Breast Cancer Histological Analysis using Neural Networks

Charlie Rettig[1], Johannes von Kleist[2], Marcus Wenau[3]

[1]Bioinformatics Student, FU Berlin, Berlin, Germany.
[2]Bioinformatics Student, FU Berlin, Berlin, Germany.
[3]Bioinformatics Student, FU Berlin, Berlin, Germany.

Abstract

Breast cancer diagnosis relies heavily on histological analysis, where advances in artificial intelligence, particularly deep learning, offer the potential to enhance efficiency and accuracy. This study explores the application of neural networks to classify breast cancer histology slide images as benign or malignant using the BreaKHis dataset. Models ranged from a shallow neural network (AUC = 0.85) to advanced pre-trained CNNs like ResNet [3], DenseNet [4], and VGG [7], achieving near-perfect performance on the training dataset (AUC = 1.00). While the shallow network demonstrated comparable performance to traditional methods, its high false-negative rate raises concerns for clinical deployment. Conversely, pre-trained models showed superior accuracy but struggled to generalize to external datasets due to overfitting to color artifacts. This underscores the need for robust training datasets and architectural refinements. Our findings highlight the promise and limitations of neural networks for breast cancer diagnosis, emphasizing the importance of external validation to ensure clinical reliability.

Keywords: Breast cancer, Neural networks, Histological analysis, Deep learning

1 Introduction

Breast cancer is one of the most common types of cancers worldwide[9], with early detection being the most important step for patient survival.

The fundamental task in breast cancer diagnosis is distinguishing between benign and malignant tumors. This binary classification problem is well-suited for Neural Network diagnosis, where the neural networks can be trained on labeled histopathological images.

The success of neural networks (NNs) in this domain can be attributed to their ability to identify subtle patterns and features in medical images that might not be immediately apparent to human observers.

The development of an automated diagnostic tools using neural networks would bring the following advantages:

- Increased efficiency
- Support for pathologists in decision-making (show images of low certainty)
- Reduction in diagnostic errors
- Standardization of diagnostic criteria
- Improved accessibility to diagnostic services in more places

Considering the advantages of neural networks in diagnostics, investing time and resources into developing neural networks to differentiate malignant and benign breast cancer tissues would be a worthwhile endeavor.

2 Goal

The goals of our methodology were the following:

1) The implementation of multiple neural networks to determine the most effective approach or even create a more effective ensemble of the Neural Networks for the detection and categorization of visual patterns that can be used to differentiate Benign and malignant tumors

2) To evaluate the accuracy and efficiency of our neural networks. To see how viable they are for clinical use

This project could prove significant in the diagnosis of malignant breast tumors.

3 Data and Preprocessing

To train our models, we used the BreaKHis dataset, consisting of histological images of breast cancer tissue [8]. For our models, we used the full initial dataset of 7,909 images of breast cancer tissue samples. The original image dimensions were 700x460 pixels. The dataset exhibits an imbalanced distribution (Figure 1), with approximately two-thirds (n=5429) of the images showing malignant tumors and one-third (n=2480) showing benign cases (Figure 1). This imbalance is present in all magnification subsets.

In hopes of counteracting potential artifacts associated with the partially subjective nature of the data collection process described by Spanhol et. al (namely, selection of regions of interest), we decided to maximize the robustness of our dataset by including all magnifications present in the original dataset (40X, 100X, 200X, and 400X).

The data was split into three groups: a training set, a test set, and an external validation set. The original class distribution ratio (benign to malignant) was maintained across all three sets to ensure representative sampling. For some models, the class distribution was balanced by either over-sampling of the minority class or under-sampling of the majority class. Model-specific dataset split information can be found in the Methods section.

We additionally used the BACH dataset [1] for external validation of our most performant model. The dataset is relatievly similar to the previous one in the sense that

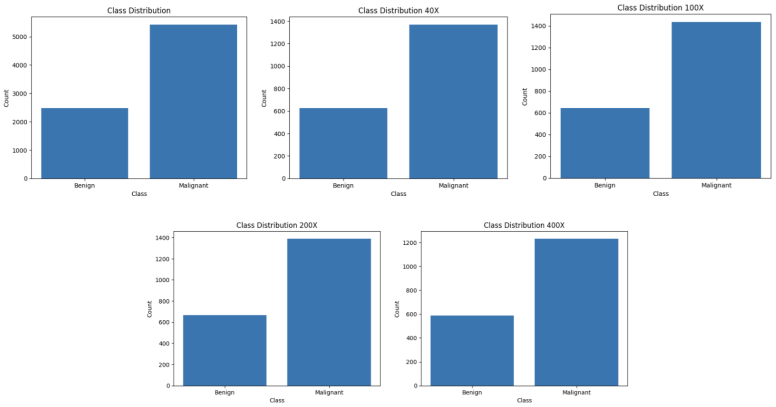

Fig. 1 Bar plots of class distribution in original dataset.

