Comparative Study of Image Classification using Machine Learning Algorithms

Maisun Al Zorgani and Hassan Ugail

July 9, 2018

Comparative Study of Image Classification using Machine Learning Algorithms

Maisun Mohamed Alzorgani
Computer Science, Engineering and Informatics,
University of Bradford
Bradford City, UK
m.m.s.alzorgani@bradford.ac.uk

Professor Hassan Ugail
Director, Centre for Visual Computing,
University of Bradford
Bradford City, UK
h.ugail@bradford.ac.uk

Abstract—**This study compared five common machine learning algorithms for performing classification included Support Vector Machines (SVM), K-Nearest Neighbor (KNN), Naïve Bayes (NB), Binary Decision Tree (BDT) and Discriminant Analysis (DA). AlexNet deep learning model was used to build these machine learning classifiers. The building classifiers were implemented and evaluated according to standard performance criteria of Accuracy (ACC), Precision (P), Sensitivity (S), Specificity (Spe) and Area Under the ROC Curve (AUC). The five methods were evaluated using 2608 histopathological images for head and neck cancer. The comparison was conducted using 2 times 10-fold cross validation. For each method, the pre-trained AlexNet network was used to extract features from the activation layer. The results illustrated that, there was no difference between the results of SVM and KNN. Both have the same and the higher accuracy than others were 99.98 %, whereas 99.81%, 97.32% and 93.68% for DA, BDT and NB, respectively. The present study shows that the SVM, KNN and DA are the best methods for classifying our dataset images.**

***Keywords-** Machine learning methods, AlexNet ConvNet, head and neck cancer.*

I. INTRODUCTION

In recent years, several research concentrate on the applications of deep learning technique for medical image Computer-aided diagnosis (CAD) methods that use deep neural networks to learn patterns of the image based on a large training data set [5]. Deep learning has been reported to significantly outperform classical machine learning methods for object classification and has been increasingly used for medical image analysis [5]. So far, the applications of deep learning for medical images include the detection and segmentation of lesions from histopathology images ([5],[13]).

The deep learning method was the convolutional neural network (CNN), which is a deep neural network dedicated for image classification, it is also named the ConvNets in some literatures ([5],[14]). According to Wang and colleagues (2017), CNN has advantages than classical machine learning methods; the advantages are [5];

- CNN does not take hand-crafted features as input; it eliminated the needs for tumor segmentation and feature selection, making the whole process much more convenient and less prone to user bias [5].
- CNN takes an image patch of n × n pixels as input and then does classification according to the pattern of the image patch; it learns the patterns of patch appearance from a large amount of training patches. The outputs of CNN are the scores for different classes, and the class with the highest score is considered as the classification result [5].
- CNN also avoids using the debated texture features which are affected by tumor size [5].
- CNN utilizes the image appearance pattern around the tumour. The appearance pattern includes information of local contrast, nearby tissues, boundary sharpness, and etc. Such information is different from but as powerful as the diagnostic features [5].
- CNN are not affected by the size of the tumour, because they are computed from the entire image patch which includes both the tumour and its surrounding tissues [5].

BDT is ensembles of decision trees. Wang and colleagues (2017) found that ensemble methods outperform other classifiers in their comparison studies. They demonstrated that the mechanism of a decision tree can utilize different features to compensate each other, and the ensemble of decision trees combines multiple weak tree classifiers into a strong classifier [5].

Wang and colleagues (2017) demonstrated that SVM belongs to the kernel-based classifier family, which implicitly maps the input features into a higher dimensional feature space using a kernel function that measures the distance between feature points in the mapped space, so SVM is able to achieve much better classification performance than conventional linear classification methods [5].

This study aimed to compare the performance of multiple machine learning methods for classifying head and neck tumour from histopathological images.

II. MATERIAL

A. Data sets

In this study, data sets were collected from two sources. The HNSCC and Normal histology images were obtained from The Ethical Tissue department at the University of Bradford, whereas Salivary Glands images from WebPathology site [8]. A total 2608 histopathological images, 1184, 1184 and 240 for HNSCC, salivary glands and normal histology, respectively. These images were stored in jpg format.

B. Data augmentation

Dosovitskiy et al. (2013) investigated the role of data augmentation in deep learning to get enough different samples which needed to train a CNN from the images [3]. The dataset images were rotated (90, 180 and 270), flipped left to right horizontally and then vertically to create a larger sample size and to make the approach recognize tumor cells in different orientations (figure 1).

