Mammogram Mass Classification Utilizing Machine Learning

Alvionna Shiergetya Sunaryo
Computer Science and Mathematics Major
Mount Holyoke College
South Hadley, MA, 01075
sunar22a@mtholyoke.edu

Claiborne Ngodinh
Computer Science Major
Mount Holyoke College
South Hadley, MA, 01075
ngodi22c@mtholyoke.edu

Drew O'Brien
Biochemistry and Computer Science Major
Mount Holyoke College
South Hadley, MA, 01075
obrie23e@mtholyoke.edu

Wenyu Zhang
Biochemistry and Data Science Major
Mount Holyoke College
South Hadley, MA, 01075
zhang26w@mtholyoke.edu

Abstract

Early detection of breast cancer is critical for treatment and ultimately, breast cancer survivability. The application of machine learning techniques to the classification of mammogram masses as malignant or benign would facilitate the early detection of breast cancer and ultimately, save lives. This paper presents four different machine learning models for application to this classification problem: a logistic regression model for binary classification, a convolutional neural network model for binary classification, simple artificial neural networks for use in both binary and multiclass classification and a one versus all multiclass classification model. With these models the authors proposed solutions to both the problem of classifying a mammographic mass as benign or malignant and a multiclass classification exploration of mammographic mass multiclass classification. The authors demonstrated the potential use of machine learning techniques to classify mammogram masses into benign, malignant, incomplete mammogram results and suspicious, a category in which the mass could either be benign or malignant and requires a biopsy for diagnosis.

1 Introduction

Breast cancer is a disease where abnormal breast cells grow uncontrollably. Cancer cells can easily spread, or metastasize, to other parts of the body through blood vessels and lymph vessels. In 2019 an estimated 41,760 people died from breast cancer (American Cancer Society, 2019a). Early detection of breast cancer is crucial as breast cancer is the most treatable in the early stages. Regular screening, typically in the form of a mammogram, is essential for early detection. The Center for Disease Control recommends that women who have an average risk of breast cancer and are 50 to 75 years years old get a mammogram every two years (Center for Disease Control, 2020).

Radiologist examine mammograms to identify abnormalities and changes over time (American Cancer Society, 2019b). Should the examining physician find a mass, which is an area of breast tissue distinct from the surrounding tissue due to its density, shape and edges, said mass could either be benign or malignant (cancerous). It is the role of the radiologist to determine whether the mass is benign or requires further testing to conclusively diagnose breast cancer. There have been numerous

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

studies building machine learning models to detect and predict breast cancer, using diagnosis datasets to train the models. The authors' first goal is to create a machine learning model to classify masses as benign or malignant. Such a diagnostic tool would enable early detection, improving a patient's prognosis significantly.

While early detection is critical for effective treatment, and as such any model should err on the side of caution, the psychological affects of a false positive can be enormous (Center for Disease Control, 2020). As such the authors also aimed to implement mutlticlass classification machine learning models. These models divided up the mammographic masses into four categories: incomplete mammogram, benign, suspicious mass that should be followed up on, and malignant. These additional categories are designed to encourage any further testing necessary without causing undue panic. The first, an incomplete mammogram classification, aims to cover cases where the mammogram needs to be redone due to incomplete or unclear imaging. The other added category, suspicious mass with recommended follow up testing aims to convey that while the mass could be benign a biopsy is required to truly determine the malignancy. Furthermore, such a category would facilitate the identification of particular results for which a closer look by a radiologist would be imperative. An efficacious multiclass classification model would promote early detection of cancerous tumors while alleviating the psychological toll of a false positive result.

2 Related Work

Some examples of previous research models include a binary logistic regression classifier model that used both the standard sigmoid function and a weighted sigmoid function. The performance metrics of the weighted function showed an improvement from the standard function, with accuracies of 96.83 percent and 95.42 percent respectively (Khairunnahar, et al., 2019). Another paper compared 3 machine learning models: Naive Bayesian classification, Radial basis function (RBF) network, and J48 decision trees. Though the highest accuracy, 97.36 percent, resulted from the Naive Bayesian classification, the RBF model's accuracy, 96.77 percent, is important to note since the authors are implementing a neural network and want to compare accuracies between the RBF model and their own proposed models (Vikas, et al., 2018).