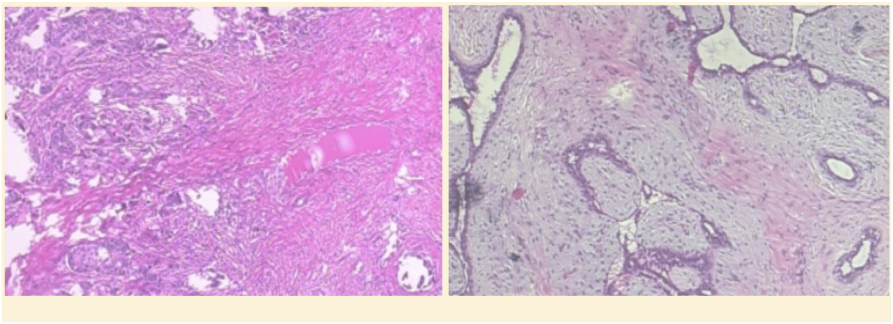

Fig. 2 Example images of malignant (left) vs. benign (right) tissue under 40X magnification.

it is made via the same staining method, however it includes only 40x magnification images. Further information can be found in the article by Aresta et al. [1].

3.1 Image Augmentation

Visual inspection of malignant and benign images revealed patterns that could give insights into what features a model could learn, which thus informed data augmentation decisions. For example, across the vast majority of images, no grayscaling was done because malignant tumors exhibited higher color intensity than benign samples, likely the result of increased stain absorption by malignant tumors.

The images were augmented in different ways. Exemplary images of each class (non-augmented) are shown in Figure 2. Augmentations made included resizing, rotation, color normalization, among other changes. Model-specific augmentation information is detailed further in the Methods section.

3.2 Missing Data and Imputations

There were no missing data identified in the dataset, as all data (tumor class, magnification, etc.) were extracted directly from encodings in the image file names. Only one image was found to be corrupted.

Binary class labels were extracted as strings from the file names ('B' for benign and 'M' for malignant samples) and converted to numeric 0 and 1 labels (0 for benign, 1 for malignant).

4 Methods

We employed five different approaches for the classifications of the tumors: a shallow neural network, a self-programmed CNN, 3 different pre-trained CNNs (ResNet, VGG and DenseNet) and finally an ensemble of the three pre-trained CNNs.

4.1 Exploratory Analysis

Exploratory analysis consisted of class distribution analysis for each of the microscope magnifications included in the BreaKHis dataset. Plots were generated using Matplotlib ([5]) version 3.9.2.

4.2 Model Construction

Data was stored using Python's NumPy ([2]) version 1.26.2 and models were constructed using TensorFLow ([6]) version 2.18.0. Plotting was performed with Matplotlib ([5]) version 3.9.2.

4.3 Model Evaluation

Each model was evaluated using accuracy, and by constructing a confusion matrix and a ROC curve. Precision and recall were subsequently derived from the confusion matrices. To gain insight into the learning process for hyperparameter tuning, learning curves (derived from overall accuracy) were additionally constructed and analyzed.

4.4 Shallow neural network

The shallow NN is a simple feed-forward neural network. It begins with an input layer that takes flattened image data and consists of 2,352 neurons (28 pixels x 28 pixels x 3 color channels). The model has one hidden layer with 64 neurons and ReLU activation, followed by a single-neuron output layer with a sigmoid activation function for binary classification. The model uses binary cross entropy for a loss function and Adam optimization.

4.4.1 Architecture design

In the following section the design of the shallow NN architecture will be described.

- Input layer: 2,352 neurons

- Hidden layer: 64 neurons, ReLU activation
- Output layer: 1 neuron, Sigmoid activation

4.4.2 Data Processing and Augmentation

The input image dimensions were reduced from their original 700x460 pixels to a small 28x28 pixels. Original color channels (RGB) were used. All pixel color intensities were normalized by scaling to a 0-1 range.

4.4.3 Training Configuration

The following parameters were used for training:

- Batch size: 128 images
- Epochs: 32
- Learning rate: 0.001
- Learning rate adaptation: Adam optimization

4.4.4 Evaluation

Evaluation was done via a test-training split, 80% of the original data belonging to the training set, 10% to a test set, and the last 10% to a further validation set. The metrics we focused on, both during training and afterward, were accuracy, precision, and recall. Additionally, ROC curves and the associated area under the curve (AUC) were also used as metrics for evaluating the performance of our classifiers. The loss function was also plotted against the progression of epochs during training for evaluation of learning efficiency.

