Machine Learning and Artificial Intelligence
A.J. Tallón-Ballesteros and C. Chen (Eds.)

doi:10.3233/FAIA200765

A Clinical Decision Support Tool to Detect Invasive Ductal Carcinoma in Histopathological Images Using Support Vector Machines, Naïve-Bayes, and K-Nearest Neighbor Classifiers

Kyra Mikaela M. LOPEZ and Ma. Sheila A. MAGBOO
University of the Philippines Manila, Manila, Philippines

Abstract. This study aims to describe a model that will apply image processing and traditional machine learning techniques specifically Support Vector Machines, Naïve-Bayes, and k-Nearest Neighbors to identify whether or not a given breast histopathological image has Invasive Ductal Carcinoma (IDC). The dataset consisted of 54,811 breast cancer image patches of size 50px x 50px, consisting of 39,148 IDC negative and 15,663 IDC positive. Feature extraction was accomplished using Oriented FAST and Rotated BRIEF (ORB) descriptors. Feature scaling was performed using Min-Max Normalization while K-Means Clustering on the ORB descriptors was used to generate the visual codebook. Automatic hyperparameter tuning using Grid Search Cross Validation was implemented although it can also accept user supplied hyperparameter values for SVM, Naïve Bayes, and K-NN models should the user want to do experimentation. Aside from computing for accuracy, the AUPRC and MCC metrics were used to address the dataset imbalance. The results showed that SVM has the best overall performance, obtaining accuracy = 0.7490, AUPRC = 0.5536, and MCC = 0.2924.

Keywords. Invasive ductal carcinoma (IDC), oriented FAST and Rotated BRIEF (ORB), Support Vector Machines, Naïve-Bayes, K-Nearest Neighbors.

1. Introduction

Breast cancer is one of the most common types of cancer worldwide with over two million new cases of breast cancer diagnosed in 2018 [1]. This represents around 12.3% of the total new cancer cases that year. In the Philippines, breast cancer had the highest number of new cases in 2015, representing 19% of the overall new cancer cases in both men and women [2]. The most common subtype of breast cancer is called Invasive Ductal Carcinoma (IDC) which makes up 80% of all invasive breast cancer cases [3, 4]. At present, there is no definite main cause of breast cancer. Aside from genetics, there are still several risk factors that have been known to influence a person's susceptibility to breast cancer [5]. Therefore, the key to improving breast cancer survival is early detection and screening [6]. In most cases, IDC can manifest as micro-calcifications or thickening of breast tissues [7, 8]. For further confirmation, doctors may recommend a breast biopsy, the only diagnostic procedure that can determine and verify the presence of cancer [9, 10].

Research has shown that the application of artificial intelligence and machine learning during diagnosis has helped further improve cancer detection and staging [11, 12]. Computer-aided diagnosis has helped with automating labor-intensive steps and reducing reader bias [13-15]. Several studies have shown that accurate breast cancer predictions may depend on the right combination of feature selection and/or ML techniques [16, 17].

2. Methodology

A literature review was conducted to determine the performance of traditional machine learning as well as deep learning approach for classification of histopathological images particularly for invasive ductal carcinoma. The dataset used in this study, the "Breast Histopathology Images", consisted of 162 whole mount slide images of Breast Cancer specimens scanned at 40x magnification. From there, 277,524 patches of size 50 x 50 were extracted (198,738 IDC negative and 78,786 IDC positive). This dataset was originally described by Cruz-Roa [18] and is now hosted in Kaggle [19]. Due to hardware limitations, for this particular study, only 54,811 images were selected of which 48,848 was used for training and 10,963 was used for training. This new set maintained the original 28:72 or approximately 3:7 ratio of positive to negative images.

