
BrCaVision: Predicting Breast Cancer Prognosis by Detecting Mitosis, Identifying Histological Tumor Subtypes, and Scoring HER2 in Whole Slide Tissue Images Using Deep Learning

Dheepthi Mohanraj

North Carolina School of Science and Mathematics

Durham, North Carolina, United States of America

Table of Contents

1. Summary	4
2. Introduction	4
3. Research Questions	8
4. Definition of Terms	8
Support Vector Machines	8
Deep Learning and Convolutional Neural Networks	8
Transfer Learning and Fine-Tuning	9
Base Model Characteristics	10
Whole Slide Images	12
Hematoxylin & Eosin Stain (H&E)	12
5. Framework	13
Technology Stack and Infrastructure	13
Data Preprocessing and Data Augmentation	14
Histological Subtype Classification Model	14
Dataset Description	14
Training Details	16
Heatmap	17
Mitosis Identification Model	18
Dataset Description	18
Training Details	19
Heatmap	19
HER2 Scoring Model	19
Dataset Description	19
Training Details	23
Heatmap	23
Web Application	23
Description	24
Architecture	24
Web Application Captures	25
6. Findings	34
Selected Performance Metrics	34
Histological Subtype Classification Model	34
Mitosis Identification Model	37
HER2 Scoring Model	38
7. Discussion	39
Histological Subtype Classification Model	39
Mitosis Identification Model	40
HER2 Scoring Model	41
8. Conclusion	41
9. References	41

1. Summary

Breast cancer is the leading cause of cancer deaths among women worldwide, but early detection and diagnosis can significantly improve a patient's prognosis. Conventionally, biopsy breast tissue is graded based on morphological and genetic features; lesion subtype, mitotic count, and HER2 (IHC) score are three of the most significant prognostic markers. However, the manual examination of breast tissue slides is time-consuming and subject to inter- and intra-pathologist variability. Also, many developing nations lack the skilled medical professionals necessary to provide thorough and timely diagnoses. Artificial intelligence innovations in healthcare can address these issues and the repercussions of late or incorrect diagnosis. In this research, I developed BrCaVision, a state-of-the-art automated system for determining malignancies, classifying histological subtypes, detecting mitosis, and scoring HER2 in a whole slide image (WSI). Using transfer learning methods, I constructed sophisticated convolutional neural network models that accomplished 85% accuracy for subtype determination and a sensitivity of 0.98 for mitosis detection in a fraction of the time required for gold-standard methodologies. Also, I developed a web application that allows medical professionals to input a WSI and obtain a HER2 score and heatmaps that mark tissue subtypes and mitotic cell locations. By deploying the application on a public web hosting service, professionals can acquire timely and accurate breast cancer diagnosis and prognosis prediction across the world. Being easily accessible and fully automated, BrCaVision serves as a step forward in the field of bioinformatics and can save lives with reliable, accurate, and efficient breast cancer detection.

2. Introduction

In the United States, about one in every eight women (13%) will develop invasive breast cancer during her lifetime. Furthermore, approximately 270,000 American women will be diagnosed with invasive breast cancer, and about 42,000 American women will die from breast cancer in 2020 alone [1]. While procedures such as breast exams and mammograms can identify abnormalities within breast tissue, the only definitive method of diagnosing breast cancer and determining a patient's prognosis is through a biopsy and evaluation by a pathologist.

Breast cancer is distinguished from benign lesions by pathologists, whose evaluations are the gold standard for determining patient prognosis. However, individual pathologist opinions often disagree with diagnoses made by a consensus of experts [2] [3]. In one study, when 6900 individual diagnoses made by 115 pathologists were compared to consensus-based diagnoses, 17% of benign cases with atypical hyperplasia (atypia) and 3% of benign cases without atypia were found to be misdiagnosed as invasive ductal carcinoma or ductal carcinoma in situ [3]. Also, 10% of malignant carcinoma or ductal carcinoma in situ cases were misdiagnosed as benign lesions [3]. The latter misdiagnoses are especially concerning, as the erroneous classification of a malignant tumor as a benign lesion can delay critical patient treatment.

Pathologists are also responsible for determining what histological subtypes of benign tissue or malignant tumor are present within a biopsied breast tissue sample, as each subtype of breast cancer responds more quickly or more effectively to different treatments. For instance, invasive lobular carcinoma is more responsive to systemic therapy than the more common invasive ductal carcinoma [4]. Also, recognizing benign lesion subtype is crucial for determining a patient's risk of developing subsequent breast cancer [5]. However, there are often discrepancies in pathologist opinion for the classification of breast tissue subtypes. In one study, where six pathologists were asked to classify benign lesions into three categories, the overall Cohen's kappa coefficient for the categorization was 0.48, which suggests only a moderate amount of agreement between the pathologists [6]. In Longacre et al. 2006 [7], inter-rater agreement was calculated for the classification of subtypes of invasive breast carcinoma; among 13 pathologists, the highest agreement rates achieved were for mucinous (96.0%), lobular (78.7%), and tubular (78.0%) subtypes.

With the emergence of digital pathology and whole slide imaging (WSI) techniques, researchers have developed computer-aided detection (CAD) tools to classify WSI of breast tissue as benign or malignant. However, despite the importance of determining a breast lesion's histological subtype(s), only a few previous studies have constructed systems for the automatic classification of lesions into subtypes. In a notable study, six subtypes of breast cancer were divided into two subgroups (cancer type I and cancer type II). The algorithm constructed throughout the study obtained a binary classification accuracy of 89% (while classifying between benign and cancerous tissue) and a multiclass accuracy of 80% (while differentiating between benign tissue, cancer type I, and cancer type II) [8].

Similar to lesion subtype, mitotic count (or the number of dividing cells visible in hematoxylin and eosin (H&E) stained histopathology slides) is a critical predictor of tumor aggressiveness and an indicator of breast cancer prognosis [9]. However, the manual detection of mitotic cells often has poor inter-pathologist agreement due to the variable texture and morphology of cells undergoing mitosis [10]; this process is also tedious and time-consuming. An automated process for detecting mitotic cells will lead to increased accuracy and consistency between pathologist counts while simultaneously reducing the time and cost of reaching a breast cancer diagnosis [10]. A fair amount of research has already been conducted on methods to automatically classify regions of breast tissues WSIs (or individual cells/complexes themselves) as mitotic [10] [11] [12] [13]. Notably, [10] presents a novel approach to mitotic nuclei detection that outperforms similar state-of-the-art systems. A similar study's deep detection mitosis classifier achieved a significantly higher F-score than competing systems using the same test dataset [13].

Receptor tyrosine-protein kinase erbB-2 (human epidermal growth factor receptor 2) is the protein encoded by the erb-b2 receptor tyrosine kinase 2 (ERBB2) gene [34]; the protein and the gene are commonly referred to as HER2/neu and HER2, respectively [35]. In healthy cells, the HER2/neu protein maintains normal cell growth [36]. However, the HER2 gene is a proto-oncogene. When HER2 is amplified within breast tissue, breast cells will produce an excessive number of HER2/neu receptors. This protein overexpression allows malignant breast cells to proliferate and spread uncontrollably; as a result, HER2-positive breast cancers tend to be more aggressive and carry a significantly higher risk of recurrence [35].

It's crucial that pathologists test every case of primary invasive breast cancer to determine the patient's HER2 status. A breast cancer patient who is HER2-positive exhibits HER2 gene amplification and an abnormally high number of HER2/neu protein receptors within her malignant breast tissue; a patient who is HER2-negative has breast cancer that doesn't express this amplification and overexpression. Modern HER2-targeted therapies that utilize tyrosine kinase inhibitors or targeted monoclonal antibodies can inhibit the HER2 gene and significantly improve a HER2-positive breast cancer patient's prognosis [35]. However, these chemotherapy supplements often produce side effects of heart damage (sometimes leading to congestive heart failure, among other serious side effects), liver problems, and lung disease [37]. Also, these therapies pose monetary drawbacks as well. Trastuzumab (Herceptin) is the most widely utilized targeted therapy for HER2-positive breast cancer patients. For the treatment of non-metastatic breast cancer, adjuvant trastuzumab is often prescribed over 52 weeks with a drug cost of approximately \$100,000 [38]. Furthermore, oncologists are required to administer 9 to 12 months of intravenous trastuzumab following the completion of adjuvant systemic chemotherapy, which further exacerbates the cost of treatment [38]. As a result, HER2-targeted therapies are currently only recommended for patients who exhibit HER2 amplification due to their cardiotoxicity and the relatively few benefits they provide for HER2-negative patients [37]. Hence, it's critical to eliminate as much uncertainty as possible from the HER2 status testing process to minimize the prevalence of false-positive results (which may lead to a patient being prescribed an unnecessary and harmful chemotherapy regimen).

Pathologists test for a breast cancer patient's HER2 status by conducting immunohistochemistry (IHC) staining or fluorescence in situ hybridization (FISH) testing (or both). Approximately 80% - 90% of all primary HER2 status testing conducted in the United States utilizes IHC assays, while only 10% - 20% of testing uses FISH; this may be because IHC is faster, less expensive, and less specialized when compared to FISH [39]. In an IHC assay, a cross-section of a patient's malignant breast tissue and a control section containing HER2/neu protein overexpression are fixed, embedded in paraffin, and stained with IHC. Then, a pathologist visually compares the staining pattern within both samples using light-microscopy and estimates HER2/neu expression levels within the patient's tissue [51]. An IHC result of 0 or 1+ corresponds to HER2-negative breast cancer, while one of 3+ denotes HER2-positive cancer;

an IHC result of 2+, which occurs in approximately 10% of cases, typically represents an 'equivocal' result, and the patient's HER2 status must be determined using a FISH assay as well [40]. On the other hand, FISH testing allows pathologists to quantitatively assess the amount of HER2 gene amplification present within a patient's malignant breast cells [42].

However, both IHC and FISH assays carry significant disadvantages. While IHC is simpler to implement, variance in tissue handling and fixation methods and test sensitivity can impact the accuracy of the tests [45]. While FISH testing is inherently less susceptible to these concerns, medical professionals must undergo specialized training to conduct FISH's more complicated procedures [45]. Also, in 2010, the benchmark cost for a FISH test was \$339, almost three times larger than the cost of an IHC test (\$118) [45]. Healthcare providers can easily circumvent these inevitable obstructions to accessible and affordable HER2 status testing by adopting a computer-aided system for scoring, which shows promise of increased accuracy and consistency while simultaneously reducing the time and cost of determining a patient's HER2 status.

The development of automated image analysis systems for HER2 scoring has been an emerging topic of interest for several years. However, a lack of publicly available and curated image datasets of prepared breast tissue slides coupled with ground truth HER2 scores has hindered research until the University of Warwick's 2016 HER2 Scoring Contest Dataset was made publically available to researchers. While only two studies into this novel approach to HER2 scoring have been published so far, current research manifests the potential and limitless boundaries of this field. The first landmark study analyzed specialized features within the dataset's IHC-stained whole slide images (WSIs), such as the connectedness of stained cell membrane regions and texture intensity descriptors, to achieve 93.86% accuracy for HER2 scoring on individual tissue image patches [44]. However, the novel system developed over the course of that research still requires an immunohistochemistry-stained slide to conduct image analysis. In order to completely eliminate reliance on IHC-based methods, another study conducted image analysis using slides stained with hematoxylin and eosin (H & E), a standard benchmark pathology stain that is commonly utilized by pathologists to distinguish between malignant and benign tissues during initial analysis. H & E is more inexpensive and accessible than IHC, so constructing an automated tool to distinguish morphological correlates of HER2-positive and HER2-negative tissue within H & E stained tissue would require less cost and time. However, this study achieved only 83.3% classification accuracy on the University of Warwick's testing dataset and 53.8% on a similarly constructed private dataset [43]. There is a critical need for a unique, reliable, and more accurate system for automated HER2 scoring through the analysis of WSIs of H & E-stained biopsied breast tissue.

This paper presents a unique approach to subtype classification, mitosis identification, and HER2 scoring within WSIs of breast tissue, named *Breast Cancer Vision* (BrCaVision).

Whole slide tissue images are often over 100 megabytes in size, making them unwieldy and, as a result, relatively unused in previous biomedical research. Almost every paper referenced in the introduction made use of pre-tiled and curated datasets of histopathological images. My research revolves around the construction of three convolutional neural network (CNN) models that provide highly accurate and sensitive subtype, mitosis, and HER2 classifications for individual tiles within breast tissue WSIs in a fraction of the time required for a pathologist to conduct a manual assessment of the images. I also developed a simple-to-use user interface (UI) that allows medical professionals to input a WSI and obtain heatmaps that demarcate tissue subtypes, denote mitotic cell locations, and score for HER2 within each region in the WSI.

3. Research Questions

- Can deep learning methods be leveraged to construct an automated classifier for histological subtypes of breast tissue, mitotic cells, and HER2 score that can determine subtypes, mitosis, and HER2 score within a fraction of the time needed for conventional manual analysis?
- Can this classification software be made accessible to remote areas where world-class pathologists, radiologists, and medical providers may not be available to analyze digital medical images and report a diagnosis?

4. Definition of Terms

Various pipelines for the architectures of the subtypes, mitosis, and HER2 scoring models were developed and evaluated before a base network and classifier were selected for each of the three classification tasks. A conventional machine learning-based Support Vector Machine (SVM), a custom convolutional neural network (CNN) model, and two pre-trained transfer learning-based CNNs (with one using ResNet-50 and the other using VGG16 as a base model) were the prototypes considered.

Support Vector Machines

SVMs are machine learning algorithms that are utilized to determine the optimal hyperplane between separable classes of image data for classification or regression purposes [48]. These versatile kernel machines operate in an implicit feature space due to their usage of kernel functions; as SVMs don't have to compute the coordinates of the data in the feature space, they are less computationally expensive than similar methods for automated classification. However, when leveraged on a relatively small dataset or in a setup where features outnumber samples, SVMs tend to overfit on the provided dataset and, as a result, generalize poorly; in these scenarios, choosing suitable kernel functions and regularization terms is crucial [49].

Deep Learning and Convolutional Neural Networks

Deep learning is a subset of Artificial Intelligence (AI) that consists of algorithms that extract progressively higher-level features (or attributes) from input data over multiple layers of

analysis [14]. Deep learning models can classify image data after being trained to distinguish which features correspond with which classes of images [50]. In particular, convolutional neural networks (CNNs) are regularized deep neural networks composed of hidden neuron layers that extract features from images.

