Mammogram Classification Using Texture Features and Multi-Layer Perceptron

Adam Gibicar, Matthew Basso

***Abstract*—Mammography is the best approach for non-invasive, early detection of breast cancer. By using image analysis and machine learning methods in biomedical images, significant diagnosis times can speed-up, and higher consistency can be achieved. This paper describes an automated feature extraction and classification approach for mammogram images. The features extracted from ROI mammogram images include breast density, mass shape, and Haralick texture features obtained from Haar, Biorthogonal discrete wavelet decomposition, and Gabor filters. Feature pruning was then performed using three different methods: feature importance (FI), principal component analysis (PCA), and pearson's correlation coefficient. Lastly, a simple 3-layer automatic neural network (ANN) was used and these techniques are evaluated based on their corresponding classification accuracy. Our results show that the Gabor-filter-GLCM-FI method classifies benign and malignant masses the best with an accuracy of 74.9% and AUC of 0.755.**

***Index Terms*—Mammogram, Classification, Medical Imaging, Image Analysis, Machine Learning**

I. Introduction

Healthcare professionals and doctors receive extensive training in order to detect, diagnose, and prevent diseases. Although these professionals are highly qualified it is important for them to have access to a wide range of technology to further improve diagnostic capabilities. To get a visual representation of a patients anatomy, different imaging modalities can be used to capture radiological images of soft tissue, bones, and vessels. Some of these modalities include: x-ray, medical resonance imaging (MRI), computer-assisted tomography (CT), and full-field digital mammography (FFDM). Biomedical images can give insights to different conditions and help healthcare professionals come to a more comprehensive diagnosis. For instance, a mammogram image of the breast can show abnormal growths or changes in breast tissue before they are large enough to be felt in a breast examination [1]. Image processing and feature extraction algorithms can be performed on medical images which can aid the physician with diagnostic information, helping the physician diagnose the condition with higher accuracy. These technologies are called computer aided diagnosis (CAD) systems and can help radiologists in interpreting mammogram images for classification between benign and malignant tumors [1].

In Canada, breast cancer is the third most common cancer, and is the first most common cancer among women occurring in 83% of women 50 plus in age [2]. Currently the most effective screening of breast cancer is mammography [1]. The Canadian task Force on Preventative Health Care recommend women aged 50 to 74 to get routine breast cancer screenings with mammography every 2 to 3 years [2]. Microcalsifications and masses are two important early signs of the disease therefore, radiologists screen each mammogram image looking for these features [3]. Although microcalsifications and masses are important pathologies to look at, masses are harder to detect since their features can be obscured or may be similar to normal breast parenchyma [3]. Therefore, radiologists face a difficult problem of the identification of masses within breast tissue. Some typical mass shapes can be seen in Figure 1. Benign masses are generally smooth and round in shape, whereas malignant tumors typically have rough contours with spiculations and concavities [4]. Also, benign masses are more homogeneous in texture whereas, malignant tumors exhibit heterogeneous texture [3].

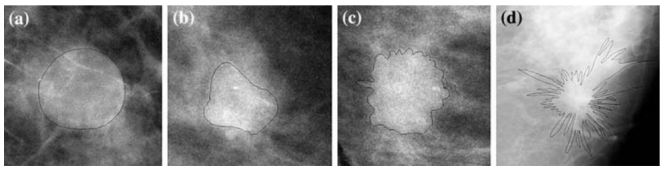

Fig. 1: Examples of benign and malignant breast masses. a) Well-circumscribed benign mass with round contour. b) Macrolobulated benign mass. c) Microlobulated malignant tumor. d) Spiculated malignant tumor [3].

Although the characteristics of benign and malignant masses are known, differentiation between the masses is still difficult since normal tissue can obscure them and other features such as breast parenchyma are much more prominent than the lesion itself [3]. To reduce the number of negative biopsy to positive biopsy which are about 5 : 1 and as high as 11 : 1 in some clinics [5], good features need to be extracted to classify the benign and malignant masses. A radiologist's workday involves a number of extensive tasks, and new streamline radiology workflows are needed to make radiologists more efficient, reduce costs, and improve consistency between diagnosis. To help with workflow consistency, precision, and efficiency, big data and computer aided diagnosis systems can be used. The ultimate goal is to create a computer aided diagnosis system that can use texture features extracted from the masses using feature extraction to correctly classify the masses. The features would then be used in a classifier to classify the images between benign or malignant tumors. Two mammograms can be seen in figure 2, the one on the left being a benign mass while the one on the right contains a malignant mass. As demonstrated, it is very hard to visually differentiate between the two images since objects are obscuring the lesion. Therefore, discriminating between benign and malignant tumors using automated feature extraction and a classification model could solve this problem.

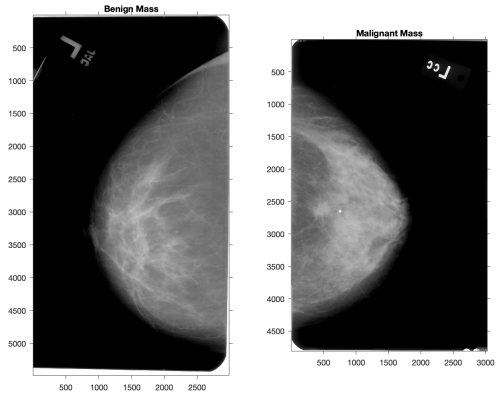

Fig. 2: Comparing benign (left) and malignant (right) mammograms.

