Evaluation of Hue Saturation Value Thresholded Breast Cancer Histopathological Image Detection on Ensemble Models

MSc Research Project
MSc in Data Analytics

Baris Anik
Student ID: x21236178

School of Computing
National College of Ireland

Supervisor: Rejwanul Haque

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Baris Anik
Student ID:	x21236178
Programme:	MSc in Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Rejwanul Haque
Submission Due Date:	14/08/2023
Project Title:	Evaluation of Hue Saturation Value Thresholded Breast Cancer Histopathological Image Detection on Ensemble Models
Word Count:	7533
Page Count:	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	14th August 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Evaluation of Hue Saturation Value Thresholded Breast Cancer Histopathological Image Detection on Ensemble Models

Baris Anik
x21236178

Abstract

Breast cancer is currently the most common type of cancer in the world. Current studies show that computer-assisted detection and diagnosis of breast cancer may be helpful. Developing algorithms to operate in this area helps reduce human errors in the diagnostic process and shortens this process. Convolutional neural networks (CNN) based models are frequently used for visual processing in breast cancer diagnosis. For this reason, in this study, CNN-based transfer learning models with potential success were first compared, and then various ensemble models were established with successful individual models. Hue saturation value (HSV) is a frequently preferred color space in medical visual processing studies. It was preferred in this study because it is closest to human perception and easier to manipulate. This study specifically examines the effect of images thresholded in HSV color space on ensemble models. The study results show that HSV thresholding negatively affects performance in individual models. However, HSV is secondary in ensemble models but shows results very close to not-thresholded ensemble models. In addition, HSV models have been observed to perform better than the original models in all magnification factors except 40X.

1 Introduction

1.1 Background and Motivation

Recently, breast cancer has surpassed lung cancer to become the most common cancer globally. (Han and Yin; 2022) As shown in figure 1, especially in Asia, one of every five women is diagnosed with breast cancer. In advanced stages, breast cancer metastasizes and causes death. Therefore, early diagnosis is critical for the treatment of the disease. In the process of regular mammography screenings and self-examination, a biopsy is performed in case of detection of a suspicious case. Small samples can be obtained from the tumor detected in the biopsy process by vacuum-assisted biopsy, fine needle aspiration, and core needle aspiration. A surgical operation needs to be performed to obtain a larger sample. The sample obtained by biopsy is stained with a solution called Hematoxylin and Eosin (H&E) before being examined under a microscope. This solution facilitates the diagnostic process by staining different components, such as tissue and cell nuclei, in purple and blue colors. (Marini et al.; 2021) The diagnosis of breast cancer requires educated radiologists. Failure to diagnose cancer on time may make the treatment process impossible, and in case of misdiagnosis, patients may suffer physical and mental harm. Machine learning models can sometimes make better predictions than experts. For instance, in the study of Turk et al. (2022), a random forest algorithm was more successful than trained radiologists in diagnosing MRI of people with diffuse glioma. While machine learning algorithms are still not entirely seen as a mechanism to carry out the diagnostic process personally, they are well suited to assisting doctors in the diagnostic process. In this way, the diagnosis process can be shortened, and doctors can be prevented from making a false diagnosis with fatigue factors. (Lee et al.; 2022)

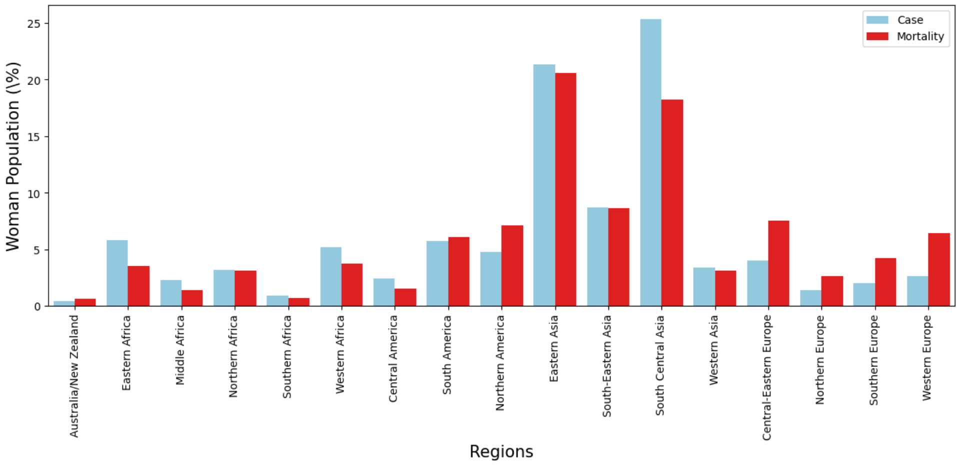


Figure 1: Case and mortality rates of breast cancer among the female population on 2020 (Arnold et al.; 2022)

Convolutional neural networks (CNN) are a frequently preferred model type in medical visual processing. For this reason, transfer learning models with CNN infrastructure are frequently preferred. Transfer learning is a method that uses optimal weights obtained by training on a ready-made model. In this way, the training period is shortened and contributes to model performance improvement. It is possible to adapt models to different data sets by changing the input and output layers of transfer learning methods and retraining particular layers of the models. This process is called fine-tuning. (Aljuaid et al.; 2022) Ensemble learning is a preferred approach to increase the models' individual performance and obtain models with high generalization capacity. It is created by combining multiple models with a voting approach. With this voting approach, the predictions of different models act with a certain weight or equally on the final prediction of the ensemble model. (Alotaibi et al.; 2023)

1.2 Research Question and Contribution

This study seeks to answer the following research question: "How well can HSV-thresholded histopathological images perform on ensemble models against ensemble models trained with original histopathological images to classify breast cancer regarding sensitivity, accuracy and specificity?"

Past studies have shown that HSV color space is compatible with histopathological visuals. In general, HSV color space is preferred for color transformations. HSV thresholding aims to remove the areas that are not necessary for diagnosis in the visual and thus to increase the success of ensemble models. If the expected performance increase is achieved, it may increase performance in computer-aided medical diagnostic classification studies that use histopathological visuals as inputs, especially in breast cancer diagnosis.

1.3 Structure of the Paper

Other parts of the study are as follows: Section 2 examines machine learning methods used in breast cancer and studies using HSV color space. Section 3 contains the methodology followed by this study. Section 4 describes the design specification. Section 5 includes the implementation and finally the results and conclusion are discussed in Sections 6 and 7.

2 Related Work

2.1 HSV Applications on Medical Image Pre-Processing

Aswathy and Jagannath (2021) have developed an approach to classifying breast cancer using support vector machine (SVM) based on visual features (color, texture and geometrical). BreakHis and The University of California Santa Barbara (UCSB) datasets were used in the study. The authors state that the contrast-limited adaptive histogram equalization (CLAHE) process used in the pre-processing phase makes the anomalies in the image more pronounced. With the k-means clustering algorithm, with a hyperparameter setting k equals 3, cell nuclei were detected in the visuals and separated from other textural components. The authors specifically explained using k-means because it performs well in detecting round-shaped objects. The k-means performs well for cell nuclei detection with 93% accuracy score. Detecting cell nuclei helps determine the region of interest (ROI) for the feature extraction phase. During the feature extraction phase, it was performed in three different phases. Hue, saturation, and value (HSV) color features were extracted during the color feature extraction phase. The authors base the selection of HSV for the extraction of color characteristics on the fact that HSV has the color scale closest to human perception. In the classification phase, in addition to SVM, artificial neural networks (ANN), k-Nearest neighbor (kNN), and random forest (RF) algorithms were also used for performance comparison. The authors state that SVM outperforms other models with an accuracy value of 91%. However, the ANN performs very close with an accuracy of 90.2%. Moreover, regarding the sensitivity metric, it outperformed SVM, which showed a 90% sensitivity result with a value of 96.1%. This study is a good example of using HSV color space in histopathological visuals. A simple process such as extracting HSV characteristics has been shown to be beneficial to model success.

