# Mammography Image BI-RADS Classification Using OHPLall

Bob Vanderheyden
IBM
Acworth, GA, USA
rvanderh@students.kennesaw.edu

Ying Xie
*Dept. of Information Technology*
*Kennesaw State University*
Kennesaw, GA, USA
ying.xie@kennesaw.edu

*Abstract*— **Medical image analysis and classification, using machine learning, particularly Convolutional Neural Networks, have demonstrated a great deal of success. Research into mammography image classification tended to focus on either binary outcome (malignancy or benign) or nominal (unordered) classification for multiclass labels [1]. The industry standard metric for radiologist's classification of mammography images is a rating scale called BI-RADS (Breast Imaging Reporting and Data System), where values 1 through 5 are a distinct progression of assessment that are intended to denote higher risk of a malignancy, based on the characteristics of anomalies within an image [1][2][3]. The development of a classifier that predicts BI-RADS 1-5, would provide radiologists with an objective second opinion on image anomalies. In this paper, we applied a novel Deep Learning method called OHPLall (Ordinal Hyperplane Loss - all centroids), which was specifically designed for data with ordinal classes, to the predictions of BI-RADS scales on mammography images. Our experimental study demonstrated promising results generated by OHPLall and great potential of using OHPLall models as a supplemental diagnostic tool.**

***Keywords—ordinal hyperplane loss, ordinal classification, deep learning, machine learning, mammography, BI-RADS***

## I. Introduction

The American Cancer Society reports that, in 2017, over 300,000 people in the United States were diagnosed with breast cancer and over 40,000 people died from the disease. Due to improvements in treatment and early detection, the death rates that are attributed to breast cancer have declined 39% from 1989 to 2015 [1]. Developing a mammography image BI-RADS classifier that provide radiologists with an additional tool for early detection of breast cancer may help save 1,000's of lives per year [2].

Much of the work in mammography classification focuses on a single binary outcome (malignant-benign) or take the approach that used to analyze "multi-class" label data, where a label with N different classes is recode into an Nx1 vector of N-1 0's and a single value of 1. The first would require an over-simplification of the BI-RADS classification problem, while the second doesn't include the ordering information of the BI-RADS labels/classes [3].

A third approach uses "Ordinal Regression" which is essentially modification of the multi-class DNN approach [9]. In this approach, a single Deep Neural Network is used to predict the classes. Their approach is very similar to a multi-label classification problem using a DNN, where multiple outputs are estimated with all elements of the output layer being the value from a sigmoid function. To set up the analysis for $k$ ordinal classes, the label value for each record is recoded into a $k-1$ length vector. For a given class value, 'a,' all index values of the vector with position value (using the standard 0 index value for the 1st position in the vector) that are less than 'a' minus the minimum ordinal value are coded with a 1. All other values are coded with a zero [9].

The three ordinal class case, with ordinal values '1', '2' and '3', is illustrated below (Table 1). For the three-class problem, the neural network essentially estimates two binary models. The first output predicts the likelihood that the label is greater than '1', and the second predicts the likelihood that the label is greater than '2.' Once the algorithm converges or reaches a predefined stopping point, a classification rule, typically whether or not the value is greater than 0.5, converts each output vector into a binary array that is similar to the one used for training. Ordinal classes are assigned based on which encoded vector that the binary output matches. If the first position is zero then the record is assigned the value of the minimum label [4].

*Table 1 Ordinal Regression Three Class Label Encoding*

| Label | Vector |
|---|---|
| 1 | [ 0, 0 ] |
| 2 | [ 1, 0 ] |
| 3 | [ 1, 1 ] |

It should be noted that, while the vast majority of class predictions will conform to one of the vector values of the encoded ordinal classes, it is possible for vector values that do not conform to exist. In the three-class problem, it is possible that a prediction of '[0, 1]' results from applying the resulting model to a data record (either in the training set, a test or validation set or to completely new data). It is left to the analyst to determine how to classify these nonconforming results.

In [5], we introduced a novel loss function called Ordinal Hyperplane Loss (OHPL) that was specifically for ordinal classification. OHPL first uses a set of parallel hyperplanes to represent samples in a feature space. Then each class can be represented using the centroid of all the hyperplanes that belong to this class (denoted as Hyperplane Centroid). Based on the definition of Hyperplane Centroid, OHPL further quantifies both the discrepancy between the ordering of hyperplane centroids in the feature space and the given ordinal relationship among the classes, and the relative closeness of each sample towards the Hyperplane Centroids of the classes that this sample doesn't belong to. We further developed a deep learning strategy called OHPLNet that learns to map data from its original feature space to an optimal feature space where the Ordinal Hyperplane Loss is minimized. In [5], experimental studies showed that OHPLNet consistently outperforms other ordinal classification methods on multiple data sets.

