International Journal of Engineering Technology and Applied Science

(ISSN: 2395 3853), Vol. 6 Issue 1 January 2020

CLASSIFICATION OF BREAST CANCER DATASET USING NEURAL NETWORK: A NOVEL AND EFFICIENT APPROACH

Jagriti Kumari[1], Anshul Sarwagi[2]

M.Tech Scholar, IES College of Technology, Bhopal (M.P.) India
Assistant Professor, IES College of Technology, Bhopal (M.P.) India
jagritijoy41109@gmail.com[1]
ANSHULSARAWAGI301@gmail.com[2]

ABSTRACT - As we have seen we have applied various traditional machine learning algorithms for classification of various datasets. But sometimes this algorithm does not meet our expectation. So in this work we have introduced the concept of neural network which will be applied on the cancer dataset and it shows how neural network concept will work better than traditional machine learning algorithm. The concept we will use will be the implementation of Convolution Neural Network for tumor analysis. We will also show the comparison of the previous machine learning algorithm with the new algorithms which we have introduced in this chapter. The implementation of deep learning algorithm on the biological data is our novel approach and we will also show hoe the accuracy has been improved by using this novel approach. The classification is mainly done to classify tumors into malignant and benign.

Keywords: Machine Learning, Classification, CNN, RNN, Neurons

I. INTRODUCTION

Cancer is one of the most dangerous diseases in the world today. The death rate is increasing day by day. It is also predicted that increasing rate of death due to cancer will also get increased in the coming days. In the year 2017, 1688780 new cases are there and 600920 cancer deaths are projected to be occurring in the United States. It is also observed that 20% cancer incidence rate is higher in man than woman and death rate is 40% higher [1]. A lot of new cases will also get arise by the year 2030. To diagnose this type of disease we need intelligent system and treatment of this type of disease can be drastically reduced. The reduction can be in the form of medical errors which leads to decrease in financial loses and human loses. We must give special attention during implementing this type of intelligent system. The main problem which we face during designing this type of system is that the increase number of parameters occurs due to increase number of patients. It is very easy to analyze when we have less number of patients or parameters but it is very difficult when there are more number of patients or parameters. In large number of parameters it is very difficult to find the relationship by observation even by many modern and expert professionals. This is the only reason we have seen in the last few decades, there is a rapid development in the area of computer diagnostic equipments which basically helps physician to remove the irregularities from the data. These types of equipments reduce the error done by human and provide information related to medical area in detail. In the last few areas the machine learning techniques has been increase drastically in the area of disease classification. Even these machine learning techniques has increase very rapidly in the area of research studies in making very important prediction [2].

There are many literature based on cancer disease and many of these literatures are based on the cancer classification [3]. Most of the time classification technique is used for classification of cancer and non cancer patient. Sometimes we apply hybrid method for classification which is basically with the help of combination of existing classification method. Even in the research paper we have observed no one discuss about the cost related to misclassification of this disease. The main motive behind writing this paper is that most of the time we have used traditional machine learning algorithm like Naïve Bayes [4] , SVM [5], KNN [6] and many more, but in this paper we have introduced the concept of deep learning for classification between the cancer and non cancer patient. Deep learning is nothing, it is just the part of machine learning but here we will use the concept of neural network [7] and artificial intelligence [8]. We will also show how this deep learning concept performs better than traditional classification algorithm.

II. LITERATURE SURVEY

Now we will discuss some of the literature based on machine learning and deep learning algorithm in the area of cancer research. According to kourou et. al. [9] has stated that early detection of cancer can be done by using machine learning algorithms and they have also shown how these algorithms work better in the area of classification between cancer and non cancer patients. Now Zhou et. al. [10] has shown how cancer classification and prediction can be done using logistic regression with Bayesian gene selection approach.

Sweilam et. al. introduced least square support vector machine and active set strategy to show the classification on breast cancer dataset[11]. Khosravi et. al describes Deep Convolutional Neural Networks Enable Discrimination of Heterogeneous Digital Pathology Images. This work is mainly focused on several computational methods based on convolutional neural networks(CNN) and build a stand-alone pipeline to effectively classify different histopathology images across different types of cancer [12] .

