**A Novel Metadata Approach to Breast Cancer Classification Using Decision Tree and Convolutional Neural Networks**

# **Abstract**

Accurate detection and assessment of breast cancer in its early stages is crucial for reducing the mortality rate of patients. While mammography screenings are integral to this process, their assessment is a complex task with significant variability due to professional experience and human errors. Deep Learning has become one of the most popular tools for assisting radiologists in this process. However, the majority of such Deep Learning Computer Aided Diagnosis (CADx) systems are overly fixated on extracting features completely independent of human insights. The drawbacks to this include the immense size of the resulting neural networks, which are costly to train, and the fact that such models are usually Black Boxes, meaning it is nearly impossible to gain insight into the problem-nature of breast cancer through machine-extracted features. In this paper, we propose a novel, light, decision-tree-based CADx system. The core of the system is a tree model generated from CBIS-DDSM (Curated Breast Imaging Subset of DDSM) dataset of pathology and metadata features identified by radiologists. Specific tree-model categories with low accuracies are then linked with Convolutional Neural Networks (CNNs) that train specifically within these categories to improve performance. Additionally, we investigate using CNNs to label lesions that exhibit these metadata features. We also experiment with different data organization techniques and data augmentation to optimize the accuracy of the tree model and the CNNs respectively. With CBIS-DDSM and its accompanying metadata, the proposed CADx system achieved 89.4% accuracy, while the CNN applied to identify the feature “Amorphous” achieved 95.9% accuracy, 94.6% precision, 94.0% sensitivity, and 0.961 AUC. The hybrid and computation-friendly modeling framework of the proposed CADx system makes it universally transferable to be applied to generic datasets, and scalable to lay the groundwork for a potential crowd-sourcing, web-based data analytics hub for the medical community.

# **1. Introduction**

Breast cancer is one of the leading causes of death for women around the world and the accurate detection and assessment of breast cancer in its early stages is crucial for reducing the mortality rate. Currently, mammography is the most useful tool for general population screening. However, it is difficult to accurately detect and diagnose breast lesion solely based on mammography. The reliability of this process is highly dependent on the expertise of the radiologist, which leads to a high number of false positives and additional examinations.

Experienced radiologists can identify signs of breast cancer from mammography images with a certain accuracy. Abnormalities found in a mammogram are broadly categorized as masses and calcification. These have distinguishing characteristics that correlate them to either benign or malignant. For example, the benign tumors often have more rounded shapes, while malignant tumors often have more jagged outlines (see Figure 1).

(a) Sample Malignant Tumor ROI

(b) Sample Benign Tumor ROI

A picture containing outdoor, cloudy, nature, clouds

Description automatically generated A picture containing night sky

Description automatically generated

Figure 1: Sample CBIS-DDSM Mammography

Computer Aided Detection and Diagnosis (CAD) systems are used to aid radiologists to improve reliability and accuracy in their decision making. Such systems can significantly reduce the amount of efforts needed for the assessment of a lesion in clinical practice, thus minimizing the number of false positives that lead to unnecessary and uncomfortable biopsies. CAD systems are divided into CADe, which serves to detect suspicious lesions in a mammogram, and CADx, which serves to diagnose detected lesions, i.e., classify as benign or malignant.