In this paper, we investigate methods for automated feature extraction and classification of mammogram images. The best method will be chosen in the final system design. Our method is based on feature engineering and machine learning in the classification of breast masses. The proposed system enables learning about mass texture and classification. We evaluated the performance of the method in breast cancer classification via receiver operating characteristic (ROC) analysis.

II. Materials and Method

A. Data

The dataset used in this study is the CBIS-DDSM (Curated Breast Imaging Subset of DDSM), an updated version of the DDSM providing easily accessible data and improved ROI segmentation [4]. The DDSM is a collection of mammograms from the following sources: Massachusetts General Hospital, Wake Forest University School of Medicine, Sacred Heart Hospital, and Washington University of St Louis School of Medicine [4]. The data was split into a training and testing set based on BI-RADS category. The split was obtained using 20% of the cases for testing and 80% for training [4]. The mammogram images were taken from craniocaudal (CC) and/or mediolateral oblique (MLO) views, which are the standard views for screening mammography [4]. From the whole breast mammogram images, abnormalities were cropped by determining the bounding rectangle of the abnormality with respect to its ROI. These cropped ROI's were then used in our classification analysis. The images were stored as Digital Imaging and Communications in Medicine (DICOM) format, which is standard for medical images. DICOM images are entirely lossless and preserve all information from the original DDSM files. Also, csv files were attached with descriptors for mass shape, mass margin, calcification type, calcification distribution, and breast density; overall BI-RADS assessment from 0 to 5; rating of the subtlety of the abnormality from 1 to 5; and patient age [4]. Table I shows characteristics of the mammogram images used for experimentation.

TABLE I: Characteristics of mammogram images used for experimentation.

Type	**Grayscale**
Dimensions	Not consistent
Pixel Depth	12bpp
File Format	Raw (.dcm)
Number of Images	1696 (912 benign, 784 malignant)

B. Experimental Design

For this research, various feature extraction techniques are used to describe the texture of mammogram images. Using a simple 3-layer automatic neural network (ANN), these techniques are evaluated based on their corresponding classification accuracy. The best method for feature extraction will be taken as the final method in the system. First, the input images are pre-processed using histogram equalization (HE). Next, three different filtering techniques are used to highlight textural information. Haar, Biorthogonal discrete wavelet decomposition and Gabor filters are used. The filtered images are then used to calculate corresponding grey-level co-occurrence matrices (GLCMs) for each. Using GLCMs, texture features such as energy, entropy, contrast, homogeneity, and correlation are computed. Once the feature extraction is complete, there are three separate feature sets, one for each filtering technique used. Each feature set is then subject to three different feature pruning techniques; feature importance (FI) via extra trees, principal component analysis (PCA) and correlation. After feature pruning, each new set of features is fed into the neural network and classification scores are generated. See figure 3 for a flowchart representation of the proposed system.

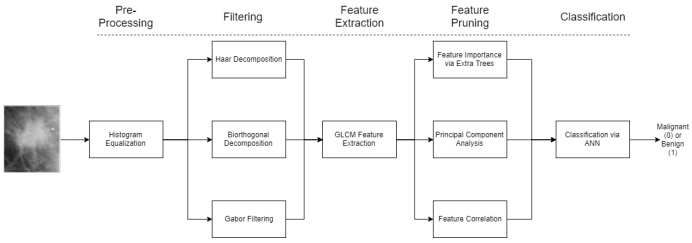

Fig. 3: Experimental design

C. Pre-Processing

1) Histogram Equalization (HE): Pre-processing is an important step prior to gathering information and characteristics from images. Mammogram images are often taken with low contrast, high background noises and artifacts, making diagnosis difficult [6]. The aim of this pre-processing step is to enhance the texture features of the breast masses. Histogram equalization is a pre-processing step that adjusts image intensities to enhance contrast in images. In histogram equalization, a nonlinear map is applied such that the pixels of the input image are mapped to an output image with a uniform histogram distribution. This enhancement in contrast can be seen in figure 4 [7].

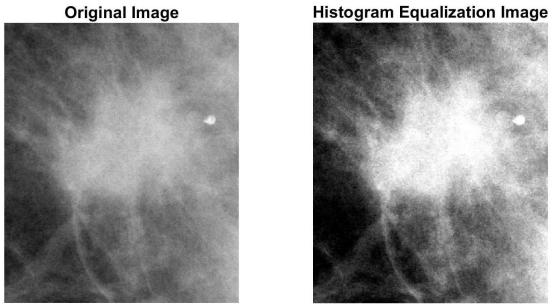

Fig. 4: Histogram equalization.

D. Filtering

1) Discrete Wavelet Transform (DWT): Multi-resolution techniques intend to transform images into a representation of both the spatial and frequency information. The use of wavelet transform was first proposed for texture analysis by Mallat (1989) [8]. This transform provides a robust methodology for texture analysis in different scales. The wavelet transform allows for the decomposition of a signal using a series of elemental functions called wavelets and scaling [8]. These are created by scaling and translating the mother wavelet:

$$\psi_{s,u}(x) = \frac{1}{\sqrt{s}}\psi\left(\frac{x-u}{s}\right) \tag{1}$$

where "s" governs the scaling and "u" the translation [8]. The mother wavelet is localized in both spatial and frequency domain and it has to satisfy the constraint of having zero mean[9].