Transfer Learning and Fine-Tuning

Transfer learning is a procedure that involves leveraging features learned for one problem on a different (but often similar) project [15]. In the context of deep learning, transfer learning often consists of the following steps: 1) obtaining the layers from a previously trained model, 2) removing the fully-connected output layers and replacing them with a set of fully-connected layers with randomly initialized hyperparameters, 3) freezing the hidden layers in the architecture to preserve previously determined neuron weights, and 4) training the network with a dataset relevant to the problem being addressed. A final step known as fine-tuning occurs after the frozen model has been ‘warmed up’ or trained for a small number of epochs on the new dataset. In fine-tuning, the entire model is unfrozen to potentially achieve improvements in model performance by allowing for the model’s pretrained features to adjust to the new data.

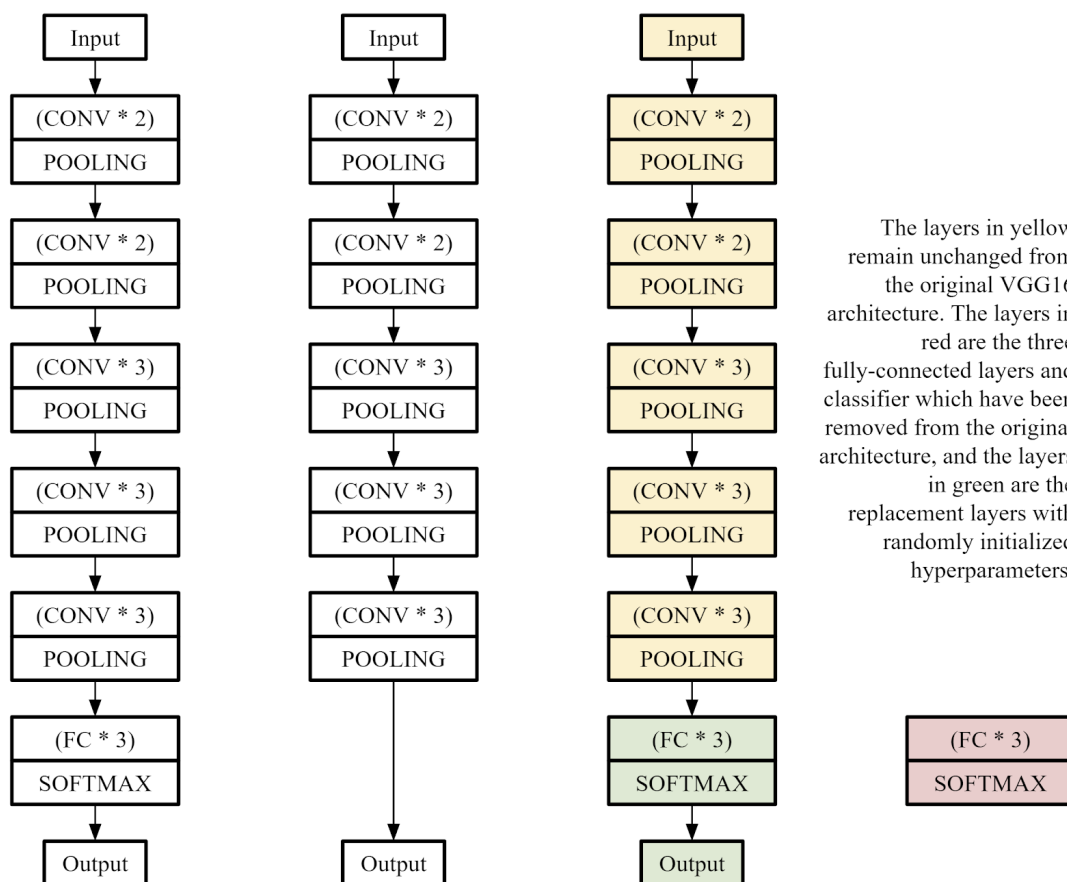


Figure 4-1: Comparison of the original VGG16 network architecture (left) with layers removed to form a feature extractor (middle) and new layers added as part of fine-tuning.
Credit: Dheepthi Mohanraj

The prototype networks were initialized with random hyperparameters and trained on a preliminary dataset; checkpoints were used to save each model and its weights upon the completion of each epoch. The best-performing model was tested using a small number of images from the BreakHis dataset, and the model's performance was compared to those of the other models. At this step, a transfer learning approach using VGG16 as a base network was determined to be the most effective prototype. Transfer learning and fine-tuning were both utilized to construct the histological subtype, mitosis identification model, and HER2 scoring models.

Base Model Characteristics

The base model for the final fine-tuned models was the VGG16 convolutional neural network. Characteristics of VGG16 include that it was trained with the training dataset for the 2014 ImageNet Large-Scale Visual Recognition Challenge (ILSVRC-2014) competition, which consists of over 1.3 million images from ImageNet (ImageNet is a database of over 14 million images representing 1000 classes) [16] [17], and that it achieved only 6.8% top-5 error with classification tasks using the competition's testing dataset of 100,000 images [16]. ResNet-50, a CNN which employs residual learning and obtained significantly lower top-1 and top-5 error than VGG16 when trained and tested on the same respective datasets [18], was explored as an alternative to VGG16. However, when fine-tuned and employed to construct the tissue subtype model, the ResNet-50-based model obtained lower accuracy and took significantly longer to train when compared with the VGG16-based model. With the technology stack and infrastructure available, VGG16 was the most viable option for a base model.

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 150, 150, 3)	0
block1_conv1 (Conv2D)	(None, 150, 150, 64)	1792
block1_conv2 (Conv2D)	(None, 150, 150, 64)	36928
block1_pool (MaxPooling2D)	(None, 75, 75, 64)	0
block2_conv1 (Conv2D)	(None, 75, 75, 128)	73856

block2_conv2 (Conv2D)	(None, 75, 75, 128)	147584
block2_pool (MaxPooling2D)	(None, 37, 37, 128)	0
block3_conv1 (Conv2D)	(None, 37, 37, 256)	295168
block3_conv2 (Conv2D)	(None, 37, 37, 256)	590080
block3_conv3 (Conv2D)	(None, 37, 37, 256)	590080
block3_pool (MaxPooling2D)	(None, 18, 18, 256)	0
block4_conv1 (Conv2D)	(None, 18, 18, 512)	1180160
block4_conv2 (Conv2D)	(None, 18, 18, 512)	2359808
block4_conv3 (Conv2D)	(None, 18, 18, 512)	2359808
block4_pool (MaxPooling2D)	(None, 9, 9, 512)	0
block5_conv1 (Conv2D)	(None, 9, 9, 512)	2359808
block5_conv2 (Conv2D)	(None, 9, 9, 512)	2359808
block5_conv3 (Conv2D)	(None, 9, 9, 512)	2359808
block5_pool (MaxPooling2D)	(None, 4, 4, 512)	0
flatten_1 (Flatten)	(None, 8192)	0
dense_1 (Dense)	(None, 256)	2097408
dropout_1 (Dropout)	(None, 256)	0
dense_2 (Dense)	(None, 8)	2056
=====		
Total params: 16,814,152		
Trainable params: 9,178,888		
Non-trainable params: 7,635,264		

Figure 4-2: The architecture of fine-tuned VGG16, a convolutional neural network, with the number of total and trainable parameters present during subtype classification model training.

Credit: Dheepthi Mohanraj

VGG16's architecture consists of a stack of convolutional layers followed by three fully-connected (FC) layers; the first two FC layers have 4096 channels each, while the latter contains 1000 channels (with 1 channel for each of the 1000 classes in the ILSVRC-2014 database) and employs the softmax classifier [16]. The input to VGG16 consists of 224×224 RGB images; each image is passed through convolutional layers that use a 3×3 filter with a stride of 1 pixel. Max-pooling is performed at regular intervals (*see* Figure 4-2) over a 2×2 window with a stride of 2 pixels. Also, every hidden layer includes the Rectified Linear Unit (ReLU) activation function [16].

Whole Slide Images

Whole slide images (WSIs) are images constructed from extremely high-resolution scans of conventional glass slides. They are developed using hardware and software which combine multiple scanned slide tiles or strips into a single seamless image [25]. A raw and uncompressed WSI can require several gigabytes of memory, making the images difficult to view, store, and manipulate [25]. As a result, WSIs are rarely utilized or explored in computational settings [53]; this paper presents a novel method of handling WSI input for a CNN model so that an entire slide can be evaluated for tissue subtype classification, mitosis identification, and HER2 scoring, rather than just individual tiles from the slide.

Hematoxylin & Eosin Stain (H&E)

The hematoxylin & eosin (H&E) stain is one of the primary stains utilized in histology; it is also one of the most widely used tissue stains for medical diagnosis. Hematoxylin stains the nuclei of cells blue [21], while an eosin counterstain colors cytoplasm, connective tissue fibers, and the extracellular matrix in various shades of red and pink [22]. The contrast in colors allows pathologists to observe the layout and distribution of the cells on a stained tissue slide. The tissue collected during a breast biopsy is stained, processed, and then whole-mounted on slides to be observed or scanned at different magnifications. Pathologists may consult digitized WSIs of the tissue for further analysis.

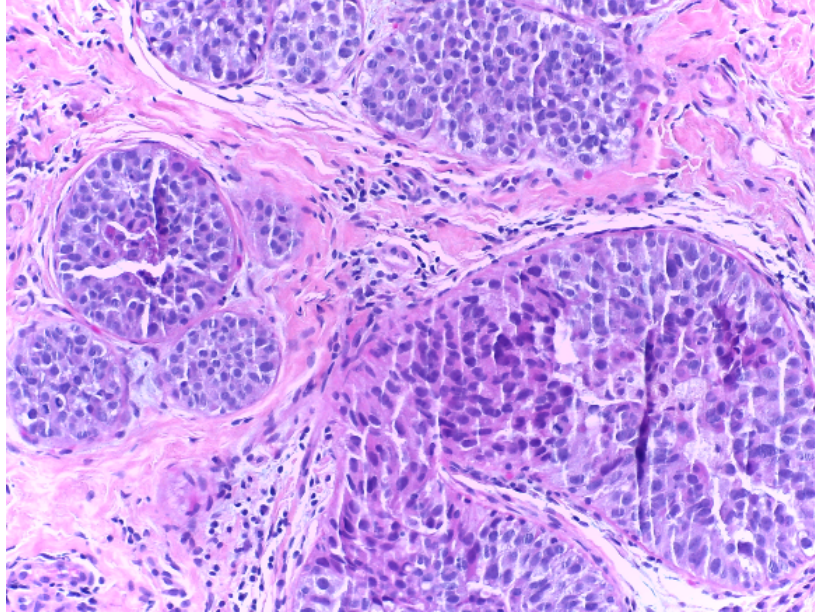


Figure 4-3: An example of a H&E stained slide of in situ carcinoma of the breast.
Credit: <https://icjar2018-challenge.grand-challenge.org/Dataset/>

5. Framework

Technology Stack and Infrastructure

An Anaconda conda virtual environment was created to maintain a workspace setup and install the necessary Python packages and libraries. NVIDIA's CUDA Toolkit and cudNN were used to conduct high-performance GPU-accelerated deep learning neural network training and optimization. These applications reduced the amount of time needed for GPU performance tuning.

The Python programming language was used to produce the histological subtype classification, mitosis identification, and HER2 scoring models. The high-level Keras API was run on top of the end-to-end Tensorflow machine learning platform to build and train the models. The Scipy and Numpy libraries for Python enabled numerical computing capabilities, and the Matplotlib and Seaborn plotting libraries were used to create graphic visualizations during exploratory data analysis (EDA) and model evaluation (AUC - ROC curve and loss/accuracy vs. epoch plots). The pandas software library was utilized for data manipulation and analysis (also, the pandas DataFrame structure was used during results storage). The scikit-learn library was leveraged to calculate statistics (including Cohen's kappa coefficient, Precision, Recall, F1-Score, and Support), and the scikit-image library was used to analyze the morphological features of the WSI image patches. The Pillow library was utilized to serialize and save each model and its weights.

The web application was constructed with numerous web development technologies (including Flask, SQLAlchemy, WTForms Javascript, CSS, and HTML) and was hosted on Amazon Web Services.

Data Preprocessing and Data Augmentation

Data preprocessing and data augmentation were performed on each of the three datasets used to train, validate, and test the histological subtype classification, mitosis identification, and HER2 scoring models.

Exploratory data analysis was performed on all three datasets using Matplotlib visualization algorithms to explore class characteristics and identify class imbalance or other patterns that may have impacted training. The data was analyzed for inconsistencies in color (which can often lead to issues during automated analysis) and preprocessed [24]. Then, the datasets for histological subtype classification and mitosis identification were split using an 80/10/10 ratio for training, validation, and testing (respectively), while the dataset used for HER2 scoring was split using a ratio of 60/20/20. Each subset of data was randomly sampled and representative of the proportions of classes in the original dataset. Finally, the data was separated into individual folders for each class, which was required to utilize the Keras ImageDataGenerator batch feed algorithm.

Data augmentation is a regularization technique that involves applying several random transformations to training images in the original training datasets for each model; this prevented model overfitting and helped the model generalize better [23]. Data augmentation can also improve model performance without the need for a larger dataset. The ImageDataGenerator class in Keras was utilized to create additional training data by applying random translations, rotations, zoom, shearing, horizontal flips, and vertical flips to original training images.

Histological Subtype Classification Model

Dataset Description

The tissue subtype model was trained using the publicly available Breast Cancer Histopathological Image Classification (BreaKHis) dataset [19] [20] [57]. The BreaKHis dataset consists of 7,909 microscopic hematoxylin-eosin (HE) stained histopathology images of breast tissue at magnifications of 40X, 100X, 200X, and 400X. The dataset consists of 2,480 benign sample images (with subtype classifications of adenosis, fibroadenoma, phyllodes tumor, and tubular adenoma) and 5,429 malignant sample images (with subtype classifications of ductal carcinoma, lobular carcinoma, mucinous carcinoma, and papillary carcinoma). The samples were collected from 82 patients through lumpectomy (a procedure in which a lump in the breast is

surgically removed and tested for cancerous cells), and the images have dimensions of 700 x 460 pixels, three channels (RGB) with 8-bit depth in each channel, and PNG format [19].