CAD systems often utilize Deep Learning, one of the most popular approaches to medical image analysis. They excel in various machine learning tasks including object detection and classification. Contrary to conventional machine learning methods, deep learning does not rely on hand-crafted feature extraction. Rather, the models adaptively learn the appropriate feature extraction process from the input data with respect to the target output. Extensive research has been done in this area comparing various deep architectures for breast cancer detection and diagnosis. The most common type of deep learning architecture is the convolutional neural network (CNN). Sahiner et al. (1996) published one of the first demonstrations of breast mass classification using CNN where they achieved a 90% true positive rate and 31% false positive rate. Carneiro et al. (2015) introduced transfer learning by using a pre-trained CNN that was fine-tuned using unregistered mammograms and concluded that the pre-trained models are superior to the randomly initialized ones. Huyuh et al. (2016) used transfer learning to extract tumor information from medical images via CNNs originally pretrained for nonmedical tasks, alleviating the need for large datasets. Their support vector machine (SVM) classifier achieved an area under the ROC curve (AUC) of 0.86 using both analytically extracted features and CNN-extracted features. Regab et al. (2019) had a similar approach where they fine-tuned the pre-trained AlexNet to classify 2 classes (benign or malignant) instead of 1000 classes and connected the last fully connected layer to a SVM classifier. The fine-tuning was done on both the Digital Database for Screening Mammography (DDSM) and the Curated Breast Imaging Subset of DDSM (CBIS-DDSM). Their SVM achieved an accuracy of 87.2% with an AUC of 0.94 using CNN-extracted features. Jiao et al. (2016) expanded on this idea by separating the features obtained from the fine-tuned CNN into high level and middle level and training one linear SVM classifiers for each group of CNN-extracted features. The prediction was then calculated by combining both SVM classifiers’ outputs. Levy and Jain (2016) compared transfer learning to from-scratch training and declared the former to be superior. Their GoogleNet model achieved an accuracy of 92.9%. Additionally, they investigated the effect of data context and concluded that cropping larger bounding boxes of fixed size around the lesion is more effective compared to cropping with proportional padding.

While Deep Learning provides an alternative to the process of investigating the discrimination ability of the features, the drawback is that meticulous training and tuning is required before deep learning models can produce any meaningful results. Additionally, it is nearly impossible to understand the decisions taken by their complex mathematical (Black Box) models. Thus, White Box systems, such as decision trees, are sometimes desired, since their results are easily interpretable by humans and offer insights of problem nature.

Amongst the CADx publications, the majority focuses on using some variation of neural networks and deep learning to learn the features of benign and malignant tumors. Here, an emphasis is usually put around the models learning which features to extract completely independent of human insights. There are several problems to this. First, without human insight to simplify the problem, the complexity involved with breast cancer classification requires immense neural network models with tens of millions of parameters, making them difficult and costly to train. This also means that they are terrible at adapting to new mammography data, as the models need to be constantly re-trained to incorporate new information. Second, as mentioned previously, deep learning models are usually black boxes, which means that while the machine-extracted features may make sense to the models, these are, at least currently, nearly impossible to decipher for humans. This way researchers forgo the possibility to gain valuable insights into the problem nature of breast cancer.

Considering the above, this paper proposes a novel, highly cost effective, light, decision-tree-based CAD system for breast cancer classification. This CAD system differs from others in several ways. Firstly, we experimented with training neural networks for the classification of metadata features and the classification of lesions under low-sample, specific categories that the tree model struggles with. In other words, CNNs are used here as a supporting class to the main tree model, which identifies specific categories with low accuracy according to metadata features derived from modern radiology. Secondly, considering that most of the CADx systems are designed to assist radiologists and labs instead of replacing them, the proposed CAD system is designed accordingly to have the advantage of being able to take both CNN-generated metadata features as well as radiologist-generated metadata features as input. With this important feature, this CAD system can still support radiologists by simplifying the diagnostic process and allow radiologists to specialize and divide-and-conquer, even without the CNNs. Together, where traditional deep learning fails this system excels. Lastly, this CAD system is also easily scalable and transferrable, owing to to its decision-tree core and the small size of its supporting CNNs.

The remainder of the paper is organized as follows: Section 2 illustrates the dataset used in this study. Section 3 details the research framework and model development of the decision-tree-based CAD. Section 4 presents the corresponding results and discussion. Finally, in Section 5 conclusions are drawn.

# **2. The dataset: CBIS–DDSM**

The Curated Breast Imaging Subset of DDSM (CBIS-DDSM) is an updated, standardized subset of DDSM containing 2620 full mammogram images. The images are linked benign, benign without callback, and malignant cases, selected and curated by a team of mammographers. The images are converted to the DICOM format, and an updated ROI (region of interest) segmentation is provided for each lesion. One benign and one malignant sample ROI from the dataset are presented below.

The dataset is split to training and testing subsets containing 80% and 20% of the cases, respectively. The images come in both MLO view and CC view, which are depicted below. Both types of images were used in this study without discrimination. For this study, only calcification cases were extracted, totaling 1238 and 309 ROIs for training and testing respectively.