One study done in 2017, analyzed the performance rates of a model used primarily for linear problems, Support Vector Machine (SVM), and a model used for non-linear problems, K-Nearest Neighbors (KNN) (Islam, et al., 2017). Interestingly, in both the training and testing phase, the SVM model performed better in almost every metric like accuracy and specificity. The SVM model had a training accuracy of 99.68 percent, and a testing accuracy of 98.57 percent. On the other hand, the KNN model had a training accuracy of 98.35 percent and a testing accuracy of 97.14 percent.

Some researchers have utilized other SVM models in the pursuit of prediction with varying results (Kourou, et al. 2015). For cancer recurrence prediction, one SVM model used a hold-out validation method, which is a method in which the dataset is divided into training and testing sets, typically with a 80-20 ratio. That SVM model had a 89 percent accuracy. The other SVM model used a 10-fold cross validation method and had a 95 percent accuracy. Though that is surprising, it is worthy to note that the first dataset focused on pathogenic, epidemiological data whereas the second dataset focused on population data. The authors collecting data also note an SVM model with a Leave-one-out cross validation method for cancer survivability prediction. This model was run on genomic data and had an accuracy of 97 percent.

In yet another study, not only were different machine learning models like SVM, KNN classifiers, and decision trees compared against each other, but different variations of those models were also analyzed to determine which model and which type of that model was the most accurate and fastest to construct (Obaid, et al., 2018). Out of the kernel functions: linear, quadratic, and cubic, the most accurate and fastest was the quadratic kernel SVM, with an accuracy of 98.1 percent and a time of 3 seconds. Out of the three KNN types: fine, medium, and coarse, the most accurate was the medium KNN classifier with an accuracy of 96.7 percent. All KNN types had a time of 2 seconds. For the decision tree types: complex, medium, and simple, the medium and complex decision trees had the same accuracy of 93.7 percent. However, the medium decision tree was faster than the complex tree, with a time of 3 seconds. Therefore, the best decision tree model was the medium decision tree. Overall, the best model for accurate predictions was the quadratic kernel SVM model.

For the majority of the research referenced above, the dataset used was the Wisconsin Breast Cancer Dataset, which contains around 699 samples, all labeled into either malignant or benign and has 11 different known features. However, some breast cancer prediction research trained models on different datasets such as the breast cancer dataset obtained from M.G Cancer Hospital and Research Institute in Visakhapatnam, India (Vaka, et al., 2020), Digital Database for Screening Mammography (DDSM) and other image datasets (Houssein, et al., 2020), as well as Mammographic Image Analysis Society (MIAS)/mini-MIAS (Rouhi, et al., 2015).

The authors that used the breast cancer dataset obtained from M.G Cancer Hospital and Research Institute compared already known and used machine learning models such as SVM and Naive Bayesian classification with a new proposed model, the Deep Neural Network with Support Value (DNNS). The DNNS works through 3 various steps of filtering out noise from an image, extracting specific features, and then isolating the breast tumor from the images. This is all done through Histo-sigmoid based fuzzy clustering, which adds the histogram data and sigmoid function to a fuzzy clustering algorithm. The DNNS model actually had the best performance out of all models, having an accuracy of of 97.21 percent (Vaka, et al., 2020).

One journal article made a comprehensive review of all the research done on breast cancer image datasets including Digital Database for Screening Mammography (DDSM) and others into one cohesive article. Though there are many in-depth explanations about results and methodology, the most relevant results are the Artificial Neural Network models done on mammogram images. The accuracies for the listed models and research range from 90.94 percent to 96 percent (Houssein, et al., 2020).

Finally, the authors who used the Mammographic Image Analysis Society (MIAS)/mini-MIAS dataset proposed two different methods to predict and diagnose breast cancer tissues. The first is an Artificial Neural Network (ANN) model and the second is a Cellular Neural Network (CNN). The authors then compared their proposed models to other models that had already been created and analyzed how the proposed model's performances matched up with existing models. Though not the best performing model out of every model analyzed, their models performed relatively well, staying in the range from 81.58 to 96.47 percent (Rouhi et. al, 2015).

