Deep Learning for Cancer Classification

Stanford CS229 Project

Mustafa Abdelrahim Haroun Fadl
ICME
Stanford University
mustafaf@stanford.edu

Anirban Chatterjee
CGOE
Stanford University
ani1991@stanford.edu

Kai Li Tan
CGOE
Stanford University
kailitan@stanford.edu

1 Introduction

Cancer is the leading cause of death for 10 million people worldwide every year. Our ability to precisely detect and treat various forms of cancers early can improve a patient's chances of survival. In most of biomedical image detection, focus is usually around one particular type of disease instead of building a model which can generalize across different diseases, morphology, and imaging techniques. This can potentially create a bottleneck for diseases with lower data availability. In our work we focused on achieving that ability to generalize across a multitude of diseases with high enough accuracy. To that end input to our algorithm was a set of biomedical images which are then passed through different set of neural network model with the predicted output being one of 7 different classes (6 malignant cancer types + 1 healthy).

2 Related Work

1. Based on our literature review CNN based architecture [1] [2] or SVM + pre-trained deep learning classifiers for feature extraction[3] [4] is still the most used method in this domain. Application of transformer architecture is relatively rare in medical imaging.
2. Data availability can be an issue for diseases like breast cancer [2].
3. Most of the work is usually focused on one type of cancer at a time or variants of the same cancer (e.g. different types of skin cancers) and it is not clear whether this learning can be transferred between diseases. There have been some recent efforts to build novel architectures of cross-attention to improve the classification accuracy through multi-task learning [5].

3 Dataset and Features

Our dataset was collated from 4 different cancer types Brain MRI [6](Meningioma, Glioma and Pituitary), Breast ultrasound [7] (Breast cancer), Skin images [8] (Melanoma) and Morphological images using microscope (Acute lymphoblastic leukemia) [9]. We assigned each category of malignant cases into their individual groups and combined all benign/non-cancer cases into one to create a total of 7 classes (see examples in Figure 1). This decision and the fact that our breast cancer dataset was smaller compared to other datasets caused some imbalances in our data (as observed in Appendix Figure 6). In total we had 18000 images available for our project. We split that into 80%-10%-10% into the training/validation/test buckets.

Preprocessing/Normalization: As we were working with different image types (.png, .jpg and .bmp) with different resolutions, we converted them into PyTorch tensors of size (256, 256) or (224,

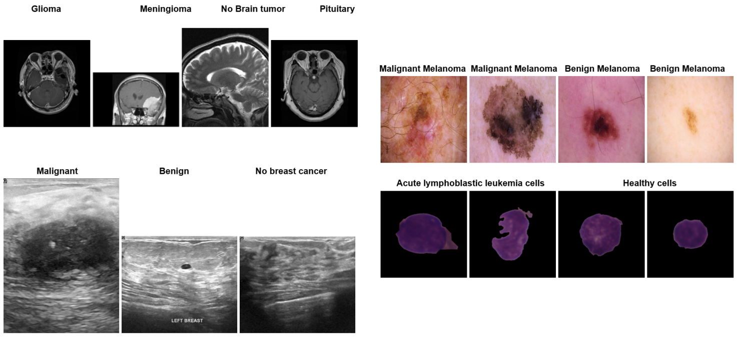

Figure 1: Examples from the dataset

224) depending on algorithmic and memory optimization necessity. Additionally, we performed algorithm-specific pre-processing as needed, for example in fine tuning using pre-trained ViT we applied normalization using mean 0.5 and standard deviation 0.5 with resampling and resizing. For Resnet-50 we normalize the image using ImageNet mean and std.

Data Augmentation: We used the Albumentation library in Python to perform a series of data augmentation to increase the validation and test accuracy, including increasing brightness, rotation, cropping, etc. We also tried augmentation to boost the number of samples in categories like breast cancer where we have a relatively smaller dataset.

4 Methods

In this section, we will describe the algorithms or models that we have used in our project. We experimented with four different models: **Multi-Layer Perception (MLP)**, **Convolutional Neural Network (CNN)**, **Resnet-50**, and **Vision Transformer (ViT)**. Each algorithm is briefly explained below together with its relevance to the task.