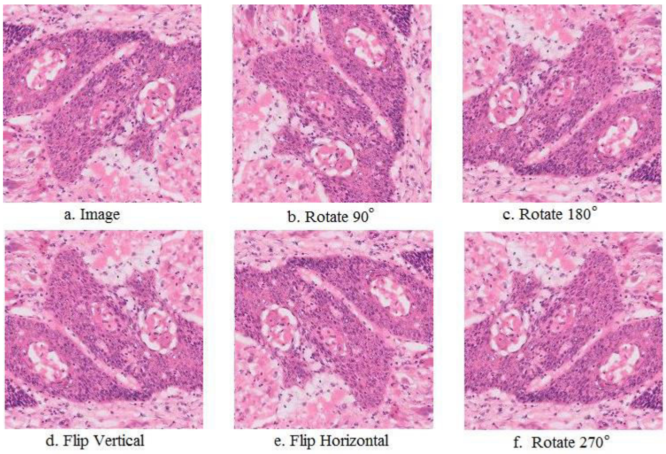

Figure 1 shows Data Augmentation of the original image

III. METHODOLOGY

A. Feature Extraction

Feature extraction is a way to use pre-trained networks without consuming time and effort into training. Learned features were extracted from a pre-trained network, and then used to train a classifier [9].

B. AlexNet Model

AlexNet model (Krizhevsky et. al.,2012) is trained on 1.2 million images of the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) 2010 dataset [1]. The ILSVRC 2010 training set contains 1000 different categories, representing everything objects such as flowers, vehicles, animals and so on [1]. As depicted in figure 2, the deep AlexNet architecture is made of a stack of eight layers [1]: the first five layers are convolutional and the remaining three are fully-connected. The input layer is configured to take a fixed sized 227 x 227 RGB image as an input; they normalize all the training images to get the same range of values for each of the input features. The first convolutional layer filters the input image with 96 kernels (size: 11x11x3) with a stride of 4 pixels. The second convolutional layer filters the output (normalized and pooled) of the first convolutional layer with 256 kernels (size: 5x5x48). The third, fourth, and fifth convolutional layers are connected without any intervening pooling or normalisation layers. The third convolutional layer has 384 kernels (size: 3x3x256), the fourth convolutional layer has 384 kernels (size: 3x3x192), and the fifth convolutional layer has 256 kernels (size: 3x3x192). The fully-connected layers have 4096 neurons each. The output of the last fully connected layer is fed to a 1000 categories. A softmax layer produces the probability distribution for the outputs of the last fully connected layer by converts them to real values between zero and one with sum one.

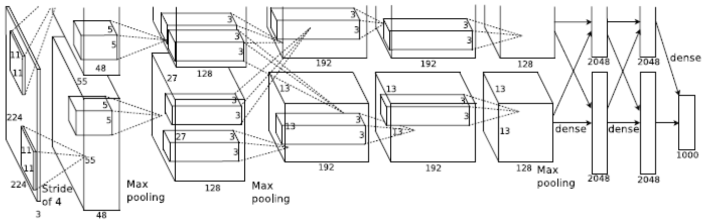

Figure 2 shows AlexNet Architecture

C. Machine Learning Algorithms

- SVM uses an *error-correcting output codes (ECOC)* algorithm (Escalera et. al., 2010) to classify multiclass models. ECOC reduces the problem of classification with three or more classes to a set of binary classifiers. ECOC models can improve classification accuracy, even compared to other multiclass models [7]. ECOC classification assigns a one-versus-one coding design, which determines the classes that the binary learners train on, and a decoding scheme, which determines how the results (predictions) of the binary classifiers are aggregated [7]. Let M be the coding design matrix with elements m_{kl}, and s_l be the predicted classification score for the positive class of learner l. A new observation is assigned to the class ($\hat{k}$) that minimizes the aggregation of the losses for the L binary learners. That is,

$$\hat{k} = \arg\min_{k} \frac{\sum_{l=1}^{L} |m_{kl}| g(m_{kl}, s_l)}{\sum_{l=1}^{L} |m_{kl}|} \qquad (1)$$

- The k-nearest neighbor algorithm estimates the predictors within each class by looks for in the observation to find the nearest points to predictor points and response values to those nearest points , and then it classifies an observation by estimating the posterior probability for each class and expected classification cost; Let y is the predicted classification, K is the number of classes, P^ (k|x) is the posterior probability of class k for observation x and C(y|k) is the cost of classifying an observation as y when its true class is k [10].