3 Dataset

The mammography dataset curated by Lee, et al. (2017) was used throughout this study. The dataset draws from the Digital Database for Screening Mammography (DDSM), a popular database for digital mammography research (Heath et al. 2001; Heath et al. 1998). The Curated Breast Imaging Subset of DDSM (CBIS-DDSM) contains 891 mass cases selected by a trained mammographer (Lee, et al. 2017). Each case consists of an identified mass seen on the craniocaudal (CC) and/or mediolateral oblique (MLO) views. The CC and MLO views are the standard views used in screening mammography. In the curated dataset each view is associated with a full image, a binary mask image defining the region of interest (ROI) and a version cropped around the region of interest. The region of interest is the area of the mammogram that contains the abnormality, as identified by Lee et al. using a segmentation algorithm. The images were made available in the form of Digital Imaging and Communications in Medicine (DICOM) files, which contain both the image and associated metadata, as is standard in the field of medicine.

The dataset includes the physician inputted metadata for each abnormality in an associated CSV file. The CSV file contains different rows for each view (CC or MLO) of an identified abnormality. The associated metadata for each view of an abnormality is as follows: Patient ID (the first 8 characters of images in the case file), breast density category, view (CC or MLO), associated abnormality number (1 except in cases containing multiple abnormalities), mass shape, mass margin, BI-RADS assessment, pathology, subtlety rating and the paths to each of the associated images files (full, ROI mask and cropped). The patient ID could be used to identify both the associated CC view and MLO view for a single patient. The mass shape for each abnormality view was characterized as falling into one or more of eight categories: irregular, oval, architectural distortion, lymph node, lobulated, focal asymmetric density, round and asymmetric breast tissue. Similarly, the mass margins were identified by one or more of five categories: spiculated, ill defined, circumscribed, obscured and microlobulated. The subtlety rating refers to the difficulty in viewing the image, on a scale of one to five, as reported by the radiologist who uploaded the case.

The Breast Imaging and Data System (BI-RADS) assessment is used by radiologists to classify mammogram, and follow-up test, findings into categories numbered zero to six (American Cancer Society). Category zero indicates that the screening mammogram produced incomplete information, it may be difficult to read or interpret. Typically this means that a different or magnified mammogram view or ultrasound may be needed to determine if any abnormalities are present. A BI-RADS score of zero could also indicate that the radiologist needs to compare the image to previous screenings to identify any differences. A BI-RADS score of one to three indicates normal or (probably) benign findings with at most, and only for a score of three, more frequent imagining recommended. A BI-RADS score of four indicates that a suspicious abnormality was found and biopsy is recommended, but the biopsy may still reveal that the abnormality is benign. Category five and six indicate that there is at least a 95% chance that the abnormality is cancerous.

The creators of the CBIS-DDSM dataset provide a standardized training/test split (Lee et al. 2017). 20% of the mass cases were separated into the test set and the reset into the training set. The split was performed so that the difficulty level and ratio of benign to malignant cases of the training and test sets were equal.

4 Methods

4.1 Feature Engineering and Extraction

The training and test sets provided in the CBIS-DDSM dataset were kept the same except for exclusion of cases missing either the MLO or CC view and cases where the dimensions of the full mammogram image did not match those of the provided binary mask. This resulted in a training set size of 455 cases and 139 test cases. While the sample size was less than ideal it was in fact larger than that used by most of the publications in this field that were reviewed by the authors (a bunch of citations). This is in fact, a well known limitation for studies into machine learning applications across medicine (Shaikhina et al., 2017). Due to the researchers' lack of expertise in radiology and oncology, as well as time and monetary limitations, the authors opted not to attempt to integrate multiple sources of data.

There were ultimately two forms of feature engineering and extraction performed. The first combined the physician inputted metadata with six numerical features extracted from each of the masked mammogram views. These features were used in the binary logistic classification, one-vs-all multi-class classification and both (binary and multi-class classifying) simple neural networks. The pydicom library was used to work with the DICOM files. For each mammogram view (CC and MLO), the binary mask was used to select the pixels in the region of interest from the pixel array of the full image (Figure 1). From this masked array of pixels, the height, width, area, as well as the maximum, minimum and average intensity of the region of interest were determined. Due to the large size of the original DICOM files, the associated metadata and the features extracted from the images were combined and exported in a CSV file, providing a much easier and faster way to load the newly engineered features into each model. The data extracted from the CSV file for both the training and test set were normalized before use in the models.