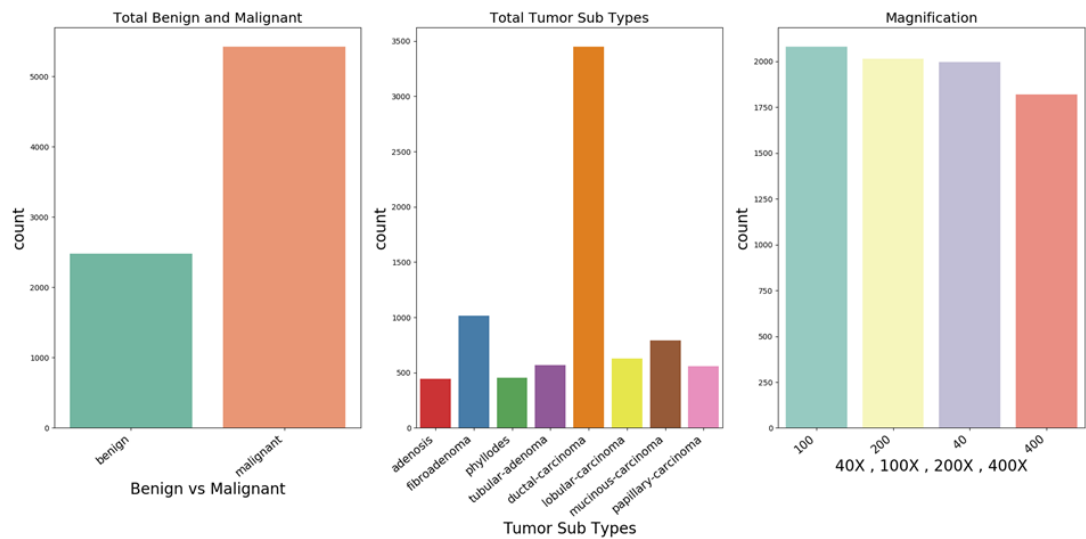


Figure 5-1: Exploratory data analysis on the BreaKHis dataset.
Credit: Dheepthi Mohanraj

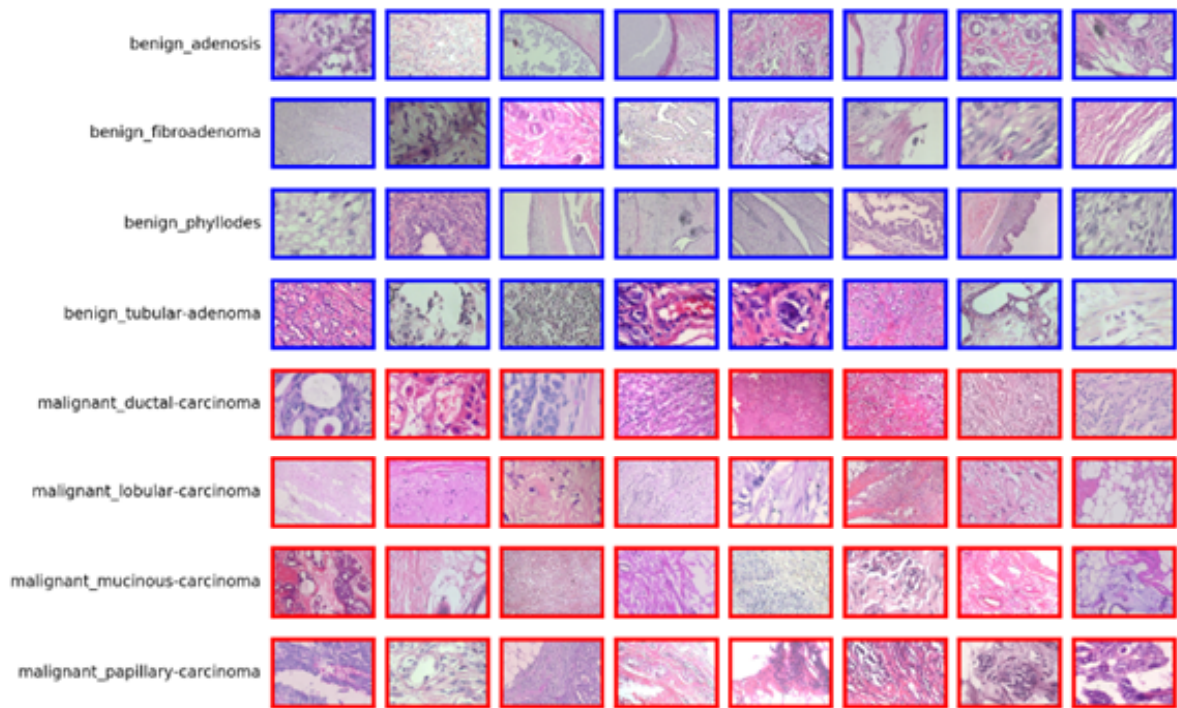


Figure 5-2: Benign and malignant WSI image patches from the BreaKHis dataset.
Credit: Dheepthi Mohanraj

Training Details

The VGG16 convolutional neural network (CNN) accepts image input. Dimensionally, the ‘width’ and ‘height’ of the input correspond to the dimensions of the input image, while the ‘depth’ represents the three red, green, and blue (RGB) color channels. CNN architectures (including VGG16) are typically composed of combinations of three main types of layers: convolutional layers, pooling layers, and fully-connected layers.

A convolutional layer utilizes filters that perform convolution operations while scanning the input, developing a feature map that is taken as input by subsequent layers. For VGG16, each 3×3 filter is convolved across the width and height of the input volume; over time, the CNN will learn filters that are ‘activated’ (or respond) when they encounter particular morphological features [51]. Each neuron in a convolutional layer is only connected to a small section of the input volume.

After convolution, an activation function is used to introduce non-linearities to the network [52]. Every hidden layer was equipped with the Rectified Linear Unit (ReLU) activation function. In earlier neural network research, a neuron’s output f as a function of received input x was often modeled with $f(x) = \tanh(x)$ (the hyperbolic tangent activation function) or with $f(x) = \frac{1}{1 + e^{-x}}$ (the sigmoid activation function), both of which have ranges of $(-1, 1)$ and are saturating nonlinearities. However, as $x \rightarrow \infty$ or $x \rightarrow -\infty$, the derivatives of both the hyperbolic tangent and sigmoid activation functions approach zero. This property gives rise to the *vanishing gradient problem*, which prevents the weights present in relatively deep neural networks from changing (therefore stunting training progress and extending the time needed for training). On the other hand, networks with ReLU (which are neurons with the non-saturating $f(x) = \max(0, x)$ nonlinearity), overcome this vanishing gradient problem; as a result, they tend to train faster than equivalent networks with sigmoid or hyperbolic tangent-activated neurons [54].

After an activation function is applied, pooling layers perform downsampling to account for the spatial invariance introduced by convolutional layers; by reducing the number of parameters being considered in the CNN, pooling layers prevent overfitting [51]. In VGG16, max-pooling is performed at regular intervals (see Figure 4-2) over a 2×2 window with a stride of 2 pixels.

The final fully-connected (FC) layers compute class scores; the neurons in each of these layers are fully connected to all of the neurons in the preceding layer. Neurons with an activation function known as the softmax classifier determine class probabilities for an input image (the class with the largest probability of corresponding to the input is often taken to be the network’s ‘prediction’). During the fine-tuning process, the set of three fully-connected layers and softmax

classifier were removed from the top of the VGG16 architecture and replaced with a new set of three fully-connected layers (with randomly initialized hyperparameters) and the classifier.

Mini-batch gradient descent was used to train the model. The loss function was categorical_crossentropy, one of the most common loss functions used in multi-class classification problems. Adaptive Moment Estimation (Adam) was chosen as the optimizer over similar algorithms such as RMSprop and Adadelta since Adam implements bias-correction and momentum and often outperforms other optimizers towards the end of optimization (as gradients become sparser) [16].

During warm-up training, the model's hidden layers were frozen to prevent neuron weights (learned from training with the ImageNet database) from being altered. Data augmentation was not employed during this initial phase of training. The model was warmed-up using the BreakHis training dataset after its images were rescaled and resized to 150×150 . The model was pre-trained over 50 epochs using the resized images with a batch size of 20.

While conducting fine-tuning, the layers in the conv5 block (*see* Figure 4-2) were unfrozen and set to be trainable. Keras ImageDataGenerator was utilized to perform data augmentation (*see* **Data Preprocessing and Data Augmentation**), and the training images were rescaled and resized into 150×150 images again. Class weights were incorporated into the loss function to address the class imbalance. The Keras EarlyStopping callback was added to halt training when there is little to no change in validation loss for more than 15 epochs to conserve time and resources when the model showed little to no progress. Also, the ReduceLROnPlateau callback adjusts the model's learning rate hyperparameter from $1e-4$ to $1e-5$ if the validation loss started to plateau, and the ModelCheckpoint callback was utilized to only save the best model during training. Finally, the model trained for 100 epochs with batch size 20 and an initial learning rate of $1e-4$. Training history was visualized by plotting the training accuracy, training loss, validation accuracy, and validation loss for each epoch.

Finally, the fine-tuned model was evaluated with the BreakHis testing dataset. The performance metrics, confusion matrix, and classification report for testing were calculated, and the metrics were visualized by plotting the AUC-ROC curve.

After multiple trials to determine which batch size, learning rate, and loss function would increase model performance, the best performing model was saved and used to create a histological subtype heatmap for an entire WSI.

Heatmap

A heatmap of histological subtypes can be constructed for WSIs with any dimensions and file size; however, the program will take longer to process images that require a substantial

amount of memory. The heatmap algorithm takes an input of a single WSI of H&E stained breast tissue, and the Openslide Python library was utilized to handle the WSI .svs file. The numbers of rows and columns of image patches are calculated based on the image's dimensions, and the coordinates of the top-left corner of each patch are determined. Based on those coordinates, the image tiles are read from the whole WSI. If an image patch contains at least 10% tissue, the image is considered for subtype classification. The model assigns a class prediction and corresponding label color to each image. Once all the valid tile image data were read, classified, and assigned a label color, a heatmap was created for the original WSI. This entire process takes approximately 10 minutes to complete for a 100 - 200 megabyte WSI and a single GPU system configured with tensor-gpu.

Mitosis Identification Model

Dataset Description

The mitosis identification model was trained using the publicly available MITOS-ATYPIA-14 Challenge Dataset curated by the Pathology Department at Pitié-Salpêtrière Hospital in Paris, France [26]. The MITOS-ATYPIA-14 dataset contains 1,136 microscopic hematoxylin-eosin (HE) stained histopathology images of breast tissue at X40 magnification. The dataset was initially divided into two folders based on the scanner (Aperio Scanscope XT or Hamamatsu Nanozoomer 2.0-HT) that developed the digitized slide images. The Aperio images have dimensions of 1539×1376 pixels, while the Hamamatsu images have dimensions of 1663×1485 pixels; the images have three channels (RGB) and TIFF format [26]. Each image was annotated by two pathologists and was classified into 'mitosis' and 'not mitosis' classes. The dataset has a CSV file of ground truth data that contains x- and y-coordinates of the approximate center of each set of mitotic cells in each image.

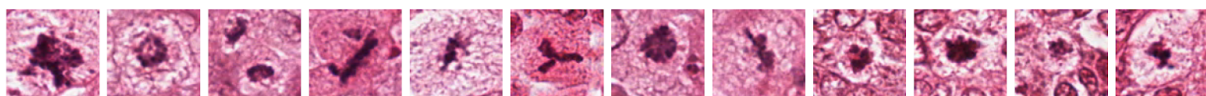


Figure 5-3: Mitosis image patches created from the original MITOS-ATYPIA-14 Challenge dataset.

Credit: Dheepthi Mohanraj

Using the MITOS-ATYPIA-14 Challenge Dataset, a mitosis patch dataset for training the mitosis identification model was developed. The coordinates of mitosis location from the ground truth CSV file were utilized to generate 'mitosis' class image patches of size 64×64 , while 'not mitosis' class images from the original dataset were tiled to create 64×64 'not mitosis' class image patches. A final dataset consisting of 53,928 mitosis and 53,928 non-mitosis image patches was split in an 80/20 ratio for training and validation.

Training Details

Before fine-tuning, the three fully-connected layers and the classifier were removed from the top of the VGG16 architecture. The remaining layers were flattened, and a dense layer with single-class classification was added to the top of the network; this layer was initialized with the Glorot normal initializer. The convolutional layers were frozen during warm-up training to prevent changes to the model's initial weights. The model was compiled with an Adam optimizer with a learning rate of $1e-3$ and L2 weight regularization of $1e-3$. Finally, the model was pre-trained with a mitosis dataset of 64×64 image patches with batch size 20 for 50 epochs.

During fine-tuning, the layers in the conv5 block (*see* Figure 4-2) were unfrozen and set to be trainable. With Adam, fine-tuning was conducted with a batch size of 256, a learning rate of $1e-4$, momentum of 0.9, and L2 weight regularization of $1e-3$. Finally, the model was trained with batch size 256 for 50 epochs. Training history was visualized by plotting the training accuracy, training loss, validation accuracy, and validation loss for each epoch; other statistics (namely precision, recall, and F1 score) were also recorded and collected for the model at each epoch. The same checkpoints utilized for the training of the subtype classification model were used to train the mitosis identification model. The best performing model was saved and used to identify mitosis in an image patch and create a mitosis heatmap for an entire WSI.

Heatmap

Similar to subtype classification, a mitosis heatmap can be constructed for WSIs with any dimensions and file size. Using the OpenSlide Deep Zoom generator, an input WSI is tiled into patches of 64×64 , which were processed in batches. Each generated tile was analyzed based on morphological features and tissue percentage. First, dense tissue was separated from the background, and small dark spots or general noise in the image background were removed using binary closing. Bright areas were enlarged and dark areas were shrunk using binary dilation, remedying holes in tissue regions that may have interfered with my mitosis identification model's capabilities. Then, each tile is validated against a threshold tissue percentage; tissues within the threshold are considered for classification. Tiles predicted to contain mitosis were visualized with a mark to specify mitosis location within the WSI. Once all the valid tile image data were read, classified, and assigned a label color, a heatmap was created for the original WSI. This process takes 10 minutes to complete for a 200 megabyte WSI and a single GPU system configured with tensor-gpu.

HER2 Scoring Model

Dataset Description

The HER2 scoring model was trained using the 2016 Her2 Scoring Contest Dataset, which was curated by the University of Warwick and The Pathological Society of Great Britain

& Ireland [46] [47] [55]. The Her2 Scoring Contest Dataset's training set consists of 52 pairs of whole slide images (WSIs), with each pair including one hematoxylin and eosin (H&E) stained tissue WSI and one immunohistochemistry (IHC) stained tissue WSI from a single sample (a single case). The testing subset consists of 28 pairs of WSIs in a similar format. Each set has equally distributed cases for all 4 possible stages of HER2 scoring (0/1+/2+/3+); the manually-determined HER2 score for each image is recorded in a ground truth .csv file for the training set. As the Her2 Scoring Contest Dataset was curated for a contest, the ground truth data for the testing set has been redacted and removed from the dataset; as a result, the WSIs from the testing set could not be utilized to train a CNN. Each case in the training set was evaluated by at least two specialized pathologists at Nottingham University Hospitals, an institution that regularly participates in the United Kingdom National External Quality Assessment Scheme (NEQAS) for Immunocytochemistry and In-situ Hybridization, before being issued a HER2 score [46]. The digitized slide images were developed with a Hamamatsu NanoZoomer C9600. As a result, this dataset's WSIs were stored in .ndpi format (while the heatmap algorithms developed for the histological subtype classification and mitosis identification models accept WSIs in .svs format).

