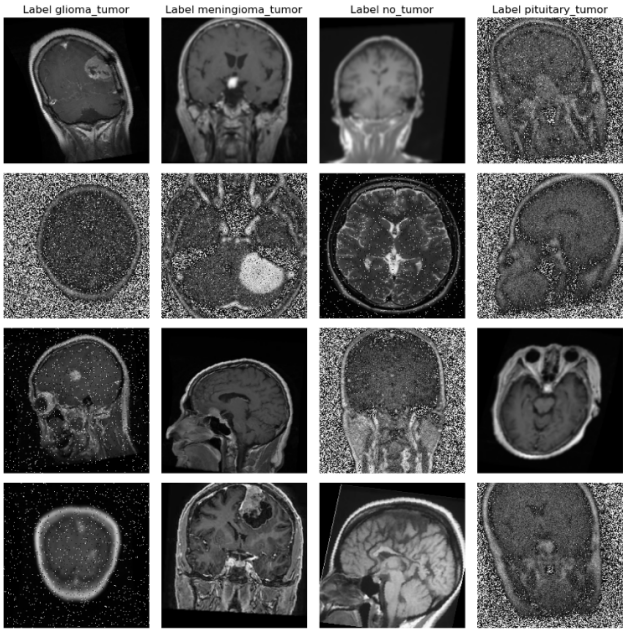

Figure 3: Example Augmented Images from the Dataset

3 Approach

The quantitative results can be seen in Table 1. The confusion matrices can be seen in Figure 10 and the ROC curves can be seen in Figure 11, which are provided in the Appendix. Since the main intention in this experiment was to compare multiple models and Explainability methods, a lot of experimentation was conducted. As far as architectures are concerned, 2 Custom CNNs were used and compared to Google's Visual Transformer (Dosovitskiy et al., 2021) and Resnet (He et al., 2015).

Model Architecture	Augmentation + Resample	Accuracy	F1 Weighted Average
2 Layer CNN	No	94%	94%
2 Layer CNN	Yes	83%	83%
5 Layer CNN	No	95%	95%
5 Layer CNN	Yes	94%	94%
Google ViT	Yes	99%	99%
Resnet50	Yes	99%	99%

Table 1: Quantitative Results

The architectures for both models can be seen in the Appendix, figures 7 and 8 respectively. The first custom model is a 2 Layer CNN and the second custom model is a 5 Layer CNN. The custom models were trained both on the original dataset and on the resampled+augmented dataset. Google's Vision transformer and Resnet were fine-tuned only on the resampled+augmented dataset. The validation and the test set are the same in every case. The small model was trained for 40 epochs, while all the other ones were trained for 20 epochs. The learning rate was set to 0.0001 in every case and the optimizer was chosen to be AdamW. After every training epoch, the model was tested on the validation set. The model with the least validation loss (and best validation accuracy) was kept in the end. The training and validation loss curves can be seen in the Appendix. In some cases (such as the 5 Layer no augment case), the training should have been continued for a few more epochs since the 2 losses have not diverged yet, but the losses seem to have reached a plateau anyway, so continuing the training would not have led to a significantly better result. Also interesting is the fact that the fine-tuned models only need 1 or 2 epochs to reach the 99% performance metric.

There are many interesting insights that can be discussed from these results. First of all, the fine-tuned state-of-the-art models ViT and Resnet50 achieve the best results metrics-wise. This is to be expected, since they are trained on huge and difficult datasets, which means that they are able to capture the intricacies of this dataset effectively. However, it is the Custom Models that provide the most interesting results. It can be seen that the simple 2 Layer CNN trained on the original dataset achieves an initially impressive 94%, while the one trained on the augmented data falls off to 83%. This could be to a number of reasons, the most prominent being the increased difficulty of the augmented data is simply too hard to capture with such a small architecture. It should be noted this result was tested on some different splits of the data as well, and the best result gained was 90%, which goes to show that indeed the model underperforms when compared to the one trained on the original data. It will be shown in the next section, that these numbers do not capture the "true" performance of the models. The 5 Layer CNN exhibits different behaviour to the smaller model. The one trained on the original data and the one trained on the augmented data have almost the same performance (95% vs 94% respectively). Again, it will be shown that the model trained on the augmented data, while having slightly worse performance, will better distinguish the features of the data. In any case, there is another thing that should be mentioned. The numbers shown here are obtained after training on a single train-val-test split. It could be that these change when the splits are changed. So indeed the 5 Layer CNN could be performing better on the augmented dataset than on the non-augmented one, and the difference shown here is just statistically unimportant. This would require cross-validation to discover, which would take more training time. However, it should be highlighted that the 2 Layer CNN's performance is much worse on the augmented data than on the original, and that is not expected to change that much when changing the splits.