The BI-RADS scores were used to sort the data samples into four categories for use in the multiclass classification models. Patients with a BI-RADS score of zero were sorted into the first category, consisting of inconclusive or incomplete mammogram results which would necessitate either redoing the mammogram or comparing the results to the patient's previous ones. The second category, comprised of benign masses, was created from patients with a BI-RADS score of one to three. Patients with a BI-RADS score of four were placed in the third category, suspicious masses that may or may not be malignant but for which follow up in the form of a biopsy is recommended. Finally, the fourth category, malignant masses, covered all cases with a BI-RADS score of five or six.

The second form of feature engineering simply involved the extraction of the pixel array from the DICOM files containing the cropped versions of the CC and MLO views. To ensure standardization the images were resized and cropped as necessary to ensure that each image had the same dimensions. This was possible to implement without obfuscating the features of the images due to only minimal modifications being necessary. These extracted pixel arrays were used in the convolutional neural network model.

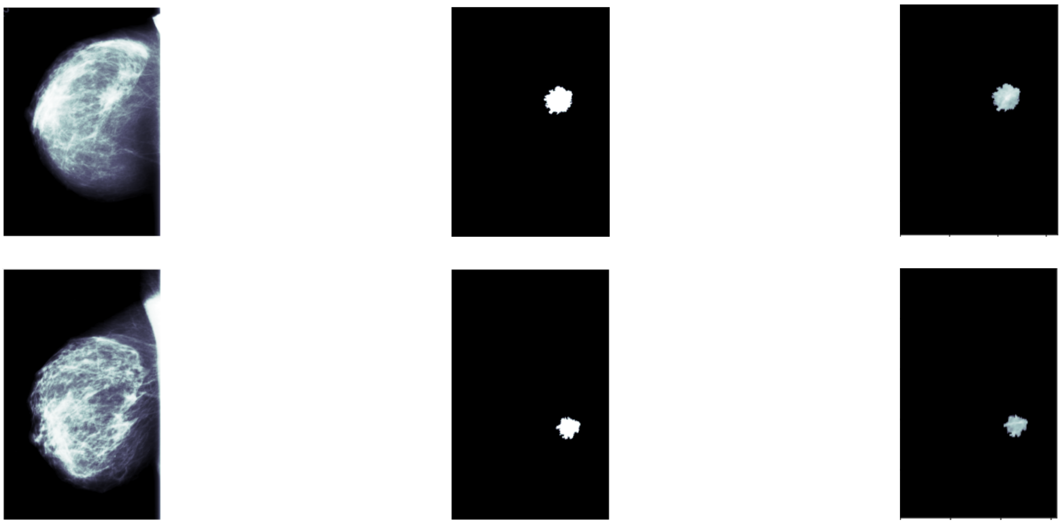

Figure 1: Region of interest masking of the CC and MLO views of one case: (upper left) full CC image; (upper middle) binary mask delineating the CC ROI; (upper right) masked CC image (lower left) full MLO image; (lower middle) binary mask delineating the MLO ROI; (lower right) masked MLO image.

4.2 Logistic Regression

Mammogram masses can be classified into two categories: benign and malignant. Due to the classifying nature of this problem, Logistic regression is an appropriate model to accomplish this binary classification task. The logistic regression model was implemented using the sklearn library, specifically the LogisticRegression method. The authors used L2 regularization to combat overfitting, particularly given the small sample size.

4.3 One vs. All Multi-Class Classification

As stated previously, binary classification may lead to unnecessary stress and obscure the convoluted nature of the problem of mammographic mass classification. With this in mind, it was necessary to explore multiclass classification models. With this in mind, the authors used the four categories created during the feature engineering phase to train and test a one-versus-all mutliclass classification model.

The authors implemented one-versus-all classification from scratch. This implementation began with the construction of a method which was used to create a binary classifier for each of the four categories. The classifiers were trained using logistic regression, resulting in a model capable of classifying data samples into one of the four categories.

4.4 Artificial Neural Network Model

Preliminary research pointed to the suitability of an Artificial Neural Network (ANN) modeling approach for the classification of masses observed on mammograms (Kourou et al. 2015). The complicated nature of the data and nature of the classification problem persuaded the authors to implement and analyze a neural network model. The neural network structure enabled its use in both the binary and multiclass classification problems. Furthermore, ANN model is a more flexibility of an artificial neural network, as opposed to that of logistic regression and one-vs-all multi-class classification, is particularly suited to the complex features and parameters inherent to the problem of mammogram mass classification.