The wavelet decomposition of an 2-D image can be obtained by performing the filtering consecutively along horizontal and vertical directions (separable filter bank). This is depicted schematically in figure 5 [9]. In discrete wavelet decomposition, the output detail images give the detail coefficients of the original image [10]. It is found that the wavelet detail coefficients provide the texture descriptors of the mammographic ROI [10]. In the proposed method, 2 different wavelet families are used: Haar and Biorthogonal (bior1.3). The Haar filter is one of the pioneer methods for wavelet decomposition. Biorthogonal is used because they are known to be

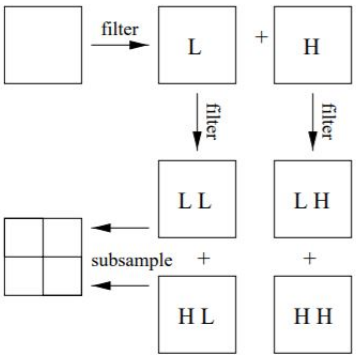

Fig. 5: 1 level of a wavelet decomposition in 3 steps: 1) Low and High pass filtering in horizontal direction, 2) The same in vertical direction, 3) Sub-sampling [9]

better descriptors of texture [11]. For each method, the ROI mammogram image is decomposed into its horizontal, vertical and diagonal components (figures 6 and 7).

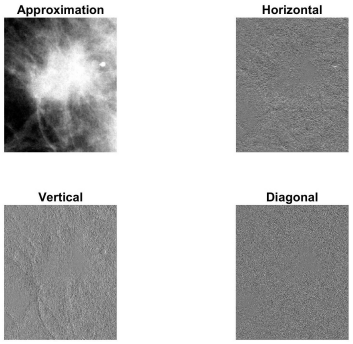

Fig. 6: Haar wavelet decomposition of ROI Image.

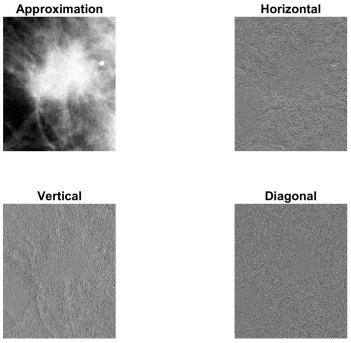

Fig. 7: Biorthogonal wavelet decomposition of ROI Image.

2) Gabor Filtering: Gabor filters have become a standard for texture analysis and have been used to analyze texture of mammogram images [12]. Masses in an ROI contain micro-patterns at different scales and orientations. These micro-patterns are helpful in the recognition of cancerous regions. Gabor filters can effectively be used to detect these micro-patterns [12]. Gabor filters are based on multi-channel filtering, which is intended to be a representation of the human visual system. The human visual system decomposes an image formed in the retina into several filtered images, each of them having variations in intensity within a limited range of frequencies and orientations [8]. A Gabor filter bank consists of Gaussian filters that cover the frequency domain with different radial frequencies and orientations [8]. In the spatial domain, a Gabor filter *h(x,y)* is a Gaussian function modeled by:

$$h(x,y) = \frac{1}{2\pi\sigma_g^2} \cdot \exp\left[-\frac{(x^2+y^2)}{2\sigma_g^2}\right] \cdot \exp(j2\pi F(x\cos\theta + y\sin\theta)) \tag{2}$$

where σ_g determines the spatial bandwidth of the filter. In frequency domain, the Gabor filter is defined by:

$$H(u,v) = \exp\left[-2\pi^2\sigma_g^2\left((u-F\cos\theta)^2 + (v-F\sin\theta)^2\right)\right] \tag{3}$$

The parameters that define each of the filers are:

1) The radial frequency (F) where the filter is centered in the frequency domain.
2) The standard deviation σ of the Gaussian curve.
3) The orientation θ

The filter bank used in this paper is based on the implementation from [8]. The filter bank was created with 6 orientations (0^o, 30^o, 60^o, 90^o, 120^o and 150^o) and 3 combinations of frequency: F=0.3536, F=0.1768, and F=0.0884, as seen in [8]. The difference in this implementation is that the bandwidth for each filter was kept constant, with σ= 1. This creates a total of 18 different filters (figures 8 and 9) at varying orientations and radial frequencies. The magnitude was computed and the synchronized average of the images was taken to combine the information from all filters, while reducing noise, as seen in figure 10.