However, in the formulation of OHPL that is described in [5], the hyperplane centroid ordering is applied to the full training set within each learning iteration, which makes this approach difficult to scale to large data sets, especially in a computing environment with limited computing resources. In one of our experimental studies, we laid OHPLNet upon a simple CNN to perform ordinal classification on medical images using a computer with a NVIDIA 1080 Ti GPU that has 10 GB of GPU memory. In this experiment, the maximum number of images that can be processed in a training batch is 500. In order to apply OHPL strategy to mammography image classification, we proposed in this paper an enhanced OHPL version that is called OHPLall. OHPLall is able to effectively assess the loss that is caused by improper ordering of Hyperplane Centroids in the feature space by using mini-batch of data that most likely only contains small numbers of samples from a given class. Our experimental results showed that the performance of mammography image classification using OHPLall is promising and better than ordinal regression.

The rest of the paper is organized as follows. In section II, we review the basic concepts of our OHPL. Then we briefly describe OHPLNet, a deep learning strategy using OHPL in section III. In section IV, our new development of the OHPLall strategy is presented. We further applied OHPLall to ordinal classification on mammography image in section V and analyzed the experimental results in section VI. Finally, we conclude our paper in section VII.

## II. Basic Concepts of OHPL

As the name implies that Ordinal Hyperplane Loss (OHPL) uses ordered linear hyperplanes, as the basis for calculating the loss for data distribution in a feature space. The loss function is designed to utilize simple scalar distance calculations, combined with a standard application of large margin loss. The loss function enables the use of stochastic gradient descent, in optimizing data transformations.

A linear hyperplane can be expressed as a simple mathematical equation of the form: $\boldsymbol{w}^T\boldsymbol{x} + c = 0\mathrm{w}^\mathrm{T}\mathrm{x} + \mathrm{c} = 0$, where $\boldsymbol{w}$ w and $\boldsymbol{x}$x are vector valued and c is a scalar constant. A set of parallel hyperplanes of this form differ in their c values. As a direct consequence, the 'distance' between two parallel hyperplanes can be defined to be the absolute value of the difference in their c values divided by $|\boldsymbol{w}|$. Given $\boldsymbol{w}$ w**,** we further denote the hyperplane that goes through the ith data point $\boldsymbol{x_i}\mathrm{x_i}$ can be defines as the set of points satisfying:

$$\boldsymbol{w^T x} + c_i = 0 \quad (1)$$

then bring $\boldsymbol{x_i}$ into (1), we have

$$\boldsymbol{w^T x_i} + c_i = 0 \quad (2a)$$

$$c_i = -\boldsymbol{w^T x_i} \quad (2)$$

further bring (2) into (1), we have the expression of the hyperplane that goes through $\boldsymbol{x_i}$

$$\boldsymbol{w^T x} - \boldsymbol{w^T x_i} = 0 \ (\mathbf{3})$$

Given the hyperplanes going through each data point in a feature space, we can now represent a class in that feature space by calculating its **Hyperplane Centroid (HC)**. For instance, the hyperplane centroid for the kth class, denoted as $HC_k\mathrm{HC_k}$, can be expressed as

$$HC_k\!: \boldsymbol{w^T x} - \frac{1}{n_k}\sum_{y_i=k} \boldsymbol{w^T x_i} = 0 \quad (4)$$

Given the definition in (4), all ordinal classes are represented as a group of hyperplane centroids, which are parallel to each other, in the feature space. Now we define OHPL, such that we can quantify the loss in a data distribution that is produced by a data transformation$\boldsymbol{\varphi(x)}$ $\varphi(x)\varphi(x)$ with respect to a given vector $\boldsymbol{w}$ w. According to the intuitive criteria of an optimal data distribution that are described in section 3.1, OHPL consists two components, namely Hyperplane Centroid Loss and Hyperplane Point Loss. Hyperplane Centroid Loss reflects the loss caused by non-optimal ordering of Hyperplane Centroids per the ordinal relationship of the classes, while Hyperplane Point Loss reflects the loss caused by non-optimal relationship between individual data points and the hyperplane centroids of their classes.

### *1) Hyperplane Centroid Loss(HCL)*

Hyperplane Centroid Loss (HCL), the first component of OHPL, ensures that the hyperplane centroids are properly ordered, per the ordering of the classes. This ordering can be expressed as a difference in adjacent HCs. If the adjacent HCs are properly ordered, then the transitive property ensures that all HC's are properly ordered. Therefore, we require that the HCs for adjacent classes *k* and *k+1* adhere to: $\boldsymbol{HC_{k+1} - HC_k > \delta}$ $HC_k - HC_{k+1} > \delta$ , for $\boldsymbol{\delta > 0.}$ This means, if $HC_{k+1}$ is at least $\delta$ from $HC_k$

$HC_{k+1}$ is at least $\delta$ $d$istance from $HC_k$, then the ordering is correct with sufficient distance between the adjacent classes. Since the difference is unbounded from above, this formulation doesn't introduce a distance assumption. Given adjacent classes *k* and *k+1*, and $\delta > 0$ $\delta > 0$,the Hyperplane Centroid Loss contribution of $HC_k$$\mathrm{HC_k}$ relative to $HC_{k+1}$$\mathrm{HC_{k+1}}$ is defined as:

$$HCL = \sum_{i=1}^{k-1} \max(HC_i - HC_{i+1} + \delta, 0) \quad (5)$$

*2) Hyperplane-Point Loss (HPL)*

The second component of OHPL is "Hyperplane-Point Loss" (HPL). In calculating this loss component, individual data points are compared to a specific set of Hyperplane Centroids, to access the point's contribution to the loss of the data distribution. HPPL is actually, the sum of two analogous loss functions, that work in different "directions" a la the formulation of (5).