Consequently, ANN models were implemented for both the binary and multi-class classification problems. The tensorflow and keras libraries were used to implement and optimize the ANN models. The Scikit-learn library's randomized search algorithm was adapted for hyper-parameter optimization for both ANN models. The use of 10-fold cross-validation in the randomized search implementation was particularly useful for tuning the model due to the low sample size. A variety of optimizers, activation functions, epochs, batch sizes, numbers of hidden layers and numbers of neurons per hidden layer were tested. With the use of randomized search, the list of potential working hyperparamers was successfully narrowed down.

Ultimately, the authors discovered that different parameters and hyper-parameters must be enacted for binary and multi-class classification problems to attain the optimized result. For binary classification problem, the Nadam optimizer, the softsign activation function for the hidden layers, and the softplus activation function for the output layer, were selected for use in the final model for binary classification. Additionally, three dense hidden layers were used with 24 neurons in each layer. The model was trained in 120 epochs with a batch size of 24 with a learning rate of 0.001. For the multi-class classification model, the Adamax optimizer, tanh and sigmoid activation functions performed the best in the randomized search. This in combination with four hidden layers with 24 nodes each was used to train the model in 6005 epochs with a batch size of 22 with a learning rate of 0.0001.

4.5 Convolutional Neural Network

With the goal of improving the accuracy of the classification models, the authors conducted a preliminary investigation into the use of convolutional neural networks (CNNs). The potential use of this model was suggested by the imaging nature of the problem. While feature engineering was used to obtain numerical features from the mammogram images for the other models, a CNN could potentially extract more.

The authors implemented CNNs with the tensorflow and keras libraries for the MLO and CC views separately. The inputs of these models were the product of the second form of feature engineering, images cropped around the region of interest and transformed to a standard size. Due to the high information storage of DICOM files, these images remained quite large, with a width and height of over 350 pixels. With the goal of maximizing sample size, the samples u-excluded from the training and test sets of the other models due to differing full image and mask sizes, were included for the CNNs. This resulted in 486 training cases and 146 test cases. With the goal of combining the two CNNs in future research in mind, the cases missing one of the views were still excluded.

Due to the long runtime for training the CNN and time-limitations for this study the authors were unable to optimize the hyperparamers as was done for the previous neural network models. As such, the structure of the CNN models, aside from minor changes (less than five pixels each) to the width and height of the input were identical. Both consisted of three convolutional layers, two pooling layers and after flattening the data, one densely connected hidden layer.

4.6 Confusion Matrix

With the goal of facilitating the evaluation of the various models, the authors implemented confusion matrices for each of the implemented models except for the convolutional neural networks. The sklearn and mlxtend libraries were used to create confusion matrices for both the binary and multiclass classification models. This resulted in an easy to understand display of the true positive, false positive, true negative and false negative rates for those models.

5 Results and Evaluation

The authors focused on two methods of evaluating the performance of the implemented machine learning models, the confusion matrix and the accuracy. These metrics were determined to be particularly critical for evaluating the performance of mammogram mass classification for a few reasons. The use of a confusion matrix visualization for each model allowed the authors to single out the rate of false positives and false negatives for each model. For the problem of mammogram mass classification we define a false positive as a case where the mass is either classified as malignant or as requiring further testing when it, is in fact, benign. On the other hand, a false negative case would be one in which a mass is classified as benign but is in fact malignant (or requires further testing). While a false positive would result in unnecessary medical bills and psychological stress, the case of a false negative is far more dire. The false classification of a mass as benign would lead to delayed treatment and potentially an avoidable death.