$$\hat{y} = \arg \min_{y=1,..,K} \frac{\sum_{k=1}^{K} P\langle k \mid x \rangle C\langle y \mid k \rangle}{\sum_{k=1}^{K} |m_{kl}|} \tag{2}$$

- The discriminant analysis algorithm estimates the predictors within each class by using a multivariate normal distribution. It assumes that predictor has a Gaussian mixture distribution. For linear discriminant analysis, the model has the same covariance matrix for each class, only the means vary. It use expected classification cost (equation 2) for prediction [11].
- Naive Bayes algorithm (Hastie et. al., 2008) estimates the densities of the predictors within each class, models posterior probabilities according to Bayes rule, and then it classifies an observation by estimating the posterior probability for each class, and then assigns the observation to the class yielding the maximum posterior probability [12]. Let Y is the random variable corresponding to the class index of an observation, $X_1,...,X_P$ are the random predictors of an observation and $\pi(Y=k)$ is the prior probability that a class index is k

$$\hat{P}(Y = k|X_1, \dots, X_P) = \frac{\pi(Y = k) \prod_{j=1}^{P} P(X_j|Y = k)}{\sum_{k=1}^{K} \pi(Y = k) \prod_{j=1}^{P} P(X_j|Y = k)} \quad \textit{for all } k = 1, \dots, K \tag{3}$$

- The Binary Decision Tree algorithm (Coppersmith et. Al., 1999) represents the observation in binary tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. It estimates the predictors within each class by following the branches of observation until it reaches a leaf node If reaches a leaf node, it returns the classification of that node. Predicted Class Label was calculate be minimize the expected classification cost (Equation 1). For trees, the score of a classification of a leaf node is the posterior probability of the classification at that node. The posterior probability of the classification at a node is the number of training sequences that lead to that node with the classification, divided by the number of training sequences that lead to that node [4].

D. Pre-processing Step

The study was performed in a MATLAB environment. For classification, it should be set down that 80% of data were randomly chosen for training. The remaining 20% were used for testing. Bukar & Ugail (2017) demonstrated that training set of data should be considerably larger in order to give more accurate results [2]. AlexNet model requires input images of size 227-by-227-by-3, but the dataset images have different sizes. So the training and test images were resized to height 227 and width 227 before they are input to the pre-training network.

E. Standard Performance

The machine learning classifiers were implemented and evaluated according to standard performance such as Accuracy (ACC), Precision (P) and Sensitivity (S). Sokolova and Lapalme (2009) extracted these terms from the confusion matrix [6]. The standard performance measurements were formulated as depicted in the following equations [6]; where accuracy evaluates the overall effectiveness of a classifier; Precision evaluates the class agreement of data labels with the positive labels given by the classifier; Sensitivity evaluates the effectiveness of a classifier to identify positive labels; Specificity evaluate the effectiveness of a classifier identifies negative labels; Area Under Curve evaluates Classifier's ability to avoid false classification.

$$\text{Accuracy} = \frac{(\text{tp} + \text{tn})}{(\text{tp} + \text{fn} + \text{fp} + \text{tn})} \tag{4}$$

$$\text{Precision} = \frac{\text{tp}}{\text{tp} + \text{fp}} \tag{5}$$

$$\text{Sensitivity} = \frac{\text{tp}}{\text{tp} + \text{fn}} \tag{6}$$

$$\text{Specificity} = \frac{\text{tn}}{\text{tn} + \text{fp}} \tag{7}$$

$$\text{AUC} = \frac{1}{2}\left(\frac{\text{tp}}{\text{tp} + \text{fp}} + \frac{\text{tn}}{\text{tn} + \text{fp}}\right) \tag{8}$$

IV. RESULTS

The performance results of using AlexNet model to build machine learning classifier were reported in the table 1. The five methods were implemented using the functions of MATLAB R2018a. These results used 2 times 10-fold cross-validation to evaluate the classifiers. For each of cross-validation, the performance values were calculated for each feature set based on the nine folds of training samples, via grid search in the parameter space. Therefore, each cross-validation might have slightly different values, and the average optimal value was reported. The Classifier used a linear function as the kernel function, Layer 'fc7' to extract features from dataset images, and the sequential minimal optimization method to find the separating hyper plane, its average kernel size was 2.0. A 3 X 3 confusion matrix was shown in figure 3.