4 Explainability

For the explainability part of this report, a lot of different methods were considered. For the custom models, both saliency maps and Grad-CAM (Selvaraju et al., 2019) were calculated. For Resnet50, only the Grad-CAM was calculated, and for the Vision Transformer, the Attention Heatmap was calculated. The saliency maps are of course calculated for the whole model. while the Grad-CAM needs a specific convolutional layer. This was chosen to be the last convolutional layer in every case, because this is the layer that captures the most important information, since it contains inputs from all the previous layers. Of course, some visualizations in the code can be seen for the second-to-last convolutional layer, which typically shows less important features for the models. The attention heatmap for the ViT is also calculated for the last attention layer, because the previous attentions layers tend to focus more on segmentation rather than feature extraction. This, again, can be seen in the code. Picking interesting test instances involves some manual work. There are 2 different test scenarios which should yield interesting insights. The first scenario is identifying instances where the model disagrees with the true label. These instances are easy to identify separately for each model, but unfortunately there are no instances on which every model disagrees with the true label, so some instances where the "worst" models disagree with the true label are picked. The second scenario involves picking instances where the model makes the correct classification decision but based on focusing on the wrong features. This could be happening due to luck (the 4 label scores are very close to each other, but the correct label has a slightly higher score) or because the model is certain that some features lead to that particular label, but in reality they do not. The second scenario is shown to be very common in the 2 custom architectures which are trained on the original dataset. This section will also highlight how the 94% metric obtained from the small 2 Layer CNN with no augment is misleading, and this is obvious looking at all the following example images that compare the explainability methods for the same test instance. Figure 4 shows the first test instance.