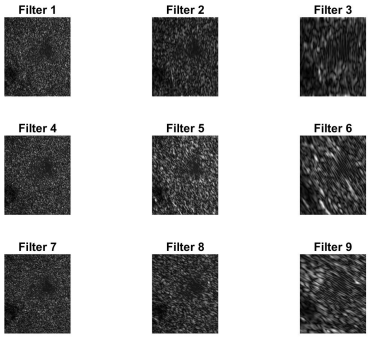

Fig. 8: Magnitude image of Gabor filters 1-9

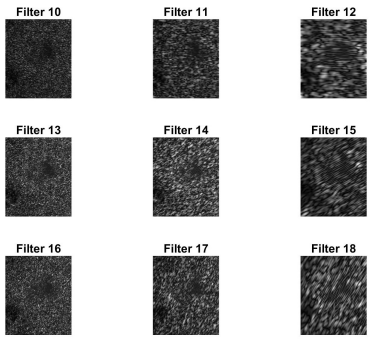

Fig. 9: Magnitude image of Gabor filters 10-18

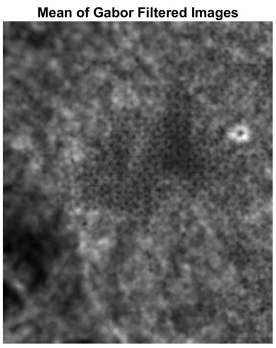

Fig. 10: Average image of Gabor filter magnitudes

E. Feature Extraction

1) Haralick Texture Descriptor: Texture in its simplest term describes the roughness of an object, and measures the characteristics of the intensity fluctuations between groups of neighbouring pixels, similar to the human visual system [7]. Julesz found that textons are defined as being the elementary units of texture, and various texture regions can be decomposed using textons [13]. Julesz also found that the frequency content, orientation and periodicity of these textons can provide important clues on how to differentiate between two or more textured areas [13]. Texture information in images gives spatial relationships between gray levels. Statistical analysis of the spatial distribution can be used to discriminate between different classes [7]. Since discriminatory information about specific regions or objects can be gathered from textural characteristics, such measure can be used as an essential tool in biomedical image analysis.

In texture analysis, Haralick et al. purposed Gray-Level Co-occurrence Matrices (GLCM) as a tool to extract textural features. GLCM is a statistical method of examining texture that considers the spatial relationship of pixels is the gray-level co-occurrence matrix, also referred to as the gray-level spatial dependence matrix [14]. GLCM characterizes the texture of an image by calculating how often pairs of pixel with specific values and in a specified spatial relationship occur in an image, creating a GLCM [14]. The normalized GLCM can be represented where *P(l1,l2)* is the number of occurrences of grey level *l1*, and *l2* at a distance d and angle θ.

$$p\left(l_1, l_2, d, \theta\right) = \frac{P\left(l_1, l_2\right)}{\sum_{l_1=0}^{L-1} \sum_{l_2=0}^{L-1} P\left(l_1, l_2\right)} \qquad (4)$$

The co-occurence matrix is then analyzed by extracting Haralick statistical measures which describe the texture. Some examples are contrast, energy, homogeneity, dissimilarity, correlation and many more. To fully extract texture characteristics relating to mammogram images, it is necessary to extract texture features which describe relative homogeneity or non-uniformity as these texture properties can be used to differentiate between normal and abnormal images.

When looking at the texture of benign and malignant breast masses there is a noticeable difference in texture. Benign breast masses have a rounded smooth appearance with a defined boundary between mass and background, while the inside of the mass is relatively uniform in texture. However, malignant masses have indistinct boundaries between mass and background, and have a higher density and an overall non-uniform appearance. Since benign and malignant masses have different textural characteristics, image analysis can be used to gather their textural characteristics so they can be utilized in a classifier.

After performing the discrete wavelet transform on the mammogram images, GLCM was computed on the output images for each method. The GLCM was computed at a four different directions which are named, horizontal, vertical, left diagonal, and right diagonal respectively, that is, $0°$, $90°$, $45°$, and $135°$ shown in figure 11. Four different GLCM's were outputted and the mean features were taken. The distance *d* taken for each direction was 1, with 8 grey levels. Lastly, the Haralick features used that best described homogeneity and non-uniformity were: contrast, correlation, energy, entropy, and homogeneity.

Fig. 11: Four different directions GLCM is calculated.

2) Additional Features: Descriptors for mass shape, mass margin, calcification type, calcification distribution, and breast density; overall BI-RADS assessment from 0 to 5 were gathered from the CBIS-DDSM dataset. From these features mass shape and breast density were choose to be added to the texture feature vector. Mass shape and breast density are important features as they describe the abnormality and breast tissue.

Mass shape is an important feature as they describe the shape of the lesion. Table II shows the different categories of mass shapes and their description. These mass shapes are determined when the mass is segmented from the ROI.

Breast density is a measure of the amount of glandular, fibrous connective , and fatty tissue in the breast [15]. Radiologist use the Breast Imaging Reporting and Data System, called BI-RADS, to group different types of breast density [15]. BI-RADS classifies breast density into four categories. The first category is almost entirely fatty breast tissue which can be found in about 10% of women causing the breast to look radiolucent on mammogram [15]. The second category is scattered areas of dense glandular tissue and fibrous connective tissue which can be found in about 40% of women [15]. The third category is heterogeneously dense breast tissue with many areas of glandular tissue and fibrous connective tissue, found in about 40% of women [15]. The last category is extremely dense breast tissue, found in about 10% of women causing the breast to look radiopaque on mammogram [15]. Figure 12 shows the different categories of breast density. Breast density is an important feature as in many cases normal breast tissue, due to high breast density can obscure breast masses making the mammogram hard to read. In many cases women with dense breasts may be called back for follow-up tests more often than women with fatty breasts [15].

TABLE II: Mammogram mass shape.

Mass Shape	**Description**
Irregular-Architectural Distortion	Looks like an irregular arrangement of tissue strands, but without any associated mass seen.
Architectural Distortion	Tissue strands, but without any associated mass seen.
Oval	Mass looks to be oval in shape.
Irregular	Mass looks to be irregular in shape.
Lymph Node	Lymph node is enlarged.
Lobulated-Lymph Node	Lobulated-lymph node is enlarged.
Lobulated	Mass is labulated.
Focal Asymmetric Density	Cannot identify if there is a mass.
Round	Mass is round.
Lobulated-Architectural Distortion	Lobulated tissue strands.
Asymmetric Breast Tissue	Asymmetric mass.
Lobulated-Irregular	Lobulated and irregular mass.
N/A	No mass shape.
Oval-Lymph Node	Oval lymph node enlarged.
Lobulated-Oval	Lobulated oval mass.
Round-Oval	Round oval mass.
Irregular-Focal Asymmetric Density	Irregular focal density but cannot identify if there is a mass.
Round-Irregular-Architectural Distortion	Round and irregular distortion mass.
Round-Lobulated	Round lobulated mass.