Extensive preprocessing methods were used to prepare the dataset before training the model. First, each of the 52 H&E-stained WSIs from the Her2 Scoring Contest Dataset's training set was manually evaluated for major blemishes, stains, marks, or color inconsistencies that may have interfered with automated image preprocessing methods and model training/testing. 7 WSIs were discarded during this step, leaving 45 WSIs for consideration. The remaining WSIs were slightly downsampled and converted to .png format to facilitate image analysis procedures.

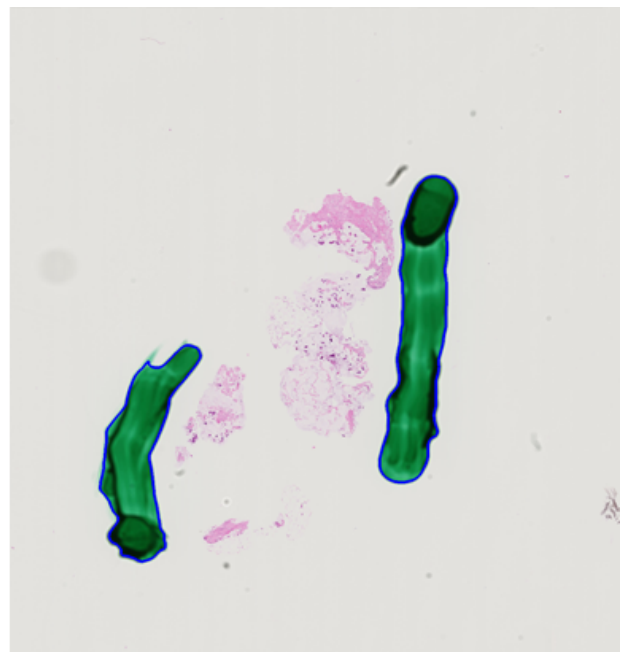
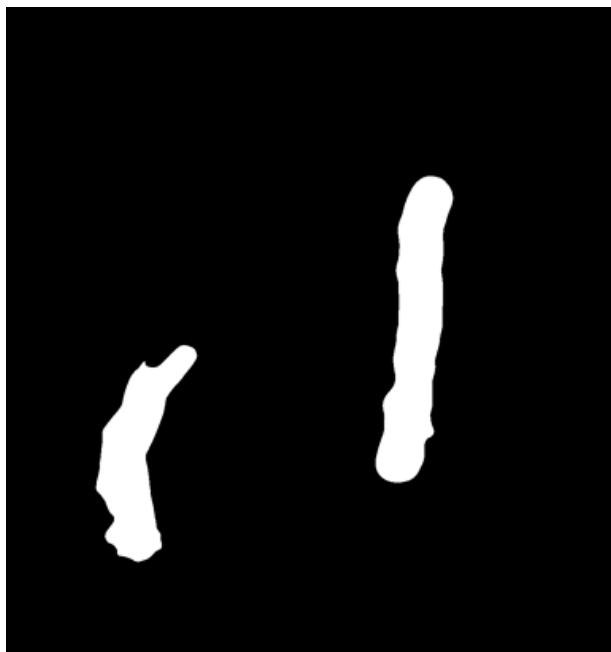


Figure 5-4: Otsu's thresholding and contouring on a downsampled WSI containing visible marker stains. This slide was discarded from consideration because the automated image preprocessing procedures mistook markings on the slide to be tissue-containing regions.

Credit: Dheepthi Mohanraj

Subsequently, an image processing technique known as *Otsu's thresholding* was leveraged to extract breast tissue material from the slide background. Otsu's thresholding algorithms automatically select an optimal threshold on the saturation (S) channel of a color image's Hue-Saturation-Value (HSV) colorspace to distinguish between two 'classes' of pixels in the image (belonging to the foreground or the background) [41]. The results of the thresholding were visualized with a binary mask of the downsampled WSIs; foreground tissue regions were labeled in white, while the slide background was designated by black. A contouring algorithm was also employed on the downsampled images to approximate the regions of interest denoted by Otsu's thresholding. Finally, each of the original WSIs was tiled into 150×150 pixel image patches; the final training, validation, and testing datasets for training the HER2 scoring model were constructed with patches that consisted of at least 50% foreground (tissue). The comprehensive dataset of image patches was split using a ratio of 60/20/20 for training, validation, and testing (respectively); each subset of data was randomly sampled and representative of the proportions of classes in the original dataset.

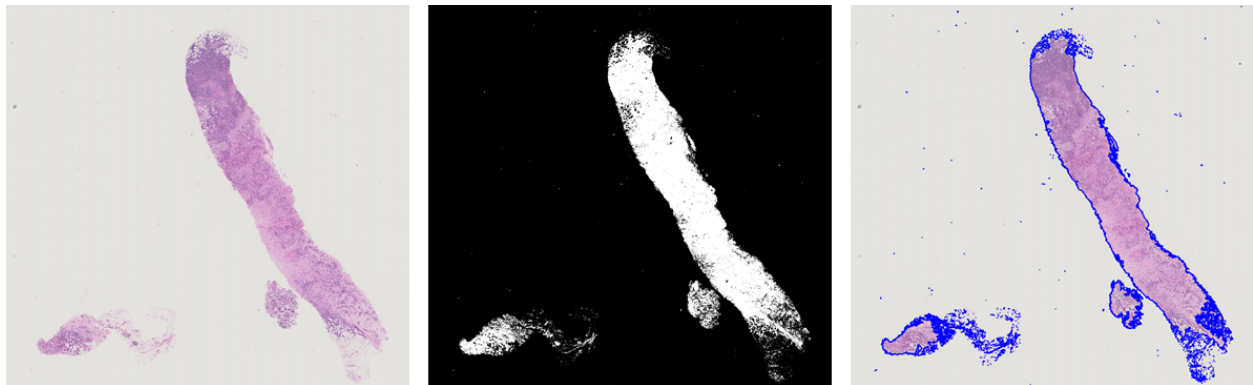


Figure 5-5: (left) A downsampled WSI, (center) its Otsu's thresholding visualization, and (right) the results of contouring.

Credit: Dheepthi Mohanraj

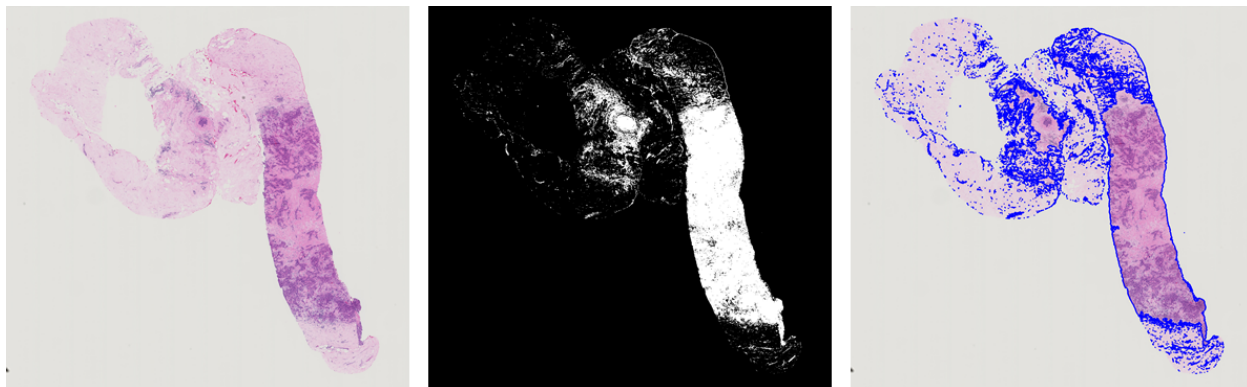


Figure 5-6: (*left*) Another downsampled WSI, (*center*) its Otsu's thresholding visualization, and (*right*) the results of contouring. Note that thinner and lighter tissue regions were excluded during the thresholding process.
Credit: Dheepthi Mohanraj

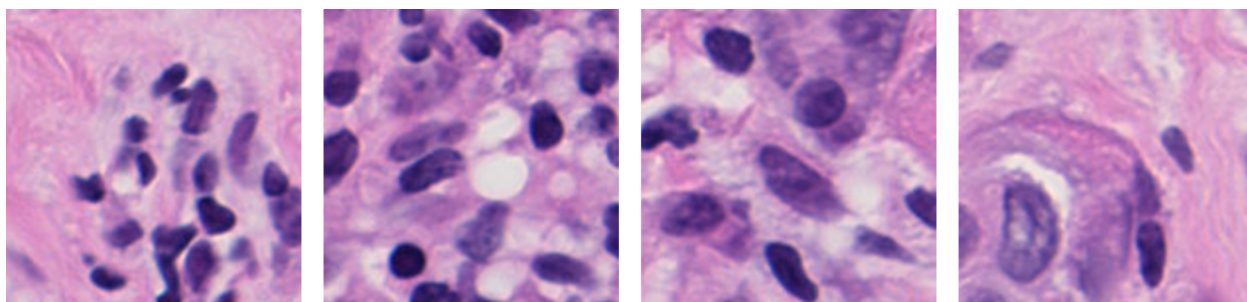


Figure 5-7: Image patches with HER2 scores of 0, 1+, 2+, and 3+, from left to right.
Credit: Dheepthi Mohanraj

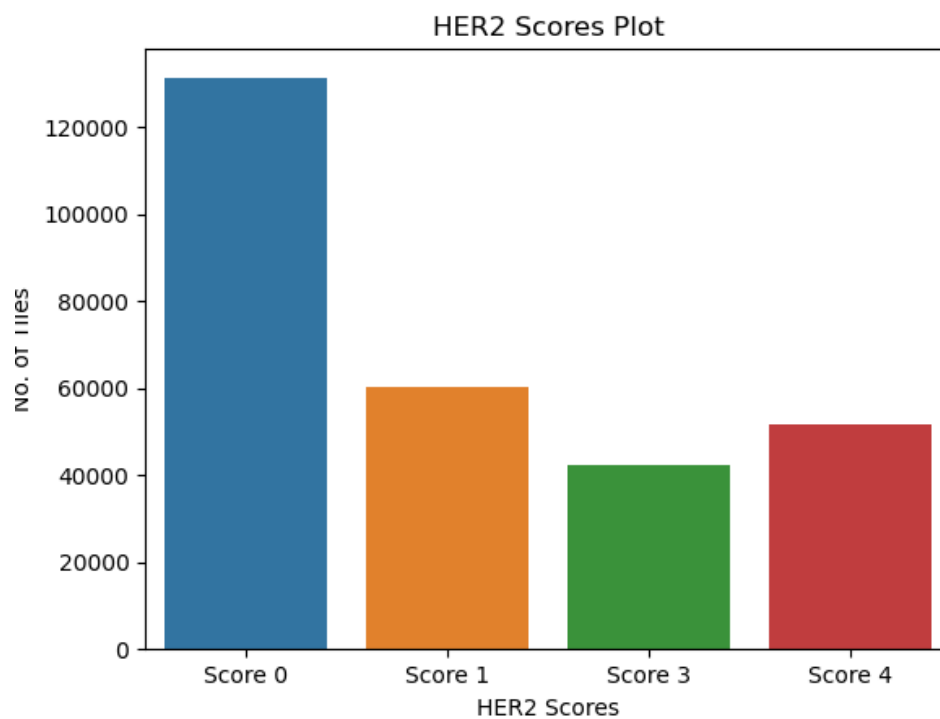


Figure 5-8: Exploratory data analysis on the final image patch dataset constructed with Her2 Scoring Contest data.
Credit: Dheepthi Mohanraj

Training Details

Before fine-tuning, the set of three fully-connected layers and softmax classifier were removed from the top of the VGG16 architecture and replaced with a new set of three fully-connected layers (with randomly initialized hyperparameters) and the softmax classifier. The network's hidden layers were frozen during warm-up training to prevent changes to the model's initial weights. The model was compiled with an Adam optimizer, the categorical_crossentropy loss function, and a learning rate of 1e-4. Finally, the model was pre-trained with a HER2 scoring dataset of 150 x 150 image patches with batch size 32 for 50 epochs.

During fine-tuning, the layers in the conv5 block (see Figure 4-2) were unfrozen and set to be trainable. With Adam, fine-tuning was conducted with a batch size of 32 and a learning rate of 1e-3. Finally, the model was trained with batch size 32 for 150 epochs. Training history was visualized by plotting the training accuracy, training loss, validation accuracy, and validation loss for each epoch; other statistics (namely precision, recall, and F1 score) were also recorded and collected for the model at the end of training. The same checkpoints utilized for the training of the histological subtype classification model were used to train the HER2 scoring model. The best performing model was saved and used to score HER2 in an image patch and create a HER2 scoring heatmap for an entire WSI.

Heatmap

A heatmap of HER2 scores can be constructed for WSIs with any dimensions and file size; however, the program will take longer to process images that require a substantial amount of memory. The heatmap algorithm takes an input of a single WSI of H&E stained breast tissue, and the Openslide Python library was utilized to handle the WSI .svs file. The numbers of rows and columns of image patches are calculated based on the image's dimensions, and the coordinates of the top-left corner of each patch are determined. Based on those coordinates, the image tiles are read from the whole WSI. Image patches that are chosen through the Otsu's thresholding and preprocessing methods are considered for HER2 scoring. The model assigns a class prediction and corresponding label color to each image. Once all the valid tile image data were read, classified, and assigned a label color, a heatmap was created for the original WSI. This entire process takes approximately 15 minutes to complete for a 100 - 200 megabyte WSI and a single GPU system configured with tensor-gpu.

Web Application

Description

The BrCaVision web application allows physicians to access, upload, and analyze breast tissue tiles and whole slide images. Physicians can run the models on whole slide images to produce heatmaps of the classification results. Through the web application, physicians can also run the deep learning models to generate classification results for region of interest tissue image patches. Physicians can then log in and access the specific records they uploaded previously along with any generated outputs. The app is hosted on Amazon Web Services, which provides powerful web servers to run Python web applications.

Architecture

The BrCaVision web application is built using Python with many different tools, including Flask, SQLAlchemy, and WTForms. This application is based on the MVC (Model-View-Controller) design pattern. The models allow for accessing and changing the database, the controllers allow for choosing and manipulating specific data, and the views allow for displaying this data in a specific way.