III. DATA SETS

In this work we have taken a dataset related to tumors which is basically used by classifier to classify tumors into malignant and benign. The dataset size of 570 tumors with around 32 features. Here we have attribute related to tumor that are id, diagnosis, radius_mean, texture_mean, perimeter_mean, area_mean, smoothness_mean, compactness_mean concavity_mean, concave, points_mean, symmetry_mean , fractal_dimension_mean, radius_se, texture_se, perimeter_se, area_se, smoothness_se, compactness_se, concavity_se, concave points_se, symmetry_se, fractal_dimension_se, radius_worst, texture_worst, perimeter_worst, area_worst, smoothness_worst, compactness_worst, concavity_worst, concave points_worst, symmetry_worst, fractal_dimension_worst. Each attribute is representing one feature which is related to breast cancer, they all have their own value related to patient health status.

IV. METHODS

In this section we will first discuss the classification method used in machine learning. The classification algorithm which we will discuss will be Naïve Bayes [13], Random Forest [14], Logistic regression [15] and SVM [16] after that we will move to neural network or deep learning algorithm [17] which will include CNN i.e. Convolution Neural Network.

4.1 Naïve Bayes

It is one of the well known classification algorithm. It is mostly used when probability prediction belongs to particular class. It always give improved accuracy and it is one of the fastest algorithm used for train data. Normally it is used for large datasets. It is sequential algorithm that follows steps of execution followed by classification, estimation and prediction. There are different types of data mining algorithm for finding relations between the normal person and diseased person, but many of the algorithm have their own limitations like numerous iterations, heavy computational time and binning of the continuous arguments etc. Naïve Bayes has overcome various limitations and one of the best one to apply on large dataset. It considers probabilities as its factors for predicting the class if the test set is given. The following steps are involved for prediction:

1. Naïve Bayes consider the training set without their class values.
2. Now the test set is provided for prediction. A test set T can be predicted to its respective class Cx and Cy only if it fulfills the given condition
   $P(Cx|T)> P(Cy|T)$ for $1=y=m$, $y!=x$

Here P is the probability and the higher of the both probabilities P(Cx|T) and P(Cy|T) will be considered as the predicted class for T.

4.2 K- Nearest Neighbors (KNN)

The K-NN is one of the most popular classification algorithms in machine learning algorithm that is based on distances. It does not need any learning phase mainly based on instances. The input sample which we considered as training set has been combined with distance function and the choice or prediction of the new object depend on how much it is nearest to the given classes. In this we take the value of neighbors it can be taken as n, which is a integer number and on the basis of the value of n the new objects get classified to different classes. Before classifying the new object its distance measure is taken from the other object belonging to different classes, wherever the distance between the objects is less the particular object will belongs to the class of that previous resident object. Most of the time we take Euclidean method for distance measurement [18]. As the figure shown below:

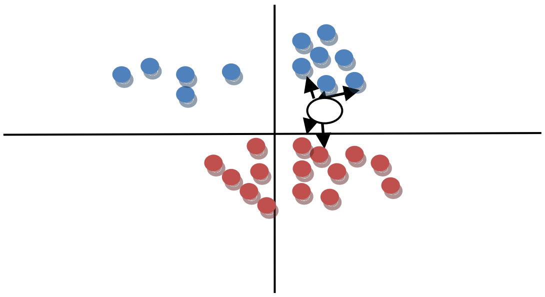


Fig 1 Representation of K-NN

In the given figure it has been shown that there are two classes Class A and Class B, Class A contains blue dots and class B contains red dots. X is the new object which has to be predicted in which class it belongs. Let us take an example that n=3, means we will calculate the distance from the nearest 3 object to this particular new object. Out of this calculation whose measured distance is less that will be taken as our predicted class and new object x will become object of that class. So in the above figure it is clearly stated that x will be in Class A.

4.3 Support Vector Machine

It is based on linear classification and it act as a binary classifier. Firstly It was introduced by HAvA Siegelmann and Vladimir Vapnik and he has shown its effectiveness mainly in the area of pattern recognition problem. Many times it has shown better classification than other classifiers mainly in case of small dataset. Let us discuss how it works, it segregate a pair of training vectors for two dissimilar groups (x_1, y_1), (x_2, y_2),…, (x_m, y_m), where x_i €R^d Represent vectors in d-dimensional attribute space and y_i € {-1, +1} is a group label.