4.1 Multi-Layer Perceptron (MLP)

MLP is a neural network that consists of an input layer, multiple hidden layers, and an output layer. Each neuron in a layer is connected to all the neurons in the next layer. The output of each layer is given by the following equation:

$$z^{(l)} = W^{(l)} a^{(l-1)} + b^{(l)}$$
$$a^{(l)} = \sigma(z^{(l)})$$

where $W^{(l)}$ and $b^{(l)}$ are the weights and biases for layer l respectively, and σ is an activation function. MLP serves as a fundamental baseline for image classification tasks even though it doesn't have the ability to capture spatial dependencies in images. In our project, we used 2 hidden layers with sizes 256 and 128 respectively followed by the final output layer of 7 units.

4.2 Convolutional Neural Network (CNN)

These are networks that are specifically designed to process spatial data, which makes them suitable for image classification tasks. Unlike the MLP, CNNs extracts spacial information by applying learnable filters across the input. The convolution operation is defined as follows:

$$(I * K)(i, j) = \sum_m \sum_n I(i + m, j + n) K(m, n)$$

where I is the input image and K is the kernel or filter. In addition, CNNs also contain pooling layers, such as max pooling, which is used to reduce the spacial dimensions of the input. After a series of convolutional and pooling layers, the extracted features are passed through fully connected layers to produce the final classification output. Our model consisted of three blocks of convolution, ReLU, and pooling layers with kernel sizes of 3 and 2, respectively, followed by flattening and passing the output through a small MLP with a dropout of 0.5.

4.3 ResNet-50

This is a model that belongs to a subclass of CNNs known as residual networks. Residual networks are CNNs that has a special type of connection called residual connection the purpose of which is to address the problem of vanishing gradients in deep CNNs. The output of a single residual block can be written as follows [10]:

$$\mathbf{y} = \mathcal{F}(\mathbf{x}, \{W_i\}) + \mathbf{x}$$

where **x** and **y** are the input and output vectors of the layer considered and the function $\mathcal{F}(\mathbf{x}, \{W_i\})$ is the residual mapping to be learned. The residual neural networks are divide based on the number of layers, and the specific one that we have used in this project is the ResNet-50, which has 50 layers and its exact architecture is shown in Figure 4 in the Appendix section. For fine-tuning we removed the last output layer and added two linear layers. [11]. In addition, we already tried an ensemble model in which we combined the ResNet-50 and the MLP models to get better performance.

4.4 Vision Transformer (ViT)

This is a deep learning architecture that applies a transformer model. ViT uses the self-attention mechanism applying it to image patches rather than using convolutional filters. The input image is divided into patches which are embedded into feature vectors and then processed by the transformer layers. The self-attention mechanism is defined as follows [12]:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V$$

where Q, K, and V are the query, key, and value matrices, and d_k is the dimension of the keys and queries. The exact architecture for the vision transformer that we have used in this project is the one that have been introduced by Google in this paper [13], and the transformer encoder layer is shown in Figure 5 in the appendix [13].

where the Multi-Head Attention is just a multi-layer self-attention, and the norm is a layer normalization which normalize the activations along the feature dimension. We used a pretrained model from Google (via HuggingFace) to fine-tune using our data.

5 Experiments / Results / Discussion

We used F1-score and confusion matrix as our primary metrics because we have a highly imbalanced dataset. F1-score is found useful as it provides the harmonic mean of both precision and recall, and it can be calculated as such:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{1}$$

where precision and accuracy are defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

F1-Scores from our experiments are shown in Table 1.

Model	**F1-Score**
MLP	81.93%
CNN	82.81%
ResNet-50	87.93%
Ensemble (ResNet-50 + MLP)	88.56%
Vision Transformer	94.00%

Table 1: F1-Score of Models

As shown in the table, the F1-Score for MLP, which is our baseline model did not perform well on the dataset. Although the results were only slightly better after implementing CNN and ResNet-50 models, they still fell short of our expectations. To improvise performance further, we combined the MLP and ResNet-50 models, into an ensemble model which resulted in a better F1-Score. However, the F1-Score is still not to a satisfactory level. After exploring the vision transformer, we achieved our highest F1-Score of 94.00% which best fits our dataset. The vision transformer exhibited a balanced performance, with both precision and recall at 94.00%.