Luz et al. (2022) presented a study investigating the combination of CNN models with different color adjustment settings within the ensemble model. The main focus of the study is to prevent the color variation problem in the images from adversely affecting the prediction of the models. The PatchCamelyon (PCam) dataset is a balanced dataset that contains 327680 histopathological images of 96 x 96-pixel size was used in the study. The ensemble model includes the ResNet50, VGG19, InceptionV3, Xception, DenseNet121 and DenseNet201 models. These models are primarily fine-tuned. Reinhard and Macenko color normalization techniques were used individually on images. These visual methods also need a target image as input. For this reason, it is necessary to select a reference image from expert pathologists. The post-processing image of the sample image created a brighter visual with the Reinhard technique, while the Macenko technique produced a visual with darker tones. The study experimented with eight color spaces: CIE RGB, HED, HSV, LAB, LUV, RGB, YIQ and YUV. The training was carried out on a one-to-one model with each color space. The authors interpreted the table containing the individual performance of the sub-models of the ensemble model by stating that the VGG19 performed better than the other models with 86.80% accuracy and 94.74% sensitivity. In addition, DenseNet201 achieved a remarkable sensitivity value with a score of 96.23%. In the ensemble model results, it is seen that the images with HSV color space show the best results with 96.12% sensitivity and 91.93% accuracy.

Luz et al. (2022) has the closest methodology to this work but also contains significant differences. The authors based their strategy on the use of color spaces on the implementation of colors on a different scale. This study aimed to use the HSV color scale in the thresholding process. The Pcam dataset used by the authors is extensive and contains different examples, so it is reasonable for them to determine a more complex strategy. The BreakHis dataset used in this study has lower visual variation because it is obtained from a single source. Therefore data augmentation was preferred in this study. The study results show that in most combinations, a minimum accuracy of 80% and a minimum sensitivity value of 90% are achieved. Even in this case, HSV achieves a positive difference compared to other color scales. In addition, the positive effect of ensembling these models has been observed.

Although HSV color space is generally applied to histopathological visuals, there are studies in which HSV color space is used on mammograms. Ural (2022) carried out a breast cancer classification study with a dataset of 50 benign and 50 malignant breast cancer mammograms from Yildirim Beyazit University Ataturk Research Hospital. A test dataset of 30 visuals, 15 benign and 15 malignant samples, was also created from the MIAS dataset. Since mammograms have colder tones, they are first converted into LAB color space. The rendered images are then returned to the HSV color space. Thanks to the HSV transformation, it has been noticed that benign images are included in colder tones while malignant images are included in the hotter color scale. This way, HSV information was transmitted to Learning Vector Quantization and Support Vector Regression models. Both machine learning methods appear to achieve a sensitivity value of over 90%. In this context, HSV conversion also performs well with mammograms.

Wahyuni et al. (2022) conducted a machine learning study on detecting invasive ductal carcinoma (IDC) subtype due to the prevalence of (IDC) in breast cancer at around 70%-80%. The study used a dataset containing 227524 IDC samples from Kaggle. Feature extraction was performed with an image histogram, HSV, RGB, and GLCM combination. With GLCM, correlation, contrast, homogeneity and energy properties were extracted from the image. Image histogram was used to extract variance, skewness, standard deviation, mean and entropy properties. Backpropagation and SVM algorithms were tried individually for classification. In order to ensure that the models do not have over-fitting problems, 5, 7, and 10-fold cross-validation processes were applied. With the combined features and backpropagation algorithm, a sensitivity value of 96.51% was achieved in the 7-fold cross-validation process. Backpropagation and SVM models trained only with HSV features also achieved remarkable success with 95.64% sensitivity and 97.54% specificity values. This study shows that HSV is beneficial in terms of feature extraction as well as CNN models.

2.2 Transfer Learning Approaches on Breast Cancer Visual Classification

One of the older studies testing the transfer learning method on breast cancer was the Chang et al. (2017) study. Although the entire BreakHis dataset was used in the study, the 40X magnification factor subset was used for training. The authors state that the reason for choosing this magnification factor subset is to verify the ability to identify the region of interest (ROI). In order to meet the high number of visual needs of CNN for training purposes, data augmentation was applied to benign and malignant visuals with rotation, flip, and random distortion operations. Although data augmentation has been applied, benign (3504 images) and malignant (7680 images) classes are not balanced. The original InceptionV3 model was used in the study as the transfer learning method. Cross-entropy loss function was used in the evaluation process of the model. Cut-off optimization was applied in the model. That choice is because the wrong classification of malignant samples has a higher cost than benign ones. For this reason, benign and malignant accuracy values were also measured within the stages during the model's training process. At a cut-off value of 0.4, 83% benign and 89% malignant accuracy were obtained. Since the malignant classification accuracy value decreased after this stage, this stage was preferred as the final version. In addition, an area under curve (AUC) score of 93% was achieved for both classes at a cut-off value of 0.4. This study states that monitoring the training of the models at every stage may be helpful. Therefore, Chang et al. (2017) contributed to using the early stopping method in this study.

One of the studies focusing on the class imbalance problem in breast cancer classification studies is Singh et al. (2020). The VGG19 model was applied to 277524 histopathologic images as 198738 IDC and 78786 healthy tissue (non-IDC) histopathological images. The authors state that this study differs from previous studies because the entire dataset used is included in the process. The difference in the number of visuals between classes is remarkably high. In addition, the coefficients of the VGG19 model trained with the ImageNet dataset were kept. However, various modifications have been applied to the model. Two different trials were carried out. In the first attempt, the dense and softmax layers were left after the five blocks of VGG19. In the other experiment, the dense and softmax layers were removed, and a classifier was placed at the end of the five blocks of the model. Random forest was preferred as the classifier. The reason for this preference is that the random forest algorithm performs better on high-dimensional data and has the ability to handle over-fitting. The extracted features from the five blocks of VGG19 are transmitted to the random forest, and a binary classification is performed. Consisting of the random forest structure added to the modified VGG19, the model achieved 93.31% sensitivity, 94.66% specificity, and 90.30% accuracy. Despite the class imbalance in the dataset, an impressive result has been obtained, especially regarding sensitivity metric. This study is valuable because it shows that using all layers of transfer learning models may only sometimes be beneficial.

Unlike other studies, the Saber et al. (2021) study works on mammogram analysis with the MIAS dataset. The dataset consists of 61 benign, 52 malignant, and 209 non-tumor samples. However, the CNN-based transfer learning model is also used in this study. With the contrast adjustment carried out during the pre-processing phase, it was seen that the anomalies in the images became apparent. Then, non-breast regions are removed by morphological analysis. Data augmentation has been applied to all dataset. In the study, ResNet50, VGG16, VGG19, Inception V3, and ResNet V2 models were preferred for transfer learning. The last three layers of the models have been removed. Instead of these layers, a dense layer was added, and fine-tuning was performed. However, a CNN model has been retrained for comparison. The resulting extracted features were passed to an SVM algorithm. In this process, stochastic gradient descent with momentum (SGDM) was used for optimization. When the results are examined, it is seen that the most successful model in terms of sensitivity in the classification performed without pre-processing is VGG16, with a value of 55.76%. Regarding accuracy, Inception V3 (62.50%) and ResNet V2 (64.06%) models are successful. In the classification performed after the pre-processing process is applied, a significant difference is observed in the sensitivity and accuracy metrics with results over 90%. The most successful model in terms of sensitivity is the VGG16 with a value of 96%, while the most successful model in terms of accuracy is the Inception V3 with 96.19%. In the experiments where transfer learning methods were combined with SVM, VGG16 and Inception V3 models obtained sensitivity values of 97.27% and 97.2% with 10-fold cross-validation, respectively. One of the study's significant findings is that the 10-fold cross-validation process increases performance in every model except VGG16. The study shows that CNN-based transfer learning models also perform well as feature extractors in combination with another model. In contrast, it emphasizes how effective the pre-processing process is. Although the dataset used in the study differs from histopathological studies in terms of format, color scale, and content differences, it shows that simple procedures applied, especially in the data augmentation stage, can be effective.

More recent academic studies include model comparisons are made rather than a model being tried. Aljuaid et al. (2022)'s study is a typical example of this kind of study. The authors presented a transfer learning model comparison to classify histopathologic breast cancer images in binary and multi-class ways with the BreakHis dataset. The transfer learning models evaluated in the study are ResNet18, ShuffleNet, and InceptionV3-Net. In order to be classified accurately, pre-processing and simple data augmentation was used. It is not intended to balance the number of images between classes. The size of the augmented images is small to keep the processing process fast. The individual performance of these methods was measured. As a result, it is stated that the ResNet18 model shows a higher success in both binary classification and multi-class classification methods. It scored 97.5% in the binary classification on the ResNet18 and InceptionV3-Net sensitivity metrics, and surpassed the score of 97.5% on the specificity metric. In the Accuracy metric, the ResNet18 model is more successful. In the case of multi-class classification, the ResNet18 excels over other models in all metrics.