Flask is the primary web framework used in the application; it powers the routing and session management aspects. When a web server runs the Flask application, the server performs a Python function whenever a request is made. These functions return a specific HTML response. For example, when a physician chooses to upload a tile, a Flask view function returns the HTML web page that displays the upload form (which is what a user sees when they access the website through their browser).

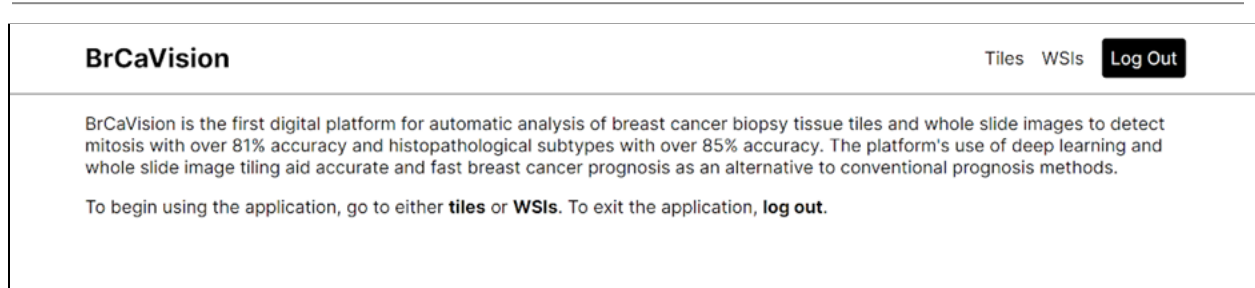
SQLAlchemy is the database toolkit used to store, read, and update specific database values. This application uses the object-relational mapper (ORM) part of SQLAlchemy, which is an abstraction layer that allows rows in database tables to be represented as Python objects. The database tables themselves are stored in a file using SQLite, a popular SQL relational database management system (RDBMS) that allows data to be stored in a file on the server.

The records that are created by physicians are stored in the SQL database. When a physician uploads a specific image (such as an SVS whole slide image), the Flask view function takes the uploaded file data and stores it in the "uploads" folder on the server. If the uploaded file was a whole slide image, the image is converted into a more lightweight JPEG, which allows a version of the image to be displayed on the web. This conversion is managed through Python's Pillow image processing library.

The records database itself includes a path to the uploaded file, patient ID, and patient name. When a physician runs the predict functions on a whole slide image, a new image displays the results map. This image is also converted into a more lightweight JPEG, and the path to the generated file and generated JPEG are stored. When a physician runs the mitosis or

subtypes model on an individual tile, the result (which is generated as text) is stored in the database.

Web Application Captures



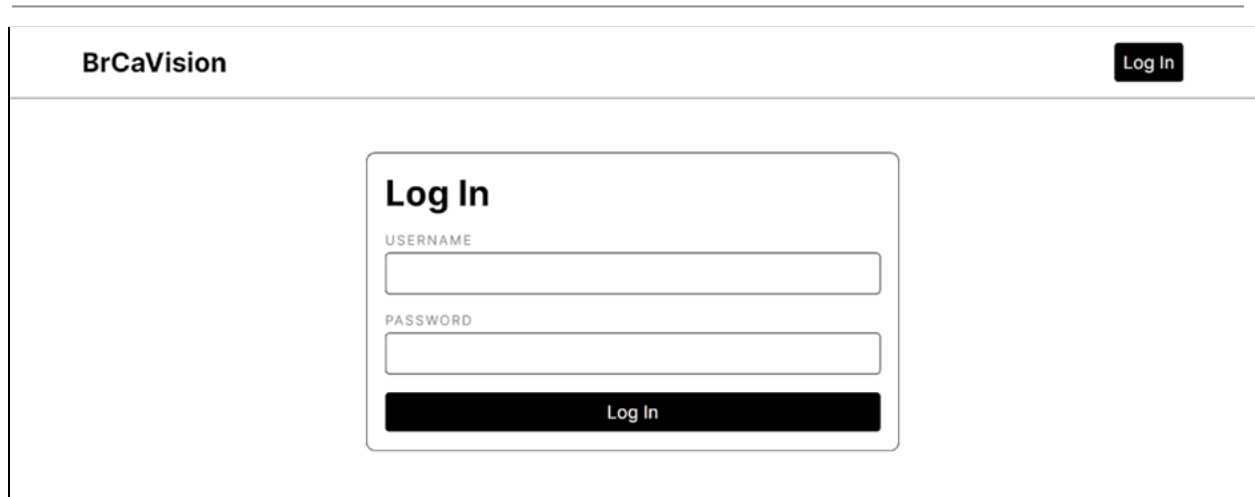
The screenshot shows the BrCaVision Home Page. At the top left is the logo "BrCaVision". At the top right are links for "Tiles", "WSIs", and a "Log Out" button. The main content area contains a paragraph describing the platform's capabilities and a instruction to begin using the application.

BrCaVision Tiles WSIs **Log Out**

BrCaVision is the first digital platform for automatic analysis of breast cancer biopsy tissue tiles and whole slide images to detect mitosis with over 81% accuracy and histopathological subtypes with over 85% accuracy. The platform's use of deep learning and whole slide image tiling aid accurate and fast breast cancer prognosis as an alternative to conventional prognosis methods.

To begin using the application, go to either **tiles** or **WSIs**. To exit the application, **log out**.

Figure 5-9: BrCaVision Home Page
Credit: Dheepthi Mohanraj



The screenshot shows the BrCaVision Login Page. At the top left is the logo "BrCaVision". At the top right is a "Log In" button. The main content area contains a "Log In" form with fields for "USERNAME" and "PASSWORD", and a "Log In" button.

BrCaVision **Log In**

Log In

USERNAME

PASSWORD

Log In

Figure 5-10: BrCaVision Login Page
Credit: Dheepthi Mohanraj

BrCaVision

TilesWSIsLog Out

Upload a Tile

PATIENT NAME

TYPE

Subtypes

TILE

Choose FileNo file chosen

Add Record

Figure 5-11: BrCaVision ‘Upload a Tile’ Page
Credit: Dheepthi Mohanraj

BrCaVision

TilesWSIsLog Out

Upload a WSI

PATIENT NAME

WHOLE SLIDE IMAGE

Choose FileNo file chosen

Add Record

Figure 5-12: BrCaVision ‘Upload a WSI’ Page
Credit: Dheepthi Mohanraj

BrCaVision

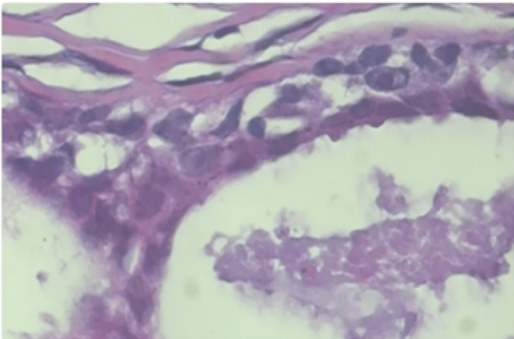
Tiles

WSIs

Log Out

Add a New Tile

Patient A



Prediction: According to the model, this tissue image displays benign tissue cells with approximate probability 100.0%. The predicted cancer subtype is Benign Adenosis.

Subtype	Probability
Benign Adenosis	100.0%
Benign Fibroadenoma	0.0%
Benign Phyllodes	0.0%
Benign Tubularadenoma	0.0%

Predict

Reset Results

Figure 5-13: BrCaVision ‘Predict a Tile’ Page. Here, the histological subtype classification model accurately classified a tile depicting benign adenosis.
Credit: Dheepthi Mohanraj

BrCaVision

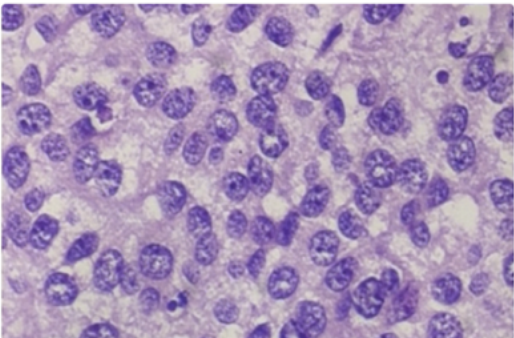
Tiles

WSIs

Log Out

Add a New Tile

Patient A



Prediction: According to the model, this tissue image displays malignant cancer with probability 100.0%. The predicted cancer subtype is Malignant Ductal Carcinoma.

Subtype	Probability
Malignant Ductal Carcinoma	99.87%
Malignant Lobular Carcinoma	0.0%
Malignant Mucinous Carcinoma	0.04%
Malignant Papillary Carcinoma	0.09%

Predict

Reset Results

Figure 5-14: BrCaVision ‘Predict a Tile’ Page. Here, the histological subtype classification model accurately classified a tile depicting malignant ductal carcinoma.
Credit: Dheepthi Mohanraj

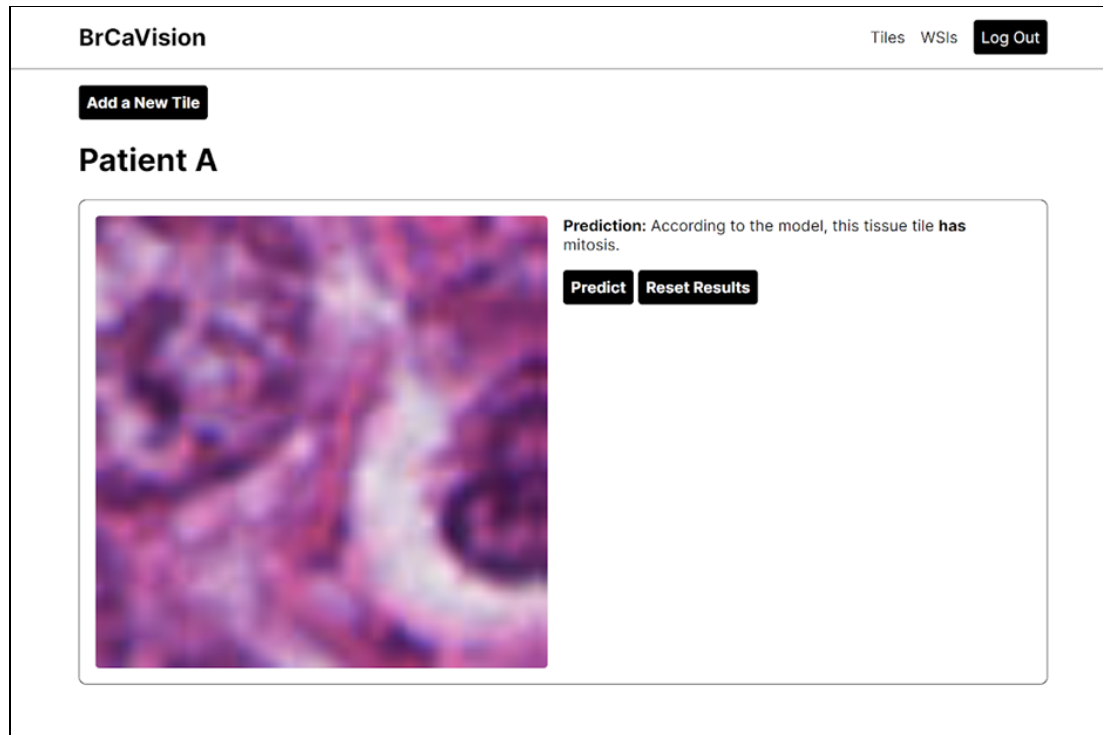


Figure 5-15: BrCaVision 'Predict a Tile' Page. Here, the mitosis identification model accurately classified a tile depicting mitosis.

Credit: Dheepthi Mohanraj

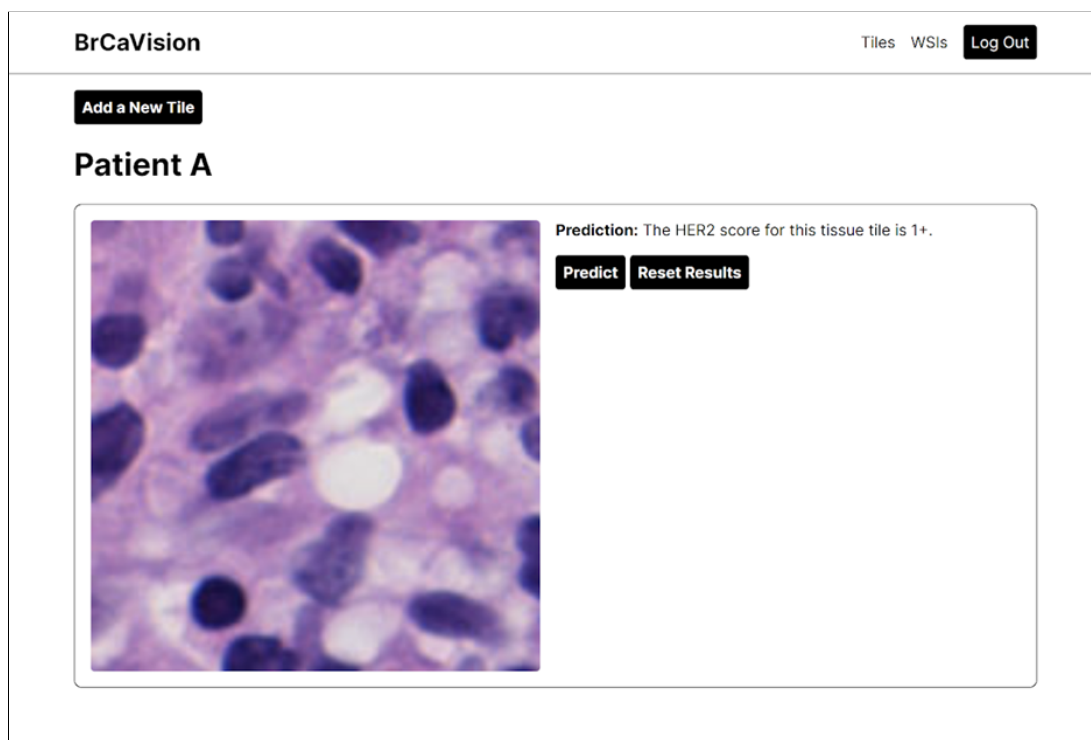


Figure 5-16: BrCaVision 'Predict a Tile' Page. Here, the HER2 scoring model accurately classified a tile depicting tissue with a HER2 score of 1+.