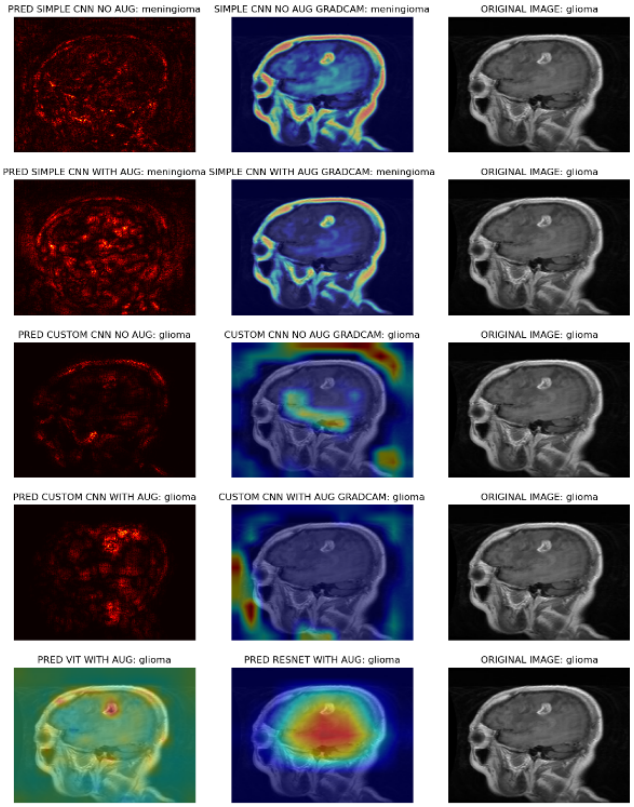

Figure 4: First XAI Example

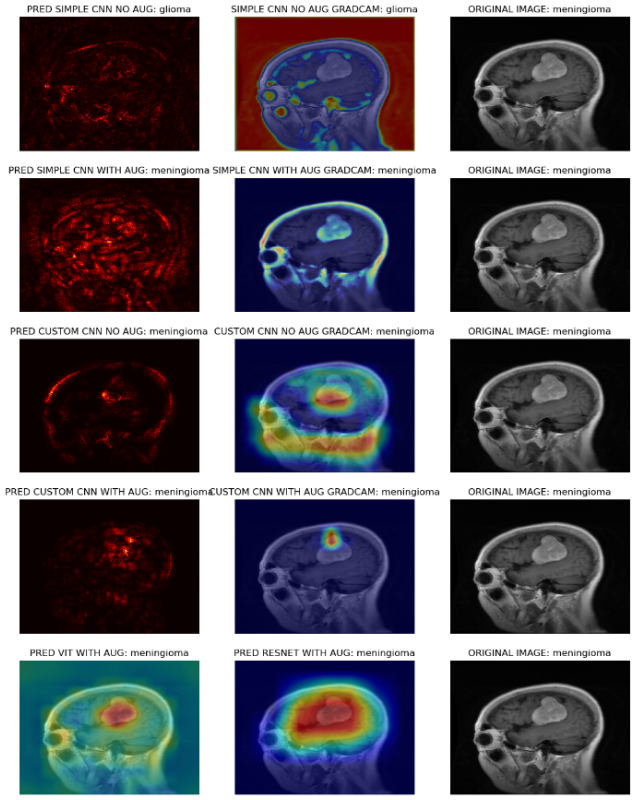

Figure 5: Second XAI Example

In this case, the small model completely misses the label, both when augmented and when not. However, through saliency maps and Grad-CAM, it is obvious that the model in both cases actually detects the tumor, but it also detects other unimportant features, which are throwing the prediction off. The 5 Layer CNN in this case predicts the correct label in both cases, but Grad-CAM actually shows that the features considered important to the model are actually not related to the tumor. The saliency map for the augmented case shows that the tumor is detected somewhat strongly, and the saliency map for the non-augmented case shows that the tumor is largely missed. In both the 2 Layer and the 5 Layer CNN, the saliency map is much more "intense" for the augmented case, which shows that the augmentation enhances the model's decision, even if that decision is wrong in the end. Grad-CAM for Resnet shows that the model detects and considers the tumor as the most important feature to the model's decision (not surprising considering its 99% performance metrics). The attention heatmap for the Vision Transformer shows that many parts of the image are considered as features important to the decision of the model, but the tumor is indeed the most "intense" feature driving the decision. Figure 5 shows the second test instance.

A lot of interesting conclusions can be drawn from this instance as well. Just as the first example, the saliency maps of the augmented models show that indeed augmentation can help the models focus better on the important features. Every model except the 2 Layer non augmented model correctly label the instance. The Grad-CAM for this model immediately shows us the reason behind this, which is that the model focused on the eyes and the base of the skull instead of the tumor in the middle of the brain. The augmented version of the simple model focused on the outside area of the brain, where typically a meningioma tumor would be found, instead of focusing on the tumor itself. This would be a typical "scenario 2" instance. The non augmented 5 Layer CNN indeed considers the actual tumor a strong feature but also focuses on the base of the skull as well, which is irrelevant to the true label, while the augmented version focuses solely on the tumor itself. This is exactly why the 95% vs 94% metrics of those models respectively do not tell the whole story of the "true" performance of these models. While it is true that the non-augmented model has higher accuracy in the test set, the augmented model still "sees" the true features more accurately, which is why it should be preferred in such applications. The Vision Transformer and the Resnet models follow the same trend as before: ViT focuses strongly on the tumor itself but also on other uncorrelated features, while Resnet focuses solely on the true tumor. Figure 6 shows the third test instance.

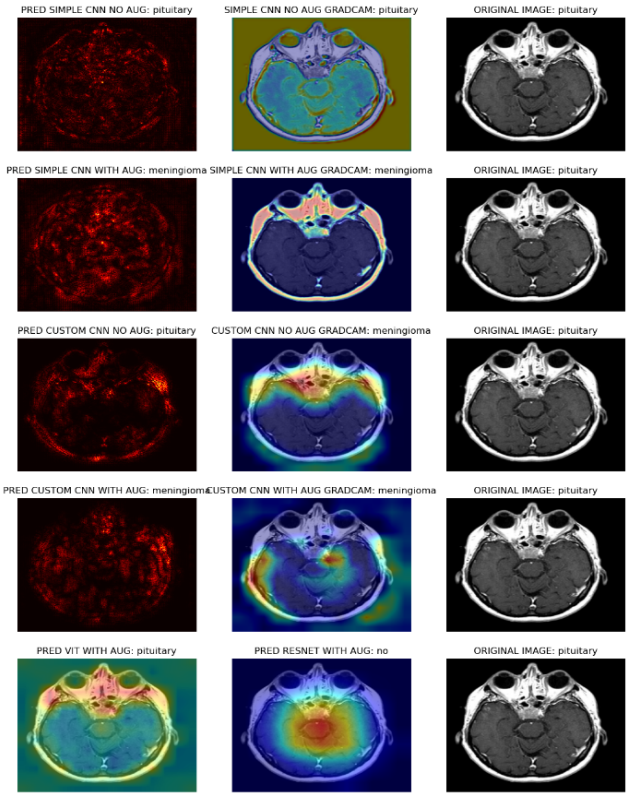

Figure 6: Third XAI Example

This is one of the few cases where Resnet50 fails to identify the correct label. Looking at its Grad-CAM, it is unclear why, since it seems to be focusing on the tumor itself. However, this is one of the drawbacks of using Grad-CAM on the last convolutional layers. While these tend to hold the most information about the features of the model, they are also the lowest resolution, so the Grad-CAM gives an approximation heatmap rather than an exact one, so it is possible that Resnet is focusing on the non-tumor space below the pituitary tumor rather than on the tumor itself. Also interesting is the fact that the attention heatmap of the Visual Transformer focuses on many irrelevant features rather on just the tumor itself. There is a possibility that the model actually "sees" some features there, like an inflammation of the glands, but this would still be a secondary feature compared to the tumor itself. The saliency maps on all cases fail to provide any meaningful insight, as they show that the models do not focus on the correct features even when predicting the final label correctly. However, in both augmented cases, the saliency maps show "stronger" feature focus than on the non-augmented cases. Grad-CAM also shows a better representation of the models here, as it shows a less noisy feature importance than the saliency maps, which is true for almost every case.