Since the lowest metric score in both classification types is 95.5%, it is seen that the transfer learning methods used individually achieve successful results. In addition, the success of the ResNet models on the dataset is remarkable, and it has inspired the experimentation of variations of the ResNet models in this study.

Saini and Susan (2023) proposed a novel approach with customized VGG16 pre-trained model. The authors suggest that properly designed deep learning model architecture will also solve the problem of class imbalance. In the study, the authors concatenated all layers until block four pool layer with a naive Inception block in the original structure of the VGG16 model and added batch normalization, flatten, dropout (with a dropout rate of 0.4), and dense layers to this structure. The trainability of the layers in the concatenated block has been deactivated. The use of VGG16 as a base model in the study is based on the fact that this model gives successful results in many image classification scenarios. In addition, It is stated that the structure of VGG16 is suitable for handling different magnification factors. As the reason why the block four pool layer was accepted as the limit in the original VGG16 structure, the authors mentioned observation in the experimental results that the layers after the block four pool layer complicated the training process and did not positively affect the result. The authors also used the BreakHis dataset during the training and testing phase of the study. This is very important in measuring the model's generalization ability and significantly increases the reliability of the test results of the model. In terms of training metrics of the model, the Adam optimizer and categorical cross-entropy loss function were used. The model achieved 96% accuracy in 40X, 100X and 200X magnification factor images. For the 400X magnification factor, an accuracy value of 93% was achieved. With the fine-tuning process, the success rate is increased up to 98%.

The authors state that modifying transfer learning models will be efficient regarding processing time, class imbalance problem, and model accuracy. Since an ensemble approach is discussed in this study, the success of the interaction of model combinations with datasets is focused rather than the customization of the models used. Of course, the one-to-one modification of the models used through a series of experimental studies can make the study more successful, but this also brings about a process that can be pretty time-consuming. The study shows the positive effect of fine-tuning on the VGG16. This is quite a remarkable result as it creates ideas that using fine-tuning in other CNN models can also be helpful.

2.3 Ensemble Learning Approaches on Breast Cancer Visual Classification

Das et al. (2022) conducted a study in which six different methods for histopathological classification of breast cancer were included in an ensemble model. BreakHis and BreCaHAD datasets were used in the study. The algorithms used in the study can be listed as follows: OTSU, Local Adaptive Threshold, K-Means, Fuzzy C-Means, Watershed, Level set and U-Net. The authors compare the individual performances of the models and state that the U-Net and Fuzzy C-Means methods were found to be successful. Since the study was based on clustering algorithms, it used Jaccard Index and Dice Similarity Index as performance metrics. It is observed that the proposed method is more performant than similar studies, especially in the Jaccard Index results. This study shows that ensemble models can be used not only in CNN-based models, but also in cases where text-based feature extraction is performed.

Abbasniya et al. (2022) provides a model to classify histopathological breast cancer images using deep features with gradient boosting methods. The variety of models of the study is quite extensive. A total of 16 transfer learning methods were compared. The authors state that the Inception-ReNet-v2 model shows the most successful results. This study examined convolutional models with Gradient Class Activation Map (Grad-CAM) on a channel basis. In addition to Inception-ResNet-v2, Light Gradient Boosting Machine (LightGBM), Categorical Boosting (CatBoost), and Extreme Gradient Boosting (XGBoost) classifiers were used in the study. Various ensemble models were tried in the study and the best results were obtained in an ensemble model consisting of Inception-ResNet-v2 and gradient boosted decision trees with an accuracy range of 95-97%. This model uses a soft voting approach. This study shows that soft voting is suitable for working in this domain. However, the structure of the hybrid model is novelty because it is not very common in the literature.

Alotaibi et al. (2023) created an ensemble model with vision transformer (ViT) and data-efficient image transformer (DeiT) models to classify histopathological breast cancer visuals in a multi-class manner. Unlike other studies, the undersampling technique was preferred instead of oversampling. The ensemble predicts the model with a soft (average) voting operation. The study results show that the balanced dataset is more successful, with 98,08% sensitivity and 98.17% accuracy scores. In addition, the proposed model achieves a more successful result than the individual models. However, the fact that the study's results contain high values may show over-fitting. This may be a result of the undersampling preference. This study shows that the strategy determined before the training of the dataset is as important as the selected models.

Zheng et al. (2023) proceeds with a multi-model approach. The authors uses an ensemble model which contains DenseNet201, ResNet50, InceptionV3, Xception, VGG16 and VGG19 to make binary classification on BreakHis breast cancer dataset. In order to eliminate the class imbalance problem, horizontal and vertical flip operations were applied to the visuals in the benign class, which had 2480 examples. 469 of the augmented benign images were randomly selected. In the study, it has emphasized that transfer learning shortens training time and increases success. Moreover, the authors state that the methods included in ensemble learning were chosen because the relevant methods were compared with their competitors and obtained higher results. During the training process, the trainability properties of the weights of the transfer learning models are deactivated, ensuring that the remaining weights from ImageNet are preserved. The authors stated that they consider it more important the accuracy than processing time in the study and discussed traditional ensemble learning approaches rather than fast ensemble learning approaches. They used the voting approach to achieve results in the ensemble learning structure they used. The reason why this preference is made in favor of the weighted voting application is that this method is usually the most preferred and easily manipulated in the case of binary classification. During the training phase of the model, Adam was preferred as the optimizer. The input size of the model is 224x224 pixels. Different batch size values were used in the training of the models that make up the ensemble model. Training lasted 60 Epoch. Looking at the individual performance of the models, VGG19, ResNet50 and DenseNet201 achieved sensitivity values in the range of 97% to 99% in the transfer learning approach. InceptionV3 had a sensitivity value of 99.26% with its training from scratch approach. It is seen that the models trained with the transfer learning approach have a sensitivity value of at least 90%. The two most successful models in individual scores were ResNet50 and DenseNet201. The ensemble model has a sensitivity and accuracy value of 98.90%. In general, it can be said that the model provides a good performance. If the InceptionV3 model, which gives the lowest results in the ensemble model, is removed from the ensemble model, the results may change positively.

This study has been inspiring in many subjects such as the choice of transfer learning models used, the benefits of transfer learning approach, the advantages of ensemble models. The authors conveyed their operations in a very transparent manner and simply conveyed the essays they carried out with a correct visualization and summarization strategy to the readers. Therefore, it has been a very useful example in the preparation of an academic study.

3 Methodology

3.1 Data Gathering

In this study, BreakHis dataset (Spanhol et al.; 2016) is used. BreakHis is a data set published by the Federal University of Parana, The Laboratory of Vision, Robotics and Imaging, containing 7909 700x460 pixel PNG histopathological images from 82 patients. The images have eight classes, four benign subspecies and four sub-malignant species. The dataset includes four magnification factors: 40X, 100X, 200X and 400X. However, the dataset is class-imbalanced because it contains 68.64% malignant samples. The visual distribution of the dataset is given in the table below:

Table 1: Image distribution per class and sub-class.

Main Class	Subtype	40X	100X	200X	400X	Subtype Total	Main Class Total
Benign	Adenosis	114	113	111	106	444	2480
	Fibroadenoma	253	260	264	237	1014	
	Phyllodes Tumor	109	121	108	115	453	
	Tubular Adenoma	149	150	140	130	569	
Malignant	Ductal Carcinoma	864	903	896	788	3451	5429
	Lobular Carcinoma	156	170	163	137	626	
	Mucinous Carcinoma	205	222	196	169	792	
	Papillary Carcinoma	145	142	135	138	560	
TOTAL	-	1995	2081	2013	1820	-	7909

The dataset was obtained from the website of the Federal University of Parana by filling out the form on February 7, 2023. Data owners consent to use the dataset in non-commercial research, provided that citation can be demonstrated. Therefore, there is no ethical obstacle to the use of the data set.