Below figure shows the procedure of linear kernel based SVM, which maps the non linear input space into new linearly separable space. It is shown that all vectors lying on one side of the hyper plane are labelled as -1, and the vectors which are aligned on the another side are labelled as +1.The training points which are near to hyperplane in the transformed space are considered as support vectors. Comparing to training set, the size of support vectors are less, these support vectors determine the margin of the hyperplane and the decision surface.

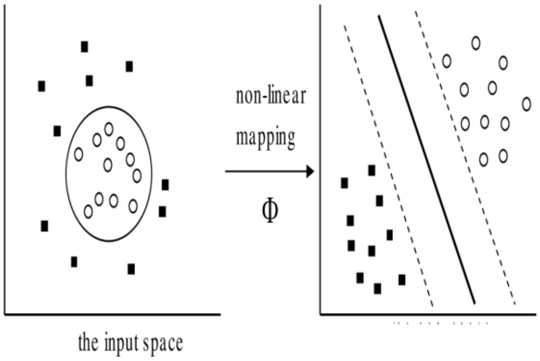


Fig 2 SVM Classification Scheme

4.4 Logistic Regression

It is one of the best model for prediction [9] , classification [3] and regression [19]. It is also known as logit method. It is also probabilistic linear classifier [20]. It mainly considers two factors weight matrix W and bias vector b. An input vector is projected into hyperplanes which are considered as classes during classification. So during classification the distance of the new elements to a hyperplane is calculated and it reflects the new element is a member of the corresponding class. To predict the class of new element, let take the mathematical calculation. The probability that the new input vector m will belong to class c and a stochastic variable can be written as S, the equation will be:

$$P(S ==c \mid m, W, b) = \text{softmax}_c(W_m + b)$$

$$= \frac{e^{W_{cm}+b_c}}{\sum_j e^{W_{jm}+b_j}}$$

The prediction of the class S_{pred} having probability maximal, particularly

S_{pred}=argmax$_c$P(S= c | m, W, b)

Logistic regression works in two phases . In the first phase which is actually known as training phase the partial sample is taken from the dataset to get the best values of the parameters W and b. Next phase is the test phase where the model which has been trained is tested to see the response.

V. PROPOSED WORK

The ANN algorithm can handle wide variety of tasks, nowadays it is mostly used for pattern recognition sentiment analysis, vision enhancement system and many more. They consist of mainly three layers namely input layer, hidden layers and output layer. They are basically to generate an output which is the combination of different inputs. Many times it contains multiple hidden layers depending on the dataset. Sometimes it is very useful for classification. There is drawback also in case of ANN that is mainly

time consumption. ANN is time consuming because there are different layers for processing and each layer will take some time during processing. Due to its genetic layer architecture it can lead to poor accuracy. Given below figure will show the architecture of ANN with interconnected group of nodes.

ANNs are a type of biologically-inspired feed-forward networks characterized by a sparse local connectivity and weight sharing among its neurons.

A ANN can also be seen as a sequence of convolutional and subsampling layers in which the input is a set of H * W * D images, where H is the height, W is the width and D is the number of channels which, in the case of RGB images corresponds to D = 3

