Term Project Report:Applied Machine Learning Algorithm On Histopathology Images For Breast Cancer Prediction

Qingyang Li
School of computing and information,
University of Pittsburgh
Pittsburgh, PA 15213
Email:qil77@pitt.edu

***Abstract*—Breast cancer is the second most common cancer among women in the United States, and the major cause of death in women worldwide. With the rapid population growth, the risk of death cause by breast cancer can rise exponentially. Classification methods in machine learning has proven to be an effect way to classify data in medic field, those methods are widely used in diagnosis and analysis to make decision. An diagnosis system can aids doctors and easing the strain on medic resources. With the help of machine learning algorithms, it is non-trivial that the early detection of breast cancer can be improved and more intervention can be guaranteed before the cancer turn to the late stage. In this experiment, I compare five supervised machine learning techniques named logistic regression, support vector machine(SVM), K-nearest neighbors, random forests, convolutional neural network(CNN) in classifying breast cancer images. To counter the effect of class imbalance, I oversample the training as an attempt to achieve better model performance. Finally, I evaluate the models' performance with respect to area under receiver operating characteristic, F1-scores,recall and false negative.**

I. Introduction

Breast cancer is the second leading cause(after lung cancer) of cancer death in women. Incidence rates have increased by 0.5% per year, about 287,850 of women's new cases of invasive breast cancer are expected to be diagnosed in the US during 2022, and 43,250 of women's death is estimated.However,since 2007, breast cancer death rates have continued to decrease in older women. From 2013 to 2018, the death rate went down by 1% per year.[10] These decreases are believed give to the credit of screening and the increase awareness, as well as the improvement medic procedures in treatment.

Biopsy is one of the process in screening to diagnosis the cancer. It is a test that removes tissue or fluid from the breast and to be observed by a specialist under a microscope. And with the development of digital imaging techniques in recently years, much more attention has driven to binary classification of caners in whole mount slide images of breast cancer specimens. Artificial neural network(ANN) such as CNN has always be a trending and effective method in image classification and be widely used in radiology due to their automatic feature extraction and representation learning ability.[13]

Other than ANN, there are amount of supervised algorithms are very popular classification methods and classifies instances based on the feature's values, traditional methods such as SVM can achieve high accuracy and good performance in breast cancer prediction with texture features[1]. however, histopathology image classification are often lack of annotation in dataset.

In this research project, I focus on traditional supervised algorithms techniques compared with convolutional neural network, analysis their performance based on histopathology image classification, and derive the dataset with oversampled instances, applied to the models and observe the influences.

II. Related work

There are numerous methods on applying classification methods to the histopathology images, most of them are deep learning or neural network based, which is already popular in this fields. Thus, my search on related work mostly focus on the viability of traditional supervised algorithms on histopathology images classification.

A. Penalized logistic regression with data enhancement

Decompose the origin images data to 28x3 sub-images, Then, nine of the traditional statistical standards (mean, mean absolute deviation, median absolute deviation, standard deviation, entropy, energy, skewness, kurtosis, root mean square) are extracted from every sub-image. As a result, 756 features have been obtained from each histopathology images ,applied filter to the features matrix and as the input of penalized logistic regression,the penalized is to add a non-negative penalty term to the log-likelihood of logistic regression function. Final result achieve the overall 90% accuracy rate on the test data.[7]

B. Using Bag of Features and Kernel Functions with SVM

Bag of features is a feature detection framework and an adaption of the bag of words scheme, original used for text categorization and text retrieval, to extract the features from

image representation, and applied kernel functions for non-linearly to the SVM. Has the overall 0.673 precision, 0.115 recall and 0.237 F-measure in model performances.[3]

C. Detect Invasive Ductal Carcinoma in Histopathology Images Using K-Nearest Neighbor Classifiers

ORB (Oriented FAST and Rotated BRIEF) for feature extraction,ORB, also known as Oriented FAST Rotated BRIEF, was first presented in 2011 for computer vision tasks such as object recognition, detection, and matching. After feature extraction, use the features as input. In this study K = 3 resulting to Accuracy = 0.7071, Precision = 0.2563, and Recall = 0.0131. [9]

Based on the above relative works, it is possible to use traditional supervised algorithms for image classification, and with the help of specific feature extraction algorithms, traditional supervised algorithms can achieve non-trivial results on image classification.