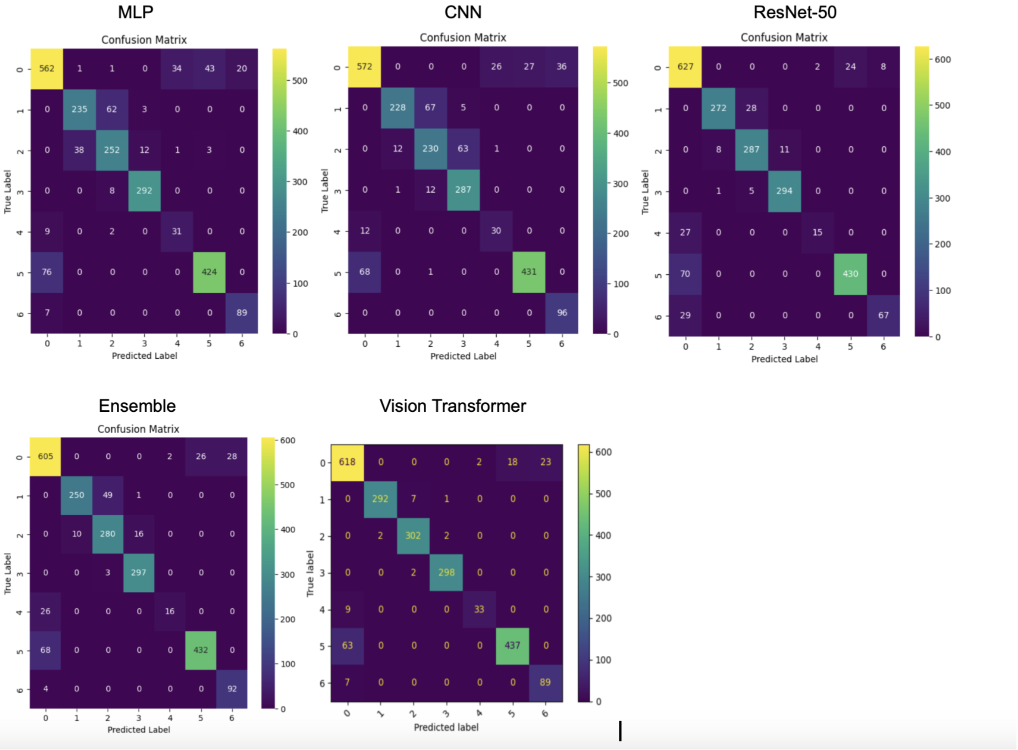

Figure 2: Confusion Matrix for All Models

We have plotted the confusion matrix for each model in Figure 2. According to the confusion matrix across models, brain cancer (class 1 and 2) classification improved as model complexity increases. Conversely, breast cancer (class 4) and leukemia (class 6) performed badly in general.

We captured the training and validation loss for all models in Figure 3. The training and validation loss for our baseline MLP model shows potential overfitting as there is a gap between training and validation loss noticeably after epoch 20. The validation loss initially declined but soon plateaued with fluctuations. Moving on to the CNN model, the gap between training and validation loss is smaller as compared to MLP, which suggest better generalization. The ResNet-50 training and validation loss show a generic downward trend with small gaps in between. They both seem to plateau out towards the last few epochs, meaning further training might not improve the performance. As for the ensemble model of MLP and ResNet-50, the training loss drops sharply and fluctuates heavily in both losses, but they are not visible on the graph shown due to large scaling. The vision transformer shows both training and validation losses decreasing and eventually level off, and the gaps between them are small. Towards the last few steps, there is a bigger gap which shows slight overfitting due to using a transformer.

Initially, we had stronger overfitting on the vision transformer due to limited data augmentation (brightness adjustments). To address this, we introduced targeted data augmentation, focusing on classes with poor performance by the confusion matrix. Specifically, the breast cancer (class 4) images were augmented by brightness, elastic transform, grid distortion, rotation, and horizontal flip. Leukemia (class 6) cell images were augmented with brightness, color jitter, sharpness, and elastic transform. Another mitigation method to overcome overfitting is using the cosine learning rate scheduler on the transformer, which improves convergence and performance. Overall, these combined