The CBIS-DDSM dataset also comes with a supporting metadata dataset that provides additional details for each case, which includes its corresponding patient ID, density category, left or right breast, CC or MLO view (see Figure 2), number of abnormality in the image, mass shape and margin (mass cases only), calcification type and distribution (calcification cases only), BI-RADS assessment, pathology, subtlety rating, and path to image files (see Figure 3). Preliminary statistical analysis indicates that the most useful metadata features for calcification cases that do not directly correlate with pathology were “calcification type” and “calcification distribution”. Thus, this study was centered around using these features for classification of benign and malignant lesions.

Diagram

Description automatically generated

Figure 2: MLO vs CC view

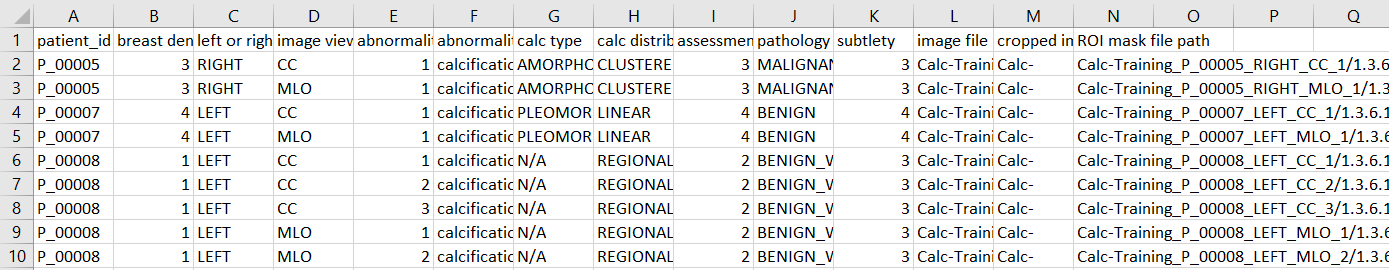


Figure 3: Sample CBIS-DDSM Metadata

As a part of exploratory data analysis, we investigated the correlation between malignant cases and the calcification types. As summarized in Table 1, one can find that only 12 out of the 47 calcification types have non-zero correlations with malignant cases. This means a notable portion of the cases could be given a confident diagnosis purely based on the calcification type. This important discovery laid the foundation for the development of our methodology.

Table 1: Summary of Correlation Analysis Results of Malignant Cases and Calcification Types

|  |  |  |  |
| --- | --- | --- | --- |
| Calcification Types | Malignant  Cases | Total Cases | %  of Malignant Cases |
| Amorphous-Round and Regular | 3 | 3 | 100.00% |
| Punctate-Fine Linear Branching | 6 | 6 | 100.00% |
| Pleomorphic-Fine Linear Branching | 25 | 28 | 89.29% |
| Fine Linear Branching | 56 | 77 | 72.72% |
| Pleomorphic | 357 | 664 | 53.77% |
| Amorphous-Pleomorphic | 6 | 12 | 50.00% |
| Punctate-Pleomorphic | 10 | 21 | 47.62% |
| Amorphous | 55 | 138 | 39.86% |
| Punctate | 21 | 106 | 19.81% |
| Round And Regular-Punctate | 1 | 5 | 20.00% |
| Round And Regular | 3 | 17 | 17.65% |
| N/A | 1 | 20 | 5.00% |
| Total | 544 | 1097 |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Round and Regular-Lucent Center-Dystrophic | 0 | 2 | 0.00% |
| Coarse | 0 | 35 | 0.00% |
| Vascular | 0 | 82 | 0.00% |
| Large Rodlike | 0 | 4 | 0.00% |
| Punctate-Lucent Center | 0 | 3 | 0.00% |
| Vascular-Coarse-Lucent Center-Round and Regular | 0 | 2 | 0.00% |
| Round and Regular-Eggshell | 0 | 23 | 0.00% |
| Dystrophic | 0 | 20 | 0.00% |
| Lucent Center | 0 | 93 | 0.00% |
| Vascular-Coarse-Lucent Centered | 0 | 8 | 0.00% |
| Coarse-Round and Regular | 0 | 2 | 0.00% |
| Coarse-Pleomorphic | 0 | 2 | 0.00% |
| Lucent Centered | 0 | 18 | 0.00% |
| Vascular-Coarse | 0 | 6 | 0.00% |
| Round and Regular-Lucent Center | 0 | 31 | 0.00% |
| Coarse-Round and Regular-Lucent Centered | 0 | 4 | 0.00% |
| Skin | 0 | 2 | 0.00% |
| Lucent Center-Punctate | 0 | 8 | 0.00% |
| Skin-Punctate | 0 | 4 | 0.00% |
| Skin-Punctate-Round and Regular | 0 | 4 | 0.00% |
| Milk of Calcium | 0 | 2 | 0.00% |
| Pleomorphic-Pleomorphic | 0 | 1 | 0.00% |
| Skin-Coarse-Round and Regular | 0 | 1 | 0.00% |
| Round and Regular-Pleomorphic | 0 | 7 | 0.00% |
| Round and Regular-Punctate-Amorphous | 0 | 2 | 0.00% |
| Round and Regular-Amorphous | 0 | 1 | 0.00% |
| Coarse-Round and Regular-Lucent Center | 0 | 10 | 0.00% |
| Large Rodlike-Round and Regular | 0 | 11 | 0.00% |
| Round And Regular-Lucent Center-Punctate | 0 | 24 | 0.00% |
| Coarse-Lucent Center | 0 | 2 | 0.00% |
| Punctate-Amorphous | 0 | 10 | 0.00% |
| Round And Regular-Lucent Centered | 0 | 14 | 0.00% |
| Punctate-Round and Regular | 0 | 4 | 0.00% |
| Eggshell | 0 | 7 | 0.00% |
| Total | 0 | 449 |  |