5 Conclusions

A lot of conclusions can be drawn from the results and the examples shown previously. First and foremost, the performance metrics themselves give an overestimation of the true performance of the models, since in every case, there are instances on which the model outputs the correct label but based on the wrong features. Moreover, the 2 biggest architectures which are pre-trained are of course the ones that both achieve the best true performance, based also on the explainability methods. This was to be expected, as Visual Transformers and models like Resnet and VGG (Simonyan and Zisserman, 2015) are typically State-of-the-Art in many Computer Vision scenarios. Furthermore, augmentation and resampling techniques do not always increase performance metrics, but help the model to better understand the good features of the data. In almost every case, the augmented models tend to focus more strongly on features than the non-augmented models. Also interesting is the fact that saliency maps and Grad-CAM do not always focus on the same features, which can be explained by the fact that Grad-Cam is always computed on a specific layer, while saliency maps compute the gradient of the output with respect to the input. The goal of this report was to provide a somewhat deep dive into the explainability of brain tumor classifiers, and how different parameters of the experiment lead to different results, both in pure performance metrics, but also in what the models truly capture as features. It was ultimately shown that explainability should be the backbone of Deep Learning models for medical applications, and that transfer learning leads to overall better results, both in performance and in explainability. It should be, however, underlined that when dealing with scenarios that involve medical data, the experiments should always include medical experts, who can recognise the true features of the data and can give relevant advice.

6 Future Work

There is still quite a lot of work and comparisons that would need to be done in order to safely generalize the conclusions drawn from this experiment. This section highlights the most important ones, as they are identified during the entire experimentation process. One of the most important experiments that has to be carried out is test this process on another dataset to see if the same conclusions can be drawn. The best method would probably be to merge different MRI brain tumor datasets and repeat the whole experimentation process on this merged dataset, in order to get more generalizable results. It is advisable to get data from as many different sources as possible in order to see the true performance of these models, and test the quality of the experimentation process. Another interesting idea is to obviously experiment with the hyper-parameters of the experiment and see how much and in what way they influence the final results, both in metrics and in explainability. The most important ones are deemed to be the augmentation parameters, such as the intensity of the salt-pepper noise and the blurring. It would be interesting to see how different augmentation techniques lead to different explainability results, and also, if only resampling or only augmenting leads to significantly lower performance. Another insightful experiment would be to perform the same process for differ-

ent train-test-validation splits of the data. The best thing would be to perform cross-validation and then take the average performance metric of each model on all the different splits. The next step would be to change small architecture elements, like the activation functions and the convolution kernel sizes to see if the tuned models can capture more or less nuances from the data. While these experiments have been carried out in other datasets, it would still be interesting to see how they affect the final explainability of the models, and how differently the models would then focus on features according to saliency maps and Grad-Cam. Another important experiment that has to be carried out is applying different explainability techniques to the models and see what they focus on. Other local methods like LIME (Ribeiro et al., 2016), SHAP (Lundberg and Lee, 2017), counterfactuals (Looveren and Klaise, 2020) and integrated gradients (Sundararajan et al., 2017) would still yield important results, but a global explainability method like TCAV (Kim et al., 2018) would be even more interesting, since it would explain the general performance of the model quantitatively instead of explaining instances, which can sometimes be misleading. TCAV, however, works by defining a concepts to test with, so careful consideration of concepts would need to be carried out.

References

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale.

Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24:381–395.

Harris, C. G. and Stephens, M. J. (1988). A combined corner and edge detector. In *Alvey Vision Conference*.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., and Sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav).

Looveren, A. V. and Klaise, J. (2020). Interpretable counterfactual explanations guided by prototypes.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints.

Lundberg, S. and Lee, S.-I. (2017). A unified approach to interpreting model predictions.

Mercaldo, F., Brunese, L., Martinelli, F., Santone, A., and Cesarelli, M. (2023). Explainable convolutional neural networks for brain cancer detection and localisation. *Sensors*, 23(17).

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2019). Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359.

Shaimaa, N., Ibrahim, Y., Hanan, A., and Mohamed, M. (2023). A robust mri-based brain tumor classification via a hybrid deep learning technique. *springer*.

Shanthi, S., Saradha, S., Smitha, J., Prasath, N., and Anandakumar, H. (2022). An efficient automatic brain tumor classification using optimized hybrid deep neural network. *International Journal of Intelligent Networks*, 3:188–196.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks.