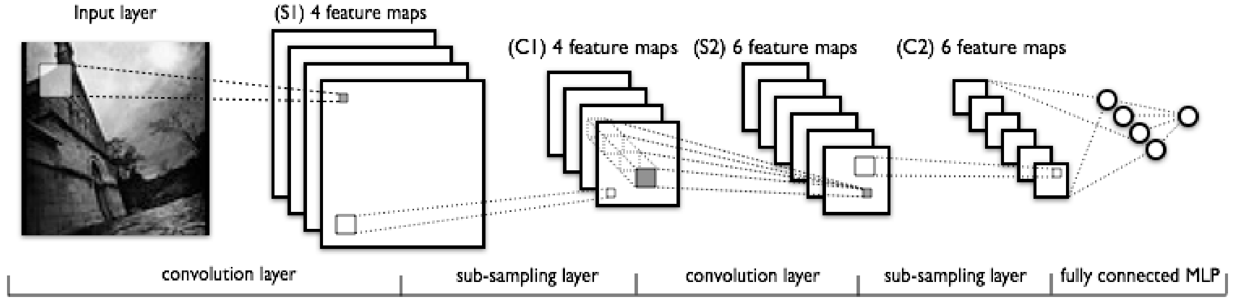


Fig 3 Architecture of Convolution Neural Network

A typical convolutional layer (volume) is formed by K filters (kernels) of size F *F * D, where F ≤ H and F ≤ W. These filters are usually randomly initialized and are the parameters to be tuned in the training process. Since the size of the filter is generally strictly smaller than the dimensions of the image, this leads to a local connectivity structure among the neurons. Each of this convolutional volumes has an additional hyper-parameter, S, which corresponds with the stride that the filter is going to slide spatially in the image. Let’s denote a particular training example as XHˆWˆD and a convolution filter WF ˆF ˆD. As it is familiar from the usual Multi-Layer Perceptron, it is customary to add a bias term b to each of the linear combinations formed. Finally, a (commonly non-linear) activation function, for example ReLU, is applied to the convolution between the input image and the kernels, which yields an activation map A of the dimensions $1+\frac{N-F}{S}*1+\frac{N-F}{S}*1$:

A=f(X *W+b)

Where * represents the valid convolution between the operands and f is the activation function. Appending

the activation maps found by applying K different kernels to the input example, an activation volume of dimensions $1+\frac{N-F}{S}*1+\frac{N-F}{S}*K$ is obtained. Note that depending on the dimensions of the image, the filter and the size of the stride, the resulting activation volume may reduce its spatial dimensions very quickly. An alternative to control this situation in advance is the use of padding techniques to the original image [21] . Finally, in order to perform dimensionality reduction directly on the data, pooling layers are applied to an activation volume or even the input image itself. These layers subsample its inputs, typically with mean or max pooling, over contiguous regions of size P * P. Figure 1 shows an example of a typical architecture for a CNN in which two convolutional and two pooling layers are applied to the original image. In this case, the extracted features obtained as are fed into a fully connected layer to perform the classification task. Note that it is possible to change the classifier set up at the end of the network with, for example, a SVM or a softmax classifier.

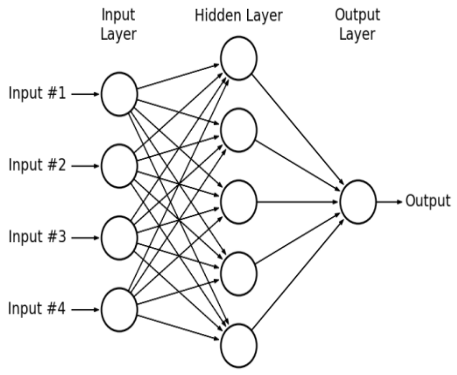
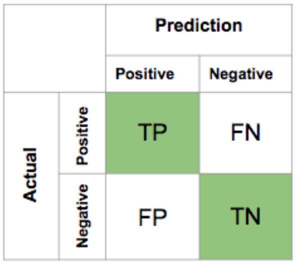


Fig 4 Architecture of Neural Network

VI. CONFUSION MATRIX

Consider a classification problem with only two classes: positive (P) and negative (N). For binary class classification training example, there can be four possible outcomes. If the training example is positive and the prediction is positive, we call it a true positive; and if the prediction is negative, it is called a false negative. On the other hand, if the training example is negative and it is classified as negative, it is called a true negative; otherwise, it is a false positive. Table 1 displays an example of a confusion matrix for binary classification.

Table 1: Confusion Matrix



		Prediction	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

A confusion matrix is a tool that allows to visualize the performance of a classifier in a supervised learning problem. By means of this matrix it is possible to access whether the system is commonly confusing pairs of classes. In the aforementioned problem, the confusion matrix summarizes the four possible outcomes from the classifier.

VII. RESULTS

The implementation is done in Python 3.6 with 4GB RAM on 64 bit processor. Here we have shown how the above discussed algorithm performs on the breast cancer data set. We have shown the output by considering their confusion matrix, precision score, recall score, F1 score and accuracy.