# **3. Methodology**

The novelty of the proposed CAD system lies in using metadata features generated from human insights. This allows for the usage of a decision-tree core to identify cases which can be given a confident diagnosis without going through a CNN. This way, CNNs can be trained to target specific categories of lesions with similar appearances. We hypothesized that a hybrid model architecture integrating several categorized CNNs with each focusing on its specialized patterns would outperform a single unified CNN that needs to handle all different patterns across all categories. The architecture of the proposed CADx system is illustrated in Figure 4.



Figure 4: Architecture of the proposed CADx system

In addition to this, we also aim to address the following critical research questions related to best application of the proposed model through extensive experimentation:

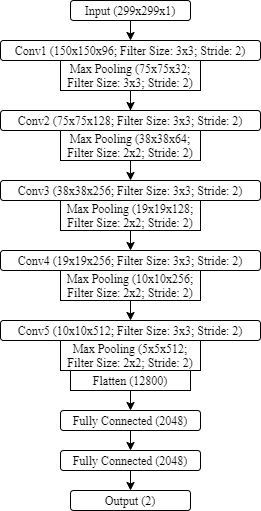
* What is the highest achievable accuracy of diagnosis purely based on the metadata features identified by radiologists? (to be addressed in 4.1)
* How to optimize the performance of classification algorithms using various data cleaning and organization methods? (to be addressed in 4.2)
* How is the CNN performance affected from specializing in specific categories where the cases are relatively homogenous? (to be addressed in 4.3)
* How do complementary CNNs boost the accuracy of the decision tree? (to be addressed in 4.4)
* How does the CNN perform when tasked with identifying a metadata feature? (to be addressed in 4.5)

## 3.1. Decision Tree

A decision tree is a predictive model that is useful as an exploratory technique. The advantages of a decision tree lie in its simplicity and transparency. Decision trees are usually represented graphically as a hierarchical structure that makes them easier to be interpreted than other techniques. This structure mainly consists of a starting node (root) and a group of branches (conditions) that lead to subsequent nodes until the leaf node that contains the final decision of the route is reached. The decision tree is a self-explanatory model because its representation is very simple. Each internal node tests an attribute while each branch corresponds to an attribute value (or a range of values). Each leaf of the tree is labeled with a probability distribution over the classes, signifying that the dataset has been classified into a particular probability distribution. In our case, each leaf of the tree is labeled with a probability of being benign and being malignant.

A tree is built by splitting the source set, constituting the root node of the tree, into subsets – which constitute the successor children. The splitting is based on a set of splitting rules based on classification features. This process is repeated on each derived subset in a recursive manner. The recursion is completed when the subset at a node has all the same values of the target variable, or when splitting no longer adds value to the predictions.