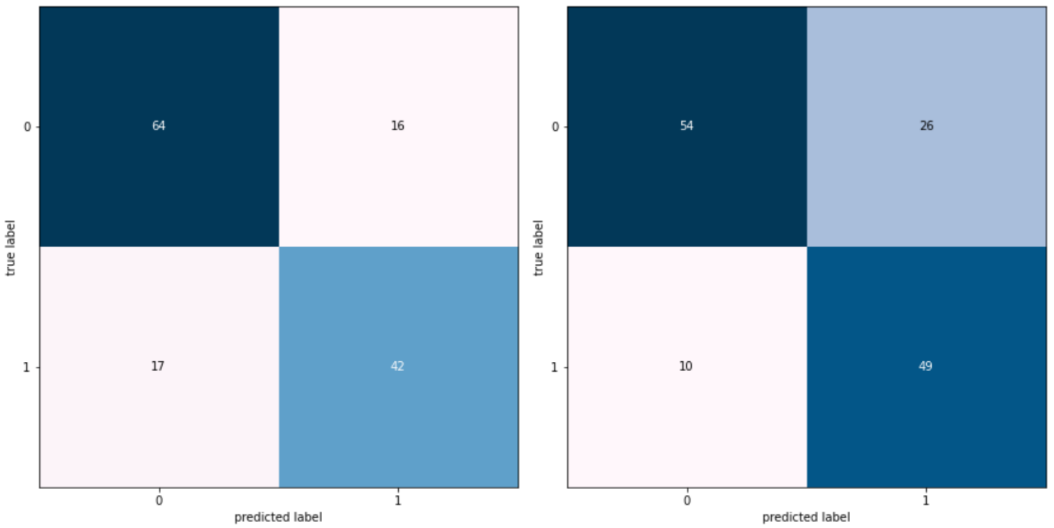

Figure 2: Confusion matrices for the (left) artificial neural network binary classification model and (right) logistic regression, binary classification, model.

The artificial neural network binary classification model resulted in 16 false positive cases and 17 false negative cases (Figure 2). In total, 106 of the cases were classified correctly while 33 were classified incorrectly. The logistic regression model had a slightly lower number of false negatives, 10. Overall, due to the higher number of false positives, 26, the logistic regression model resulted in slightly fewer correctly classified cases. Given the particular importance of false negatives however, these results indicate that logistic regression performed better then this implementation of an artificial neural network binary classification model. However, due to the small testing dataset size, 139 cases, such small differences may not hold much significance.

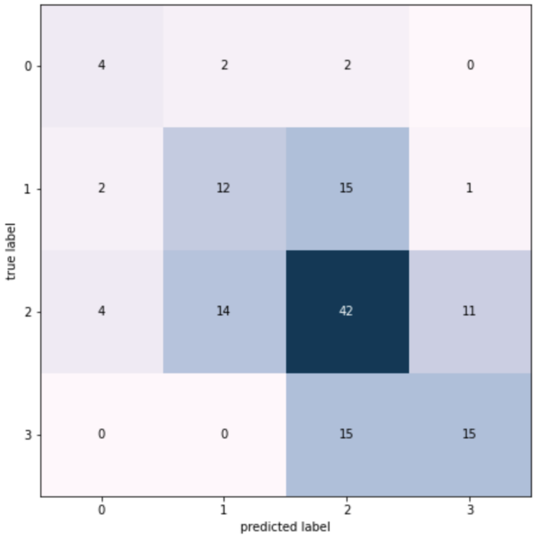

Figure 3: Confusion matrix of ANN Multiclass Classification

The confusion matrix of the ANN multiclass classification model (Figure 3) illustrates the 74 correctly classified cases and 65 incorrectly classified cases. The most impactful false positive case, benign mass cases that were classified as either malignant or requiring further testing (or re-imaging) occurred in 28 cases. This is larger than the false positive rate of the binary ANN classification model and

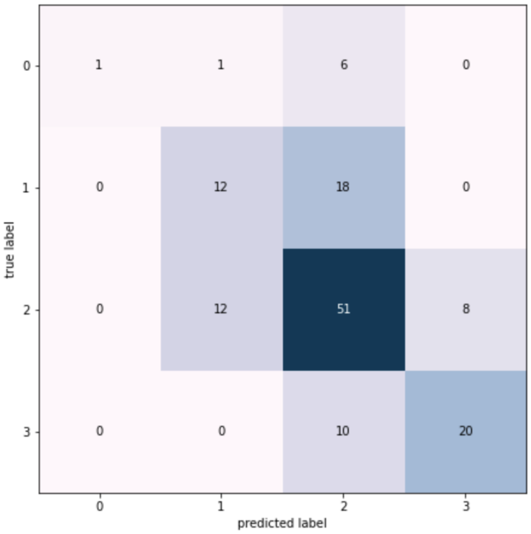

Figure 4: Confusion matrix of One-vs-all Multi-class Classification

about the same as that of the logistic regression binary classification model. Of key importance, 16 cases were mistakenly classified as benign. This is about the same rate as for false negatives in the binary classifying ANN model. Since cases classified in the third category are recommended for a biopsy anyway, the 15 cases of malignant masses being classified as just suspicious and the 11 cases of suspicious masses being classified as malignant, while not ideal, is not a very large cause for concern.