4.5 Convolutional neural network

For our self-programmed CNN implementation, we designed a sequential architecture optimized for processing histopathological images of breast tissue. The network consists of two convolutional blocks followed by three dense layers, structured to balance computational efficiency with feature extraction capability.

4.5.1 Architecture design

In the following section, the design of the CNN architecture will be described.

The network begins with two convolutional blocks:

- First block: 16 filters with 3×3 windows, followed by ReLU activation and max pooling
- Second block: 32 filters with 3×3 windows, also followed by ReLU activation and max pooling

The convolutional layers are followed by a flattening operation and three dense layers:

- First dense layer: 128 neurons with ReLU activation
- Second dense layer: 32 neurons with ReLU activation

- Output layer: Single neuron with sigmoid activation for binary classification

4.5.2 Data Processing and Augmentation

Input images were processed at a resolution of 700×460 pixels with 3 color channels (RGB). To enhance model robustness and prevent overfitting, we implemented data augmentation with the following parameters:

- Rotation: up to 20% rotation
- Zoom: 10% range
- Shear: 10% range
- Horizontal flip: disabled
- Width and height shifts: disabled
  All images were normalized by rescaling pixel values to the range [0,1].

4.5.3 Training Configuration

The model was trained using batch processing with:

- Batch size: 32 images
- Learning rate adaptation:
- Monitored on the accuracy on the validation test
- Patience of 2 epochs
- A Learning rate reduction factor of 0.5
- The Minimum learning rate was 0.00001
- Early stopping: Implemented to prevent overfitting
- Patience of 5 epochs
- Monitored on the accuracy on the validation test

4.5.4 Evaluation

Evaluation was done via a test-training split, 80% of the original data belonging to the training set, 10% to a test set, and the last 10% to a further validation set. The evaluation was done by calculating the precision accuracy and specificity, as well as an ROC curve. These metrics we calculated based on the test set which wasn't used in training and consist of 12.5% of the data.

4.6 Pre-trained models

On top of the de novo models we trained, we also used three pre-trained models as bases for classifiers: ResNet (He et al. [3]), DenseNet (Huang et al. [4]), and VGG19 (Simonyan et al. [7]). These were accessed through Keras and Tensorflow ([6]). These models were then also combined into an ensemble model, which determined classification via majority vote.

4.6.1 Architecture design

On top of these pre-trained models, we built some very simple classification layers. These generally consisted of a pooling layer, a dense layer of 512 neurons in most cases,

a dropout layer with 0.5 dropout set and an output layer with a sigmoid activation function.

4.6.2 Data Processing and Augmentation

The input image dimensions were reduced from their original 700x460 pixels to a small 224x224 pixels. Original color channels (RGB) were used. All pixel color intensities were normalized by scaling to a 0-1 range. Augmentation was undertaken using the same parameters from the CNN.

4.6.3 Training Configuration

These models were trained using batch processing with:

- Batch size: 32 images
- Learning rate adaptation: Monitored on the accuracy on the validation test
- Patience of 3 epochs
- A Learning rate reduction factor of 0.5
- The Minimum learning rate was 0.00001
- Early stopping: Implemented to prevent overfitting
- Patience of 10 epochs
- Monitored on the accuracy on the validation test

4.6.4 Evaluation

Evaluation was done via a test-training split, with 10% split between training and test data, and a further 10% split between training and validation data, i.e. data used between epochs. The metrics we focused on, both during training and afterward, were accuracy, precision, and recall. Additionally, ROC curves and the associated area under the curve (AUC) were also used as metrics for evaluating the performance of our classifiers.

5 Results

5.1 Shallow NN

The performance of the binary classification model for breast cancer histology images was evaluated using a confusion matrix and a Receiver Operating Characteristic (ROC) curve. The confusion matrix (Figure 4, left) demonstrated that the model correctly classified 185 negative samples (True Negatives) and 211 positive samples (True Positives). However, the model also misclassified 63 negative samples as positive (False Positives) and 37 positive samples as negative (False Negatives). These results suggest the model has a moderate capacity to distinguish between malignant and benign histology samples. However, there is clinically significant risk associated with the model's struggle with false negatives, which could result in missed cancer diagnoses.