In this paper, due to the limited number of metadata features that can form branches in the decision tree and the limited number of cases overall, some leaves are not accurate enough to predict whether a lesion is benign or malignant. Our solution to this is complementary CNNs that leverage the homogeneity of cases within these categories to improve accuracy.



## 3.2. Simplified VGG-16

Compared to VGG-16’s 134 million parameters, the simplified VGG-16 used in this paper is much more compact, with only 34 million parameters. Moreover, 299x299 greyscale images are used as input. The architecture of the CNN is shown in Figure 5. The first 5 layers are convolutional. These layers filter the input image with kernels of various sizes (as described in Figure 5) with a stride of 2 pixels. The max pooling layers perform a down sampling operation along the spatial dimensions to reduce the amount of computation. Together, the convolutional and pooling layers extract learned features from the original image input. The fully connected layers are feedforward neural networks that uses the extracted features to classify the input image as either benign or malignant.

Having only a quarter of the parameters of a full VGG-16, the modified CNN trains exponentially faster. This also means much less computing power is required for the CNN. This is an important advantage for a model that needs to be constantly retrained to stay concurrent with additions to the breast cancer dataset.

Figure 5: Simplified VGG16 architecture

## 3.3. ROI Extraction

The CBIS-DDSM scans are relatively large. We used the provided the ROI masks to define individual squares that completely enclosed each of the ROIs. Some padding was added to the bounding box to provide context and then the ROIs were extracted at 598x598 and then resized down to 299x299 so they could be input into the CNN. Conveniently, 598x598 was just large enough that the majority of the ROIs could fit. The ROIs larger than 598x598 were cropped to fit. This way we minimized the element of artificiality, which would help the model generalize to classifying raw scans.

## 3.4. Transfer Learning

The CNNs were initialized with weights from Escuccim, an open-source model created by Scuccimarra (2018) for detecting (rather than diagnosing) abnormalities in mammograms. This way we were able to save time and resources by bypassing the training of the convolutional layers. The fully connected layers were re-trained with extracted ROIs to either diagnose lesions or classify metadata features. The learning rate, momentum, batch size, and other hyperparameters were all carried over from Escuccim as hyperparameter tuning is not the focus of this study.

## 3.5. Data Augmentation

Training on a large number of samples elevates performance of neural network models. However, biomedical datasets typically contain a relatively small number of samples. This can be attributed to both limited patient volume and sensitivity of the information. CBIS-DDSM suffers the same problem. Our solution to this is data augmentation. We rotated each image by 0, 90, 180, and 270 degrees, augmenting each image to 4 images. This has the added benefit of helping the model generalize to raw scans of different orientations.

# **4. Results**

## 4.1. Preliminary Testing

Table 2 summarizes the performance of various classification algorithms using the metadata features calcification type and calcification distribution. We also found that other metadata features, including breast density, left or right breast, image view, and abnormality number, have little to no correlation to either a benign or malignant diagnosis. Finally, Table 2 shows that the CHAID decision tree had the best performance out of all the algorithms tested. The results also confirm our hypothesis that the metadata features themselves already provide a semi-accurate diagnosis.

Table 2: Performance of Classification Algorithms (No CNN)

|  |  |  |
| --- | --- | --- |
| Model Type | Accuracy (%) | AUC |
| CHAID | 75.2 | 0.851 |
| SVM | 75.1 | 0.843 |
| Logistic Regression | 73.8 | 0.848 |
| Bayesian Network | 73.3 | 0.843 |
| Neural Net | 73.3 | 0.837 |
| C5 | 70.8 | 0.651 |
| KNN Algorithm | 70.5 | 0.786 |
| Quest | 69.7 | 0.758 |
| Decision List | 66.7 | 0.735 |
| C&R Tree | 62.8 | 0.500 |

## 4.2. Effect of Different Data Cleaning and Organization Methods

At this point, only two metadata features, calcification type and calcification distribution, were used for classification algorithms. Also worth noting was that while there were 47 calcification types, most of these are combinations of 13 distinct calcification types. For example, “Pleomorphic-Fine Linear Branching” is a combination of the categories “Pleomorphic” and “Fine Linear Branching”. Based on our understanding of classification algorithms in general, we hypothesized that having more features and less categories under specific features would boost the classification models’ performance by reducing noise. Thus, we experimented with 3 ways of grouping calcification types based on their correlations with the malignant cases. We also tested making each of the calcification type an individual, binary feature.