3.2 Pre-Processing

The original dataset went through some pre-processing stages to increase the throughput of the machine-learning models. The BreakHis data set contains 68.64% of the images belonging to the malignant class. For this reason, the benign class has undergone data augmentation in binary class data sets. The malignant class with the highest visual number was determined in the five-class data sets, and enough visuals were created to close this difference in other classes. HSV thresholding was applied to remove the parts of the images that were not related to the texture. In this way, the areas where the white color is found are removed from the visuals.

3.2.1 Data Augmentation

Since the training process was performed using only 40X magnification factor visuals, data augmentation was not used in other magnification factors. A function named *classBalancer* has been created for data augmentation. This function takes a file directory entry. First, it determines the number of images in the folders in the input sequence. The number of images that should be created in other classes is determined by referencing the highest number of visuals. Finally, this information creates the necessary amount of images in other folders. The ImageDataGenerator method of the Keras library was used to create the image. Horizontal flip, vertical flip, zoom (range = 0.2), and random rotation (range = 20) operations were used to create the visual. The images were created in 256 x 256 pixels to be compatible with CNN models.

3.2.2 HSV Thresholding

Two different strategies have been tried with the cv2 library for HSV thresholding. The first strategy is to maintain the blue-purple color scale, on which the nuclei and bound tissues are stained, and to remove other elements. The second strategy is generally based on removing white-toned areas in the visuals. With the *thresholderDemonstrator* function created, the effects of the HSV thresholding process on the sample images were observed. As a result of the visual examination, it was seen that the second strategy gave more successful outcomes. The second strategy reached the result faster because it is easier to determine a single color than it is to detect a scale with more than one color. The generated *thresholder* function saves the images in the input path directory to the export path by applying HSV thresholding.

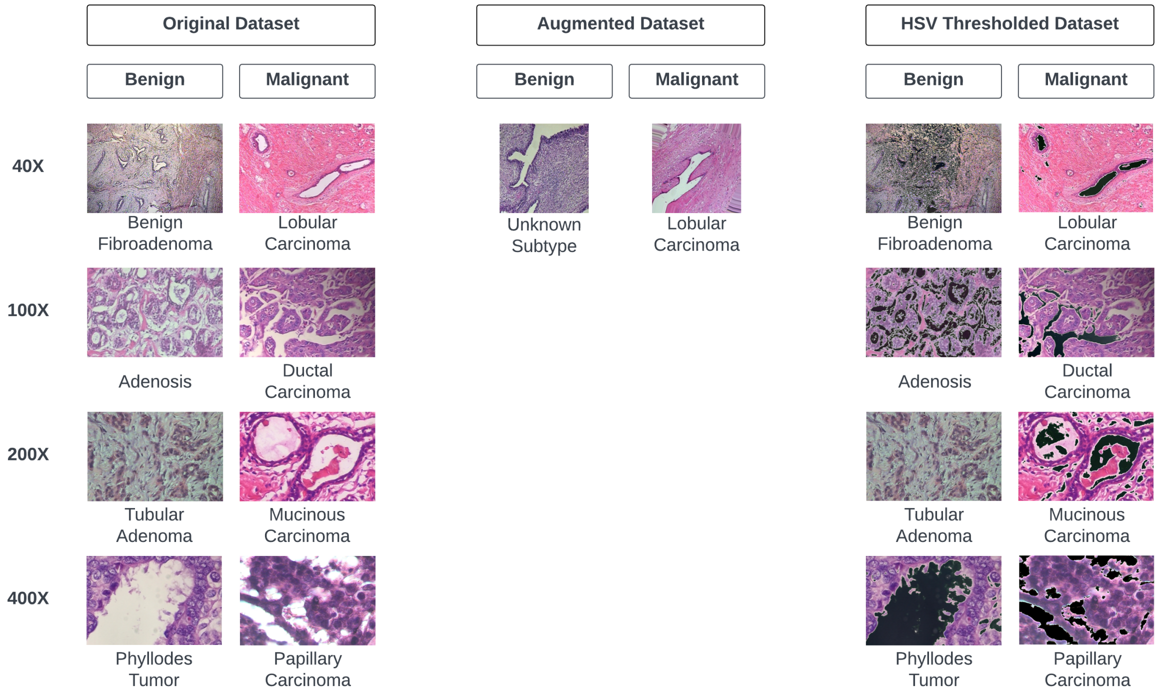


Figure 2: Examples from pre-processing stage.

3.3 Creation of Individual Models with Fine-Tuning

This study aims to create an ensemble model by combining transfer learning models. First, the individual performances of the models were compared. ResNet50, ResNet101, ResNet201, InceptionV3, Xception, VGG16, VGG19 and DenseNet201 models, which were found to perform well in the literature research, were selected for that phase. Since transfer learning models are trained with imagenet visuals, they are incompatible with the BreakHis dataset regarding input and output layers. For this reason, the models were fine-tuned. The function named *createModel* was defined and used to create the models to be used in the fine-tuning process. In the fine-tuning process, first of all, the input and output layers of the model are removed. Then a flatten layer is added after the model with a dense layer with 256 nodes and a dense layer with a value of 2 or 5 according to the expected output. The weights of all layers of the base model are preserved. In this way, both the optimal weights of the model were utilized, and the model was adapted to the BreakHis data set with extra layers. Different model variations are tested for each transfer learning model in fine-tuning. The details and differences of these variations are described in Section 3.4. Each version has different requirements. The *createModel* function has been created to create these models without errors. In addition, the *model_trainer* function is defined to organize the data sets to be selected in the model's training process and to keep the records related to the model error-free. In this way, the creation of models and errors that may occur in the training process are prevented.

In this study, two types of classification approaches are tried. The first approach is based on binary classification. This procedure is performed to classify breast cancer samples as binary or malignant. The other approach is a five-class classification. The benign subtypes of the BreakHis dataset, which normally have eight subclasses, are combined in the Benign class, while the malignant subtypes are not. The reason for trying this approach is that the correct detection of malignant samples in the study is much more critical. In this way, it will be observed whether a better performance will be obtained than the binary classification in the five-class approach.

3.4 Experimental Setups

3.4.1 Version 1

For the first versions all models are fine-tuned with a simple way. The input and output layers of the model were removed, and one flatten and two dense layers were added to the end of the models. The first dense layer contains 256 nodes with the ReLU activation function, while the last dense layer contains a number of nodes that vary according to the total number of classes with the softmax activation function. The trained weights of all layers of the transfer learning models used are preserved. Adam was chosen as the optimizer of the model with a learning rate of 10^{-3}. In the cases of five-class classification, the categorical crossentropy loss function is used, while in the binary classification the binary crossentropy loss function is used. The training process of the model was monitored with the accuracy and confusion matrix elements true positive, true negative, false positive and false negative metrics.

3.4.2 Version 2

The second version has all the features of the first version but has a data augmentation layer added before the transfer learning model. This layer includes random flip, random rotation and random zoom. The factor parameter of random rotation and random zoom operations is 0.2. It is aimed to take precautions against over-fitting by increasing the variety of training material of the model.

3.4.3 Version 3

The third version models are structurally identical to the second version models. No extra layers were added, the hyperparameters were left the same. The difference of the third version is found in the training and test data sets given to the model. The third version of the models were trained with data augmentation and class-balanced datasets.

A function is defined for the data augmentation process. This function first determines the number of classes in the specified index and determines how many images are linked to those classes. Stores class names and image numbers in a dictionary variable. Records the visual count of the class with the highest number of visuals. By subtracting the number of visuals of other classes from the maximum number, it determines the number of visuals needed according to the classes and saves this information in a dictionary variable. It navigates through each class in a for loop and creates the needed amount of visuals. Rotation, horizontal flip, vertical flip and zoom operations are performed to create diversity while creating visuals.

3.4.4 Version 4

The fourth version of the models follows all the features of the third model versions and the data set used for training. In addition, an early stopping tracker is being added. The early stop monitor that follows the validation loss value has a patience value of 5. In this way, it is aimed to obtain an optimal result without compromising the validation success at the point where the development of the model stops.

3.4.5 Version 5

The structure of the fifth version is same with version 4. However, the difference in this model is that HSV thresholding is used as an input to the data set as learning material.

3.4.6 Version 6

The sixth version is only being tested on the successful transfer learning models included in the first ensemble model. It has been added when the an ensemble model was suspected of over-fitting. It is structurally similar to the fourth model but includes a dropout layer with a ratio of 0.2.