Credit: Dheepthi Mohanraj

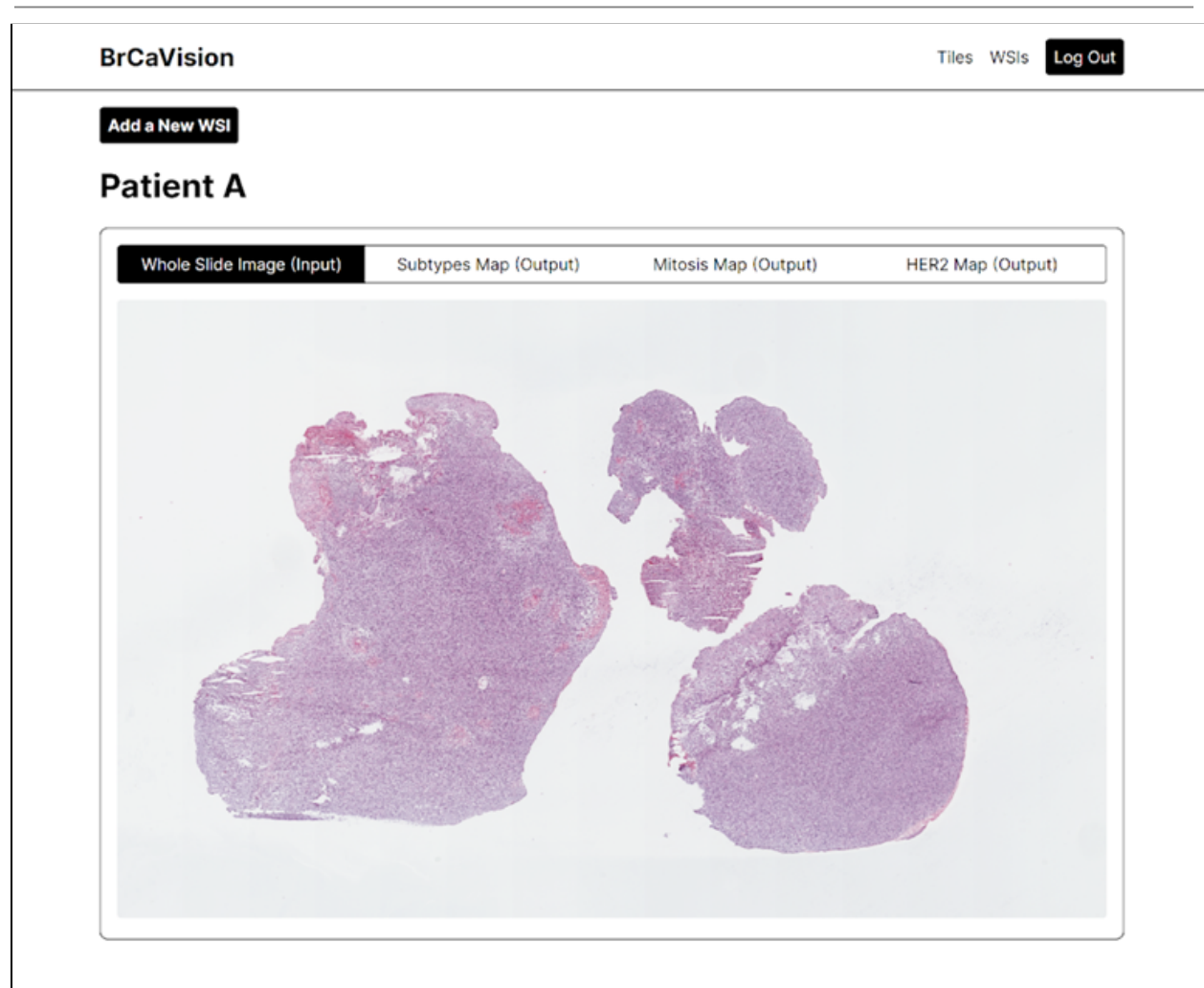


Figure 5-17: BrCaVision 'WSIs' Page. An input whole slide image has been uploaded to the application.
Credit: Dheepthi Mohanraj

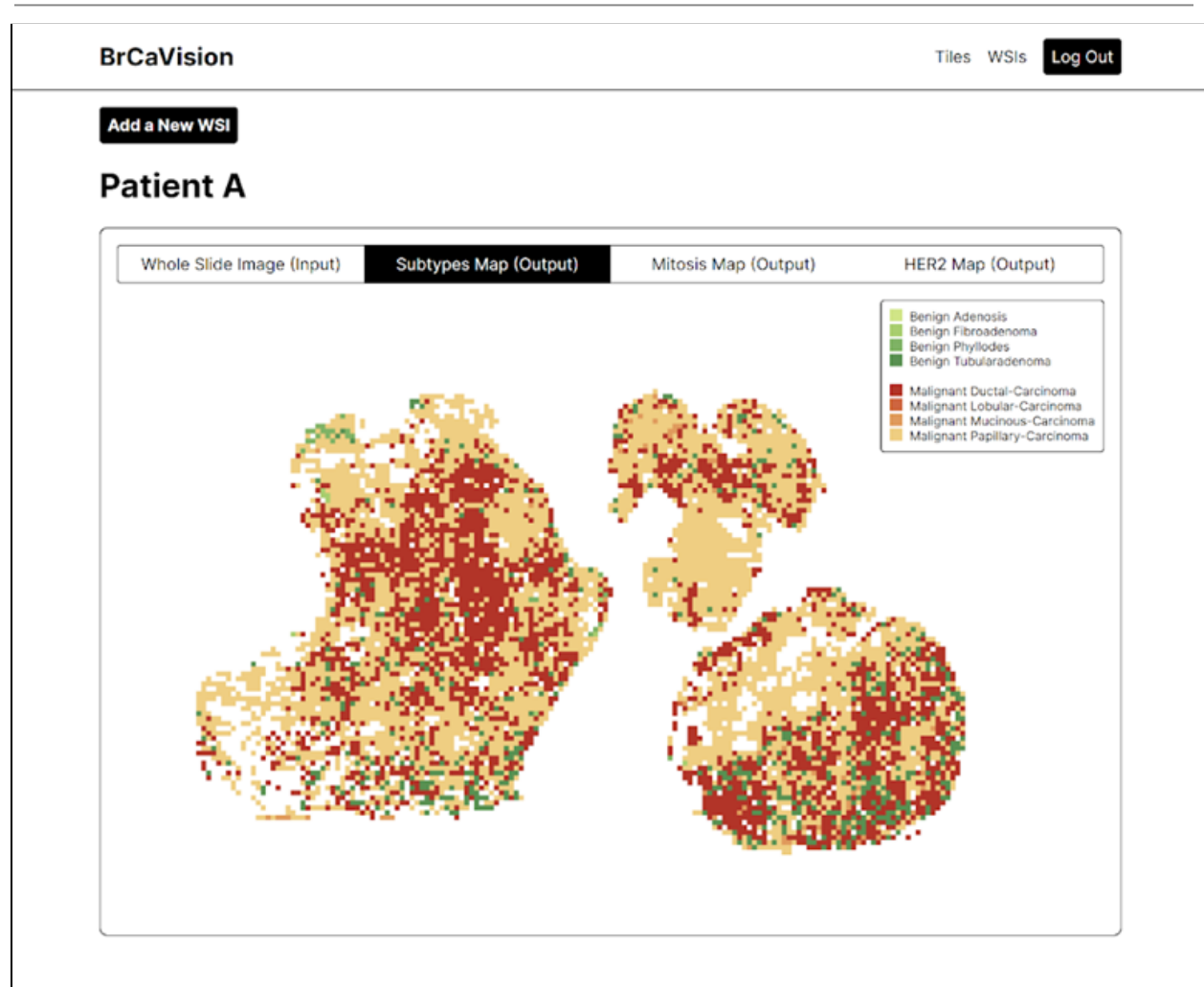


Figure 5-18: BrCaVision ‘WSIs’ Page, depicting a completed histological subtypes heatmap.
Credit: Dheepthi Mohanraj



Figure 5-19: BrCaVision 'WSIs' Page, depicting a completed mitosis identification heatmap.
Credit: Dheepthi Mohanraj



Figure 5-20: BrCaVision ‘WSIs’ Page, depicting a completed HER2 scoring heatmap.
Credit: Dheepthi Mohanraj

6. Findings

Selected Performance Metrics

The classification results of each of the three models were evaluated using certain common benchmark metrics, including:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

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

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

$$\text{F1-score/F-measure} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

TP = true positives

FP = false positives

TN = true negatives

FN = false negatives

$$\text{Cohen's kappa coefficient } \kappa = \frac{p_o - p_e}{1 - p_e} *$$

p_o = the observed agreement between two independent raters**

p_e = the theoretical probability of chance agreement between two independent raters**

* Typically only calculated for multi-class classification.

** In this scenario, the two independent raters are represented by a classification model's predictions and by the ground truth values of the data that was utilized to train and test the model.

Histological Subtype Classification Model

Warm-up training was conducted to extract features; during this phase, training loss was decreased from 2.1555 to 0.1838 over 50 epochs.

During the first epoch of fine-tuning training, the model had a training accuracy of 0.3718, training loss of 1.8914, validation accuracy of 0.4263, and validation loss of 1.6013.

The best performing model was saved at the 92nd epoch; this model obtained a training accuracy of 0.9340, training loss of 0.1875, validation accuracy of 0.8526, and validation loss of 0.5743.

During the 100th epoch of the fine-tuning training phase, the final model acquired a training accuracy of 0.9342, training loss of 0.1989, validation accuracy of 0.7579, and validation loss of 0.7647.

The 92nd epoch model achieved a testing accuracy of approximately 81.3% and a Cohen's kappa coefficient of 0.748, signifying substantial model effectiveness.

The weighted averages obtained for model precision, recall, F1-score, and support were 0.81, 0.81, 0.81, and 177, respectively; weighted averages were considered because they took the BreakHis dataset's class imbalance into account.

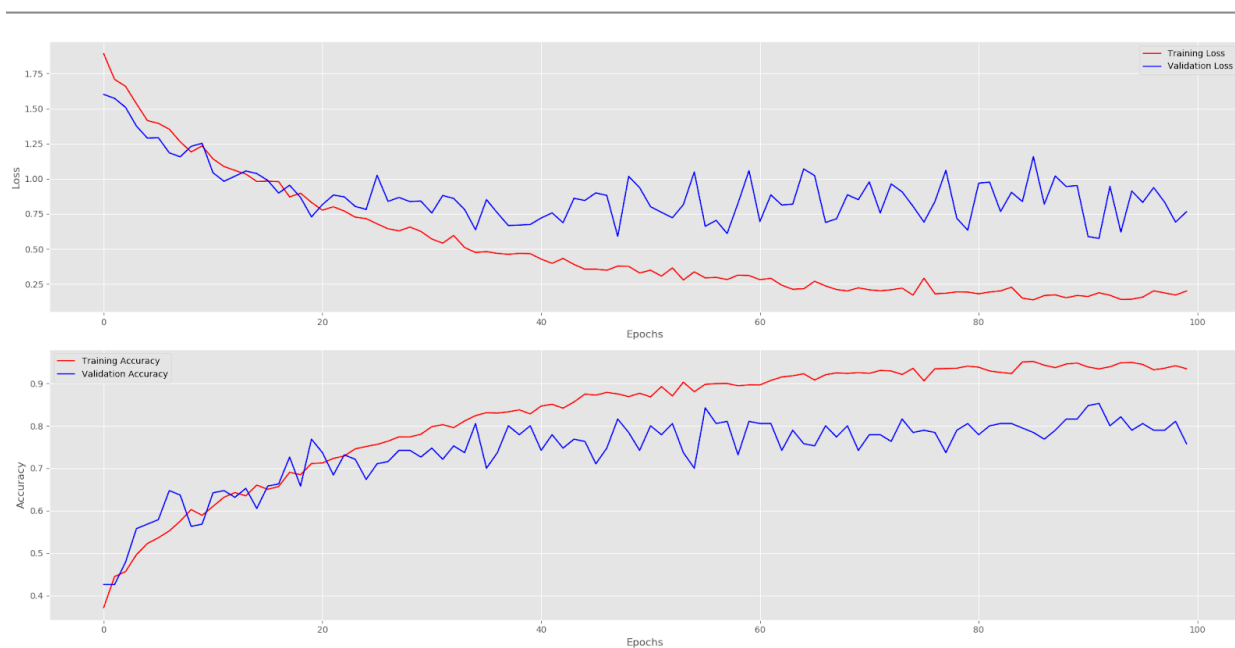


Figure 6-1: The training and validation accuracy vs. epoch plot (*top*) and the training and validation loss vs. epoch plot (*bottom*) during the fine-tuning training phase for the subtype classification model.

Credit: Dheepthi Mohanraj

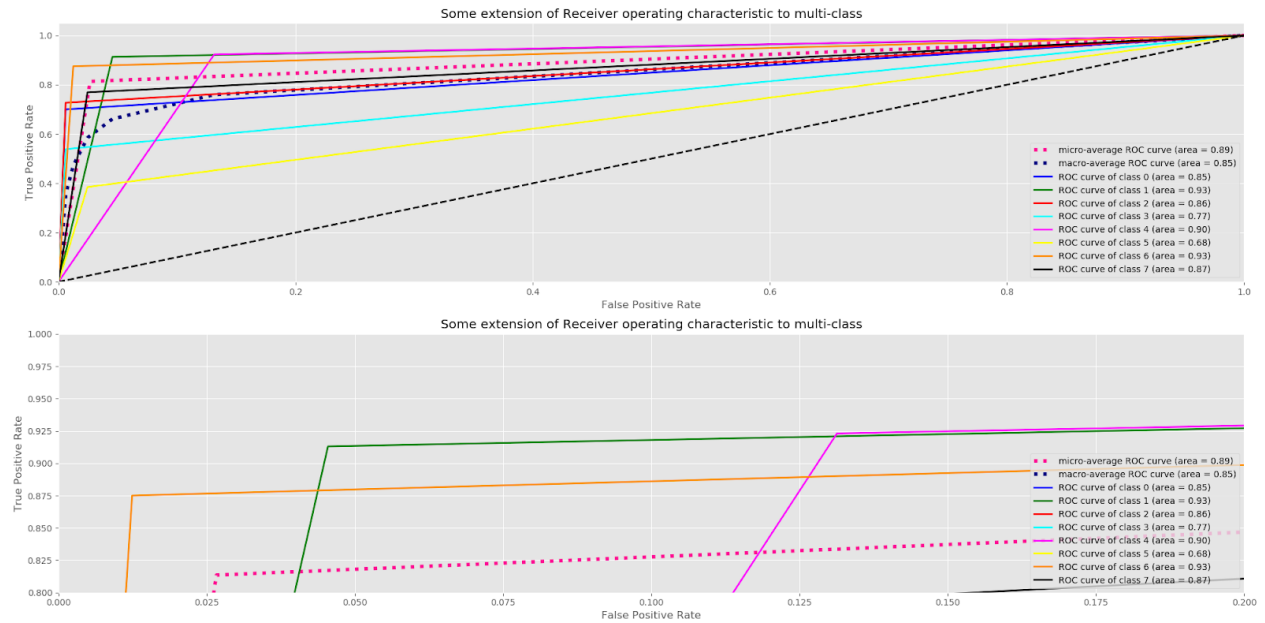


Figure 6-2: The AUC-ROC curve (*top*) and a zoomed subsection of the curve (*bottom*) for the best performing subtype classification model.
Credit: Dheepthi Mohanraj