Our approach for classification with Neural Network

Train on 455 samples, validate on 114 samples, we have

show the last 4 epoch value out of 200 epochs.

```
Epoch 197/200

455/455 [==============================] - 1s
2ms/step - loss: 0.0217 - acc: 0.9956

Epoch 198/200

455/455 [==============================] - 1s
2ms/step - loss: 0.0352 - acc: 0.9934

Epoch 199/200

455/455 [==============================] - 1s
2ms/step - loss: 0.0406 - acc: 0.9934

Epoch 200/200

455/455 [==============================] - 1s
2ms/step - loss: 0.0199 - acc: 0.9956
```

Table 2: Precision score of Different Classifiers

Classifiers	Precision score
Naive Bayes	0.875
Logistic Regression	0.957447
SVM	0.918367
Random Forest	0.956522
Proposed ANN	0.978723

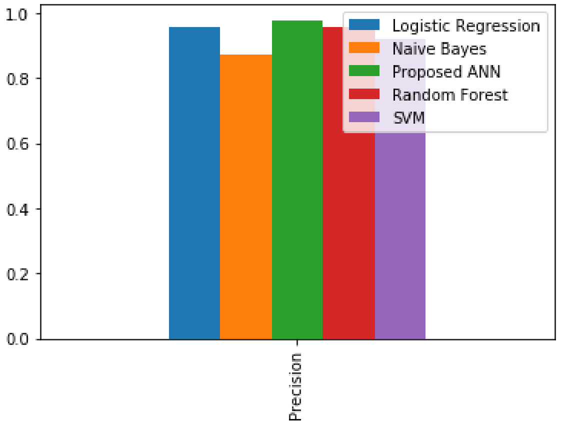


Fig 5. Comparison of Precision score of various machine Learning algorithm with proposed ANN algorithm

Table 3: Recall score of Different Classifiers

Classifiers	Precision score
Naive Bayes	0.893617
Logistic Regression	0.957447
SVM	0.957447
Random Forest	0.93617
Proposed ANN	0.978723

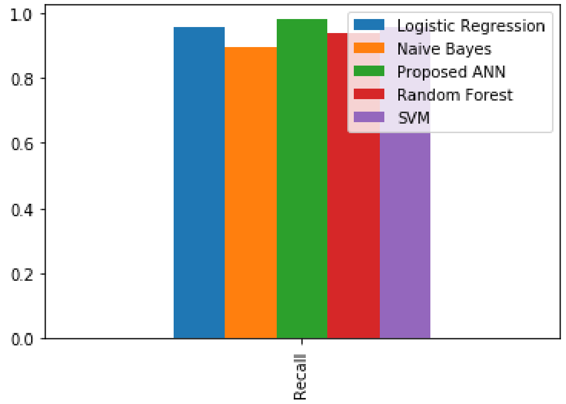


Fig 6. Comparison of Recall score of various machine Learning algorithm with proposed ANN algorithm

Table 4: Recall score of Different Classifiers

Classifiers	F1_score
Naive Bayes	0.884211
Logistic Regression	0.957447
SVM	0.9375
Random Forest	0.946237
Proposed ANN	0.978723

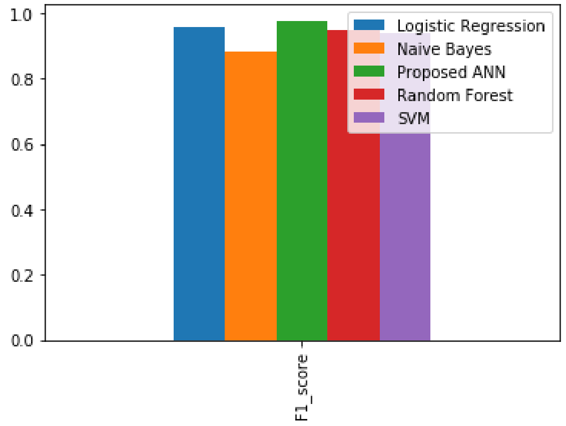


Fig 7. Comparison of F1 score of various machine Learning algorithm with proposed ANN algorithm

Table 5: Accuracy of Different Classifiers

Classifiers	Accuracy
Naive Bayes	90.350877
Logistic Regression	96.491228
SVM	94.736842
Random Forest	95.614035
Proposed ANN	98.245614