Table 3 shows the effect of the three grouping methods as well as that of extracting features from calcification types. The results reflected no significant benefit to reducing the number of categories under calcification type. Extracting binary features proved to be counterproductive. This can be explained by the fact that when each calcification type was extracted to become individual features, most of them become obsolete because the vast majority of cases do not exhibit these features – as shown in Table 1, most of these features are exhibited by less than 50 out of the 1546 cases.

We opted to use Grouping 1 for all following experiments because even though it did not boost the accuracy of the models, it did reduce the number of categories under calcification types which made the graphic representation of the tree model more readable (see Figure 6).

We also noticed that many of the cases were different abnormalities in the mammogram of the same patients (a significant group of patients had more than 1 abnormalities documented in CBIS-DDSM), and many of the different abnormalities of the same patients exhibited the same set of metadata features. Thus, we doubted whether the existence of such duplicates has inflated the accuracies of the models. Table 4 shows the effect of the removal of the duplicates on the accuracies of the models, which dismissed our suspicion because the accuracies did not change significantly. This suggests that overall, there were just as many duplicates of benign cases as duplicates of malignant cases. Thus, we kept the duplicates for all following experiments.

Table 3: Effect of Data Cleaning Methods on the Performance of Classification Algorithms

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | CHAID | | SVM | | Logistic Regression | | Neural Net | |
| Cleaning Method | Accuracy (%) | AUC | Accuracy (%) | AUC | Accuracy (%) | AUC | Accuracy (%) | AUC |
| Grouping 1 | 75.3 | 0.846 | 75.2 | 0.800 | 73.2 | 0.841 | 74.2 | 0.845 |
| Grouping 2 | 75.0 | 0.827 | 75.2 | 0.830 | 72.3 | 0.835 | 73.9 | 0.845 |
| Grouping 3 | 74.5 | 0.846 | 74.5 | 0.820 | 74.8 | 0.840 | 73.6 | 0.840 |
| Extracting Descriptive Features | 67.2 | 0.705 | 68.6 | 0.706 | 67.2 | 0.708 | 67.7 | 0.712 |

Table 4: Effect of Removal of Duplicates on the Performance of Classification Algorithms

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | CHAID | | SVM | | Logistic Regression | | Neural Net | |
| Cleaning Method | Accuracy (%) | AUC | Accuracy (%) | AUC | Accuracy (%) | AUC | Accuracy (%) | AUC |
| Grouping 1 | 73.1 | 0.829 | 72.8 | 0.823 | 71.1 | 0.805 | 71.7 | 0.808 |
| Grouping 2 | 69.0 | 0.766 | 75.3 | 0.813 | 72.4 | 0.797 | 71.3 | 0.793 |
| Grouping 3 | 71.1 | 0.788 | 72.8 | 0.804 | 71.1 | 0.803 | 70.5 | 0.804 |
| Extracting Descriptive Features | 63.5 | 0.689 | 62.8 | 0.615 | 63.5 | 0.661 | 64.2 | 0.680 |

Figure 6 is a graphic representation of the decision tree model. Other than Nodes 1, 6, and 8, all other Nodes lead to fairly accurate diagnoses. Cases that are part of the weak nodes can be fed into complementary CNNs to improve the predicting accuracy of the decision tree.

Diagram, schematic

Description automatically generated

Figure 6: Graphic Representation of the Decision Tree Model

## 4.3. Complementary CNN for Classification within Weak Categories

Table 5 illustrates the performance of 4 CNNs separately trained for classification for Node 1 (Amorphous), Node 6 (Pleomorphic 1), Node 8 (Pleomorphic 2), and over all categories. The results support our hypothesis that the CNNs that specialize in specific categories where the cases are relatively homogenous yield better performance than the CNN that is tasked with learning all the patterns across all categories. The superior performance of the CNN for Node 6 could be attributed to it having significantly more training data than the other two CNNs, as Node 6 contained the most cases out of the 3. On the other hand, the inferior performance of the CNN for Node 1 could be attributed to amorphous literally meaning “without a clearly defined shape or form” and is therefore difficult to diagnose even for the CNN.