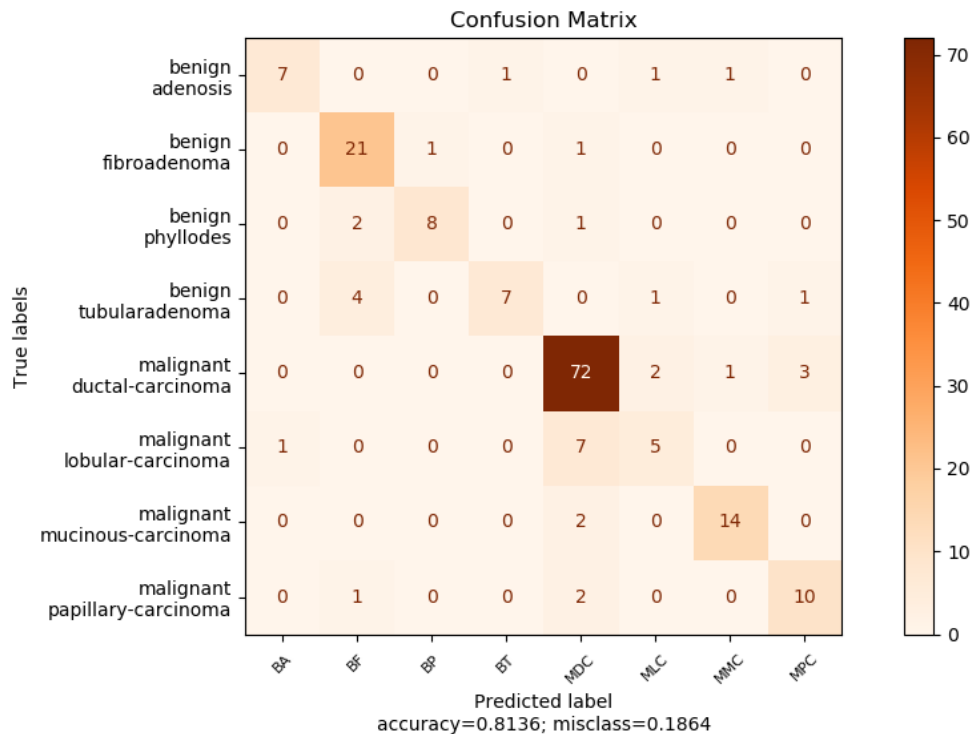


Figure 6-3: The confusion matrix obtained for the testing of the best performing subtype classification model.
Credit: Dheepthi Mohanraj

Subtypes	Precision	Recall	F1-Score	Support
benign_adenosis	0.88	0.70	0.78	10
benign_fibroadenoma	0.75	0.91	0.82	23
benign_phyllodes	0.89	0.73	0.80	11
benign_tubularadenoma	0.88	0.54	0.67	13
malignant_ductal-carcinoma	0.85	0.92	0.88	78
malignant_lobular-carcinoma	0.56	0.38	0.45	13
malignant_mucinous-carcinoma	0.88	0.88	0.88	16
malignant_papillary-carcinoma	0.71	0.77	0.74	13
Overall Statistics				
	Precision	Recall	F1-Score	Support
macro avg	0.80	0.73	0.75	177
weighted avg	0.81	0.81	0.81	177
Accuracy				
accuracy	0.81			

Figure 6-4: The model classification report obtained using scikit-learn methods.
Credit: Dheepthi Mohanraj

Mitosis Identification Model

Warm-up training was conducted to extract features; during this phase, training loss was decreased from 1.204 to 0.841 over 50 epochs.

During the first epoch of fine-tuning training, the model had a validation accuracy of 0.759, validation loss of 2.589, F1 score of 0.758, precision of 0.619, and a recall of 0.977.

The best performing model was saved at the end of fine tuning; this model acquired a validation accuracy of 0.8045, validation loss of 2.844, f1 score of 0.813, precision of 0.667, and a recall of 0.982.

HER2 Scoring Model

The best performing HER2 scoring model was obtained during the final epoch of the fine-tuning training phase. The model obtained a testing accuracy of approximately 85.7%, signifying substantial model effectiveness.

The weighted averages obtained for model precision, recall, F1-score, and support were 0.86, 0.86, 0.86, and 57185, respectively; weighted averages were considered because they took the HER2 image patch dataset's class imbalance into account.

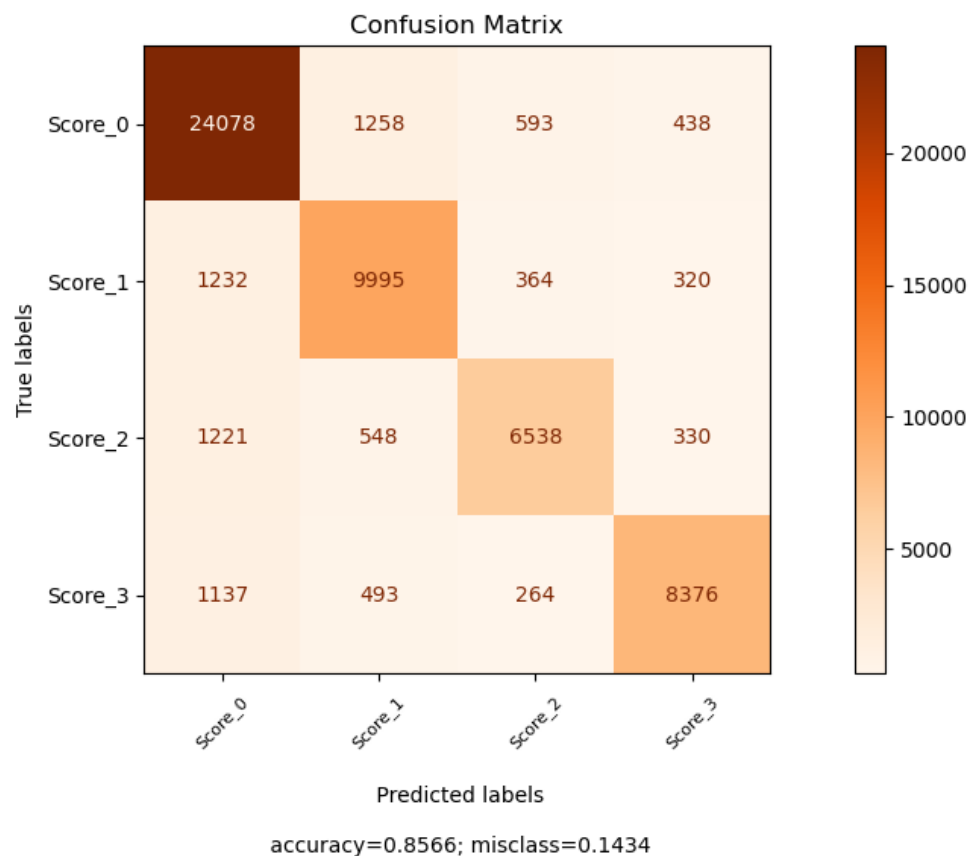


Figure 6-5: The confusion matrix obtained for the testing of the best performing HER2 scoring model.

Credit: Dheepthi Mohanraj

HER2 Score	Precision	Recall	F1-Score	Support
0	0.87	0.91	0.89	26367
1	0.81	0.84	0.83	11911

2	0.84	0.76	0.80	8637
3	0.89	0.82	0.85	10270
Overall Statistics				
	Precision	Recall	F1-Score	Support
macro avg	0.85	0.83	0.84	57185
weighted avg	0.86	0.86	0.86	57185
Accuracy				
accuracy	0.86			

Figure 6-6: The HER2 scoring model classification report obtained using scikit-learn methods.
Credit: Dheepthi Mohanraj

7. Discussion

Histological Subtype Classification Model

Since subtype classification was a multi-class task, BrCaVision's accuracy was calculated and compared against those of similar models from previous studies. The testing accuracy of my BrCaVision Histological Subtype Classification Model and similar models on the BreakHis dataset (for binary or multi-class tasks) are shown in Figure 7-1 below.

Methods	Accuracy
<i>LeNet_Raw [56]</i>	0.382
<i>AlexNet_Raw [56]</i>	0.673
<i>ORB [31][20]</i>	0.676
<i>CLBP [30][20]</i>	0.728
<i>LBP [32][20]</i>	0.731
<i>LPQ [33][20]</i>	0.737
<i>BrCaVision Histological Subtype Classification Model</i>	<i>0.813</i>

Figure 7-1: Comparing the accuracies of BrCaVision and similar models from previous studies.
Credit: Dheepthi Mohanraj

Of subtype classification models from recent studies, none obtained a higher accuracy than BrCaVision.

Mitosis Identification Model

Because the Mitosis Identification Model was used to solve a binary classification problem ('mitosis' versus 'not mitosis'), precision, recall, and F1-score were preferable statistics for evaluating BrCaVision's performance. The testing performance of my BrCaVision Mitosis Identification Model and other existing models on the MITOS-ATYPIA-14 dataset are shown in Figure 7-2 below.

Methods	Precision	Recall	F1-Score
<i>LRCNN + out group [29]</i>	0.400	0.450	0.427
<i>DeepMitosis [13]</i>	0.431	0.443	0.437
<i>CasNN [28]</i>	0.411	0.478	0.442
<i>LRCNN + in group [29]</i>	0.654	0.663	0.659
<i>SegMitos-r15R30 [27]</i>	0.495	0.785	0.667
<i>BrCaVision Mitosis Identification Model</i>	<i>0.667</i>	<i>0.982</i>	<i>0.813</i>

Figure 7-2: Comparing the precision, recall, and F1-score of BrCaVision and similar models from previous studies.
Credit: Dheepthi Mohanraj

BrCaVision obtained an F1-score of 0.813, an over 21% improvement from the SegMitos-r15R30 model with the next highest F1-score. My model also achieved a recall of 0.982, meaning that BrCaVision correctly identified 98.2% of the 'mitosis' class images from the MITOS-ATYPIA-14 dataset as containing mitosis. BrCaVision also achieved a recall of 0.982, meaning that the model correctly differentiated 98.2% of the 'mitosis' class images from the MITOS-ATYPIA-14 dataset; this recall was significantly higher than that achieved by any of the other architectures.

HER2 Scoring Model

Since HER2 scoring was a multi-class task, BrCaVision's accuracy was calculated and compared against those of similar models from previous studies. The testing accuracy of my BrCaVision HER2 scoring Classification Model and other existing models on the Her2 Scoring Contest dataset are shown in Figure 7-1 below.

Methods	Accuracy
<i>Oliveira et al. (2020) [43]</i>	0.833
<i>BrCaVision HER2 Scoring Model</i>	<i>0.857</i>

Figure 7-3: Comparing the accuracies of BrCaVision and similar models from previous studies.
Credit: Dheepthi Mohanraj

Of HER2 scoring models from recent studies, none obtained a higher accuracy than BrCaVision.

8. Conclusion

In this research, I have proposed BrCaVision, a state-of-the-art breast lesion subtype classification, mitosis detection, and HER2 scoring framework. Apart from effectively utilizing transfer learning on existing CNN architecture, my work proposes a novel method of handling high-resolution whole slide images through patch extraction and sliding window mechanisms to create a 'heatmap' of classification results for an entire WSI. My histological subtype and HER2 scoring models exhibited increased accuracy when compared to traditional machine learning models trained on the same respective dataset(s). My mitosis identification model was evaluated on sample images from various slide scanners and magnifications and similarly outperformed existing research. WSI analysis using my models is more precise, efficient, and timely than the conventional analysis of WSI by pathologists. Furthermore, my web application was designed so that my models' prediction capabilities can be seamlessly integrated with traditional clinical approaches, enhancing a medical provider's services. Hosting the application on a cloud server also helps address the global shortage of skilled medical professionals. In remote regions that do not possess the resources necessary to diagnose breast cancer, my system may be a life-saving system on its own (without pathologist aid). Being easily accessible and fully automated, BrCaVision serves as a large step forward in bioinformatics.