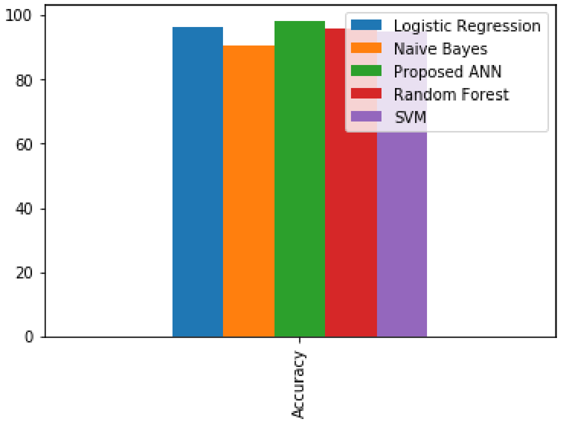


Fig 8. Comparison of Accuracy score of various machine Learning algorithm with proposed ANN algorithm

VIII. CONCLUSION AND FUTURE WORK

In this work we have discussed about the various machine learning algorithm and we have show how they perform on tumor dataset. The main focus is to apply deep learning concept for predictive analysis. Mainly it is used for predicting outcomes. Here we have shown how classification through neural network will improve the accuracy of classification. It will be helpful for prediction of cancer diagonsis. In this work we have implemented the classifiction algorith with our proposed Neural Network but in the future we will also work on other deep learning algorithms like Recurrent Neural Network, Hopfield Neural Network and many more.

REFERENCES

[1] R. Siegel, K. D. Miller, and J. Ahmedin, “Cáncer Statistics,” *Ca Cáncer J.*, vol. 67, no. 1, pp. 7–30, 2017.

[2] M. L. Spring, *Machine Learning in Action*. 2015.

[3] E. Alpaydın, *Introduction to machine learning*, vol. 1107. 2014.

[4] G. D. Rashmi, A. Lekha, and N. Bawane, “Analysis of efficiency of classification and prediction algorithms (Naïve Bayes) for Breast Cancer dataset,” *2015 Int. Conf. Emerg. Res. Electron. Comput. Sci. Technol.*, pp. 108–113, 2015.

[5] C. Cortes and V. Vapnik, “Support-Vector Networks,” *Mach. Learn.*, 1995.

[6] S. A. Medjahed, T. A. Saadi, and A. Benyettou, “Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules,” *Int. J. Comput. Appl.*, vol. 62, no. 1, p. 5, 2013.

[7] T. G. Clarkson, “Introduction to neural networks,” *Neural Netw. World*, 1996.

[8] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. 2010.

[9] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V Karamouzis, and D. I. Fotiadis, “Machine learning applications in cancer prognosis and prediction,” *CSBJ*, vol. 13, pp. 8–17, 2015.

[10] X. Zhou, K. Liu, and S. T. C. Wong, “Cancer classification and prediction using logistic regression with Bayesian gene selection,” vol. 37, pp. 249–259, 2004.

[11] N. H. Sweilam, A. A. Tharwat, and N. K. A. Moniem, “Support vector machine for diagnosis cancer disease : A comparative study,” *Egypt. Informatics J.*, vol. 11, no. 2, pp. 81–92, 2010.

[12] P. Khosravi, E. Kazemi, and M. Imielinski, “EBioMedicine Deep Convolutional Neural Networks Enable Discrimination of Heterogeneous Digital Pathology Images,” *EBioMedicine*, vol. 27, pp. 317–328, 2018.

[13] G. I. Webb, *Naïve Bayes*. 2010.

[14] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[15] A. F. Seddik, “Logistic Regression Model for Breast Cancer Automatic Diagnosis,” pp. 150–154, 2015.

[16] A. Pradhan, “SUPPORT VECTOR MACHINE-A Survey,” *Int. J. Emerg. Technol. Adv. Eng.*, vol. 2, no. 8, pp. 82–85, 2012.

[17] Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*. 2015.

[18] H. Abdi, “Distance,” *Encycl. Meas. Stat.*, 2007.

[19] D. P. R. G and R. T. Sriramaneni, “Literature Survey on Various Software Cost,” vol. 4, no. Iv, pp. 868–874, 2016.

[20] D. Mladenić, J. Brank, M. Grobelnik, and I. Natasa Milic-Frayling, “Feature Selection using Linear Classifier Weights: Interaction with Classification Models,” in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, 2004.

[21] L. Zheng, S. Wang, Z. Liu, and Q. Tian, “Packing and padding: Coupled multi-index for accurate image

retrieval," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1947–1954.