Appendix

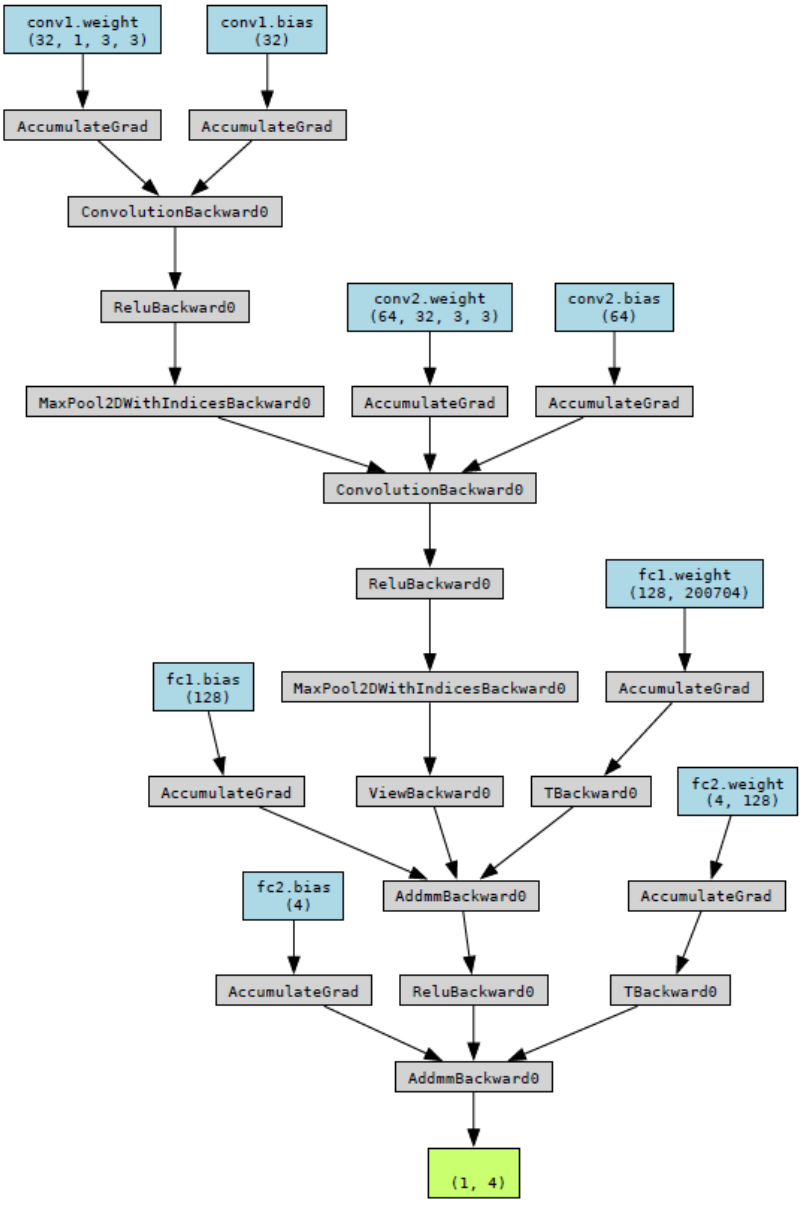

Figure 7: Simple 2 Layer CNN Architecture