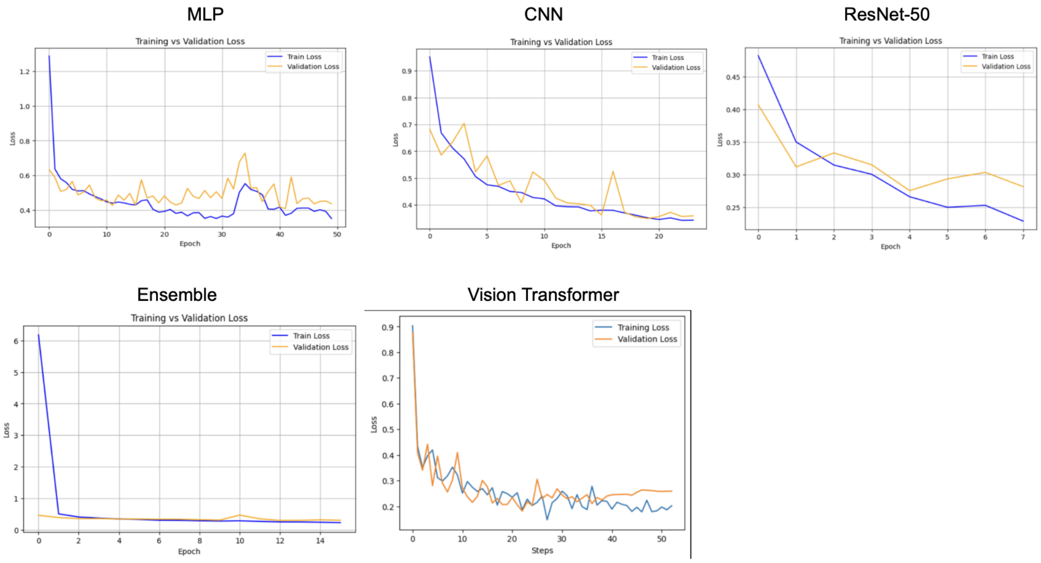

Figure 3: Training and Validation Loss for All Models

strategies resulted in enhanced model generalization as shown displayed in the classification report (Table 2).

Table 2: Classification Performance Metrics

Class Name	Class ID	Precision	Recall	F1-Score	Support
No cancer	0	0.89	0.93	0.91	661
Glioma	1	0.99	0.97	0.98	300
Meningioma	2	0.97	0.99	0.98	306
Pituitary	3	0.99	0.99	0.99	300
Breast cancer	4	0.94	0.79	0.86	42
Melanoma	5	0.96	0.87	0.92	500
Leukemia	6	0.79	0.93	0.86	96

6 Conclusion / Future Work

In this project, we explored different approaches to image classification to find the best-performing model for multimodal cancer classification. Our work showed that finetuning of comparatively simpler models like ResNet-50 can fail to generalize, whereas an ensemble or Vision Transformer can be a more effective architecture to achieve high accuracy across multiple diseases where data is generated using different imaging techniques. Future work may include training models on individual classes and then ensembling the models, or try a feature-wise transformation technique to handle the multi-modal nature of the problem, and try to increase and even out the amount of data for better performance.