Fig. 12: Four categories of breast density. a) Breast almost entirely fatty breast tissue (radiolucent). b) Dense glandular tissue and fibrous connective tissue. c) Heterogeneously dense breast tissue. d) Extremely dense breast tissue [15].

F. Feature Selection

To select features, three different feature pruning techniques were used and the best performing set

was selected. Before this was done, feature normalization was performed using z-score. Z-score subtracts the data by its mean which center the data, and is then divided by its standard deviation which performs variance scaling.

1) Feature Importance via the Extra Trees Algorithm: The first method used to determine the most informative features is the extra trees algorithm. It is a tree-based ensemble method for supervised classification and regression problems [16]. It consists of randomizing strongly both attribute and cut-point choice while splitting a tree node. For extreme cases, it builds totally randomized trees whose structures are independent of the output values of the learning sample [16]. The rank of each feature used as a decision node in a tree is used to access the relative importance of that feature with respect to the the predictability of the target variable. The expected fraction of the samples they contribute to is used as an estimate of the relative importance of the features. In this implementation, the sklearn - ExtraTrees package was used. In this implementation of the algorithm, the fraction of the samples a feature contributes to is combined with the decrease in impurity from splitting them to create a normalized estimate of the predictive power of that feature [17]. Once the features are ranked, the top five features are extracted and used for classification.

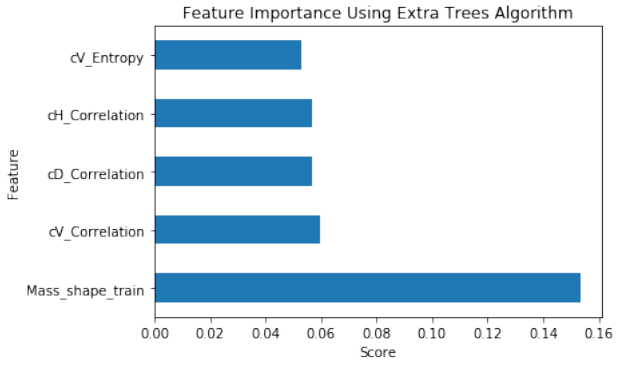

Fig. 13: Most important features using the Extra Trees algorithm

2) Principal Component Analysis (PCA): Principal component analysis (PCA) is a common technique in the field of computer science. It is used to decompose multivariate data sets into a set of successive orthogonal components that explain a maximum amount of variance [17]. It has been seen to aid in texture feature selection for ultrasound imaging of liver, to differentiate between different liver abnormalities [18]. In this paper, the same concepts of texture feature selection are applied to mammogram images. PCA is used as a feature pruning technique to reduce the complexity of the feature vector while maintaining 95 % variance in the data (figure 14). Reduction of feature complexity allows the model to also have a reduced complexity. Also, using PCA minimizes computational expense.

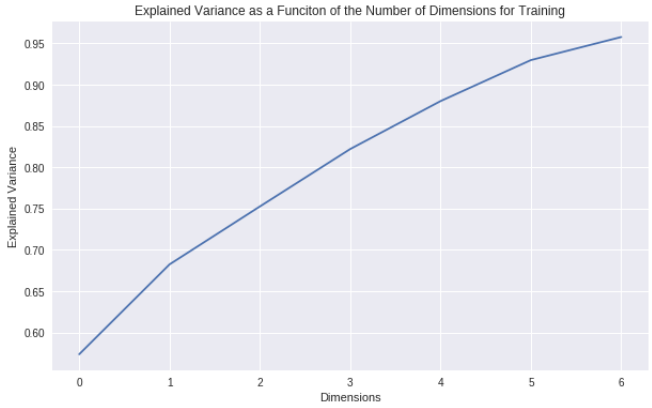

Fig. 14: Explained variance of PCA.

3) Pearson's Correlation (PCC): Pearson's correlation coefficient (PCC) is a measure of linear correlation between two variables X and Y developed by Karl Pearson. PCC looks at correlation between data and returns a value of 1 if the data is strongly correlated, -1 if the data is strongly negatively correlated, and 0 for no correlation [19]. These correlations can be seen in figure 15. When observing features in a dataset, some features might not be useful in building a machine learning model to make a necessary prediction. Therefore, feature selection plays a vital role in building a machine learning model. In this paper, PCC is performed on texture features and a correlation heatmap is produced which can be seen in figure 16. The heatmap is then analyzed and features with high correlation are more linearly dependent and hence have almost the same effect on the dependent variable. So, when two features have high correlation, one of the two features can be removed. Ideally, it is best to have features with a correlation closest to zero as this means that there is no correlation. By using PPC to reduce the data, this allows the model to decrease in complexity, decrease training time, and becomes less computational expensive.