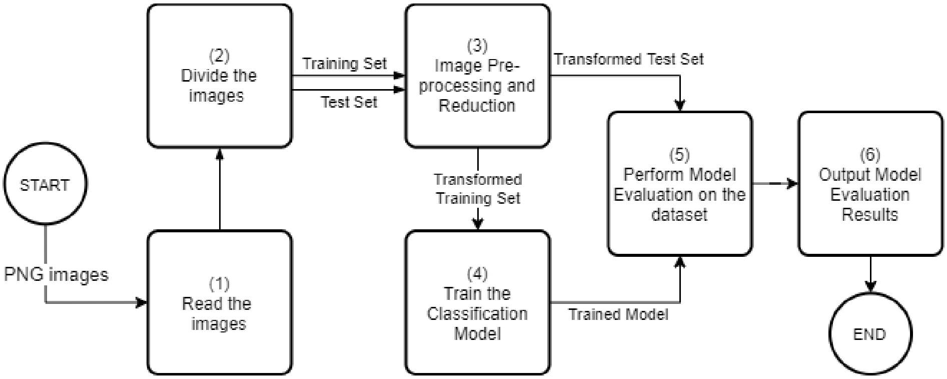

Figure 1. General Workflow

The general workflow is common to all ML applications and is illustrated in Figure 1. The images are loaded and then split into two for training and testing. The model is built after undergoing a series of pre-processing steps which includes feature extraction via ORB, feature scaling via Min-Max normalization, then clustering via K-Means over the training set. The model is then evaluated using the test set.

To train the classifier, hyperparameter values for each of the classifiers (Support Vector Machines, Naïve-Bayes, and k-Nearest Neighbors) must be specified. The user may experiment by providing these values or they may opt to use automatic hyperparameter tuning. This uses GridSearchCV, which runs a 10-fold cross validation to determine which hyperparameter values will produce the best performing estimator (classifier), ranked by the mean test score. The trained model is then saved using Joblib to be used for image classification of new biopsy images. This process will result in a trained model and the average classification accuracy attained by that model.

The machine learning classifiers are implemented via the scikit-learn libraries: SVC for SVM classifier, MultinomialNB for Naïve-Bayes classifier, knn for the k-Nearest Neighbor classifier. The performance metrics to be used is also implemented using scikit-learn and include the accuracy, precision, recall, average precision, and Matthew's Correlation Coefficient (MCC).

The ORB features of the training and test sets were extracted on a local machine while the remaining steps (codebook generation, training, testing, and performance evaluation) were performed in Google Colaboratory.

2.1. Bag of Visual Words

The Bag-of-Visual-Words (BoVW) is a technique used in image classification [20]. It represents an image as a set of features consisting of corner points, edges, and flat regions. Since an image can have multiple features, for each image, features are extracted then a visual dictionary or bag of visual words is generated using k-means clustering [21]. This visual dictionary, represented by a collection of histograms, will be used in training machine learning algorithms in order to classify a new input image [22-24]. In this study, we extracted the features of each image in the training set using ORB (Oriented FAST and Rotated BRIEF), applied normalization on each feature and then reduced the number of features via KMeans clustering. The clusters are then collected into a visual dictionary or codebook. The generated codebook is then used to build a histogram of features for each of the training images. These histograms will serve as the input for training the classifier.

2.1.1. ORB (Oriented FAST and Rotated BRIEF) for Feature Extraction

ORB, also known as Oriented FAST Rotated BRIEF, was first presented in 2011 by Ethan Rublee et. al. [25] for computer vision tasks such as object recognition, detection, and matching. ORB was developed by OpenLabs as an open source alternative to SIFT and SURF. ORB uses FAST or Features from Accelerated Segment Test to create a sequence of images, all of which are versions of the image at different resolutions [26-28]. Next, it extracts keypoints or regions in the image which are points of interest. BRIEF or Binary Robust Independent Elementary Feature then takes all keypoints found by the FAST algorithm then converts each keypoint into a binary feature vector [29]. BRIEF uses a randomly-selected distribution of point-pairs relative to a central point to create the descriptor [25, 30]. Since BRIEF is sensitive to rotation, ORB used the rBRIEF (rotation aware BRIEF) in order to make it invariant to rotation.