7 Appendix

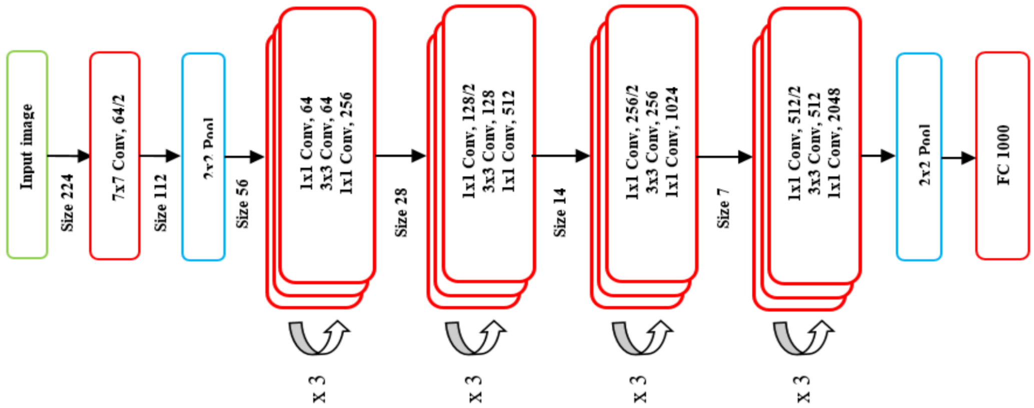

Figure 4: ResNet-50 architecture

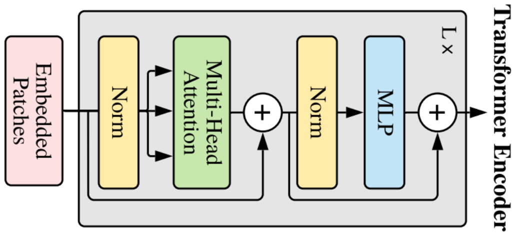

Figure 5: Transformer Architecture

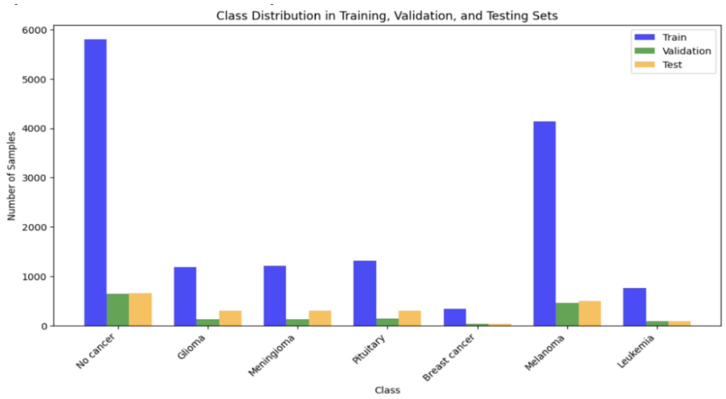

Figure 6: Class Distribution

8 Contributions

Mustafa: Contributed to data pre-processing, baseline MLP, CNN, ResNet-50, and Ensemble model.
Anirban: Contributed to data pre-processing, augmentation, MLP, CNN, Vision Transformer.
Kai: Contributed to data pre-processing, MLP, Vision Transformer, and augmentation.

References

[1] Y. N. Fu'adah, N. C. Pratiwi, M. A. Pramudito, and N. Ibrahim. Convolutional neural network (cnn) for automatic skin cancer classification system. In *IOP Conference Series: Materials Science and Engineering*, volume 982, page 012005. IOP Publishing, 2020.

[2] A. Raza, N. Ullah, J. A. Khan, M. Assam, A. Guzzo, and H. Aljuaid. Deepbreastcancernet: A novel deep learning model for breast cancer detection using ultrasound images. *Applied Sciences*, 13(4):2082, 2023.

[3] J. Kang, Z. Ullah, and J. Gwak. Mri-based brain tumor classification using ensemble of deep features and machine learning classifiers. *Sensors*, 21(6):2222, 2021.

[4] P. K. Das, V. A. Diya, S. Meher, R. Panda, and A. Abraham. A systematic review on recent advancements in deep and machine learning based detection and classification of acute lymphoblastic leukemia. *IEEE Access*, 10:81741–81763, 2022.

[5] S. Kim, T. G. Purdie, and C. McIntosh. Cross-task attention network: Improving multi-task learning for medical imaging applications. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 119–128. Springer Nature Switzerland, 2023.

[6] Msoud Nickparvar. Brain tumor mri dataset. `https://doi.org/10.34740/KAGGLE/DSV/2645886`, 2021. Accessed: 2025-03-14.

[7] Arya Shah. Breast ultrasound images dataset. `https://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset`, 2021. Accessed: 2025-03-14.

[8] Muhammad Hasnain Javid. Melanoma skin cancer dataset of 10000 images. `https://doi.org/10.34740/KAGGLE/DSV/3376422`, 2022. Accessed: 2025-03-14.

[9] A. Gupta and R. Gupta. All challenge dataset of isbi 2019. `https://doi.org/10.7937/tcia.2019.dc64i46r`, 2019. Accessed: 2025-03-14.

[10] 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 (CVPR)*, pages 770–778, 2016.

[11] Ridha Ilyas Bendjillali, Mohammed Beladgham, Khaled Merit, and Abdelmalik Taleb-Ahmed. Illumination-robust face recognition based on deep convolutional neural networks architectures. *Indonesian Journal of Electrical Engineering and Computer Science*, 18(2):1015–1027, 2020.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Daniel Weissenborn, Xinyi Zhai, Li Hou, Pei Zhai, Lutz Schubert, Florian Rombach, Michael Müller, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.