The ROC curve (Figure 4, right) yielded an Area Under the Curve (AUC) of 0.85, indicating good discriminative ability between the two classes. While the model outperformed random classification, the AUC value highlights room for improvement

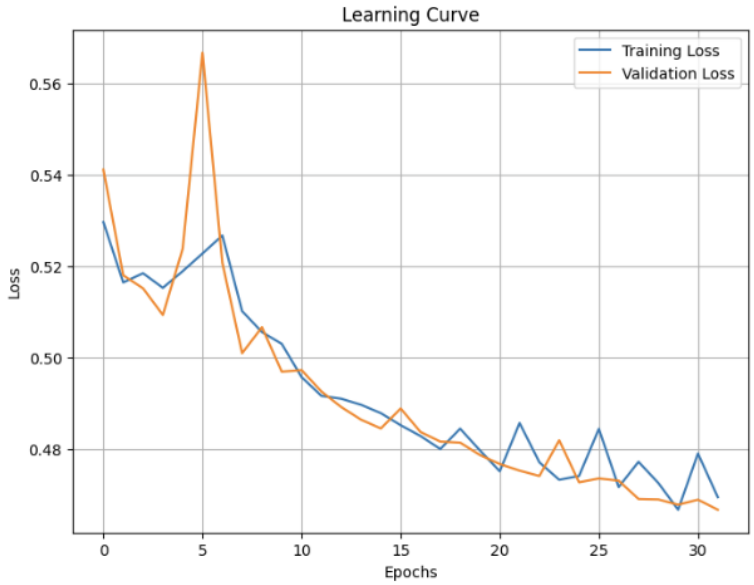

Fig. 3 Learning curve (training and validation loss over 32 epochs) for the Shallow NN.

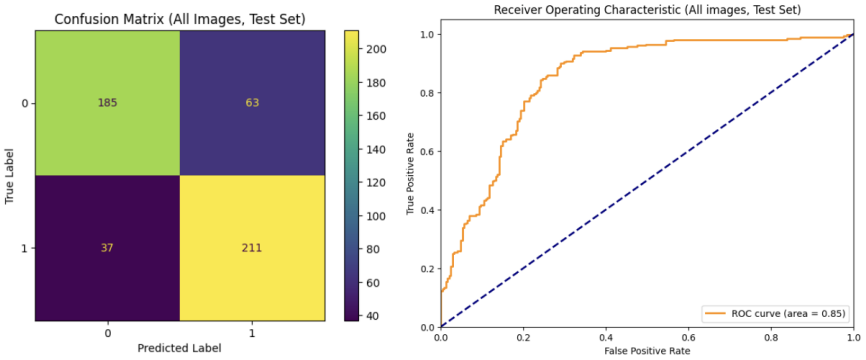

Fig. 4 Confusion matrix and ROC curve of shallow NN

in overall sensitivity and specificity. The relatively high number of false negatives emphasizes the need to optimize the model further, particularly to minimize missed cancer cases, which are critical in clinical settings. Future work will focus on improving model sensitivity through threshold tuning, enhanced data preprocessing, and potential implementation of more sophisticated architectures.

5.2 CNN

The confusion matrix (Figure 6, left) showed that the model correctly identified 250 malignant samples (True Positives) and 84 benign samples (True Negatives). The model misclassified 45 benign samples as malignant (False Positives) and 32 malignant

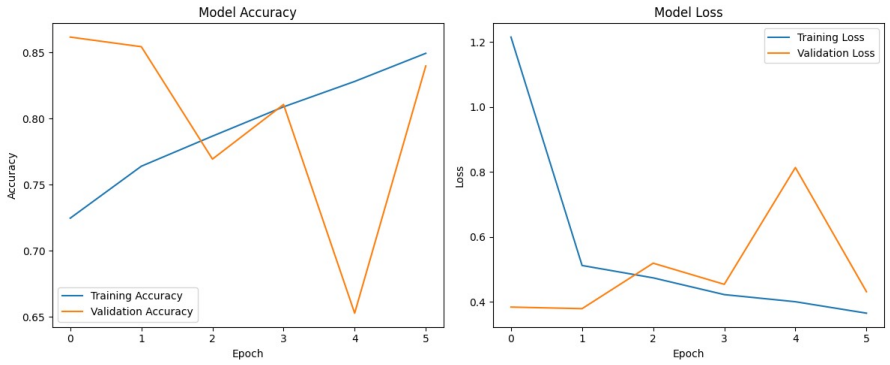

Fig. 5 Learning curve (training and validation loss over 5 epochs) for the CNN.