3.5 Creation and Comparison of Possible Ensemble Models

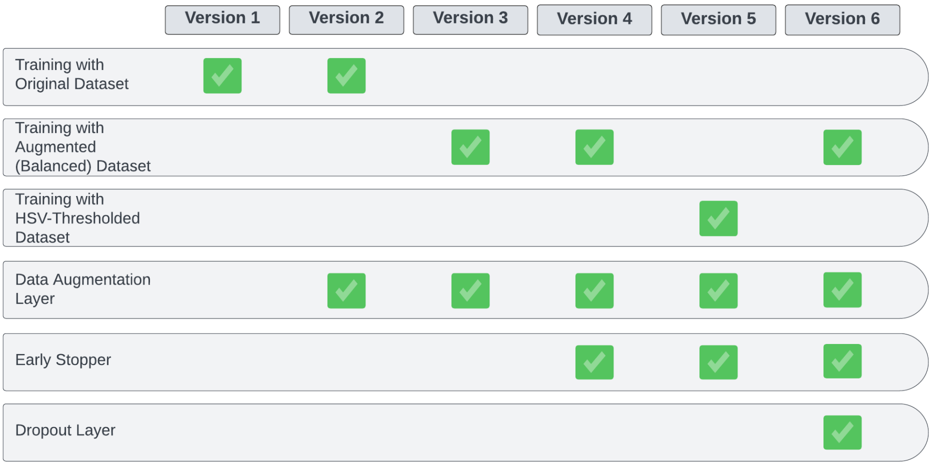
At this stage, ensemble models were created with combinations of models whose success stood out from various versions. Average voting approach was applied in all ensemble models.

4 Design Specification

4.1 Techniques Adopted

4.1.1 HSV Color Space Thresholding

HSV color space is one of the most preferred color spaces in medical visual processing studies. Aswathy and Jagannath (2021) stated that HSV color space is the closest color



	Version 1	Version 2	Version 3	Version 4	Version 5	Version 6
Training with Original Dataset	✓	✓				
Training with Augmented (Balanced) Dataset			✓	✓		✓
Training with HSV-Thresholded Dataset					✓	
Data Augmentation Layer		✓	✓	✓	✓	✓
Early Stopper				✓	✓	✓
Dropout Layer						✓

Figure 3: Visualization of individual model version comparison.

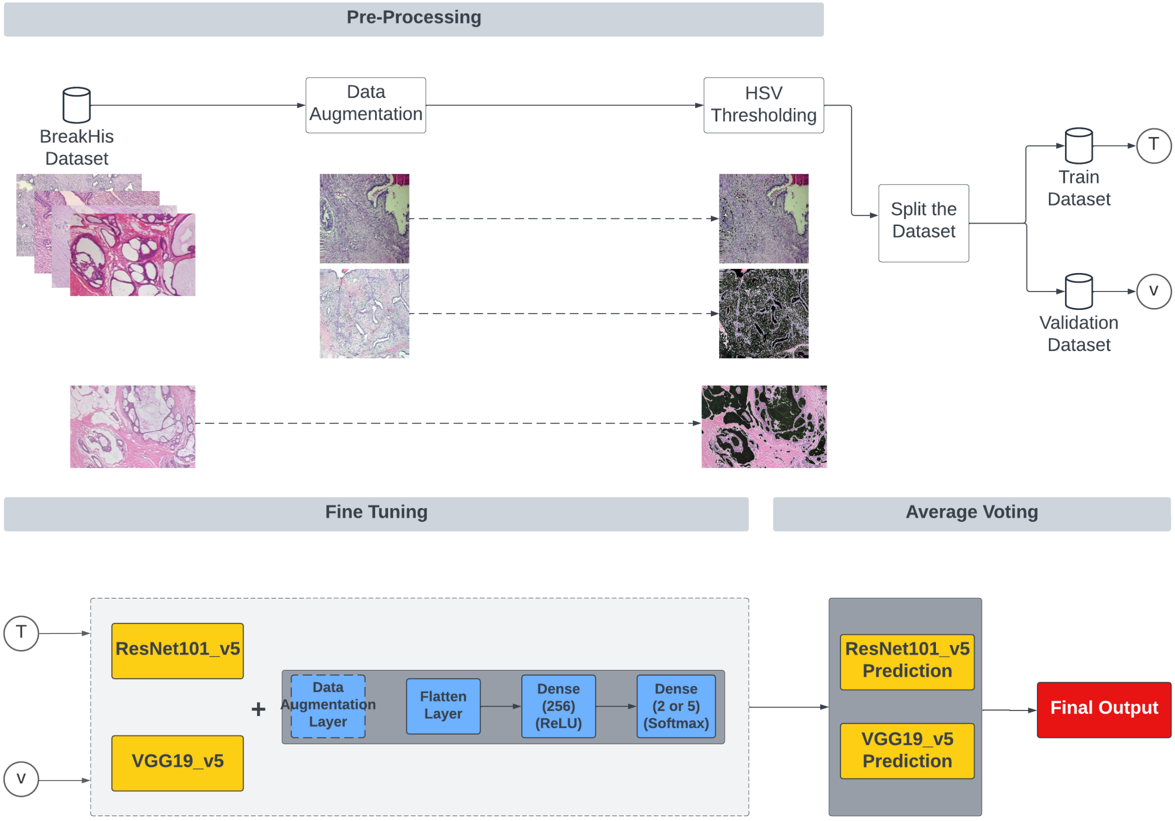


Figure 4: Architecture of the study.

space to human perception. In addition, it is a useful method for detecting, holding, or subtracting a specific range of colors. In summary, this method offers a simple and effective approach to be preferred.

4.1.2 Transfer Learning

Transfer learning is one of the frequently used methods in visual classification. It is mostly used when developing a new model to keep the model success high and to shorten the training process. (Aljuaid et al.; 2022) Indeed, they produce more efficient results that are easier to train than individual models. Optimized models generally trained on the ImageNet data set (Deng et al.; 2009), which mostly contains 3.2 million images and 12 subtrees of images, are fine-tuned and adapted to the current classification situation. In this study, CNN-based transfer learning models, which were found to perform better as a result of literature search, were used.

4.1.3 Ensemble Learning

Ensemble learning is a frequently preferred approach for analyzing histopathological visuals. Generally, the bagging sub approach is preferred instead of boosting in the breast cancer classification subdomain. After a fine-tuning process is applied to ensure that various transfer learning methods are compatible with the current data set, classification is carried out with the results that pass through the voting process. While some studies uses average voting to achieve higher results, some other studies used weighted average technique to get benefit of more successful models. (Zheng et al.; 2023)

5 Implementation

Python language and Jupyter Notebook environment were used to classify breast cancer visuals for study purposes. The Tensorflow library was used to load transfer learning models and weights and to perform data augmentation. For HSV thresholding, the glob and cv2 library was used. For the visualization process, matplotlib and seaborn libraries were used. Some functions are defined for the process to proceed within the standards and without errors.

5.1 createModel Function

The *createModel* function is used to create and compile models. It takes three parameters: base_model, version, and class_size. The function first creates a temporary sequential model in itself. According to the class_size argument, the activation function to be included in the last layer and the loss function to be used in the compile process are determined. The sigmoid activation function and binary cross entropy loss function are used for binary classification. The softmax activation function and cross-entropy loss function are used for multi-class classification. Then, it creates the model specified in the base_model parameter and deactivates the trainable property of all model layers. Then, according to the version of the model, the required features are added to the model. Details of the transactions performed here are mentioned in Section 3.4. Finally, this function returns the name of the base model and the generated model.

5.2 model_trainer Function

The *model_trainer* function is a function that allows models to be trained with the necessary datasets according to their versions. The function has parameters model_name, model, version, dataType, debug_mode, epochs, and patience. First, the defined datasets are placed in a list structure. This process aims to prevent the mixing of binary and five-class datasets. It is then added to a dictionary structure to improve the accessibility of these lists. Then the model is included in the fitting process with the dataset and early stopper following the version requirement. The history and model weights are saved when the fitting process is finished. Then, tests are performed on datasets with different magnification factors using the model, and the test results are recorded. Finally, the results obtained are visualized. The debug_mode parameter has been added to the function to perform extra experiments during the training process. In this way, the epoch number and patience value can be changed.