For the points, in a given class, if we "look" in the "increasing" direction (direction of larger ordinal class value), we only want the points that are higher than the HC for the point to potentially contribute to the loss (those below will be examined later). For points that are above their HC, but are already sufficiently close to their HC, there isn't much benefit in drawing them closer, so we want their loss contribution to be zero. Therefore, the HPL uses a margin to ensure that points that do not contribute to loss are closer to their HC than the midpoint between the HC. In *Figure 1*, the circled points are higher than the margin above its HC, so they contribute to the total HPL value. Note that the dotted margin line/threshold is closer to the HC, than to the adjacent HC.

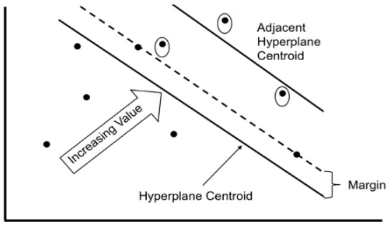


Figure 1. Computing HPL in Increasing Direction

Similarly, when we look in the decreasing direction, points that are further from their HC than the margin, will contribute to the HPL total. In *Figure 2*, the three circled points contribute to HPL.

The two components of the HPL (an increasing and a decreasing) that are summed to arrive at the total HPL. Formally, given a dataset $\mathrm{S}$$\mathrm{S}$, let $\gamma$ to be the proportion of distance between adjacent HCs, HC be the hyperplane centroid that represents the class that $x_i \in S$ $\mathbf{x_i} \in \mathrm{S}$ belongs to, $\mathrm{HC_{+1}}$$\mathrm{HC_{+1}}$ is the higher hyperplane centroid that is adjacent to HC, and $\mathrm{HPL_i^+}$ $\mathrm{HCL_i^+}$ be the HPL for the point $\mathbf{x_i} \in \mathrm{S}$ $x_i \in S$in the increasing direction, then we have:

$$0.5 < \gamma < 1.0$$

$$point\ margin = \gamma(HC_{+1} - HC)$$

$$HPL_i^+ = \max\big((f(\boldsymbol{x}_i) - HC) - (HC_{+1} - HC) + \gamma(HC_{+1} - HC), 0\big)$$

$$= \max\,(f(\boldsymbol{x}_i) - \gamma HC - (1-\gamma)HC_{+1}, 0)$$

Similarly, in the decreasing direction,

$$HPL_i^- = \max\,(\gamma HC - f(\boldsymbol{x}_i) + (1-\gamma)HC_{-1}, 0)$$

Then, the overall HPL will be the aggregation of $(HPL^+ + HPL^-)$ over all data points in $S$.

$$HPL = \sum_{x_i \in S} HPL^+ + HPL^- \quad (6)$$

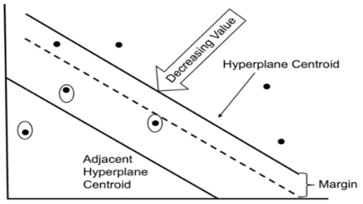


Figure 2. Computing HPL in Decreasing Direction

*3) Ordinal Hyperplane Loss (OHPL)*

Finally, the Ordinal Hyperplane Loss (OHPL) is defined as the weighted aggregation of HCL and HPL, as shown below, where $\alpha \geq 1$ reflects the importance of HCL in OHPL with respect to HPL.

$$OHPL = \alpha HCL + HPL \quad (7)$$

## III. OHPLNet: The deep learning strategy using OHPL

Given the definition of OHPL, this section describes a deep NN architecture for ordinal classification based on OHPL. Figure 3 shows a simple deep neural network (DNN) model that represents a non-linear transformation $\phi$ that maps input data from their original space to a n-dimensional space. We further add the last layer $\boldsymbol{w^T\phi(x)}$$w^T\phi(x)$ on the top of the transformation $\boldsymbol{\phi(x)}$.$\phi(x)$. Then we use the weights of the last layer, namely $\boldsymbol{w}$ $\mathrm{w}$ , to define $m$ parallel hyperplanes to represent $\mathrm{m}$ ordinal classes, such that the *kth* class will be represented by the hyperplane whoes expression is shown in (4).

Based on the hyperplane representations of the ordinal classes, we can calculate the Ordinal Hyperplane Loss (OHPL) based on the formula (7). Then the DNN can learn both an optimal transformation $\boldsymbol{\phi}$$\phi$ and an optimal vector $\mathrm{w}$ by minimizing the OHPL (recall that $\mathrm{w}$ determines the direction of those parallel hyperplanes in the feature space that is mapped by $\boldsymbol{\phi}$.