The confusion matrix for the one-versus-all multiclass classification model (Figure 4) displays a greater number, 84, of correctly classified masses. 10 fewer benign cases were mistakenly classified as requiring follow up or malignant than for the ANN multiclass classification model. The difference in masses mistakenly classified as benign was negligible.

Table 1: Accuracy with Different Models

Model	Accuracy (MoB)	Accuracy (RADS)
Logistic Regression.	74.1%	-
One vs. All	-	60.43%
Artificial Neural Network	76.26%	53.24%
	Accuracy (CC View)	Accuracy (MLO View)
Convolutional Neural Network	69.86%	68.49%

The results of the second key metric, accuracy, for each model, including the convolutional neural networks, are summarized in Table 1. Comparison of the two convolutional neural network models to the other binary classifiers, the feed forward artificial neural network and the logistic regression models, reveals a relatively low accuracy. The accuracy of the simple artificial neural network model in comparison with logistic regression reveals a slightly, less than 2.5% higher accuracy for the artificial neural network in solving the binary classification problem. However, the slightly fewer, by 7, number of false negative cases, likely the most important metric, would indicate that overall the logistic regression model is slightly preferable. On the other hand, due to the very small differences and small test set size, it would be unwise to draw any definitive conclusions as to the relative efficacy of the artificial neural network and logistic regression binary classification models.

There was, excluding the CNN models due to the lack of model optimization, a significant gap, of at least 13%, between the accuracy of the binary classification models and the mutliclass classification models. Of the multiclass classification models, one versus all multiclass classification was more accurate. In combination with the lower rate of false negatives, this suggests that, in their current states, the one-vs-all multiclass classification model is more efficacious than the multiclass artificial neural network model.

6 Conclusion

This paper describes an initial implementation of four types of models for the classification of mammogram masses: a simple artificial neural network for both binary and multiclass classification, logistic regression for binary classification, one versus all multiclass classification and a convolutional neural network. While initial results indicate that of the modeling implementations, the logistic regression binary classification model performs the best, for multiple reasons it would be ill-advised to draw broader conclusions as to the best model for mammogram mass classification from these results. First, even when taking these results at face value, the difference between the performance of logistic regression and artificial neural network binary classification is arguably far too small to draw any conclusions from. Furthermore, the extremely limited scope of this study ensures that even the more significant differences, such as the difference between the performance of the one vs all and artificial neural network multiclass classification models, are questionable.

One of the major limitations of this study was the small sample size. After the exclusion of cases with incomplete data, the training set size was 445 cases and the test set was composed of 139 cases. While not outside the norm for research into machine learning applications in medicine, particularly for research performed by smaller institutions, the small sample size is unarguably problematic. This may have caused many of the performance problems with the multiclass classification models. With such a small sample size, the number of cases in each of the four categories likely became too small for proper training or testing. For example, there were only 110 cases in the training set that fell into the benign category and only 30 in the test set. This suggests that with a large enough dataset and appropriate hyperparameter optimization, a multiclass classification approach remains a viable solution to the problem of mammogram classification.

Another potential area in which these results may misrepresent the larger problem is the efficacy of the artificial neural network models. Due to the authors' extremely limited experience with artificial neural networks it remains likely, that the overall structure and hyperparameters of the ANN models could be altered to greatly increase the accuracy of the model, even with the limited dataset used here. This means that there is a strong chance that the ANN models would, if given the appropriate expertise and time, outperform the other models. By the relatively simple nature of the logistic regression binary classification model and one-versus-all multiclass classification models, those models have less room to improve. For example, the authors hypothesize that proper use of regularization could significantly improve the efficacy of the ANN models. ANN models are generally prone to overfitting, a trend which the authors observed in the difference between the accuracy of the models on the training set, which in both cases approached perfect accuracy and the much lower accuracy of the models on the test set. The authors were able to effectively implement L2 regularization for both the logistic regression binary classification and one versus all multiclass classification models. This likely had a significant impact on the test set accuracy of these models. All attempts to implement regularization with a L2 regularizer for the bias, waits or activation function as well as with the use of dropout resulted in lowered accuracy. However, these attempts were limited, to put it simply, the authors ran out of time to successfully implement regularization for the ANN models. The issue of limited time was in fact, the other major limitation of this study.