2.1.2. Normalization of Extracted Features

Normalization is a feature scaling technique used to ensure that each feature contributes approximately proportionately to a measure. In this study, we used the min-max normalization scaled to unit range to linearly transform the extracted features to values ranging from 0 to 1. This will ensure that each data point will be on the same scale making each feature equally important while preserving the distribution of the original data. [31, 32]

2.1.3. Feature Reduction through K-Means Clustering then Codebook Generation

K-Means clustering is an algorithm used to find groups in data where k represents the number of groups or clusters. It is typically used as an unsupervised learning algorithm but is also commonly used as a vector quantization step in codebook generation in the BoVW Model [24]. Each cluster contains a centroid, a data point at the center of a cluster representing a multi-dimensional average of the cluster [32]. Now that the clusters are formed, the next step is to quantify and represent an image as a histogram by counting the number of times each visual word appears. This histogram is our actual bag of visual words. In this study, $k = 500$ is the number of clusters used.

2.2. Model Training

2.2.1. Support Vector Machines

Support Vector Machine or SVM is a supervised machine learning algorithm that aims to determine the optimal separating hyperplane to correctly classify the data in a given space [33, 34]. SVM has parameters that may be tuned to increase accuracy, especially if given non-linearly separable data points. These parameters include regularization parameter, gamma, and the kernel. The regularization parameter, known as the lambda (λ) parameter, represents the degree of importance that is given to misclassifications. The higher the value of λ, the smaller the max-margin and the lesser incorrect classifications are allowed. A lower λ value allows the classifier to find a larger max-margin but with a greater tendency to misclassify data points. The gamma (γ) parameter describes the degree of influence a single data point has over the decision boundary. For a higher gamma value, the closer data points to the hyperplane are considered which can help handle more complexity in data, but if it is considerably high, it may have a tendency to overfit the data. A lower gamma value considers farther data points but may lead to underfitting to the data, making less stable classification. The kernel trick makes use of a kernel function φ, another parameter in SVM which transforms the data into a higher dimensional feature space so that a linear separation is possible [35]. The different types of SVM kernels include: Linear kernel, Sigmoid kernel, and Radial Basis Function (RBF) kernel [36]. In this study, we used the RBF kernel with $\lambda = 1$, and gamma = 0.01 and resulting to accuracy = 0.7490, Precision = 0.6991, and Recall = 0.2135.

2.2.2. Naïve-Bayes Classifier

The Naïve-Bayes algorithm is commonly used for classification problems and is suitable for high dimensional input. Based on Bayes' Theorem of Probability [37], the goal of Naïve -Bayes is to maximize the posterior probability from the training data to formulate a decision rule for new data [38]. For variables that have categories not observable in the training set, the Naïve-Bayes model may use the alpha (α) value, also known as the additive smoothing parameter or the Laplace correction. This hyperparameter is more commonly applied in histogram steps of text classification. Certain instances that are not encountered in the training set have zero frequency, thus having zero probability. The smoothing parameter prevents the model from assigning this null probability by converting the instance count into a "pseudo-count". In this study, we used $\alpha = 1.0$ resulting to accuracy = 0.6485, Precision = 0.421, and Recall = 0.6135.

2.2.3. K-Nearest Neighbor Classifier

K-Nearest Neighbor or K-NN is another supervised learning algorithm known for its simple implementation and low calculation time. This is commonly used in statistical estimations and pattern recognition. This algorithm stores the entire training dataset, making use of all the data while classifying a new data point or instance [39].

The value of K affects the shape of the decision boundaries and is usually an odd numbered integer if the number of classes is even. A small K results in a flexible but less stable decision boundary having low bias and high variance with a tendency to overfit data. When K is relatively large, the classifier is more resilient to outliers, making smoother decision boundaries but can consequently have higher bias. Some methods like ten-fold cross-validation can be used to estimate the optimal K value. In this study K = 3 resulting to Accuracy = 0.7071, Precision = 0.2563, and Recall = 0.0131.