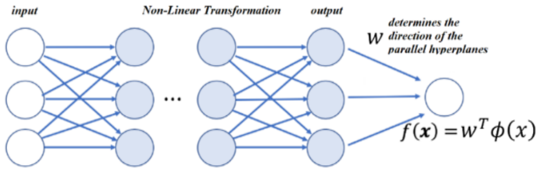

input
Non-Linear Transformation
output
W
determines the direction of the parallel hyperplanes
$f(\boldsymbol{x}) = w^T\phi(x)$

Figure 3: OHPLNet

In order to facilitate the application of OHPL deep learning strategy on different types of data for ordinal classification, we further brand this architecture as OHPLNet, a deep architecture that users can directly apply to their ordinal classification problems. Formally, an OHPL-Net contains two components. The first component is called $\boldsymbol{\phi}$ layers, which are fully connected deep nets that represents a non-linear transformation of the input data. The second component is called Hyperplane layer, which is a one-layer one-output neuron network representing the direction of Hyperplane Centroids. Again OHPLNet uses OHPL to learn optimal $\boldsymbol{\phi}$ and optimal parallel hyperplanes. If users' classification tasks involve unstructured data, OHPLNet can be put upon those deep neuron architectures that are built on specific unstructured data, such as Convolutional Neuron Network (CNN) [6] [7] on image data as shown in Figure 3 and Recurrent Neural Network (RNN) [8] on text data as shown in Figure 4.

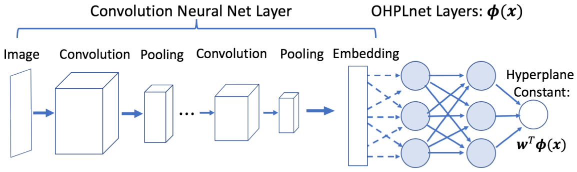

Convolution Neural Net Layer
OHPLnet Layers: $\boldsymbol{\phi}(\boldsymbol{x})$
Image
Convolution
Pooling
Convolution
Pooling
Embedding
Hyperplane Constant:
$\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x})$

Figure 3. OHPLNet upon CNN for Image Ordinal Classification

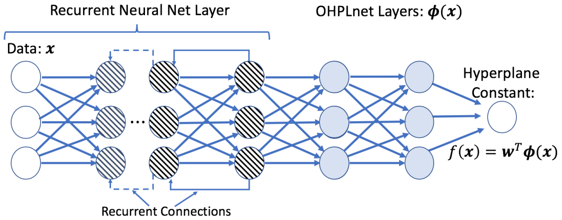

Recurrent Neural Net Layer
OHPLnet Layers: $\boldsymbol{\phi}(\boldsymbol{x})$
Data: $\boldsymbol{x}$
Hyperplane Constant:
$f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x})$
Recurrent Connections

Figure 4. OHPLNet upon RNN for Sequential Data Ordinal Classification

## IV. Further Development of OHPLall

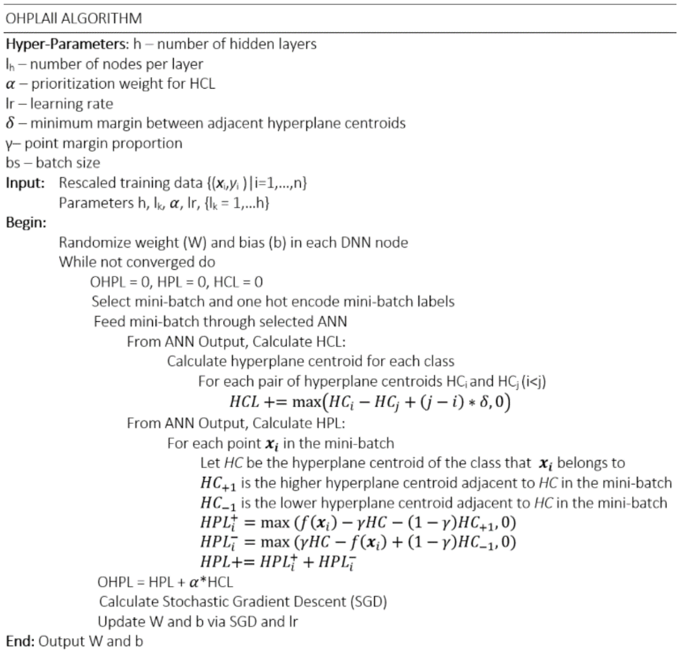
The initial work on OHPL provided a meaningful improvement over the best ordinal classifiers that are available today [5], but the methodology had some concerns that needed to be addressed. All of the benchmark data sets were relatively small in size, so the initial algorithm design was able to use the entire dataset, for calculating the hyperplane centroids for each batch submission. Since the design for that part of the algorithm used straightforward matrix operations on structured data, the conceptual investigation could be conducted without concern for that the standard benchmark datasets were too large to run in a single pass. To apply the OHPL strategy to large dataset, algorithmic changes were going to be required. Possible directions of changes include 1) incorporating efficient matrix multiplication that could be distributed to multiple computing nodes; 2) developing effective mini-batch variant of OHPL such that the loss caused by improper ordering of Hyperplane Centroids in the feature space can be assessed using mini batches of data that most likely only contains small amount of samples from partial classes. In this development, we adopted the second direction, not only because mini-batch based deep neural nets have a solid history of providing improved generalizability over large data set, but also the scalability brought by mini-batch learning strategy is far less subject to the restriction of computing powers. In order to develop mini-batch variant of OHPL, we made the following changes on the HCL estimation.