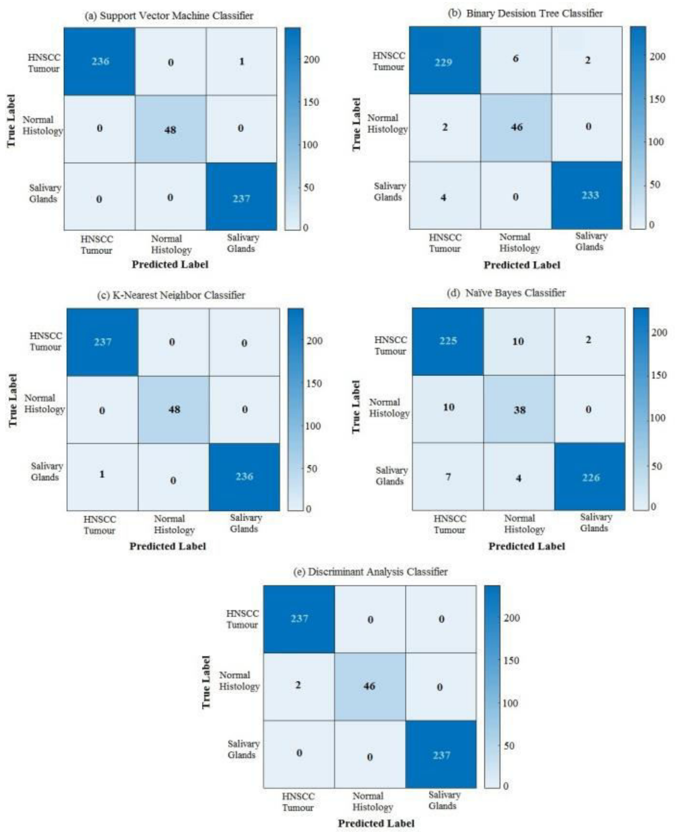

Figure 3 shows the confusion matrix for Classifiers

Classifier	ACC(%)	P	S	Spe	AUC
SVM	99.86	.9986	.9986	.9986	.9986
KNN	99.86	.9986	.9986	.9986	.9986
NB	93.68	.8982	.8839	.8839	.8911
BDT	97.32	.9692	.9502	.9502	.9597
DA	99.81	.9861	.9972	.9972	.9917

Table 1 illustrates the standard performance for Classifiers

ACKNOWLEDGMENT

We thank Professor Hassan Ugail for his encouragement and appreciation of his guidance and advice.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." Advances in neural information processing systems. 2012.

[2] A.M. Bukar, andH. Ugail. *"Convnet Features for Age Estimation*", Bradford, UK, ISBN: 9978-989-8533-66-1. 2017.

[3] A. Dosovitskiy, J. T. Springenberg and T. Brox."*Unsupervised feature learning by augmenting single images*". CoRR, abs/1312.5242. 2013.

[4] D. Coppersmith, S. J. Hong, and J. R. M. Hosking. "Partitioning Nominal Attributes in Decision Trees." Data Mining and Knowledge Discovery, Vol. 3, 1999, pp. 197–217.

[5] H. Wang, Z. Zhou, Y. Li, Z.Chen, P. Lu, W. Wang, W. Liu and L.Yu. "Comparison of machine learning methods for classifying mediastinal lymph node metastasis of non-small cell lung cancer from 18F-FDG PET/CT images", EJNMMI Research 7:11, DOI 10.1186/s13550-017-0260-9. 2017

[6] M. Sokolova and G. Lapalme. "*A systematic analysis of performance measures for classification tasks*". Inf. Process. Manage., 45(4):427–437, July 2009.

[7] S. Escalera, O. Pujol, and P. Radeva. "On the decoding process in ternary error-correcting output codes." IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 32, Issue 7, 2010, pp. 120–134.

[8] T. Google-Developers. WebPathology http://webpathology.com/index.asp

[9] T. Google-Developers. Pretrained Convolutional Neural Networks. https://uk.mathworks.com/help/nnet/ug/pretrained-convolutional-neural-networks.html, MathWorks, 2018

[10] T. Google-Developers. Classification Using k-nearest neighbor

https://uk.mathworks.com/help/stats/classification-using-nearest-neighbors.html

[11] T. Google-Developers. Discriminant Analysis Classification

https://uk.mathworks.com/help/stats/discriminant-analysis.html

[12] T. Hastie, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning, Second Edition. NY: Springer, 2008.

[13] Y. Xu, J.Y. Zhu, E.I. Chang, M. Lai and Z. Tu "Weakly supervised histopathology cancer image segmentation and classification". Med Image Anal. 2014;18:591–604.

[14] Y. LeCun, Y. Bengio, and G. Hinton ."Deep learning" Nature. 2015;521:436–44.