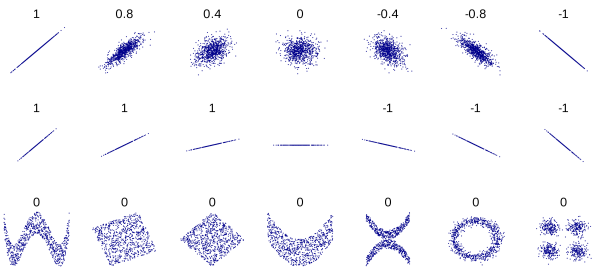

Fig. 15: Pearson's correlation coefficients of x and y. The correlation indicates the strength and direction of the linear relationship between x and y [19].

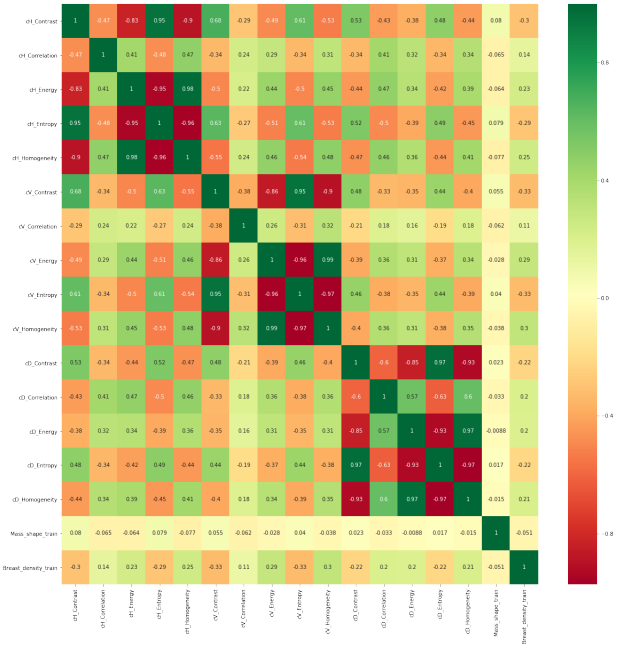

Fig. 16: Pearson's correlation heatmap of features.

G. Classification

Once the texture, shape, and breast density features are extracted and selected using feature pruning methods, the selected features are inputted into a classier to classify between benign and malignant masses. Classifiers such as artificial neural network (ANN) have performed well in mass classification [1]. The perceptron is one of the simplest artificial neural network (ANN) architectures invented in 1957 [20]. ANN's are a collection of mathematical models that behave in a similar way as biological neurons and function similar to adaptive biological learning [20] [1]. They are made of many interconnected processing elements that are weighted and act similar to synapses. A multi-layer perceptron (MLP) contains an input layer, an output layer and one or more hidden layers between them. The weights w(j,i) and w(k,j) dictate whether the inputs are either amplified or weakened to obtain the best computational solution [20]. These weights are determined by training the ANN using known samples and changing the weights through backpropagation [20]. Once the model is trained using the selected features as inputs, the model is ready to use unseen data, testing data to determine the strength of the model. Figure 17 shows a single layer neural network architecture. Highly complex computations can be performed by a network of fairly simple neurons [1]. The architecture of biological neural networks is still being highly researched [1]. ANN's use non-linear mapping functions as decision boundaries. The advantage of ANN's over other traditional machine learning methods is their capability of self learning using back propagation [20].

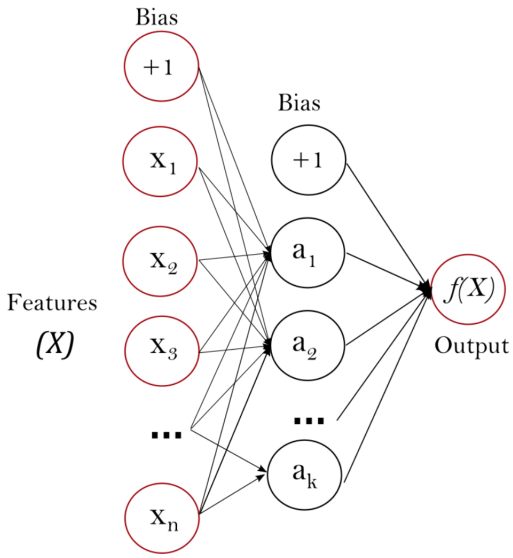

Fig. 17: One hidden layer neural network.

1) Parameter Selection: In Scikit-learn, a multilayer perceptron can be set up very easily and different parameters can be changed. The activation function is a parameter that can be changed and its purpose is to introduce non-linearity into the output of a neuron [20]. The activation function models the firing of action potentials from biological neurons [20]. The activation function can be set to either identity, logistic, tanh, or relu [21]. The alpha terms is a regularization term which helps in avoiding over-fitting by penalizing weights with large magnitudes [20]. The higher the alpha value the higher the regularization. The hidden layer size

determines the number of hidden layers and the number of hidden units in each layer. In our model we decided to use only one hidden layer to see the robustness of our features. The max iterations is the max number of times the solver iterates until it stops convergence [20]. The solver parameter is used for weight optimization [20]. Lastly, the learning rate controls how large of a step to change the model in response to the estimated error each time the model weights are updated [20]. If the learning rate is too small then it may take a long time for the model to converge and it may get stuck at a local minima. However, if the learning rate is too large then it may jump over the global minima we are trying to reach. Therefore it is important to set a proper learning rate which is usually set to 0.001 as default [21]. To select the parameters of the model, a grid search algorithm is used to select the parameters that yield the best cross validation score. The algorithm exhaustively generates candidates from a grid of parameter values [17]. The parameter values specified involved changing each parameter independently of the other (change one parameter at a time). For each candidate, 5-fold cross validation is performed. The best performing candidate is taken and the parameters saved for future training and testing. Table III shows the optimized parameter's set in our models, for each different feature set. For the number of hidden units, this is adaptively set to the number of features in the feature vector. Depending on the filtering-feature pruning technique used, the number of features changes. Thus a constant value for the number of hidden units would not be as robust. Moreover, the results were consistently better when setting the number of hidden units equal to the length of the feature vector, for all feature methods tested.