1) With the original OHPL that works on the whole data set, class labels were used to calculate an integer "distance" between adjacent hyperplane centroids. However, with mini-batch strategy, not all class labels may appear in a mini batch. For example, if the full dataset contained six distinct class labels, '0'-'6', but a mini-batch only contained records with values '2' and '4', then instead of requiring a minimum one-unit distance between the respective hyperplane centroids, the threshold will be set to $(4 - 2) * \delta$.

2) Instead of relying on the transitive property that the original OHPL uses to calculate HCL based on adjacent hyperplane centroids, we propose a new way to calculate the HCL loss component in a mini-batch by comparing all classes that were represent in the mini-batch to each the other classes within the batch. The new formulation for HCL is shown in equation (8).

$$HCL = \sum_{i<j} \max\left(HC_i - HC_j + (j - i) * \delta, 0\right) \quad (8)$$

Given that this new mini-batch strategy compares every pair of classes that appear in a mini batch for HCL estimation (instead of just comparing adjacent pairs in the original formation of OHPL), we call this strategy OHPLall. The complete OHPLall algorithm is given in Figure 5.



```
OHPLAll ALGORITHM
Hyper-Parameters: h – number of hidden layers
$l_h$ – number of nodes per layer
$\alpha$ – prioritization weight for HCL
lr – learning rate
$\delta$ – minimum margin between adjacent hyperplane centroids
γ– point margin proportion
bs – batch size
Input:  Rescaled training data $\{(\boldsymbol{x}_i, y_i) | i=1,\ldots,n\}$
        Parameters h, $l_k$, $\alpha$, lr, $\{l_k = 1,\ldots h\}$
Begin:
        Randomize weight (W) and bias (b) in each DNN node
        While not converged do
            OHPL = 0, HPL = 0, HCL = 0
            Select mini-batch and one hot encode mini-batch labels
            Feed mini-batch through selected ANN
                From ANN Output, Calculate HCL:
                    Calculate hyperplane centroid for each class
                        For each pair of hyperplane centroids $HC_i$ and $HC_j$ (i<j)
                            $HCL \mathrel{+}= \max(HC_i - HC_j + (j-i) * \delta, 0)$
                From ANN Output, Calculate HPL:
                    For each point $\boldsymbol{x}_i$ in the mini-batch
                        Let $HC$ be the hyperplane centroid of the class that $\boldsymbol{x}_i$ belongs to
                        $HC_{+1}$ is the higher hyperplane centroid adjacent to $HC$ in the mini-batch
                        $HC_{-1}$ is the lower hyperplane centroid adjacent to $HC$ in the mini-batch
                        $HPL_i^+ = \max(f(\boldsymbol{x}_i) - \gamma HC - (1-\gamma)HC_{+1}, 0)$
                        $HPL_i^- = \max(\gamma HC - f(\boldsymbol{x}_i) + (1-\gamma)HC_{-1}, 0)$
                        $HPL \mathrel{+}= HPL_i^+ + HPL_i^-$
            OHPL = HPL + $\alpha$*HCL
            Calculate Stochastic Gradient Descent (SGD)
            Update W and b via SGD and lr
End: Output W and b
```

Figure 5. OHPLall Algorithm

## V. Mammography Images Classification Using OHPLall

Radiologists use the first six categories, of the seven-point BI-RADS (Breast Imaging Reporting and Data System) rating system to classify mammography images. The seventh category is used for images that are of breasts with a known malignancy, that was confirmed via a biopsy. The zero category is used, for images where classification is uncertain and additional information is required. Categories one through six are a sequence of ordinal classes [2].

*Table 2 BI-RADS Category Scale [2]*

| Category | Definition |
|---|---|
| **0** | Additional imaging evaluation and/or comparison to prior mammograms is needed. |
| **1** | Negative |
| **2** | Benign (non-cancerous) finding |
| **3** | Probably benign finding – Follow-up in a short time frame is suggested |
| **4** | Suspicious abnormality – Biopsy should be considered |
| **5** | Highly suggestive of malignancy – Appropriate action should be taken |
| **6** | Known biopsy-proven malignancy – Appropriate action should be taken |

While the BI-RADS rating scale has seven classes, only six of the classes (1-6) are ordered (see Table ). In addition, a rating of 6 is only used when the results of a biopsy of the abnormality confirms a malignancy. As such, it wouldn't be used at the time when the radiologist was reading the images.

*Table 3: CBIS-DDSM Annotations* [9] [10]

| Annotation | Relation to Scan Event | Definition/Values |
|---|---|---|
| Side | Prior to | Left or right breast |
| View | Prior to | CC - craniocaudal<br>MLO - mediolateral oblique |
| Density Rating | Prior to | Breast density rating |
| Abnormality Type | After | Calcification (2 annotations) – Type and distribution<br>Mass (2 annotations) – shape and margin |
| Assessment | After | BI-RADS rating (0, 2-5) |
| Pathology | After Image Assessment | Benign Without Callback<br>Benign<br>Malignant |

The Cancer Imaging Archive (TCIA) contains a database of mammography images, called CBIS-DDSM (Curated Breast Imaging Subset of DDSM). The database contains over 2,600 images selected and annotated by trained mammographers (Table 3 shows the format of the annotation). Released in 1997, they remain a valid source of curated mammography data for public research into mammography classification [10] [11]. Several studies, analyzing the CBIS-DDSM data, were published in the past year or two, reporting a variety of classification algorithms, that demonstrate good success in using DNN's [12] [13]. The goal of this research is to analyze mammography images from the CBIS-DDSM database that have been classified by radiologists, to build an image classification model using OHPLall to predict BI-RADS categories two through five. The CBIS-DDSM database contains three types of images, which differ by size:

1. Full mammography images
2. Images that are cropped for standardization for use in computer-aided diagnosis and detection (CADx and CADe, respectively). Regions of interest are at the centroid of the image.
3. Regions of Interest (ROI) images are smaller images that focus more directly on the abnormality.