2.3. Performance Metrics

Although the accuracy score is reliable, it may not always be relevant to diagnosis, especially given an imbalanced dataset. Instead, the following metrics, Area Under the Precision-Recall Curve (AUPRC) or Average Precision, and Matthews Correlation Coefficient (MCC) were applied as these are the most commonly used metrics when dealing with imbalanced data [40, 41].

To know if the classifier performance is good, the performance of the random classifier must be computed first using Eq (1).

$$Performance\ of\ Random\ Classifier = \frac{Total\ Positive}{Total\ Positive + Total\ Negative} \tag{1}$$

For the dataset used in this study, the baseline performance of the random classifer is 0.2839.

Precision refers to the percentage of results that are relevant and is a good measure to determine when the cost of false-positives is high. High precision relates to the low false positive rate which is important in diagnostics so as not to subject patients without a disease to expensive and even invasive procedures. The best value for precision is 1 and the worst value is 0. Recall, on the other hand, expresses the ability to find all relevant instances in a dataset. This is commonly referred to as the true positive rate or sensitivity.

A precision-recall curve shows the relationship between precision and recall for every possible cut-off. This focuses on the minority class making it an effective measure whenever there is class imbalance [42]. The resulting score, AUPRC, also called average precision, can be used to compare performance of different classifiers. A classifier's performance is rated good if the average precision is higher than the performance of the random classifier.

The Matthew's Correlation Coefficient or MCC measures the correlation between the predicted and observed binary classification of a sample and can be directly computed from the confusion matrix. An MCC score of +1 describes a perfect prediction, a 0 is no better than a random prediction, and a -1 score represents a complete disagreement between prediction and outcome.

MCC is generally regarded as a balanced measure which can be used even if there is a class imbalance problem. In some studies, MCC is regarded as the most informative single score to establish the quality of a binary classifier prediction in a confusion matrix context since its score is high only if the classifier does well on both the *negative* and the *positive* elements [41].

3. Results and Discussion

Although deep learning using convolutional neural networks and its variants are popular techniques for image classifications, there are still a number of recent studies that still use traditional machine learning techniques such as support vector machines, Naïve-Bayes, logistic regression, k-nearest neighbors, and random forest, among others.

In some studies, the performance of traditional ML specifically SVM, is comparable to the performance of deep learning approach but requiring less resources (less number of parameters to consider, less number of training samples, less number of iterations to reach convergence, effectiveness of application of active learning techniques [43-45]. The number of available high-quality annotated images, the pre-processing techniques, the feature selection methods, and the classification algorithms employed including the selection of the best hyperparameters are factors that can affect the performance of the classification model using traditional approaches. The search for the best combination of these factors for the given dataset is the challenge ML experts are working on.

Table 1. Hyperparameter Values for Each Machine Learning Classifier

Model	Hyperparameter	Value
SVM (RBF kernel)	regularization (λ)	1
	gamma (γ)	0.01
Naïve-Bayes	additive smoothing (α)	1.0
KNN	Neighbors (K)	3

Table 1 shows the hyperparameter values used in each model. The hyperparameter values chosen were obtained by performing 10-fold cross validation on 2,500 images, a separate set of images from the training and testing dataset. K-means clustering was implemented with $k = 500$.

As summarized in Table 2, the AUPRCs of both SVM and Naïve Bayes are way above the baseline of the performance of the random classifier computed as 0.2839 based from the formula in [42] with SVM as higher among the two. K-NN was the worst in all aspects, even obtaining a negative value for MCC.

Table 2. Summary of Model Performances Based on Different Evaluation Metrics

Model	Accuracy	AUPRC	Precision	Recall	MCC
SVM	**0.7490**	**0.5536**	**0.6991**	0.2135	**0.2924**
Naïve-Bayes	0.6485	0.5089	0.421	**0.6135**	0.2538
KNN	0.7071	0.2827	0.2563	0.0131	-0.0139

SVM's precision at 69.91% means the SVM model is good at finding relevant results. MCC is also good (29.24%) indicating that the SVM model is doing well on both the *negative* and the *positive* IDC cases.