Table 5: CNN Performance Within Weak Categories vs. All Categories

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Category | Accuracy (%) | Precision (%) | Sensitivity (%) | AUC |
| Pleomorphic 1\* | 88.9 | 88.0 | 88.0 | 0.941 |
| Pleomorphic 2\* | 86.2 | 75.6 | 86.1 | 0.919 |
| Amorphous | 78.3 | 78.5 | 88.0 | 0.865 |
| All Categories | 71.0 | 71.6 | 74.3 | 0.792 |

Figure 7 further supports the results in Table 5 by showing that the specialized CNNs, in average, have higher accuracies during training than the non-specialized CNN. Note that the graphs follow the anticipated trend of converging towards their respective maximum accuracies.

Figure 7: CNN Test Results Over Epochs per Category

## 4.4. Performance of the Decision Tree Model with Complementary CNNs

Table 6 illustrates the effect of the complementary CNNs on the performance of the decision tree model. The CADx system is able to achieve a final accuracy of 89.4%.

Table 6: Decision Tree Model without vs. with complementary CNNs

|  |  |  |
| --- | --- | --- |
| Category | Accuracy Without CNN (%) | Accuracy With CNN (%) |
| Other | 100.0 | 100.0 |
| Pleomorphic | 53.6 | 88.2 |
| Punctate; Round and Regular | 80.5 | 80.5 |
| Fine Linear Branching; Amorphous Round and Regular | 78.9 | 78.9 |
| Amorphous | 60.1 | 78.3 |
| All Categories | 75.3 | 89.4 |

## 4.5. Complementary CNN for Identification of Metadata Features

Table 7 and Figure 8 describes the performance of the CNN seperately trained for identifying the feature “Amorphous” and the CNN’s training, respectively. The results show that with proper training, the CNN is capable of identifying metadata features at a very high accuracy. This means that CNNs have the potential to assist or even replace radiologists in the tedious work of identifying such metadata features in mammograms. The results also shows that the CNN is more successful with metadata feature identification than with benign-malignant classification. This can be explained by the same reason detailed in 4.3, which is that the former is a simpler task for the CNN because it is able to operate within a specific, homogenous context, where there are less and simpler patterns to be learned. The CADx system is complete with metadata-feature-identifying CNNs, the decision-tree core, and the complementary specialized CNNs.

Table 7: CNN Performance in Identification of “Amorphous” Feature vs. CNN Performance in Classification

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Category | Accuracy (%) | Precision (%) | Sensitivity (%) | AUC |
| Amorphous (Feature) | 95.9 | 94.6 | 94.0 | 0.961 |
| All Categories | 71.0 | 71.6 | 74.3 | 0.792 |

Figure 8: CNN Test Results Over Epochs (Feature Detection vs. Benign/Malignant)

# **5. Conclusions**

In this research, a novel, light, decision-tree-based CADx system was prototyped and validated for breast cancer classification to remedy the deficiency of most deep learning approaches in immense size, costly training and maintenance, intractable application, and lack of insight into the problem-nature of breast cancer. The Curated Breast Imaging Subset of DDSM (CBIS-DDSM) containing 2620 full mammogram images was used for developing, calibrating, and validating the proposed model. Specific tree-model categories with low accuracies are then linked with Convolutional Neural Networks (CNNs) that train specifically within these categories to improve performance. We further explored critical research questions related to performance optimization of the proposed model through extensive experimentation. We found that with our fine-tuned data organization, cleaning, and augmentation techniques, the proposed CADx system achieved 89.4% accuracy, while the CNN applied to identify the feature “Amorphous” achieved 95.9% accuracy, 94.6% precision, 94.0% sensitivity, and 0.961 AUC, indicating its superiority over many of its rivals. More importantly, our approach allows for maximum utilization of professional radiologist insights for a preliminary round of classification and diagnosis. This hybrid and computation-friendly feature makes it universally transferable to be applied to generic datasets and scenarios.