III. Methodology

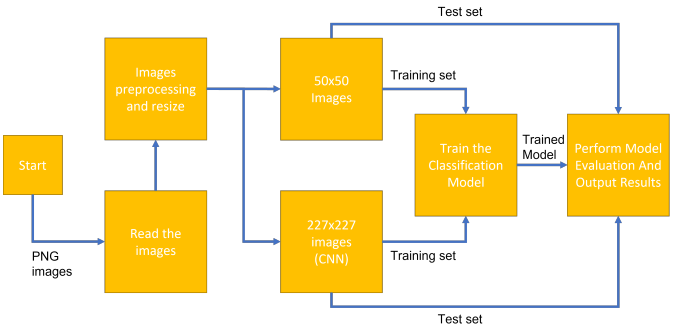

Fig. 1. General Workflow

The general workflow is commonly to all ML applications, and the workflow is illustrated in Fig 1, the images are loaded and split to training and testing, and then fit to the machine learning models. For satisfy the requirement of AlexNet, which is the CNN architecture in this experiment, the size of images for CNN input need to be resized to 277x277 rather than the original 50x50.

IV. Classifiers

The most methods that used to solve binary classification problem are logistic Regression, k-nearest neighbors, random forest, support vector machine.[5] And convolutional neural network (CNN) is most commonly applied to analyze visual imagery in deep learning. To experiment with advanced ML methods, those features are applied to basis machine learning algorithms as following:

A. Logistic Regression

In this experiment, I use logistic regression as the benchmark technique, followed by support vector machine, K nearest neighbors, random forest, and AlexNet as our more advanced models. The fitting of Logistic Regressor satisfies the most stringent assumption about the dataset, test the linearity between the images and the labels. Given the multitude of the images' pixel-level data dimensional, this quality is not likely directly observed. Thus, logistic regression is most simplistic and the weakest classification that can be used as the benchmark model. I also use Lasso regularization for its ability to filter some features by assigning them near-zero weights. Overall, I try to solve the following loss function:

$$\begin{aligned} \underset{\mathbf{w}}{argmin} - \mathbf{y} log(\frac{1}{1+e^{-\mathbf{xw}}}) \\ -log\mathbf{P}(\mathbf{D}|\mathbf{w}) + \lambda \|\mathbf{w}\|_{L_1} \end{aligned} \tag{1}$$

B. Support Vector Machine

Images classification are likely to have non-linearity features. Support vector machine (SVM) can learn to create some slightly smarter decision boundaries with the help of a kernel. I choose the radial-basis function (RBF) as the kernel and use it to process the samples [4].

$$K(\mathbf{x}, \mathbf{x}) = \exp(-\frac{\|pdist(\mathbf{x}^T, \mathbf{x})\|^2}{2\sigma^2}) \tag{2}$$

The $pdist$ represents the pairwise distance function that RBF utilizes to capture the euclidean distances between every pair in the feature vector. As a Gaussian kernel, it is used to perform transformation when there is no prior knowledge about data. With the new features created by the RBF kernel, SVM classifier can make decision boundary in a higher dimensional space.

$$\underset{\mathbf{w}}{argmin} \max(0, 1 - \mathbf{y}\mathbf{x}^T\mathbf{w})^2 + \lambda \|\mathbf{w}\|_{L_2} \tag{3}$$

C. K-nearest-neighbors

KNN is a non-parametric supervised learning model when reliable parametric estimates of probability densities are unknown or difficult to determine. It is also known for its simple implementation.[12] The geometric distance between each instances are usually calculated based on euclidean distance, the rule assign the test samples to the majority category label of its training samples, and for binary classification scenarios, the K is usually chosen to be odd number to avoid ties.

$$d(\mathbf{x_i}, \mathbf{x_l}) = \sqrt{\sum_{n=1}^{p} (\mathbf{x_{in}} - \mathbf{x_{ln}})^2} \tag{4}$$

I utilized KD-tree structure to reduce the complexity, KD-tree can balance the tree by creating a median point for high dimension features. Reduce the complexity from O(MxN) to O(Nlog(n)) where O(MxN) is the complexity of brute-force KNN.