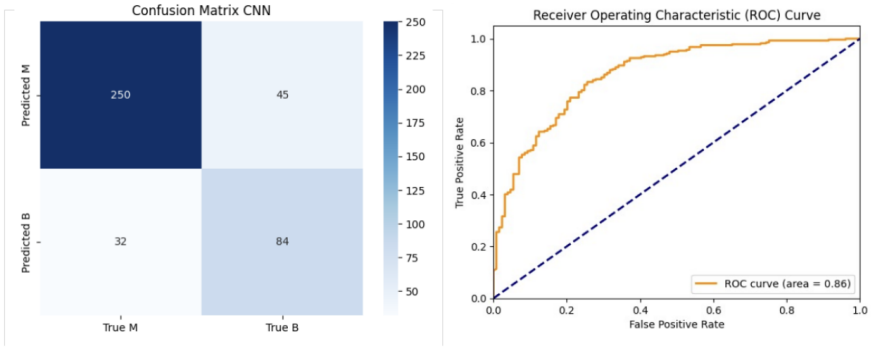

Fig. 6 Confusion matrix and ROC curve of CNN

samples as benign (False Negatives). These results indicate strong performance in identifying malignant cases, but with some limitations in correctly classifying benign samples. The relatively low number of false negatives suggests the model is conservative in its predictions, reducing the risk of missed cancer diagnosis.

The ROC curve (Figure 6, on the right) achieved an AUC of 0.87 demonstrating a sufficient ability to differentiate between malignant and benign images. The model showed a high sensitivity with 88.6%. The specificity on the other hand was rather low with 65.1% leaving much room for improvement. The precision of 84.7% suggests that when the model predicts a malignant case, it is correct in the majority of instances.

5.3 Pre-trained models

The pretrained models were quite performant on the BreakHis dataset, reaching accuracies close to 100% and AUC close to 1.00. This becomes evident looking at the plotted ROC curves of the three models leveraging pre-trained architectures as well as their confusion matrices. The ROC curve with AUCs is given below, the rest of the

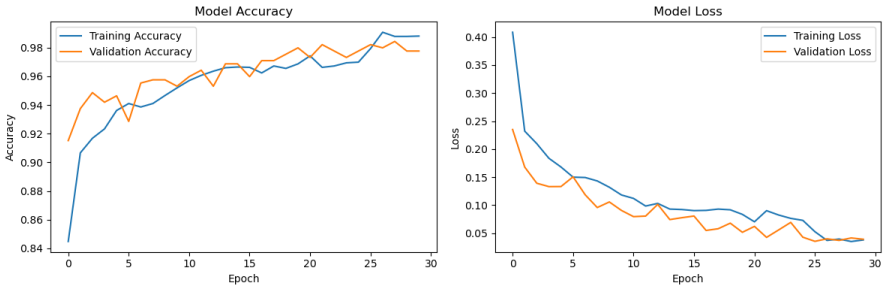

Fig. 7 Learning curve (training and validation loss over 30 epochs) for the classifier built on ResNet

confusion matrices and learning curves for DenseNet and VGG are provided in the appendix.

As is visible in Fig. 8, the performances of all models are excellent on the separate test dataset, with Resnet and the ensemble model performing the best with an AUC of 1.00, this is shared by VGG and followed up by an AUC of 0.96 of Densenet. Looking more closely at the performance of Resnet, it boasts an impressive accuracy of 98.4%, with a precision of 1 and a recall rate of 97.6% on the test data. The training history of the classifier trained on the 400x images is visible in Fig. 7.The confusion matrix of Resnet trained on 400x magnification images is provided in Fig. 9. Based on the model's higher performance, we decided to move forward with Resnet when it came to tests with generalization. At first, we generalized to all magnifications, while maintaining our simple binary classification system. This resulted in slightly lower, yet still quite high performance, as is visible in the confusion matrix in Fig. 10. Over all, Resnet performed best both on the 400x and general dataset, with the difference to the other models being, admittedly, relatively negligible.

However, testing on the external BACH validation dataset [1], even for the most performant of the models, Resnet, produced mediocre results, with an accuracy of just about 60% and exclusively malignant classifications. This was true for a specific 40x model and a generalized model over all magnifications, with the confusion matrix of the generalized model given below. The results of testing the Resnet model trained on the full BreakHis dataset are visible in Fig. 11, with the poor performance and high false positive rate quite immediately apparent.

6 Discussion

6.1 Shallow neural network

The shallow neural network achieved an area under the curve (AUC) of 0.85, demonstrating a comparable performance to the state-of-the-art classifier reported by Spanhol et al. (SVM, AUC = 0.86, trained on 200X magnification). However, the model's clinical applicability is constrained by its error profile, with a notable number of false negatives (37) and false positives (63). While false positives primarily result in additional diagnostic steps, false negatives carry a significant risk, as malignant tumors may be misclassified as benign, potentially delaying critical treatment.