The limited time available to implement this study was also the key factor in the very low accuracy of the convolutional neural networks. Due to the high runtime for training the convolutional neural network, the authors were unable to implement hyperparameter optimization. As such, these results should not be taken as an indication of the efficacy of a CNN approach but rather proof that it is possible to implement a CNN for this data.

To further improve the efficacy of a machine learning model for the classification of mammograms the authors suggest the use of a larger dataset, inclusion of an expert in radiology on the team, as well as an exploration into the implementation of regularization in the ANN models. Furthermore, optimization of the CNNs may lead to a more accurate effective model. The combination of the raw image data and the patient metadata could provide a more effective model. Finally the use of ensemble methods such as boosting and bagging could be used to combine multiple models and achieve more accurate results.

The authors have presented initial insights into the use of machine learning for the classification of mammogram masses. An effective machine learning model would promote early, lifesaving, treatment for breast cancer. Ultimately the authors propose that such a tool should be used to supplement the work of a radiologist and could potentially, with the use of a multiclass classification model aid the examining physician in focusing on the most difficult cases.

References

[1] American Cancer Society.(2019) "Breast cancer facts figures 2019-2020." Atlanta: American Cancer Society, Inc. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/breast-cancer-facts-and-figures-2019-2020.pdf

[2] American Cancer Society medical and editorial content team.(2019) "What does the doctor look for in a mammogram?" The American Cancer Society. https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/mammograms/what-does-the-doctor-look-for-on-a-mammogram.html

[3] Center for Disease Control.(2020) What is breast cancer screening? Division of Cancer Prevention and Control, Centers for Disease Control and Prevention. https://www.cdc.gov/cancer/breast/basic_info/screening.htm

[4] Heath, M., Bowyer, K., Kopans, D., Moore, R. Kegelmeyer, W.P. (2001). The digital database for screening mammography. *Proceedings of the Fifth International Workshop on Digital Mammography* pp. 212-218. Medical Physics Publishing.

[5] Houssein, E.H., Emam, M.M., Ali, A.A. Suganthan, P.N. (2020). Deep and machine learning techniques for medical imaging-based breast cancer: a comprehensive review. *Expert Systems with Applications* 114161.

[6] Islam, M., Iqbal, H., Haque, R. Hasan, K.(2017). Prediction of breast cancer using support vector machine and K-Nearest neighbors. *IEEE Region 10 Humanitarian Technology Conference (R10-HTC), Dhaka* pp. 226-229.

[7] Khairunnahar, L., Hasib, M.A., Rezanur, R.H.B., Islam, M.R. Hosain, K.(2019). “Classification of malignant and benign tissue with logistic regression. *Informatics in Medicine Unlocked* 16:100189.

[8] Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V., Fotiadis, D.I.(2014). Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal* 13:8–17.

[7] Lee, R.S., Gimenez, F., Hoogi, A., Miyake, K.K., Gorovoy, M. Rubin, D.L. (2017) A curated mammography data set for use in computer-aided detection and diagnosis research. data 4:170177.

[8] Obaid, O.I.,Mazin, A.M., Ghani, M.K.A., Mostafa, S. Al-Dhief, F.T.(2018) Evaluating the performance of machine learning techniques in the classification of Wisconsin breast cancer.” *International Journal of Engineering and Technology.* 7(4.36):160-166.

[9] Rouhi, R., Jafari, M., Kasaei, S. Keshavarzian, P.(2015). Benign and malignant breast tumors classification based on region growing and CNN segmentation. Systems with Applications 42(3):990-1002

[10] Vaka, R.A., Soli, B., Reddy S.K.(2020) Breast cancer detection by leveraging machine learning. *ICT Express*, 6(4):320-324.

[11] Vikas, C., Pal, S. Tiwari, B.B.(2018) Prediction of benign and malignant breast cancer using data mining techniques. *Journal of Algorithms and Computational Technology* 12(2):119–126.