D. Random Forest

Random forest is an ensemble learning method for classification by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees.[2] As a decision-tree based

method, random forest classifier can make much more complex decision boundaries than linear models. It segregates the dataset super-space with straight lines, which could apply to our high dimensional dataset better than a linear classifier.learn the multiple decision tree can improve the overall predict performance.As non-parametric supervised learning method, Random forest can generate reasonable predictions across a wide range of data while requiring little configuration.

E. AlexNet

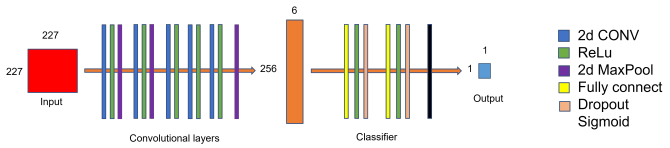

Fig. 2. AlexNet structure

As a well-constructed and regularized deep learning neural network, AlexNet has been proven to excel in image classification task consistently. The input of AlexNet should be 227x227x3 images (In the original paper, the number is 224x224x3, however that image size does not fit into the equation for convolution layer) Fig 2 outlines the structure of the AlexNet.[8]

The original AlexNet has 5 convolution layers, which is relatively insufficient for today's image classification task. But its complexity might be suitable for comparing with the traditional supervised classification algorithms without featrues extraction.

V. Experiment Settings and Results

A. dataset

The breast histopathology images in this experiment are from Kaggle.[11]. The dataset consisted of 277,524 patches of size 50x50 images that were classified as invasive ductal carcinoma (IDC) positive and negative (198,738 IDC negative and 78,786 IDC positive). Invasive ductal carcinoma is the most common subtype of all breast cancers.[6]. Loading the whole data set requires large enough memory spaces and computing powers, in order to perform experiment on my personal desktop, I decreased the samples to 10,000 patches with 7160 IDC negative and 2840 IDC positive, keep the same fraction between negative and positive labels as the whole dataset. Due to the imbalance of the original dataset, I also sampled 14,320 patches with 7160 IDC negatives and 7160 IDC positives as the oversampled data for comparing models' performance in oversampled dataset.All images are read into numpy array that contains pixel-level data by *cv2.imread()*, I resize the images to 227x227x1 as the input for AlexNet, use grey-scale images to prevent memory overflow. For the rest models, the input is flatten to one dimension array.

B. Implementation

Logistic regression is using sklearn.linear_model with lasso regularization to filter features. SVM is using sklearn.SVM with RBF kernel and ridge regularization. KNN is implemented by sklearn.neighbors with KD tree and K=101 for standard dataset,K=141 for oversampled, the K is determined by log(n) where n represents the number of instances in dataset. Random forest is implemented by sklearn.ensemble and use 250 as number of estimators, 250 is the parameter when random forest converges or has minimum changes in this experiment. And AlexNet is built by tensorflow.keras with the same structure as the AlexNet paper.

C. Analysis results

The testing methods mainly consists of extracting confusion matrix of the model tested through the test set and comparing derived results.I used Area Under Curve (AUC) of Receiver Operating Characteristic (ROC) as the main quantitative metric to compare the models. I also compare the true positive rate (recall) resulted from the model's final probability threshold and F1-score. Due to the emphasis on correctness of the cancer detection task, we additionally compared the false negatives of every model.

TABLE I
STANDARD IMAGE SET

Metrics	Models				
	Logistic Regression	**SVM**	**KNN**	**Random Forest**	**Alexnet**
AUROC	0.853	0.868	0.860	0.894	**0.9129**
Recall	0.6101	0.632	0.514	0.667	**0.7148**
F1-scores	0.648	0.683	0.629	0.689	**0.712**
False negative	354	334	441	302	**259**

The AlexNet dominant all metrics in standard image dataset and beat other models in this experiment, CNN is still the priority machine learning method to deal with images classification when facing with traditional supervised methods, due to its automatic feature extraction and representation learning techniques which is something other models do not have in native. KNN has the worst the performance, which can explain that the KNN is sensitive to the scale of data and higher dimensional data need more time to computing the data distance (In fact, running the KNN is very time consuming than other models in this experiment), even though the it still produce the poor result.