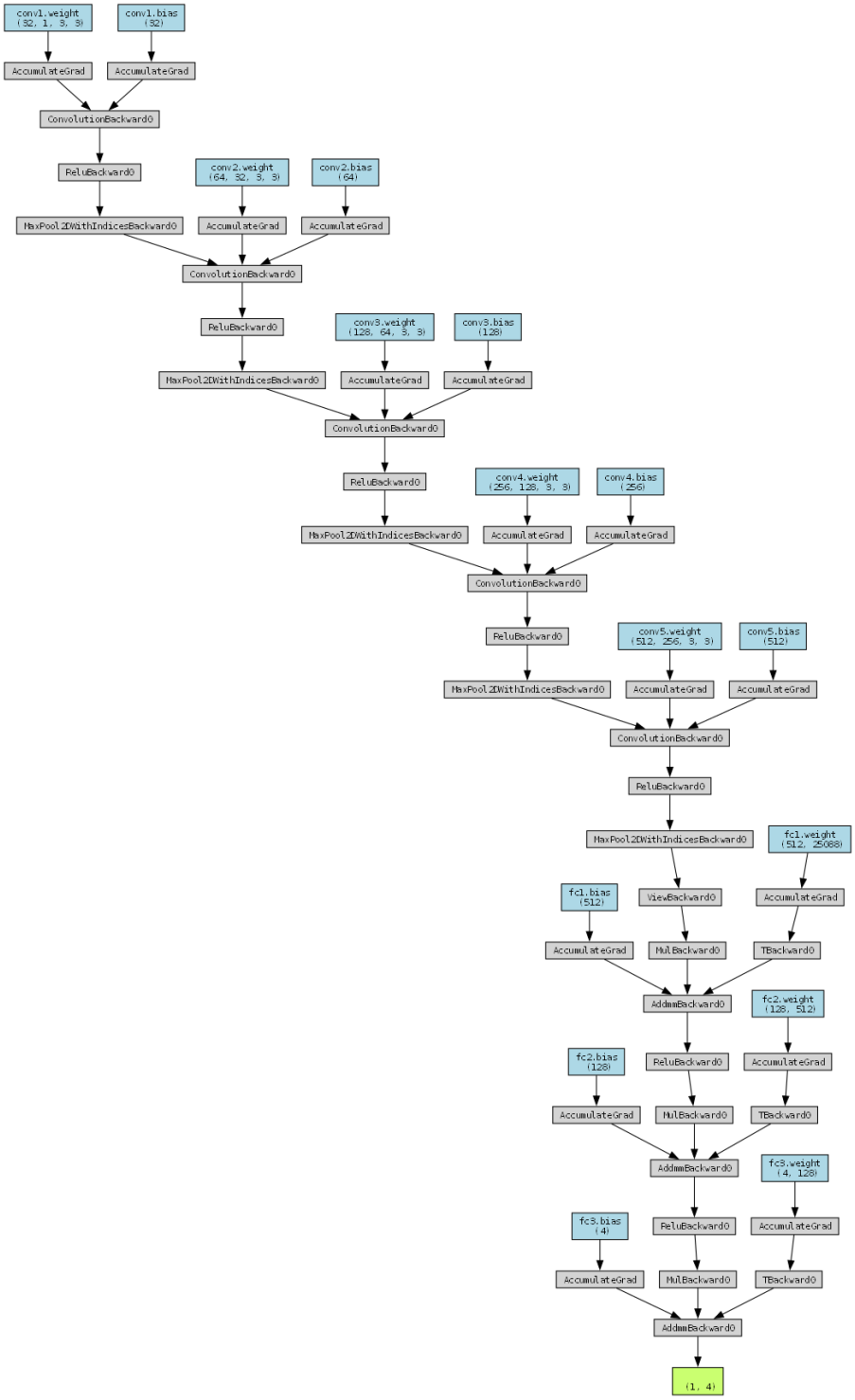

Figure 8: Custom 5 Layer CNN Architecture

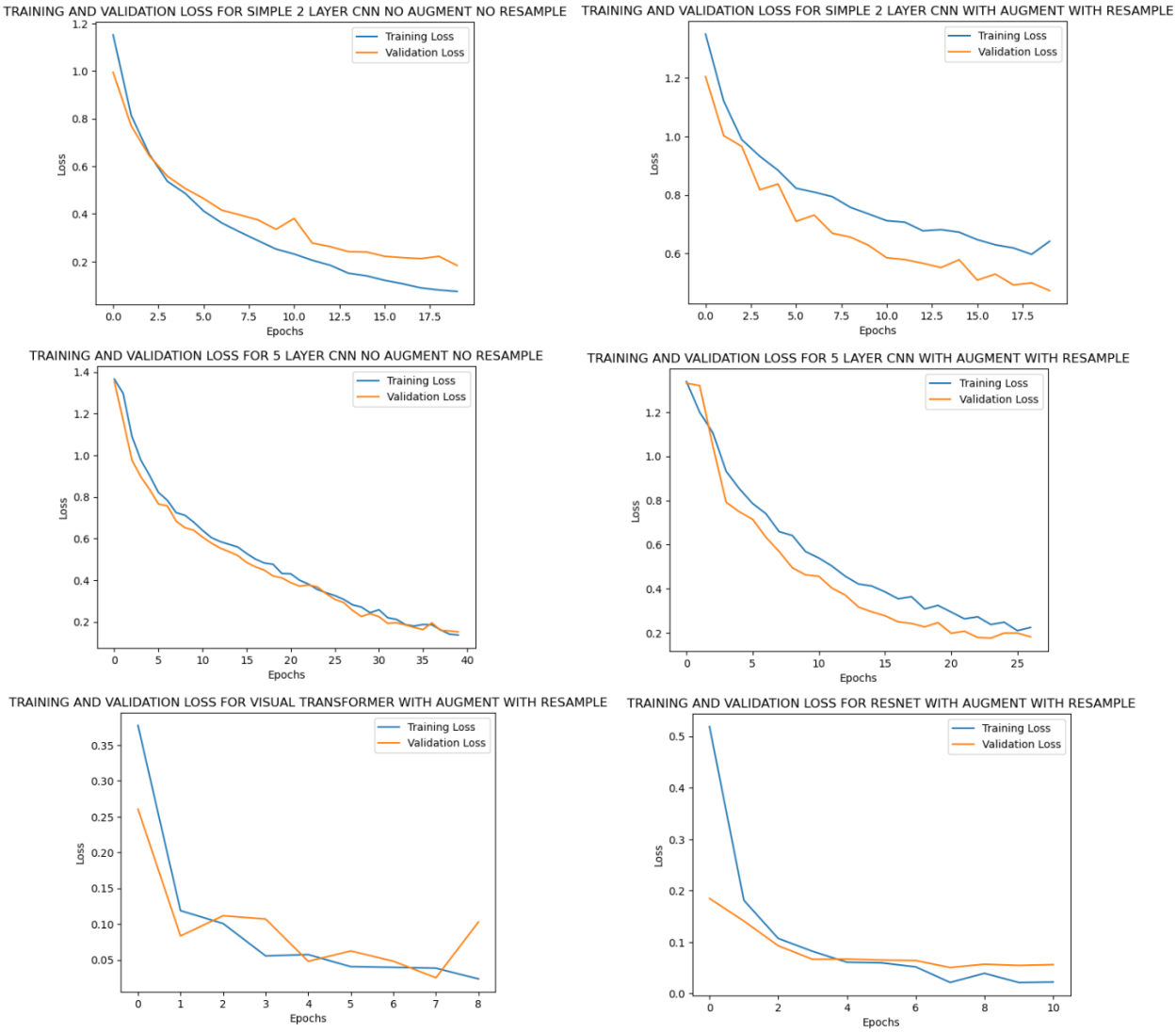