9. References

- [1]: American Cancer Society. (2021). How Common Is Breast Cancer? Retrieved April 8, 2021, from <https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html>
- [2]: Elmore, J. G., Longton, G. M., Carney, P. A., Geller, B. M., Onega, T., Tosteson, A. N. A., Nelson, H. D., Pepe, M. S., Allison, K. H., Schnitt, S. J., O'Malley, F. P., & Weaver, D. L. (2015). Diagnostic concordance among pathologists interpreting breast biopsy specimens. *JAMA - Journal of the American Medical Association*, 313(11), 1122–1132. <https://doi.org/10.1001/jama.2015.1405>
- [3]: Allison, K. H., Reisch, L. M., Carney, P. A., Weaver, D. L., Schnitt, S. J., O'Malley, F. P., Geller, B. M., & Elmore, J. G. (2014). Understanding diagnostic variability in breast pathology: lessons learned from an expert consensus review panel. *Histopathology*, 65(2), 240–251. <https://doi.org/10.1111/his.12387>
- [4]: Barroso-Sousa, R., & Metzger-Filho, O. (2016). Differences between invasive lobular and invasive ductal carcinoma of the breast: Results and therapeutic implications. In *Therapeutic Advances in Medical Oncology* (Vol. 8, Issue 4, pp. 261–266). SAGE Publications Inc. <https://doi.org/10.1177/1758834016644156>
- [5]: Guray, M., & Sahin, A. A. (2006). Benign Breast Diseases: Classification, Diagnosis, and Management. *The Oncologist*, 11(5), 435–449. <https://doi.org/10.1634/theoncologist.11-5-435>
- [6]: Cserni, G., Orosz, Z., Kulka, J., Sápi, Z., Kálmán, E., & Bori, R. (2006). Divergences in diagnosing nodular breast lesions of noncarcinomatous nature. *Pathology and Oncology Research*, 12(4), 216–221. <https://doi.org/10.1007/BF02893415>
- [7]: Longacre, T. A., Ennis, M., Quenneville, L. A., Bane, A. L., Bleiweiss, I. J., Carter, B. A., Catelano, E., Hendrickson, M. R., Hibshoosh, H., Layfield, L. J., Memeo, L., Wu, H., & O'Malley, F. P. (2006). Interobserver agreement and reproducibility in classification of invasive breast carcinoma: An NCI breast cancer family registry study. *Modern Pathology*, 19(2), 195–207. <https://doi.org/10.1038/modpathol.3800496>
- [8]: Yang, L., Chen, W., Meer, P., Salaru, G., Goodell, L. A., Berstis, V., & Foran, D. J. (2009). Virtual microscopy and grid-enabled decision support for large-scale analysis of imaged pathology specimens. *IEEE Transactions on Information Technology in Biomedicine*, 13(4), 636–644. <https://doi.org/10.1109/TITB.2009.2020159>
- [9]: Genestie, C., Zafrani, B., Asselain, B., Fourquet, A., Rozan, S., Validire, P., Vincent-Salomon, A., & Sastre-Garau, X. (1998). Comparison of the prognostic value of Scarff-Bloom-Richardson and Nottingham histological grades in a series of 825 cases of breast cancer: Major importance of the mitotic count as a component of both grading systems. *Anticancer Research*, 18(1 B), 571–576. <https://europepmc.org/article/med/9568179>
- [10]: Wang, H., Cruz-Roa, A., Basavanahally, A., Gilmore, H., Shih, N., Feldman, M., Tomaszewski, J., Gonzalez, F., & Madabhushi, A. (2014). Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features. *Journal of Medical Imaging*, 1(3), 034003. <https://doi.org/10.1117/1.JMI.1.3.034003>
- [11]: Cireşan, D. C., Giusti, A., Gambardella, L. M., & Schmidhuber, J. (2013). Mitosis detection in breast cancer histology images with deep neural networks. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8150 LNCS(PART 2), 411–418. https://doi.org/10.1007/978-3-642-40763-5_51
- [12]: Mahmood, T., Arsalan, M., Owais, M., Lee, M. B., & Park, K. R. (2020). Artificial Intelligence-Based Mitosis Detection in Breast Cancer Histopathology Images Using Faster R-CNN and Deep CNNs. *Journal of Clinical*

Medicine, 9(3), 749. <https://doi.org/10.3390/jcm9030749>

[13]: Li, C., Wang, X., Liu, W., & Latecki, L. J. (2018). DeepMitosis: Mitosis detection via deep detection, verification and segmentation networks. *Medical Image Analysis*, 45, 121–133. <https://doi.org/10.1016/j.media.2017.12.002>

[14]: SAS Institute. (n.d.). Deep Learning - What it is and why it matters. https://www.sas.com/en_us/insights/analytics/deep-learning.html

[15]: Chollet, F.. (15 April 2020). Transfer learning & fine-tuning. https://keras.io/guides/transfer_learning/

[16]: Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, 1–14.

[17]: ImageNet. (30 April 2010). Summary and Statistics. <http://image-net.org/about-stats>

[18]: He, K. et al. (10 Dec 2015). Deep Residual Learning for Image Recognition. https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf

[19]: Universidade Federal do Paraná. (n.d.). Breast Cancer Histopathological Database (BreakHis). <https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/>

[20]: Spanhol, F. A., Oliveira, L. S., Petitjean, C., & Heutte, L. (2016). A Dataset for Breast Cancer Histopathological Image Classification. *IEEE Transactions on Biomedical Engineering*, 63(7), 1455–1462. <https://doi.org/10.1109/TBME.2015.2496264>

[21]: Titford, M. (2005). The long history of hematoxylin. *Biotechnic and Histochemistry*, 80(2), 73–78. <https://doi.org/10.1080/10520290500138372>

[22]: Eosin - an overview | ScienceDirect Topics. (n.d.). Retrieved April 11, 2021, from <https://www.sciencedirect.com/topics/medicine-and-dentistry/eosin>

[23]: Chollet, F. (2016, June 5). Building powerful image classification models using very little data. <https://blog.keras.io/building-powerful-image-classification-models-using-very-little-data.html>

[24]: Gandomkar, Z., Brennan, P. C., & Mello-Thoms, C. (2018). MuDeRN: Multi-category classification of breast histopathological image using deep residual networks. *Artificial Intelligence in Medicine*, 88, 14–24. <https://doi.org/10.1016/j.artmed.2018.04.005>

[25]: Pantanowitz, L., Farahani, N., & Parwani, A. (2015). Whole slide imaging in pathology: advantages, limitations, and emerging perspectives. *Pathology and Laboratory Medicine International*, 7, 23. <https://doi.org/10.2147/PLMI.S59826>

[26]: Roux, L. (n.d.). Download - MITOS-ATYPIA-14 - Grand Challenge. Retrieved April 11, 2021, from <https://mitos-atypia-14.grand-challenge.org/Download/>

[27]: Li, C., Wang, X., Liu, W., Latecki, L. J., Wang, B., & Huang, J. (2019). Weakly supervised mitosis detection in breast histopathology images using concentric loss. *Medical Image Analysis*, 53, 165–178.

<https://doi.org/10.1016/j.media.2019.01.013>

[28]: Chen, H., Dou, Q., Wang, X., Qin, J., & Heng, P. A. (2016). Mitosis detection in breast cancer histology images via deep cascaded networks. 30th AAAI Conference on Artificial Intelligence, AAAI 2016, 1160–1166.

[29]: Li, Y., Mercan, E., Knezevitch, S., Elmore, J. G., & Shapiro, L. G. (2018). Efficient and Accurate Mitosis Detection - A Lightweight RCNN Approach. Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods, 2018-January, 69–77. <https://doi.org/10.5220/0006550700690077>

[30]: Guo, Z., Zhang, L., & Zhang, D. (2010). A completed modeling of local binary pattern operator for texture classification. IEEE Transactions on Image Processing, 19(6), 1657–1663. <https://doi.org/10.1109/TIP.2010.2044957>

[31]: Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. (2011). ORB: An efficient alternative to SIFT or SURF. Proceedings of the IEEE International Conference on Computer Vision, 2564–2571. <https://doi.org/10.1109/ICCV.2011.6126544>

[32]: Ojala, T., Pietikäinen, M., & Mäenpää, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7), 971–987. <https://doi.org/10.1109/TPAMI.2002.1017623>

[33]: Ojansivu, V., & Heikkilä, J. (2008). Blur insensitive texture classification using local phase quantization. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 5099 LNCS, 236–243. https://doi.org/10.1007/978-3-540-69905-7_27

[34]: ERBB2 - Receptor tyrosine-protein kinase erbB-2 precursor - Homo sapiens (Human) - ERBB2 gene & protein. (n.d.). Retrieved April 8, 2021, from <https://www.uniprot.org/uniprot/P04626>

[35]: Dean, L. (2012). Trastuzumab (Herceptin) Therapy and ERBB2 (HER2) Genotype. In Medical Genetics Summaries. National Center for Biotechnology Information (US). <http://www.ncbi.nlm.nih.gov/pubmed/28520362>

[36]: Definition of HER2/neu - NCI Dictionary of Cancer Terms - National Cancer Institute. (n.d.). Retrieved April 8, 2021, from <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/her2-neu>

[37]: Targeted Drug Therapy for Breast Cancer. (n.d.). Retrieved April 11, 2021, from <https://www.cancer.org/cancer/breast-cancer/treatment/targeted-therapy-for-breast-cancer.html>

[38]: Wolff, A. C., Hammond, M. E. H., Schwartz, J. N., Hagerty, K. L., Allred, D. C., Cote, R. J., Dowsett, M., Fitzgibbons, P. L., Hanna, W. M., Langer, A., McShane, L. M., Paik, S., Pegram, M. D., Perez, E. A., Press, M. F., Rhodes, A., Sturgeon, C., Taube, S. E., Tubbs, R., ... Hayes, D. F. (2007). American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. In Journal of Clinical Oncology (Vol. 25, Issue 1, pp. 118–145). J Clin Oncol. <https://doi.org/10.1200/JCO.2006.09.2775>

[39]: HER2 (Breast Cancer) Testing: MedlinePlus Medical Test. (n.d.). Retrieved April 11, 2021, from <https://medlineplus.gov/lab-tests/her2-breast-cancer-testing/>

[40]: American Cancer Society. (2021). Breast Cancer HER2 Status. Retrieved April 11, 2021, from <https://www.cancer.org/cancer/breast-cancer/understanding-a-breast-cancer-diagnosis/breast-cancer-her2-status.html>

- [41]: N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," in IEEE Transactions on Systems, Man, and Cybernetics, vol. 9, no. 1, pp. 62-66, Jan. 1979, doi: 10.1109/TSMC.1979.4310076.
- [42]: HER2 FISH Testing for Breast Cancer - Patient Corner - URM Clinical Labs - University of Rochester Medical Center. (n.d.). Retrieved April 11, 2021, from <https://www.urmc.rochester.edu/pathology-labs/clinical/anatomic-pathology/breast/her2-fish.aspx>
- [43]: Oliveira, S. P., Pinto, J. R., Gonçalves, T., Canas-Marques, R., Cardoso, M.-J., Oliveira, H. P., & Cardoso, J. S. (n.d.). Weakly-Supervised Classification of HER2 Expression in Breast Cancer Haematoxylin and Eosin Stained Slides. <https://doi.org/10.3390/app10144728>
- [44]: Mukundan, R. (2018). Analysis of Image Feature Characteristics for Automated Scoring of HER2 in Histology Slides. Medical Image Understanding and Analysis, 9–11. <https://doi.org/10.3390/jimaging5030035>
- [45]: Garrison, L. P., Lalla, D., Brammer, M., Babigumira, J. B., Wang, B., & Perez, E. A. (2013). Assessing the potential cost-effectiveness of retesting IHC0, IHC1+, or FISH-negative early stage breast cancer patients for HER2 status. Cancer, 119(17), 3113–3122. <https://doi.org/10.1002/cncr.28196>
- [46]: Qaiser, T., Mukherjee, A., Reddy PB, C., Munugoti, S. D., Tallam, V., Pitkäaho, T., Lehtimäki, T., Naughton, T., Berseth, M., Pedraza, A., Mukundan, R., Smith, M., Bhalerao, A., Rodner, E., Simon, M., Denzler, J., Huang, C.-H., Bueno, G., Snead, D., ... Rajpoot, N. (2018). HER2 challenge contest: a detailed assessment of automated HER2 scoring algorithms in whole slide images of breast cancer tissues. Histopathology, 72(2), 227–238. <https://doi.org/10.1111/his.13333>
- [47]: Qaiser, T., & Rajpoot, N. M. (2019). Learning Where to See: A Novel Attention Model for Automated Immunohistochemical Scoring. IEEE Transactions on Medical Imaging, 38(11), 2620–2631. <https://doi.org/10.1109/TMI.2019.2907049>
- [48]: Tzotsos, A., & Argialas, D. (2008). Support vector machine classification for object-based image analysis. Lecture Notes in Geoinformation and Cartography, 0(9783540770572), 663–677. https://doi.org/10.1007/978-3-540-77058-9_36
- [49]: 1.4. Support Vector Machines — scikit-learn 0.24.1 documentation. (n.d.). Retrieved April 11, 2021, from <https://scikit-learn.org/stable/modules/svm.html>
- [50]: Copeland, M. (2016, July 29). The Difference Between AI, Machine Learning, and Deep Learning? | NVIDIA Blog. <https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/>
- [51]: CS231n Convolutional Neural Networks for Visual Recognition - Convolutional Neural Networks (CNNs / ConvNets). (n.d.). Retrieved April 11, 2021, from <https://cs231n.github.io/convolutional-networks/>
- [52]: Amidi, A., & Amidi, S. (n.d.). CS 230 - Convolutional Neural Networks Cheatsheet. Retrieved April 8, 2021, from <https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-convolutional-neural-networks>
- [53]: Kumar, N., Gupta, R., & Gupta, S. (2020). Whole Slide Imaging (WSI) in Pathology: Current Perspectives and Future Directions. In Journal of Digital Imaging (Vol. 33, Issue 4, pp. 1034–1040). Springer Science and Business Media Deutschland GmbH. <https://doi.org/10.1007/s10278-020-00351-z>

[54]: Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet Classification with Deep Convolutional Neural Networks. COMMUNICATIONS OF THE ACM, 60(6). <https://doi.org/10.1145/3065386>

[55]: TIA Centre Warwick: Her2 Scoring Contest. (n.d.). Retrieved April 11, 2021, from https://warwick.ac.uk/fac/cross_fac/tia/data/her2contest

[56]: Xie, J., Liu, R., Luttrell, J., IV, & Zhang, C. (2019). Deep Learning Based Analysis of Histopathological Images of Breast Cancer. Frontiers in Genetics, 10. <https://doi.org/10.3389/FGENE.2019.00080>

[57]: Spanhol, F. A., Oliveira, L. S., Petitjean, C., & Heutte, L. (2016). Breast Cancer Histopathological Image Classification using Convolutional Neural Networks. International Joint Conference on Neural Networks 2016. <https://www.inf.ufpr.br/lesoliveira/download/IJCNN2016-BC.pdf>

Figure 4-1: Dheepthi Mohanraj

Figure 4-2: Dheepthi Mohanraj

Figure 4-3: <https://iciar2018-challenge.grand-challenge.org/Dataset/>

Figure 5-1: Dheepthi Mohanraj

Figure 5-2: Dheepthi Mohanraj

Figure 5-3: Dheepthi Mohanraj

Figure 5-4: Dheepthi Mohanraj

Figure 5-5: Dheepthi Mohanraj

Figure 5-6: Dheepthi Mohanraj

Figure 5-7: Dheepthi Mohanraj

Figure 5-8: Dheepthi Mohanraj

Figure 5-9: Dheepthi Mohanraj

Figure 5-10: Dheepthi Mohanraj

Figure 5-11: Dheepthi Mohanraj

Figure 5-12: Dheepthi Mohanraj

Figure 5-13: Dheepthi Mohanraj

Figure 5-14: Dheepthi Mohanraj

Figure 5-15: Dheepthi Mohanraj

Figure 5-16: Dheepthi Mohanraj

Figure 5-17: Dheepthi Mohanraj

Figure 5-18: Dheepthi Mohanraj

Figure 5-19: Dheepthi Mohanraj

Figure 5-20: Dheepthi Mohanraj

Figure 6-1: Dheepthi Mohanraj

Figure 6-2: Dheepthi Mohanraj

Figure 6-3: Dheepthi Mohanraj

Figure 6-4: Dheepthi Mohanraj

Figure 6-5: Dheepthi Mohanraj

Figure 6-6: Dheepthi Mohanraj

Figure 7-1: Dheepthi Mohanraj

Figure 7-2: Dheepthi Mohanraj

Figure 7-3: Dheepthi Mohanraj

Copyright © 2021 All Rights Reserved