6 Results and Evaluation

6.1 Evaluation Strategy

Sensitivity, specificity and accuracy metrics are used to evaluate the results of the models. The sensitivity metric is especially important because it is critical how much of the malignant samples are detected. The formulas for these metrics are as follows:

$$Sensitivity = \frac{TruePositive}{(TruePositive + FalseNegative)} \tag{1}$$

$$Specificity = \frac{TrueNegative}{(TrueNegative + FalsePositive)} \tag{2}$$

$$Accuracy = \frac{TruePositive + TrueNegative}{(TruePositive + TrueNegative + FalsePositive + FalseNegative)} \tag{3}$$

6.2 Comparison of Individual Performance of Models

Considering the training and validation sensitivity values in Figure 5, it is observed that ResNet50, ResNet101, VGG16 and VGG19 models are more successful. This success was seen in both binary and five-class datasets.

The results show that ResNet50, ResNet101, VGG16, and VGG19 models are more successful regarding sensitivity metric. The success of the DenseNet201 version 1 model is also observed, but this result has yet to be considered because the first version of the model is elementary and prone to over-fitting. When the results of binary and five-class models are compared, it is observed that the difference between training and validation sensitivity values increases in five-class models compared to binary models. In binary models, Xception version 2 shows high test performance on the 100X, 200X, and 400X datasets. However, since the second version of the models did not contain a balanced dataset, this version was not considered included in the ensemble models. In addition, DenseNet201 version 3 performed relatively well than the Xception and Inception versions.

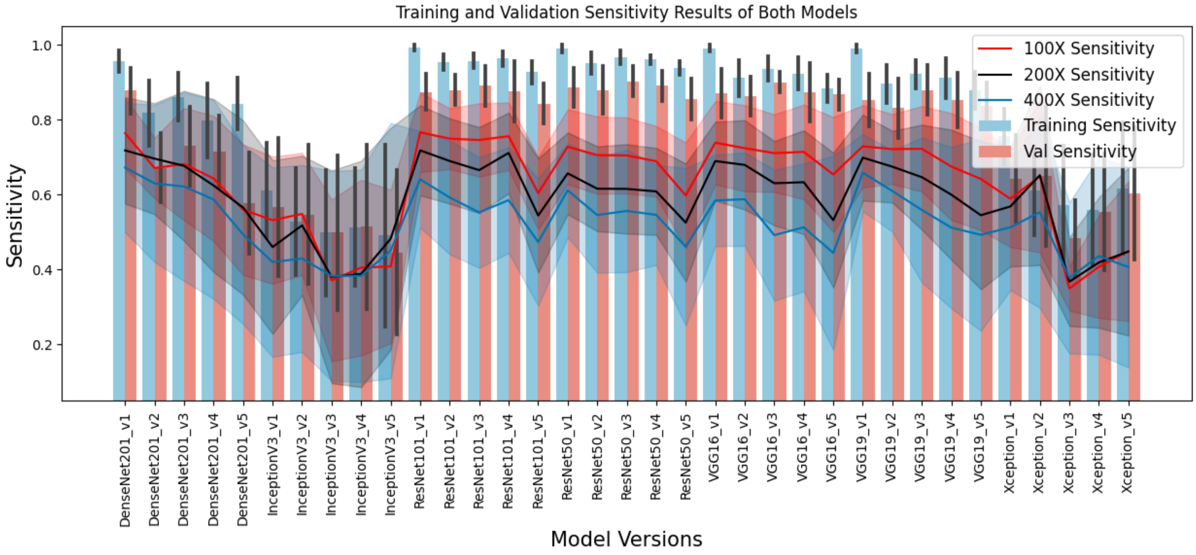


Figure 5: Training and validation sensitivity comparison of models.

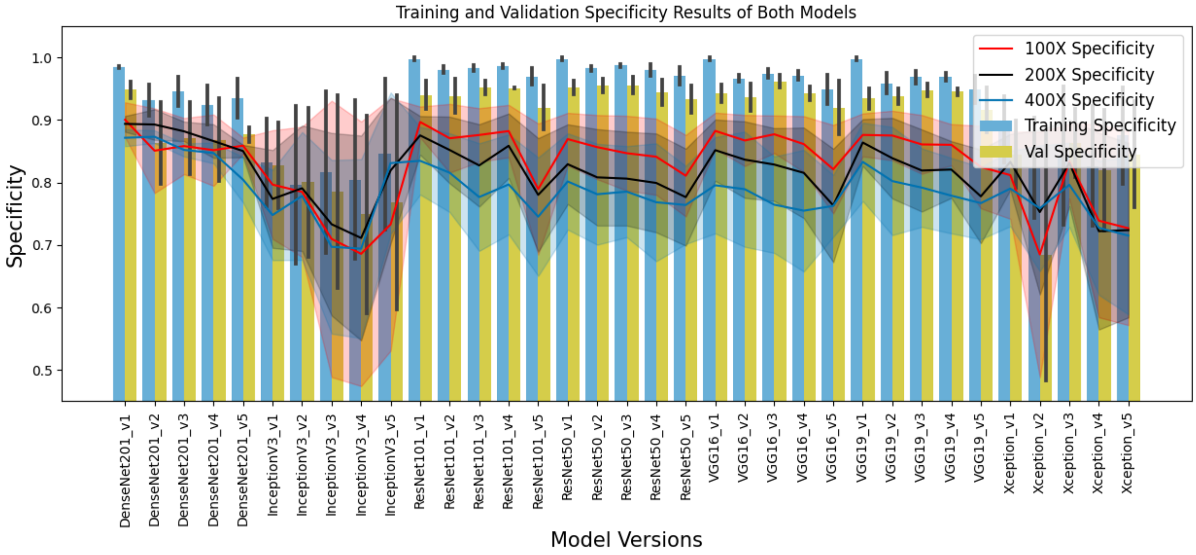


Figure 6: Training and validation specificity comparison of the all models.

The specificity results of the models also carry evidence confirming the sensitivity performance results. However, it is seen that more successful general results are obtained in terms of specificity. It is thought that the absence of subclasses of the benign class has affected the specificity metric. When the binary and five-class models are compared, it is observed that the five-class models are more successful in terms of specificity. However, the performance difference between the models shows the same results as the sensitivity results. Considering these results, it is seen that comparing model performance with a single metric may lead to incorrect results. It also appears that it is crucial to monitor the sensitivity metric on a topic such as cancer classification.

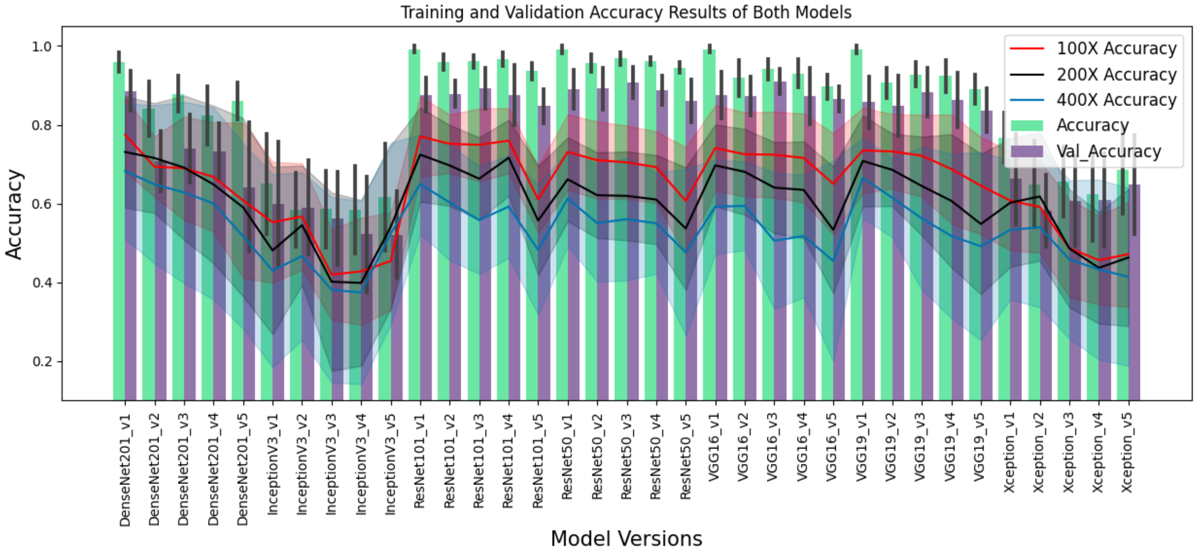


Figure 7: Training and validation accuracy comparison of the all models.