TABLE III: Parameter's set for MLP.

Method	**Activation**	**Alpha**	**Hidden Layer Size**	**Max Iteration**	**Solver**
FI	Logistic	0.001	5	1000	lbfgs
PCA	Logistic	0.001	8	1000	lbfgs
Correlation	Logistic	0.001	8	1000	lbfgs

2) Evaluation: To evaluate the model, 5 fold cross validation is used to compare both training and testing accuracy (figure 18). Moreover, specificity, sensitivity, precision and f1-score is also used to compare results on the test set. Lastly, learning curves and ROC curves were used to visualize performance.

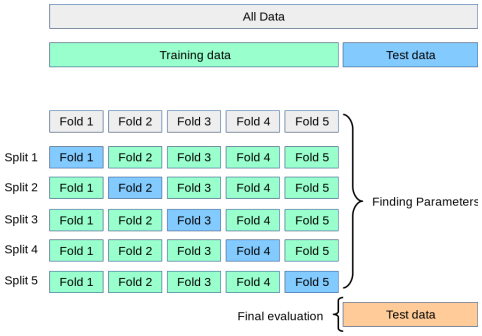

Fig. 18: 5-fold cross-validation visualization [17]

III. Results

Based on the evaluation metrics, it is seen that the combination of Gabor filtering with feature importance via extra trees performs the best. The accuracy 74.1% for training and 74.9% for testing, using 5-fold cross validation. The model also had an AUC of 0.755. It can be said that the model has good generalization as the results are repeatable through cross validation. Also, the training and testing scores are similar, indicating that the model is not overfitting the training data. Looking into the f1-score, it is seen that the model is able to classify benign cases better than malignant. One reason for this may be due to class imbalance, since there are more benign cases in the data set than malignant. Another may be due to the nature of the disease itself, since malignant tumors have more irregularities than benign. It is possible the texture features could not capture the patterns of malignant cases, nor was the model deep enough to learn patterns from them.

TABLE IV: Comparing different methods looking at training and testing accuracy's.

Texture Method	**Feature Pruning**	**Train Score**	**Test Score**
Haar Wavelet + GLCM	FI	0.726	0.698
	PCA	0.754	0.637
	Correlation	0.645	0.613
Biorthogonal + GLCM	FI	0.734	0.728
	PCA	0.737	0.625
	Correlation	0.646	0.606
Gabor Filter + GLCM	FI	**0.741**	**0.749**
	PCA	0.726	0.689
	Correlation	0.690	0.647

TABLE V: .Classification report showing the main classification metrics.

	Precision	Recall	f1-score	Support
Malignant	0.69	0.67	0.68	147
Benign	0.79	0.81	0.80	226
Accuracy			0.75	373
Macro Accuracy	0.74	0.74	0.74	373
Weighted Average	0.75	0.75	0.75	373

TABLE VI: Confusion matrix containing the number of correctly classified mammogram masses as either being benign or malignant.

	Malignant	Benign
Malignant	0.67	0.33
Benign	0.19	0.81

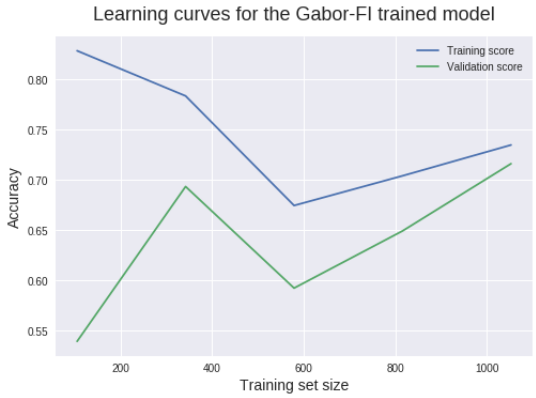

Fig. 19: Learning curve for Gabor-FI trained model.

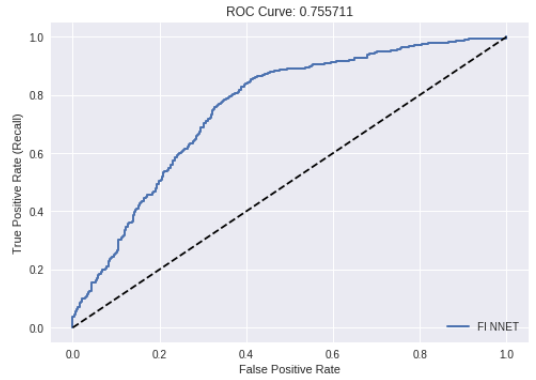

Fig. 20: ROC curve for the Gabor-FI trained model.