Figure 9: Training and Validation Loss Curves

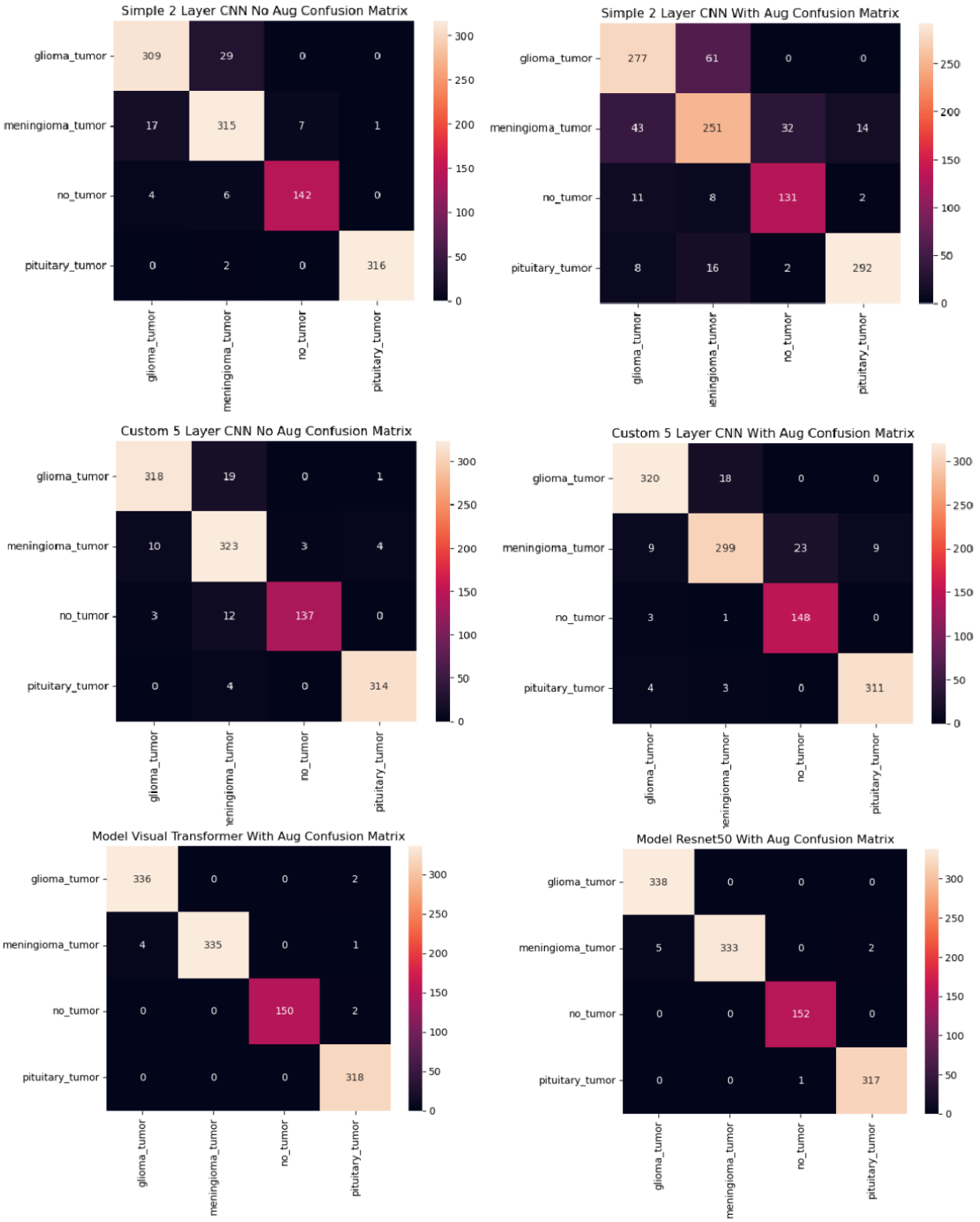