The results of the accuracy metric also confirm the results of the sensitivity metric. However, as can be seen from the specificity performance results, the five-class models are also observed to produce more successful results in general.

As a result, it was observed that the third and fourth versions of the models are seen as advantageous in two respects. The risk of over-fitting is higher because the first two versions are not trained with a balanced dataset. The fifth versions show lower results in terms of performance. The third and fourth models have measures against the problem of over-fitting because they are trained with a balanced dataset, while they have better results than the fifth version. In addition, the success of versions with HSV input was noticeably lower in individual models. However, when individual models are combined, better performance can be achieved. This will be discussed in Section 6.3.

6.3 Comparison of Performance of the Ensemble Models

Ensemble models are created with the defined *createEnsembleModel* function. A total of eight ensemble models were created from the ResNet50, ResNet101, VGG16, VGG19 and DenseNet201 versions. The first ensemble model created included ResNet50 version 4 and VGG19 version 4. The second ensemble model includes the ResNet101 and VGG19 version 4 models, as well as the DenseNet201 version 4 model. The third ensemble model includes the ResNet101 version 5 and VGG19 version 5 models. The fourth model includes the ResNet101 version 5, VGG19 version 5 and DenseNet201 version 5 models. The first two ensemble models were trained with balanced BreakHis datasets, while the third and fourth models used HSV thresholded datasets.

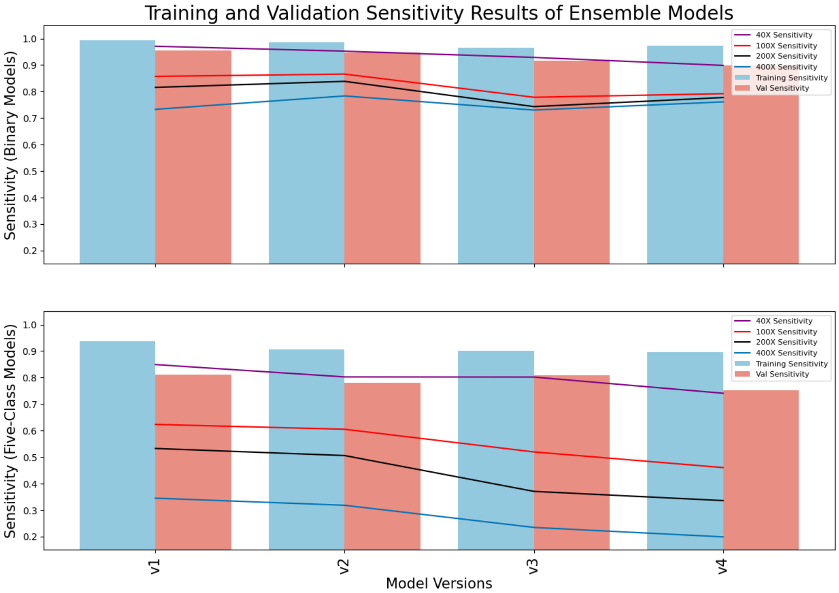


Figure 8: Training and validation specificity comparison of the all models.

Sensitivity performances show that binary models are more successful than five-class models. Ensemble version 1 binary model, including ResNet101 and VGG19, achieved a validation sensitivity performance of 95.62%.

In terms of specificity performance, it is seen that the five-class models are more successful than the binary models. Ensemble model version 2 has the most successful validation specificity performance with a score of 97.06%

When the accuracy values are examined, it is seen that binary models are more successful. It is seen that the most successful model is the ensemble model version 1, which makes binary classification with a score of 95.62%. However, even if the most successful model is the ensemble model version 1, the training accuracy value of 99% indicates that over-fitting can occur. As a precaution, a final ensemble model was created by adding a dropout layer to all ensemble model version 1 sub-models.

6.4 Experiment - Ensemble Model with Dropout

However, even if the most successful model is the ensemble model version 1, the training accuracy value of 99% indicates that over-fitting can occur. As a precaution, a final ensemble model was created by adding a dropout layer to all sub-models of ensemble model version 1. Because of the dropout layers, the new model (ensemble model version 5 with binary class structure) is significantly behind the ensemble model version 1 with 77.58% accuracy and 79.07% validation sensitivity. The performance of the new binary class model lags behind its alternatives and studies in the literature. And yet, compared to the five-class model, very positive results have been achieved. The five-class model was the most unsuccessful because it had 56.15% accuracy and 45.06% validation sensitivity.

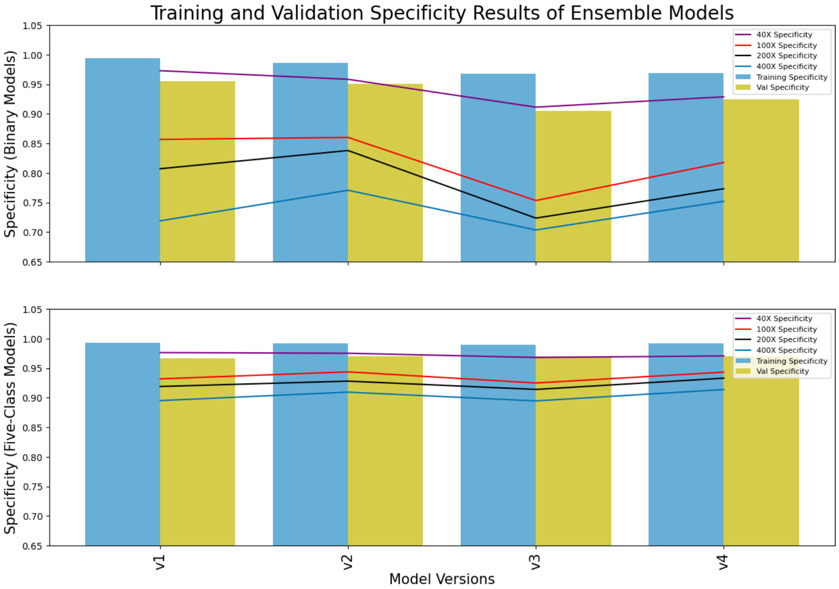


Figure 9: Training and validation specificity comparison of the all models.

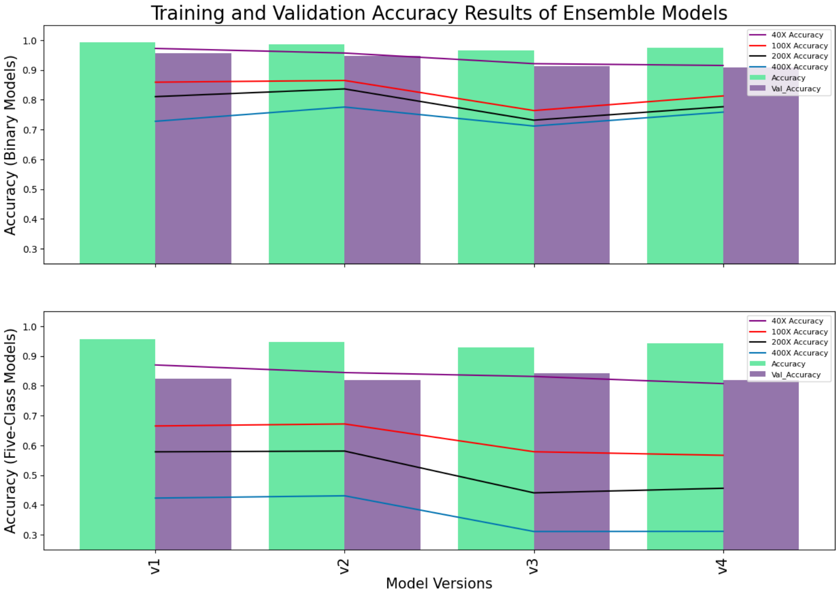


Figure 10: Training and validation accuracy comparison of the all models.