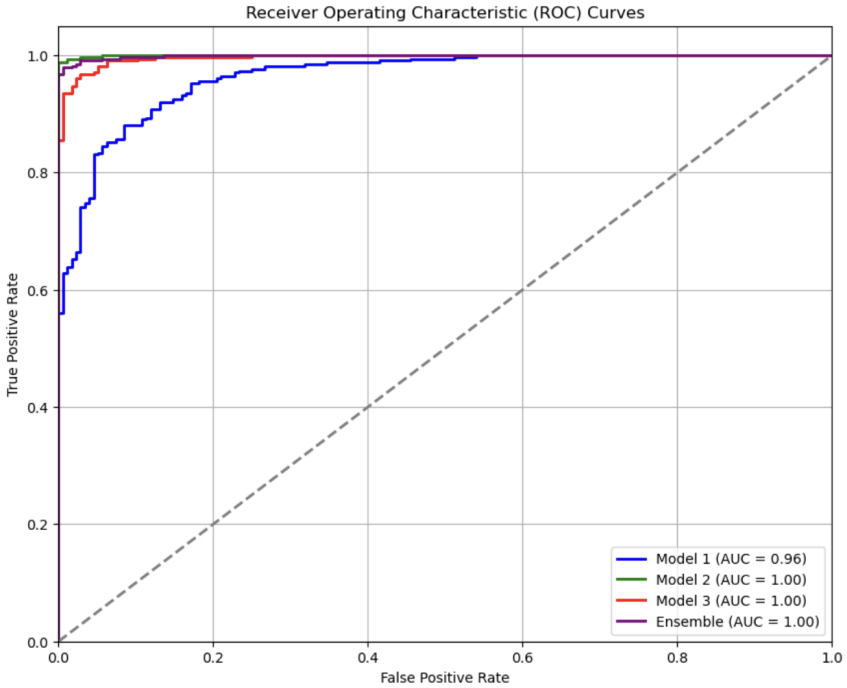

Fig. 8 ROC curves of the different pre-trained classifiers on 400x images. Model 1: Densenet, Model 2: ResNet, Model 3: VGG, Ensemble.

This highlights the need for further optimization or integration with complementary methods to reduce the false-negative rate. Despite these limitations, the comparable AUC indicates that shallow networks remain a viable, less computationally intensive alternative for histological image classification, warranting further exploration in resource-constrained settings or as a component in hybrid diagnostic pipelines.

6.2 CNN

The CNN achieved an accuracy of 81.3% falling just shy of the 85% Benchmark of the original paper. The model demonstrated strengths in identifying malignant cases (88.6% sensitivity) but struggled with specificity (65.1%), particularly in classifying benign samples. Despite not matching the original study's accuracy, our model demonstrates the promising potential of deep learning in automated breast cancer detection.

6.3 Pre-trained models

The classifiers utilizing pre-trained networks outperform the classification methods in the original paper, with the original methods obtaining accuracies of ca. 85% on training dataset and the classifiers using pre-trained models obtaining 95% or more, with high specificity and sensitivity as mentioned previously. However, these models