TABLE II
OVERSAMPLED IMAGE SET

Metrics	Models				
	Logistic Regression	**SVM**	**KNN**	**Random Forest**	**Alexnet**
AUROC	0.7625	0.800	0.6332	**0.854**	0.8496
Recall	0.7054	0.678	0.128	**0.805**	0.7600
F1-scores	0.698	0.717	0.2208	**0.7903**	0.7535
False negative	980	1778	2900	648	**564**

In the over-sampled image set experiment, random forest method surpassed all models with higher AUROC, Recall and F1-scores, outperformed AlexNet in this image set, shows the random forest is robust and stable to over-sampled data due to the multi decision-tree structure. KNN has the worst performance same with the previous image set due to the higher dimension. Logistic regression outperformed the SVM in this experiment, this can be attributed to the L1 regularization add more penalty to the features and RBF is not too ideal to this specific images set.

VI. Conclusion

I have applied breast histopathology images to a range of classification models, compare their performances based by AUROC score, Recall, F-scores, and the number of false negative. As CNN structure, AlexNet shows the good ability in feature extraction when dealing with image classification. KNN has the worst performance and longer computation time when predicting high dimensional dataset. However, the traditional machine learning algorithm showing the potential in image classification task. With additional help from feature selection algorithm, traditional algorithms such as SVM and logistic regression can complete with some advance convolutional neural network models.

In future work, I will try to applied some basic and advance feature selection algorithms to these models and compare them with the existing results to find how much improvement the models can get. Also I will try to apply these models with other images dataset to figure out if different image set can influence the models' performance in some way.

References

[1] AS Pitchumani Angayarkanni and B Dr Nadira Banu Kamal. "MRI mammogram image classification using ID3 algorithm". In: (2012).

[2] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[3] Juan C Caicedo, Angel Cruz, and Fabio A Gonzalez. "Histopathology image classification using bag of features and kernel functions". In: *Conference on Artificial Intelligence in Medicine in Europe*. Springer. 2009, pp. 126–135.

[4] Corinna Cortes and Vladimir Vapnik. "Support-vector networks". In: *Machine learning* 20.3 (1995), pp. 273–297.

[5] Destin Gong. *Top 6 machine learning algorithms for classification*. Mar. 2022. URL: https://towardsdatascience.com/top-machine-learning-algorithms-for-classification-2197870ff501.

[6] *Invasive Ductal Carcinoma*. URL: https://www.breastcancer.org/types/invasive-ductal-carcinoma.

[7] Mohammed Abdulrazaq Kahya. "Classification enhancement of breast cancer histopathological image using penalized logistic regression". In: *Indonesian Journal of Electrical Engineering and Computer Science* 13.1 (2019), pp. 405–410.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25 (2012).

[9] Kyra Mikaela M Lopez, Ma Sheila A Magboo, and A Sheila. "A Clinical Decision Support Tool to Detect Invasive Ductal Carcinoma in Histopathological Images Using Support Vector Machines, Naive-Bayes, and K-Nearest Neighbor Classifiers." In: *MLIS*. 2020, pp. 46–53.

[10] The American Cancer Society medical and editorial content team. *Key Statistics for Breast Cancer*. 2022. URL: https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html.

[11] Paul Mooney. *Breast Histopathology Images*. 2017. URL: https://www.kaggle.com/paultimothymooney/breast-histopathology-images.

[12] Leif E Peterson. "K-nearest neighbor". In: *Scholarpedia* 4.2 (2009), p. 1883.

[13] Rikiya Yamashita et al. "Convolutional neural networks: an overview and application in radiology". In: *Insights into imaging* 9.4 (2018), pp. 611–629.