Figure 10: Confusion Matrices

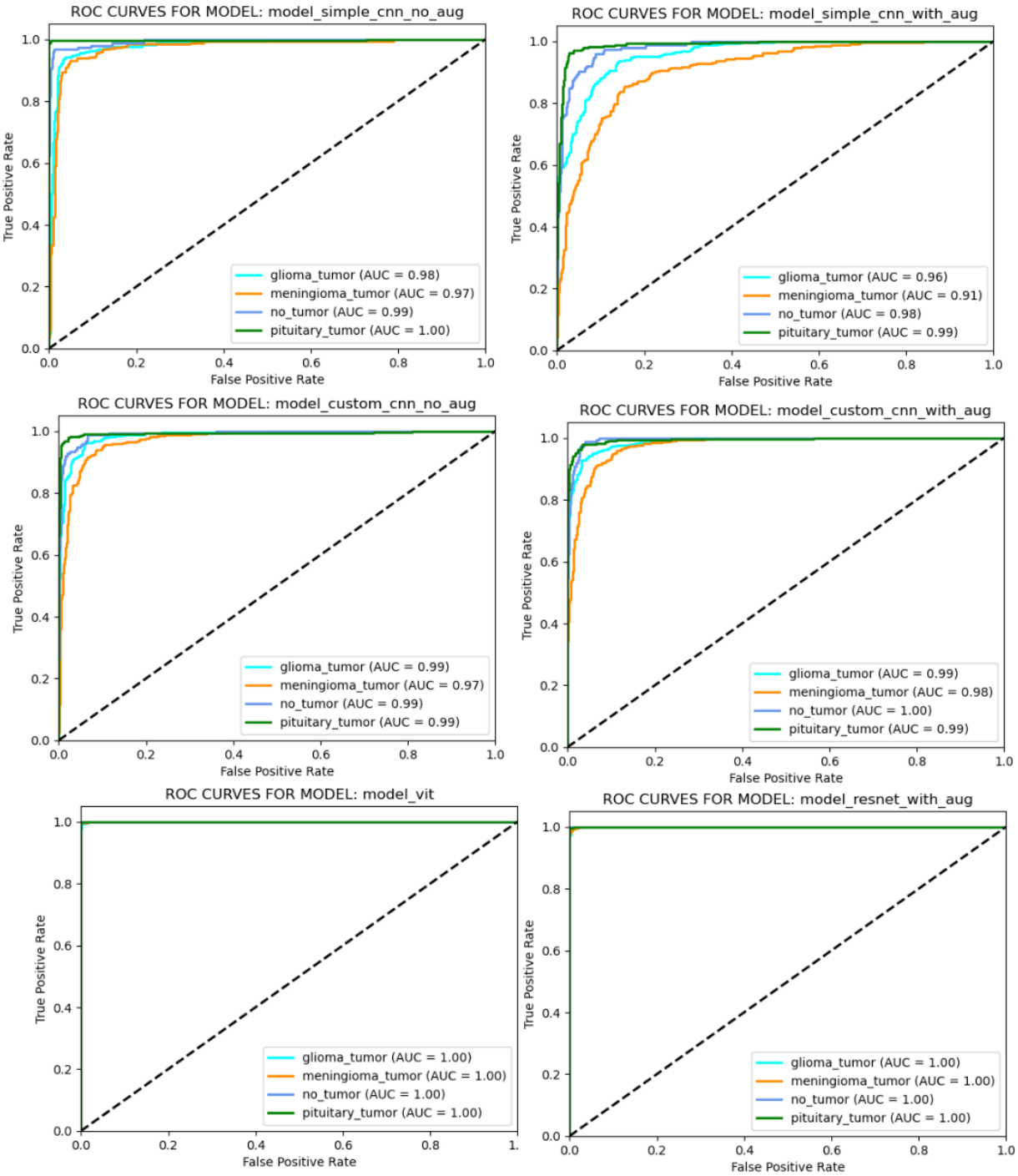

Figure 11: ROC Curves