6.5 Discussion

In this study, an individual model comparison stage was organized to determine the transfer learning models to be added to the ensemble model. Models have received specific characteristics in the framework of various versions. There was no positive effect of HSV thresholding in individual models in the trials. It has been observed that ResNet50, ResNet101, VGG16, and VGG19 stand out more than other models. Usually, the third and fourth versions of the models offer reliable performance and results. However, when the ensemble models are compared, it is seen that the models with HSV thresholding catch up with other models. The models that have undergone the HSV-thresholding process in question have the potential to pass the ensemble models trained with the original visuals after a more detailed study.

The most successful variant of binary ensemble model version 1 with HSV thresholding (ensemble model 3) experienced a performance decline of approximately 3% compared to the first version. However, trying different CNN-based models and discovering better combinations is possible. In this study, this operation could not be performed due to time constraints.

7 Conclusion and Future Work

In this study, the effect of HSV thresholding on CNN-based transfer learning models and ensemble models was observed. The HSV thresholding process is relatively simple. Thanks to this simple process, transfer learning models are intended to have a more performant training process. As seen and discussed in Section 6, HSV thresholding, which lags in individual performance, shows successful results in ensemble models.

Some methods can be tried to improve the performance of the models. As discussed in Saini and Susan (2023)'s study, each transfer learning model to be added to the ensemble model can be examined layer by layer, and performance degradation points can be removed. In addition, several precautions were taken against over-fitting in this study, but it can be ensured that over-fitting is avoided with 10-fold cross-validation, as in Saber et al. (2021)'s study. However, Aswathy and Jagannath (2021)'s work contains an inspiring method for determining the region of interest. In this study, the images were cropped to 256x256 pixels without considering the ROI. In order to improve this process, a clustering algorithm to detect cell nuclei can be included in the pre-processing stage. Finally, in some visuals, it was seen that the HSV thresholding process needed to be more successful. It is known that the reason for this is that the specified color range cannot be detected in the image. For this reason, the white areas can be clarified by adjusting the contrast or brightness first on the images, and then HSV thresholding can be applied.

References

Abbasniya, M. R., Sheikholeslamzadeh, S. A., Nasiri, H. and Emami, S. (2022). Classification of Breast Tumors Based on Histopathology Images Using Deep Features and Ensemble of Gradient Boosting Methods, *Computers and Electrical Engineering* **103**: 108382.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S0045790622005997*

Aljuaid, H., Alturki, N., Alsubaie, N., Cavallaro, L. and Liotta, A. (2022). Computer-aided diagnosis for breast cancer classification using deep neural networks and transfer learning, *Computer Methods and Programs in Biomedicine* **223**: 106951.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S0169260722003339*

Alotaibi, A., Alafif, T., Alkhilaiwi, F., Alatawi, Y., Althobaiti, H., Alrefaei, A., Hawsawi, Y. and Nguyen, T. (2023). ViT-DeiT: An Ensemble Model for Breast Cancer Histopathological Images Classification, *2023 1st International Conference on Advanced Innovations in Smart Cities (ICAISC)*, IEEE, Jeddah, Saudi Arabia, pp. 1–6.
URL: *https://ieeexplore.ieee.org/document/10085467/*

Arnold, M., Morgan, E., Rumgay, H., Mafra, A., Singh, D., Laversanne, M., Vignat, J., Gralow, J. R., Cardoso, F., Siesling, S. and Soerjomataram, I. (2022). Current and future burden of breast cancer: Global statistics for 2020 and 2040, *The Breast* **66**: 15–23.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S0960977622001448*

Aswathy, M. A. and Jagannath, M. (2021). An SVM approach towards breast cancer classification from H&E-stained histopathology images based on integrated features, *Medical & Biological Engineering & Computing* **59**(9): 1773–1783.
URL: *https://link.springer.com/10.1007/s11517-021-02403-0*

Chang, J., Yu, J., Han, T., Chang, H.-j. and Park, E. (2017). A method for classifying medical images using transfer learning: A pilot study on histopathology of breast cancer, *2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom)*, IEEE, Dalian, pp. 1–4.
URL: *http://ieeexplore.ieee.org/document/8210843/*

Das, P., Sharma, R., Dey Roy, S., Nath, N. and Bhowmik, M. K. (2022). Ensemble Segmentation of Nucleus Regions from Histopathological Images towards Breast Abnormality Detection, *2022 25th International Conference on Computer and Information Technology (ICCIT)*, IEEE, Cox's Bazar, Bangladesh, pp. 1137–1142.
URL: *https://ieeexplore.ieee.org/document/10055451/*

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database, *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255.

Han, L. and Yin, Z. (2022). A hybrid breast cancer classification algorithm based on meta-learning and artificial neural networks., *Frontiers in oncology* **12**: 1042964. Place: Switzerland.

Lee, J. H., Kim, K. H., Lee, E. H., Ahn, J. S., Ryu, J. K., Park, Y. M., Shin, G. W., Kim, Y. J. and Choi, H. Y. (2022). Improving the Performance of Radiologists Using Artificial Intelligence-Based Detection Support Software for Mammography: A Multi-Reader Study, *Korean Journal of Radiology* **23**(5): 505.
URL: *https://kjronline.org/DOIx.php?id=10.3348/kjr.2021.0476*

Luz, D. S., Lima, T. J., Silva, R. R., Magalhães, D. M. and Araujo, F. H. (2022). Automatic detection metastasis in breast histopathological images based on ensemble learning and color adjustment, *Biomedical Signal Processing and Control* **75**: 103564.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S1746809422000866*

Marini, N., Atzori, M., Otalora, S., Marchand-Maillet, S. and Muller, H. (2021). H&E-adversarial network: a convolutional neural network to learn stain-invariant features through Hematoxylin & Eosin regression, *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, IEEE, Montreal, BC, Canada, pp. 601–610.
URL: *https://ieeexplore.ieee.org/document/9607529/*

Saber, A., Sakr, M., Abo-Seida, O. M., Keshk, A. and Chen, H. (2021). A Novel Deep-Learning Model for Automatic Detection and Classification of Breast Cancer Using the Transfer-Learning Technique, *IEEE Access* **9**: 71194–71209.
URL: *https://ieeexplore.ieee.org/document/9427477/*

Saini, M. and Susan, S. (2023). VGGIN-Net: Deep Transfer Network for Imbalanced Breast Cancer Dataset, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **20**(1): 752–762.
URL: *https://ieeexplore.ieee.org/document/9744541/*

Singh, R., Ahmed, T., Kumar, A., Singh, A. K., Pandey, A. K. and Singh, S. K. (2020). Imbalanced Breast Cancer Classification Using Transfer Learning, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* pp. 1–1.
URL: *https://ieeexplore.ieee.org/document/9037082/*

Spanhol, F. A., Oliveira, L. S., Petitjean, C. and Heutte, L. (2016). A Dataset for Breast Cancer Histopathological Image Classification, *IEEE Transactions on Biomedical Engineering* **63**(7): 1455–1462.
URL: *http://ieeexplore.ieee.org/document/7312934/*

Turk, S., Wang, N. C., Kitis, O., Mohammed, S., Ma, T., Lobo, R., Kim, J., Camelo-Piragua, S., Johnson, T. D., Kim, M. M., Junck, L., Moritani, T., Srinivasan, A., Rao, A. and Bapuraj, J. R. (2022). Comparative study of radiologists vs machine learning in differentiating biopsy-proven pseudoprogression and true progression in diffuse gliomas, *Neuroscience Informatics* **2**(3): 100088.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S2772528622000504*

Ural, A. B. (2022). Systematical Analysis and Pathological Classification of Breast Cancer from Mammographic Images with Using Specific Machine Learning Methods, *Traitement du Signal* **39**(6): 2149–2156.
URL: *https://www.iieta.org/journals/ts/paper/10.18280/ts.390628*

Wahyuni, E. S., Tiyas, V. G. and Murnani, S. (2022). Analysis of Feature Extraction and Classification Methods on Histopathological Images for Diagnosing Invasive Ductal Carcinoma, *2022 9th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, IEEE, Semarang, Indonesia, pp. 263–268.
URL: *https://ieeexplore.ieee.org/document/9923967/*

Zheng, Y., Li, C., Zhou, X., Chen, H., Xu, H., Li, Y., Zhang, H., Li, X., Sun, H., Huang, X. and Grzegorzek, M. (2023). Application of transfer learning and ensemble learning in image-level classification for breast histopathology, *Intelligent Medicine* **3**(2): 115–128.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S266710262200047X*