IV. Discussion and Conclusion

In this work, a complete workflow for mammogram benign vs malignant mass classification was designed and developed. The features extracted from mammogram ROI images included breast density, mass shape, and Haralick texture features obtained from Haar, Biorthogonal discrete wavelet decomposition, and Gabor filters. In total we collected 17 features for Haar and Biorthogonal discrete wavelet decomposition individually, and 5 features for Gabor filters. Both mass structure features and texture features were then reduced using three feature pruning methods (FI, PCA, correlation) which were then compared against each other for highest accuracy. A three-layer automatic neural network (ANN) was used for the classification of the features. To select the parameters of the model, a grid search algorithm was used to select the parameters that yield the best cross validation score. The parameter values specified involved changing each parameter independently of the other (change one parameter at a time). For each candidate, 5-fold cross validation is performed. The best performing candidate is taken and the parameters saved for future training and testing. The final model was trained and final results were computed on an independent unseen test set. It was observed that out of all the models, Gabor-filter-GLCM-FI method had the highest accuracy of 74.9% and AUC of 0.755 and therefore was chosen as the model to use for the classification of benign and malignant masses. Good precision and recall were obtained too. The reported results suggest that the purposed workflow set-up is highly reliable and robust for the classification of benign and malignant masses. However, there is still a high miss-classification of malignant masses classified as benign masses. One reason for this may be due to class imbalance, since there are more benign cases in the data set than malignant. Another may be due to the nature of the disease itself, since malignant tumors have more irregularities than benign. It is possible the texture features could not capture all the unique patterns of malignant cases, nor was the model deep enough to learn patterns from them.

In the future work, we will investigate further into Gabor filters and how to adpatively select filter parameters such as frequency, bandwidth and orientation. We will also further investigate into other classification methods such as Convolutional Neural Networks (CNN's) which are a deep learning algorithms. This Deep Learning approach will be evaluated and the results will be compared with the results gathered in this paper. Finally, the designed algorithm will be integrated into a complete CAD tool for radiologists to use.

REFERENCES

[1] H.-D. Cheng, X. Shi, R. Min, L. Hu, X. Cai, and H. Du, "Approaches for automated detection and classification of masses in mammograms," *Pattern recognition*, vol. 39, no. 4, pp. 646–668, 2006.

[2] P. H. A. of Canada, "Government of canada," Dec 2019. [Online]. Available: https://www.canada.ca/en/public-health/services/chronic-diseases/cancer/breast-cancer.html

[3] R. Nandi, A. K. Nandi, R. M. Rangayyan, and D. Scutt, "Classification of breast masses in mammograms using genetic programming and feature selection," *Medical and biological engineering and computing*, vol. 44, no. 8, pp. 683–694, 2006.

[4] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin, "A curated mammography data set for use in computer-aided detection and diagnosis research," *Scientific data*, vol. 4, p. 170177, 2017.

[5] R. M. Rangayyan, N. M. El-Faramawy, J. L. Desautels, and O. A. Alim, "Measures of acutance and shape for classification of breast tumors," *IEEE Transactions on medical imaging*, vol. 16, no. 6, pp. 799–810, 1997.

[6] M. Hazarika and L. B. Mahanta, "A new breast border extraction and contrast enhancement technique with digital mammogram images for improved detection of breast cancer," *Asian Pacific journal of cancer prevention: APJCP*, vol. 19, no. 8, p. 2141, 2018.

[7] C. Solomon and T. Breckon, *Fundamentals of Digital Image Processing: A practical approach with examples in Matlab*. John Wiley & Sons, 2011.

[8] L. Ruiz, A. Fdez-Sarría, and J. Recio, "Texture feature extraction for classification of remote sensing data using wavelet decomposition: A comparative study," in *20th ISPRS Congress*, vol. 35, no. part B, 2004, pp. 1109–1114.

[9] S. Livens, P. Scheunders, G. Van de Wouwer, and D. Van Dyck, "Wavelets for texture analysis, an overview," 1997.

[10] S. Beura, B. Majhi, and R. Dash, "Mammogram classification using two dimensional discrete wavelet transform and gray-level co-occurrence matrix for detection of breast cancer," *Neurocomputing*, vol. 154, pp. 1–14, 2015.

[11] T. Greiner, "Orthogonal and biorthogonal texture-matched wavelet filterbanks for hierarchical texture analysis," *Signal Processing*, vol. 54, no. 1, pp. 1–22, 1996.

[12] S. Khan, M. Hussain, H. Aboalsamh, H. Mathkour, G. Bebis, and M. Zakariah, "Optimized gabor features for mass classification in mammography," *Applied Soft Computing*, vol. 44, pp. 267–280, 2016.

[13] B. Julesz, "Textons, the elements of texture perception, and their interactions," *Nature*, vol. 290, no. 5802, p. 91, 1981.

[14] R. M. Haralick and L. G. Shapiro, *Computer and robot vision*. Addison-wesley Reading, 1992, vol. 1.

[15] "Dense breasts: Answers to commonly asked questions." [Online]. Available: https://www.cancer.gov/types/breast/breast-changes/dense-breasts

[16] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, no. 1, pp. 3–42, 2006.

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[18] D. Balasubramanian, P. Srinivasan, and R. Gurupatham, "Automatic classification of focal lesions in ultrasound liver images using principal component analysis and neural networks," in *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2007, pp. 2134–2137.

[19] K. Yeager, "Libguides: Spss tutorials: Pearson correlation, en."

[20] A. Geron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, 2019.

[21] "sklearn.neural$_n etwork.mlpclassifier-.''[Online].Available : https : //scikit-learn.org/stable/modules/generated/sklearn.neura

ACKNOWLEDGMENT

The authors would like to thank Dr. Bruce for teaching us all this amazing stuff! We would also like to thank CBIS-DDSM for letting us use this data to conduct research on the classification of benign and malignant masses.