The ROI scans vary in size from 70 pixels to 3,000 pixels but are heavily skewed to under 1,000 pixels per side. Due to their relatively small size, these images are better choices for analysis on a desktop or laptop. Therefore, in this experimental study, we used the ROI images. The ROI images have two perspectives (Mediolateral Oblique or MLO and Craniocaudal or CC). Both perspectives were used in this research. Images are also split into Calcification and Mass subsets. For this analysis, the Calcification images are used.

For images larger than 1,024 pixels per row or column, the image outer regions are cropped to result in a maximum pixel size of 1,024 in each dimension. Images that are smaller than 1,024 pixels in their rows or columns are zero padded on the outer edges (equally on both dimensions). After pre-processing, we have1,423 images of 1,024x1,024 pixels with a BI-RADS rating of 2-5 in the training set. The database comes with an independently selected testing set. We use 306 of them with BI-RADS rating of 2-5 as our testing set.

Since BI-RADS rating (the class labels for predicting) is supposed to strongly associated with the incidents of a malignant abnormality, assessing the model results, in terms of malignancy rates, should be an interesting and useful analysis. If a model performs below expectations, but provides an appropriate distribution of malignancy cases, it may be even more useful, as a diagnostic tool, to supplement a radiologist's findings. In Table , we see that for the training set, BI-RADS 2 and 5 have expected incidence rates for malignancy. BI-RADS 3 and 4 are likely not the desired distribution. For the test set, the images seem to almost be randomly assigned in terms of malignancy incidence. This test set will be a meaningful challenge in terms of evaluating malignancy rates in the resulting classifiers.

To apply OHPLall strategy to this mammography ROI images to predict BI-RADS, we lay a 4-layer OHPLNet on the following CNN model as shown in Figure 6. For comparison, we also use the same CNN model for Ordinal Regression that is described in [4].

*Table 4 Image Counts by BI-RADS Rating*

| BI-RADS Rating | Test Pct Images with Malignancy | Number of Images | Training Pct Images with Malignancy | Number of Images |
|---|---|---|---|---|
| 2 | 35.2% | 62 | 0.2% | 473 |
| 3 | 30.4% | 81 | 35.5% | 84 |
| 4 | 40.9% | 115 | 39.9% | 742 |
| 5 | 36.1% | 48 | 98.5% | 124 |
| Total | 38.2% | 306 | 29.3% | 1,423 |

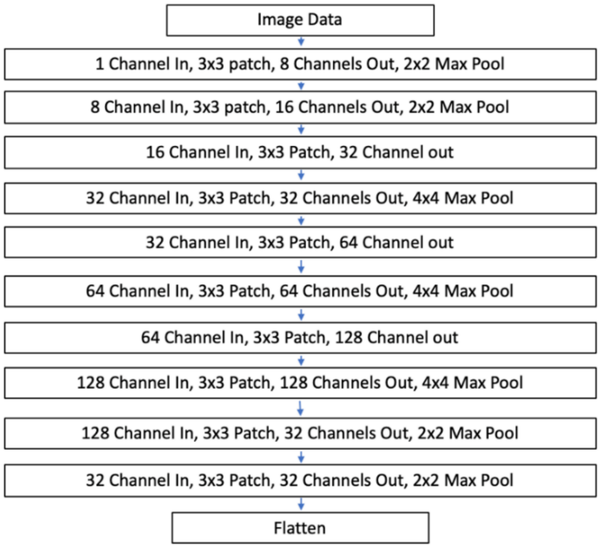


Image Data
1 Channel In, 3x3 patch, 8 Channels Out, 2x2 Max Pool
8 Channel In, 3x3 patch, 16 Channels Out, 2x2 Max Pool
16 Channel In, 3x3 Patch, 32 Channel out
32 Channel In, 3x3 Patch, 32 Channels Out, 4x4 Max Pool
32 Channel In, 3x3 Patch, 64 Channel out
64 Channel In, 3x3 Patch, 64 Channels Out, 4x4 Max Pool
64 Channel In, 3x3 Patch, 128 Channel out
128 Channel In, 3x3 Patch, 128 Channels Out, 4x4 Max Pool
128 Channel In, 3x3 Patch, 32 Channels Out, 2x2 Max Pool
32 Channel In, 3x3 Patch, 32 Channels Out, 2x2 Max Pool
Flatten

Figure 6. CNN model used in the experimental study

## VI. Experimental Results

Assessment of ordinal class labels is done using two standard methodologies. Ordinal Class problems use Mean Zero Error (MZE), instead of using traditional accuracy (proportion of records that are correctly classified). MZE is the simple ratio of the number of misclassified records and the total number of records. Mean Absolute Error (MAE) is the other key metric. In MAE is calculated by taking the sum of the absolute differences between actual label value and the predicted labels and dividing by the number of records. Two classifiers that perform comparably on MZE may not do so on MAE. In that case, the classifier with the lower MAE performs better.