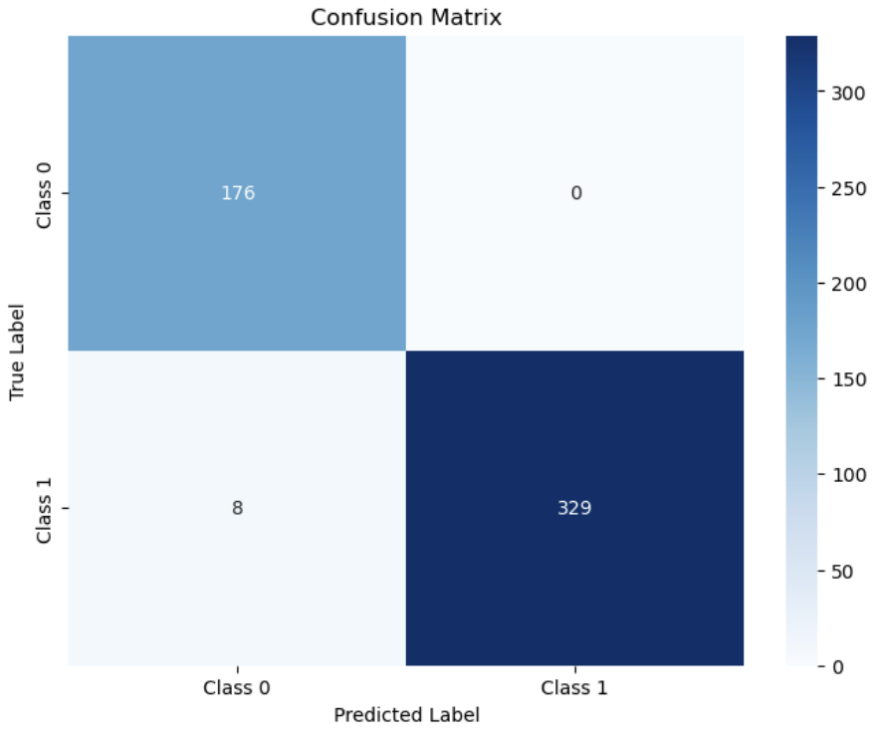

Fig. 9 Confusion matrix of the ResNet classifier, trained and tested on the 400x images.

overfit to the BreakHis [8] dataset, and more specifically, probably the color differences between benign and malignant tumors. This conclusion is supported by two pieces of evidence: the models did not learn patterns nearly as well when presented with grayscale images and the model failed to generalize to a separate external validation dataset, the BACH-dataset [1]. This overfitting could be overcome via several different avenues. One such possible avenue could be model-based, such as increasing the dropout rate further, increasing the augmentation, or more sophisticated color normalization. A further model-based solution would be to attempt to use different architectures or combinations of architectures, such as more classic ML methods such as SVMs or even more advanced deep learning methods such as visual transformers. Another separate, yet possible avenue for solutions would be to obtain more data, more specifically not only a larger amount of images, but also images from different hospitals. These would then hopefully have diverse enough "biases" in their images that the model learns to abstract away confounding influences such as uneven coloring and learns actual differentiating features between benign and malignant tumors. The problem of machine learning algorithms only learning data from a single hospital well is a well-documented problem ([10]), and as such did not come as a complete surprise in the evaluation of our classifiers.

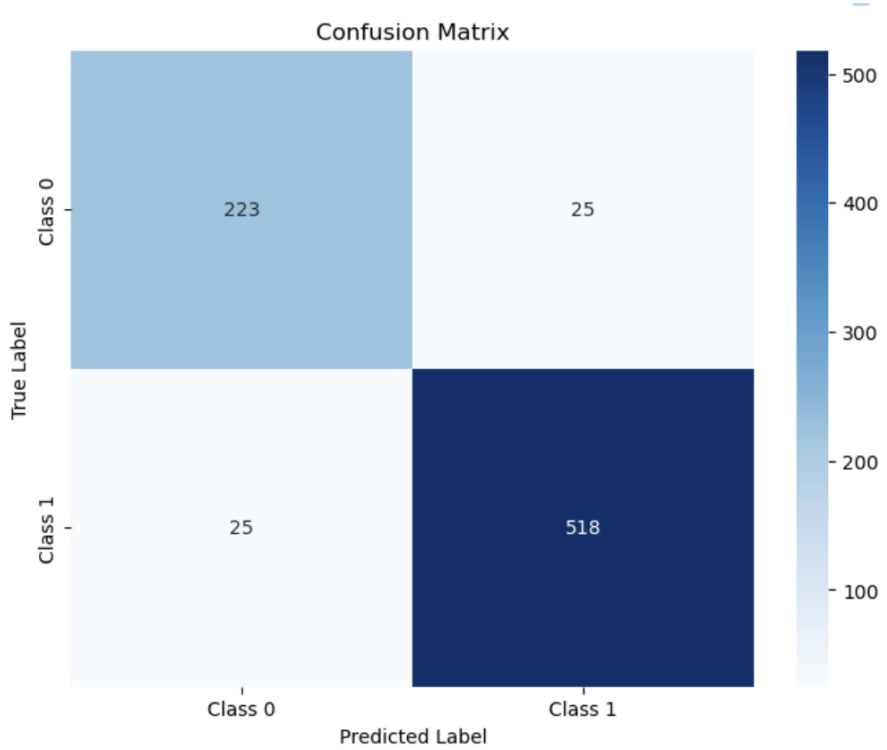

Fig. 10 Confusion matrix of the ResNet classifier, trained and tested on all BreakHis images.

Overall, deep learning methods, specifically CNNs, can provide powerful tools for image analysis, however the results on a single dataset should be treated cautiously and care must be taken when training and developing these classifiers to provide ways to circumvent overfitting.