We fully recognize that any data-oriented models (like the one developed in this study) intended for use in practice certainly can achieve its best performance if properly calibrated. The presented model and evaluation results, though still preliminary, offers the advantage of computational convenience and operational flexibility, providing a useful and actionable reference for potential users in the medical community to customize its application depending on the data availability in their target application context.

Our future research along this line is to automate and facilitate the model application and accessibility by developing a crowd-sourcing, web-based data analytics hub for the medical community.

# **References**

Almeida, Rhaylander, Dehua Chen, Agnaldo Filho, and Wladmir Brandão. "Machine Learning Algorithms for Breast Cancer Detection in Mammography Images: A Comparative Study." *In Proceedings of the 23rd International Conference on Enterprise Information Systems* 1 (2021): 660-67. Accessed November 10, 2021. https://doi.org/10.5220/0010440906600667.

Carneiro, Gustavo, Jacinto Nascimento, and Andrew P. Bradley. "Unregistered Multiview Mammogram Analysis with Pre-trained Deep Learning Models." In *Medical Image Computing and Computer-Assisted Intervention -- MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings*, 652-60. Cham: Springer International Publishing AG, 2015. Accessed November 10, 2021. https://link.springer.com/chapter/10.1007/978-3-319-24574-4\_78.

Huynh, Benjamin Q., Hui Li, and Maryellen L. Giger. "Digital Mammographic Tumor Classification Using Transfer Learning from Deep Convolutional Neural Networks." *Journal of Medical Imaging* 3, no. 3 (August 22, 2016). https://doi.org/10.1117/1.JMI.3.3.034501.

Jiao, Zhicheng, Xinbo Gao, Ying Wang, and Jie Li. "A Deep Feature Based Framework for Breast Masses Classification." *Neurocomputing* 197 (July 2016): 221-31. Accessed November 10, 2021. https://doi.org/10.1016/j.neucom.2016.02.060.

Lee, Rebecca Sawyer, Francisco Gimenez, Assaf Hoogi, and Daniel Rubin. "Curated Breast Imaging Subset of DDSM." Dataset, July 5, 2016. https://wiki.cancerimagingarchive.net/display/Public/CBIS-DDSM.

Lee, Rebecca Sawyer, Francisco Gimenez, Assaf Hoogi, Kanae Kawai Miyake, Mia Gorovoy, and Daniel L. Rubin. "A Curated Mammography Data Set for Use in Computer-aided Detection and Diagnosis Research." *Scientific Data* 4, no. 1 (December 2017). Accessed November 10, 2021. https://doi.org/10.1038/sdata.2017.177.

Lévy, Daniel, and Arzav Jain. "Breast mass classification from mammograms using deep convolutional neural networks." *arXiv preprint arXiv:1612.00542* (2016).

Ragab, Dina A., Maha Sharkas, Stephen Marshall, and Jinchang Ren. "Breast Cancer Detection Using Deep Convolutional Neural Networks and Support Vector Machines." *PeerJ* 7 (January 28, 2019). https://doi.org/10.7717/peerj.6201.

Sahiner, B., Heang Ping Chan, N. Petrick, Datong Wei, M.a. Helvie, D.d. Adler, and M.m. Goodsitt. "Classification of Mass and Normal Breast Tissue: A Convolution Neural Network Classifier with Spatial Domain and Texture Images." *IEEE Transactions on Medical Imaging* 15, no. 5 (1996): 598-610. https://doi.org/10.1109/42.538937.

Scuccimarra, Eric Antoine. "ConvNets for Detecting Abnormalities in DDSM Mammograms." Medium. Last modified May 21, 2018. Accessed November 10, 2021. https://medium.com/@ericscuccimarra/convnets-for-classifying-ddsm-mammograms-1739e0fe8028.

Smithuis, Robin, and Ruud Pijnappel. "Differential of Breast Calcifications." Radiology Assistant. Last modified May 11, 2008. Accessed November 10, 2021. https://radiologyassistant.nl/breast/calcifications/differential-of-breast-calcifications.

Tsochatzidis, Lazaros, Lena Costaridou, and Ioannis Pratikakis. "Deep Learning for Breast Cancer Diagnosis from Mammograms—A Comparative Study." *Journal of Imaging* 5, no. 3 (March 13, 2019): 37. Accessed November 10, 2021. https://doi.org/10.3390/jimaging5030037.