Twenty OHPLall models and twenty Ordinal Regression models were generated through twenty executions for each. For OHPall, the mean batch training error, for an epoch is a reasonable metric to use for as a stopping criteria. As can be seen, in Figure 7, mean batch error values that are below 0.5 results in low test set MZE and MAE. While higher mean batch error values may have low test set MZE and MAE, they may also have higher than desired test set MZE and MAE values.

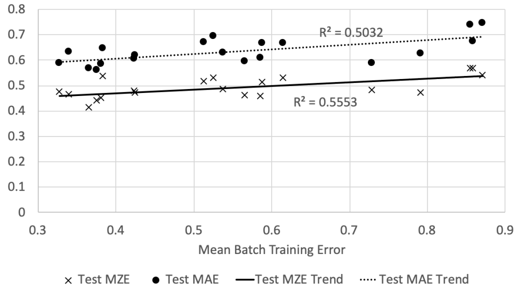


Figure 7 Training Data Mean Batch Error

As can be seen in Table 7, Ordinal Regression MZE and MAE are 25% and 43% higher, respectively, than OHPLall, on the mean values of 20 executions of each algorithm (see Table ). In addition, the MAE values for Ordinal Regression had double the standard deviation for MAE as OHPLall.

*Table 5 OHPLall vs Ordinal Regression MAE and MZE Results*

| Algorithm | Metric | MZE | MAE |
|---|---|---|---|
| OHPLall | Mean | **0.473** | **0.612** |
| | Std Dev | **0.033** | **0.046** |
| Ordinal Regression | Mean | 0.595 | 0.877 |
| | Std Dev | 0.041 | 0.099 |

In addition to assessing standard model performance metrics, it is also worthwhile to assess class predictions relative to biopsy results for the calcifications. For this evaluation, a single well performing model for each algorithm is examined. Table  reports the MAE and MZE for the selected models. If a model "struggles" to properly classify records within a given BIRAD rating, it is likely to be desirable for the errors to occur in the lower rating values and perform better in the higher ratings leading to early treatment for a malignancy. Both models perform poorly on BI-RADS '3' and '5' rated images. From the table it is clear to see that Ordinal Regression does a good job, with BI-RADS '2' rated records, but performs poorly, relative to OHPLall in the other three classes (to the point that MAE for OHPLall is roughly equal to MZE for Ordinal Regression).

*Table 6 Rating Level Assessment for OHPall Model and Ordinal Regression Model*

| BI-RADS | OHPLall MZE | OHPLall MAE | Ord Reg MZE | Ord Reg MAE |
|---|---|---|---|---|
| 2 | 0.408 | 0.732 | 0.211 | 0.338 |
| 3 | 0.696 | 0.739 | 0.739 | 0.826 |
| 4 | 0.324 | 0.386 | 0.574 | 0.767 |
| 5 | 0.750 | 0.944 | 0.889 | 1.417 |
| Total | 0.422 | 0.559 | 0.539 | 0.748 |

Per the BI-RADS definitions, it is expected that malignancy rates would increase with BI-RADS score. The algorithm that produces models that best meet this expectation would provide higher quality predictions. As shown in Table 8, Ordinal Regression predicted a significant shift in BI-RADS rating, towards the low end of the scale, resulting in very good MZE and MAE values for the '2' class, but poor results for the other classes. In addition, images classified as a '5' by OHPLall have over three times the Malignancy Rate (percent of images that were ultimately classified as malignant) as Ordinal Regression. Early identification of malignancy is critical in treating breast cancer, so this skew towards lower values versus OHPLall is less desirable for a model that is intended to be used as a diagnostic tool.

*Table 8 Malignant Counts (Ratio) for Both a High Performing OHPLall Model and a High Performing Ordinal Regression Model*

| BI-RADS | Actual Malignant Counts (Ratio) | OHPLall Malignant Counts (Ratio) | Ord Reg Malignant Counts (Ratio) |
|---|---|---|---|
| 2 | 0 (0.0%) | 7 (12.5%) | 44 (42.7%) |
| 3 | 16 (69.6%) | 28 (48.3%) | 29 (35.8% |
| 4 | 65 (36.9%) | 69 (38.8%) | 40(37.4%) |
| 5 | 36 (100.0%) | 13 (92.9%) | 4(26.7%) |

The image database also contained a number of images with a BI-RADS classification of '0'. This class is designated as "Additional imaging evaluation and/or comparison to prior mammograms is needed". While a specific rating value is not available, the models can be assessed based on the malignancy rates for the predicted classes. As was the case for the test dataset, relative to OHPLall, Ordinal Regression shifts cases to the lower end of the rating scale, as demonstrated in table 9. This skew towards the lowest available BI-RADS class includes a shift of nine malignant cases, to the '2' class, giving this Ordinal Regression a higher malignancy rate than the rates for the other three classes. OHPLall classifies two malignant cases into class '2'. OHPLall classifies over 2/3 malignant cases into classes '4' and '5', while Ordinal Regression classifies just over half of the malignant cases into class '4' and no malignant cases into class '5'. The OHPLall results are more consistent with the overall definitions of the BI-RADS measurement system.