Appendix A Contributions

The contributions of the group members were roughly as follows:

- Charlie:
  - Shallow NN construction and analysis (Methods, Results, and Discussion)
  - Exploratory data analysis
- Marcus:
  - CNN construction and analysis (Methods, Results, and Discussion)
  - Scientific background research
- Johannes:
  - Pre-trained models (Methods, Results, and Discussion)
  - Methods research

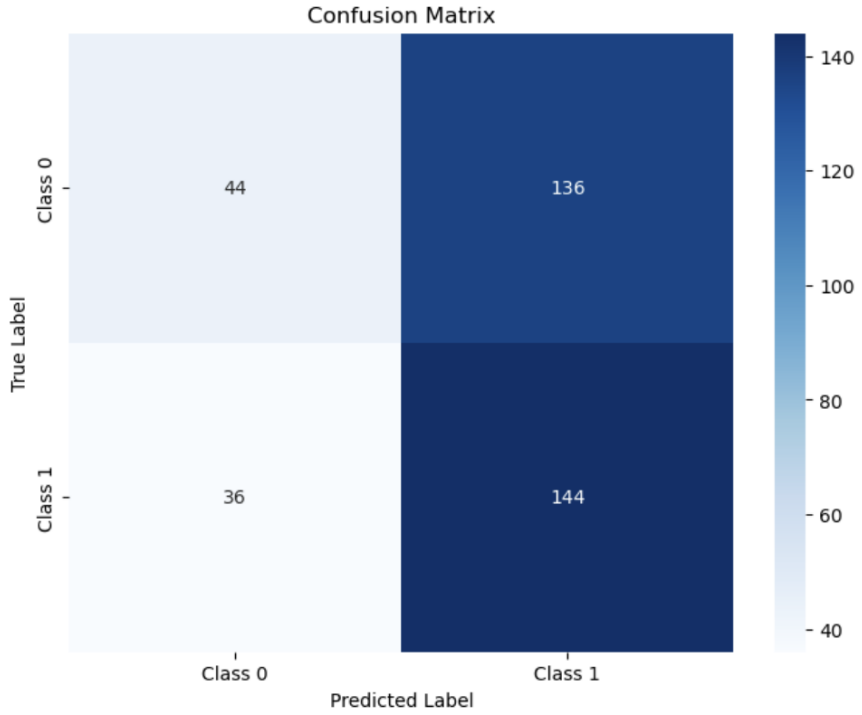

Fig. 11 Confusion matrix of the full ResNet classifier on the BACH images.

– External validation dataset

References

[1] Guilherme Aresta et al. "BACH: Grand challenge on breast cancer histology images". In: *Medical Image Analysis* 56 (Aug. 2019), pp. 122–139. ISSN: 1361-8415. DOI: 10.1016/j.media.2019.05.010. URL: https://www.sciencedirect.com/science/article/pii/S1361841518307941 (visited on 11/19/2024).

[2] Charles R. Harris et al. "Array programming with NumPy". In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. DOI: 10.1038/s41586-020-2649-2. URL: https://doi.org/10.1038/s41586-020-2649-2.

[3] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *arXiv preprint arXiv:1512.03385* (2015). Accessed: 2024-11-19. URL: https://arxiv.org/abs/1512.03385.

[4] Gao Huang et al. "Densely Connected Convolutional Networks". In: *arXiv preprint arXiv:1608.06993* (2016). Accessed: 2024-11-19. URL: https://arxiv.org/abs/1608.06993.

[5] J. D. Hunter. "Matplotlib: A 2D graphics environment". In: *Computing in Science & Engineering* 9.3 (2007), pp. 90–95. DOI: 10.1109/MCSE.2007.55.

[6] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: https://www.tensorflow.org/.

[7] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *arXiv preprint arXiv:1409.1556* (2014). Accessed: 2024-11-19. URL: https://arxiv.org/abs/1409.1556.

[8] F. A. Spanhol et al. "A Dataset for Breast Cancer Histopathological Image Classification". In: *IEEE Transactions on Biomedical Engineering* 63.7 (2016), pp. 1455–1462. DOI: 10.1109/TBME.2015.2496264. URL: https://doi.org/10.1109/TBME.2015.2496264.

[9] World Health Organization. *Breast Cancer: Fact Sheet*. Accessed: 2024-11-19. 2024. URL: https://www.who.int/news-room/fact-sheets/detail/breast-cancer.

[10] Jenny Yang, Andrew A. S. Soltan, and David A. Clifton. "Machine learning generalizability across healthcare settings: insights from multi-site COVID-19 screening". en. In: *npj Digital Medicine* 5.1 (June 2022). Publisher: Nature Publishing Group, pp. 1–8. ISSN: 2398-6352. DOI: 10.1038/s41746-022-00614-9. URL: https://www.nature.com/articles/s41746-022-00614-9 (visited on 11/19/2024).