*Table 9 Results for '0' Rated Cases*

| BI-RADS | OHPLall Counts | OHPLall: Malignant Counts | Ordinal Regression Counts | Ordinal Regression Malignant Counts |
|---|---|---|---|---|
| 2 | 2 | 2 | 15 | 9 |
| 3 | 14 | 8 | 15 | 7 |
| 4 | 42 | 16 | 40 | 17 |
| 5 | 13 | 7 | 1 | 0 |

In summary, for the classification of the available mammography images into BI-RADS rating, a Convolutional Neural Network that uses OHPLall loss provides better results than a Convolutional Neural Networks that use Ordinal Regression. Not only does it provide better overall results, but in the critical secondary assessment OHPLall works well in predicting images that have a malignancy into higher BI-RADS classes.

## VII. Conclusions

In this paper, we presented our continuous development of OHPL, a loss function specifically designed for ordinal data that enables deep learning to be applied for ordinal classification, into a new enhanced version that is called OHPLall. Instead of requiring the whole training datasets, OHPLall uses mini batches to effectively assess the ordering of the classes and the relative closeness of a sample towards its own class in a feature space. Deep learning strategy using OHPLall as the loss function is more scalable to large data sets than the original OHPL.

We further applied OHPLall to mammography image BI-RADS classification. Experimental results showed that OHPLall outperformed the Ordinal Regression approach with respect to MZE and MAE measures. By further analyzing the model results in terms of malignancy rates in each BI-RADS scale, we found that the predicted results generated by OHPLall provided a more appropriate distribution of malignancy cases among predicted BI-RADS scales than the results generated by Ordinal Regression, which demonstrated a great potential of using OHPLall as a supplemental tool in breast cancer diagnosis.

## References

[1] American Cancer Society, "Breast Cancer Facts & Figures 2017-2018," American Cancer Society, Atlanta, GA, 2017.

[2] American Cancer Society, "Understanding Your Mammogram Report," American Cancer Society, 9 October 2017. [Online]. Available: https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/mammograms/understanding-your-mammogram-report.html. [Accessed 15 March 2019].

[3] ksmith01 and Klingerc, "The Cancer Imaging Archive (TCIA) Public Access: CBIS-DDSM," The Cancer Imaging Archive (TCIA), 21 November 2018. [Online]. Available: https://wiki.cancerimagingarchive.net/display/Public/CBIS-DDSM#01fc928dcc1f420f9cd2dd80fdd37b16. [Accessed 5 March 2019].

[4] J. Cheng, "A Neural Network Approach to Ordinal Regression," 2007. [Online]. Available: http://arxiv.org/abs/0704.1028. [Accessed 5 July 2019].

[5] B. Vanderheyden and Y. Xie, "Ordinal Hyperplane Loss," in *2018 IEEE International Conference on Big Data*, Seattle, WA, 2018.

[6] G. Huang, Z. Liu , L. van der Maaten and K. Q. Weinberger, "Densely Connected Convolutional Networks,," in *2017 IEEE Conference on Computer Vision and Pattern Recognition,*, Honolulu, Hawaii, 2017.

[7] A. Karpathy, "CS231n Convolutional Neural Networks for Visual Recognition," Stanford University, Spring 2019. [Online]. Available: http://cs231n.github.io/convolutional-networks/. [Accessed 12 July 2019].

[8] A. Zhang, Z. C. Lipton, M. Li and A. J. ￼ Smola, "Dive into Deep Learning: Chapter 8.8. Gated Recurrent Units (GRU)," 2019. [Online]. Available: https://www.d2l.ai/chapter_recurrent-neural-networks/gru.html#reset-gate-in-action. [Accessed 5 July 2019].

[9] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox and F. Prior, "The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository," *Journal of Digital Imaging,* vol. 26, no. 6, pp. 1045-1057, 2013.

[10] R. S. Lee, F. Gimenez, A. Hoogi and D. Rubin, "Curated Breast Imaging Subset of DDSM," The Cancer Imaging Archive, http://dx.doi.org/10.7937/K9/TCIA.2016.7O02S9CY, 2016.

[11] 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, 2017 volume 4, Article number: 170177 (.

[12] T. Kyono, F. J. Gilbert and M. van der Schaar, "MAMMO: A Deep Learning Solution for Facilitating Radiologist-Machine Collaboration in Breast Cancer Diagnosis," *CoRR,* vol. abs/1811.02661, 2018.

[13] L. Shen, L. R. Margolies, J. H. Rothstein, R. B. McBride, E. Fluder and W. Sieh, "Deep Learning to Improve Breast Cancer Early Detection on Screening Mammography," *Clinical Orthopaedics and Related Research,* vol. abs/1708.09427, 13 August 2018.