Although a simple and traditional method has been presented, it is worth experimenting with newer methods such as [46, 47] that have shown to perform better.

4. Conclusion

This paper describes a method for Invasive Ductal Carcinoma (IDC) breast cancer classification. This is based on the Bag-of-Visual-Words (BOVW) model as a general approach and utilized the Oriented FAST and Rotated BRIEF (ORB) descriptors for feature extraction, Min-Max Normalization for feature scaling, and K-Means Clustering on the ORB descriptors to generate the visual codebook before feeding it to the SVM, Naïve-Bayes and K-Nearest Neighbor machine learning classifiers.

After evaluating the three machine learning models on various performance metrics, it was found that SVM obtained the best results in terms of accuracy (74.90%), AUPRC (55.36%), and MCC score (29.24%).

References

[1] Worldwide cancer data Global cancer statistics for the most common cancers. [Online]. Available: https://www.wcrf.org/dietandcancer/cancer-trends/worldwide-cancer-data. [Accessed: 10-Jan-2020].

[2] Laudico AV, Mirasol-Lumague MR, Medina V, Mapua CA, Valenzuela FG, Pukkala E. 2015 Philippine Cancer Facts and Estimates, 2015.

[3] Types of breast cancer. [Online]. Available: https://www.breastcancer.org/symptoms/types.

[4] Invasive ductal carcinoma (IDC). [Online]. Available: https://www.breastcancer.org/symptoms/types/idc.

[5] What are the risk factors for breast cancer? [Online]. Available: https://www.cdc.gov/cancer/breast/basic_info/risk_factors.htm.

[6] Breast cancer: prevention and control. [Online]. Available: https://www.who.int/cancer/detection/breastcancer/en/.

[7] Breast cancer signs and symptoms. [Online]. Available: https://www.acrf.com.au/support-cancer-research/types-of-cancer/breast-cancer/.

[8] Understanding breast calcifications. [Online]. Available: https://www.breastcancer.org/symptoms/testing/types/mammograms/mamm_show/calcifications.

[9] Tests for diagnosing IDC. [Online]. Available: https://www.breastcancer.org/symptoms/types/idc/tests/diagnosing.

[10] Breast biopsy. [Online]. Available: https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/breast-biopsy.html.

[11] Asri H, Mousannif H, Al Moatassime H, Noel T. Using machine learning algorithms for breast cancer risk prediction and diagnosis, Procedia Comput. Sci., 2016, 83:1064–1069.

[12] Nindrea RD, Aryandono T, Lazuardi L, Dwiprahasto I. Diagnostic accuracy of different machine learning algorithms for breast cancer risk calculation: a meta-analysis. Asian Pac. J. Cancer Prev., Jul. 2018, 19(7): 1747–1752.

[13] Dhahri H, Al Maghayreh E, Mahmood A, Elkilani W, Faisal Nagi W. Automated breast cancer diagnosis based on machine learning algorithms, J. Healthc. Eng., Nov. 2019, 2019: 1–11.

[14] Negrão de Figueiredo G, Ingrisch M, Fallenberg EM. Digital analysis in breast imaging. Breast Care (Basel)., Jun. 2019, 14(3): 142–150.

[15] Kaushal C, Bhat S, Koundal D, Singla A. Recent Trends in Computer Assisted Diagnosis (CAD) system for breast cancer diagnosis using histopathological images, IRBM, Aug. 2019, 40(4): 211–227.

[16] Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput. Struct. Biotechnol. J., 2015, 13: 8–17.

[17] Ferroni P, Zanzotto FM, Riondino S, Scarpato N, Guadagni F, Roselli M. Breast cancer prognosis using a machine learning approach. Cancers (Basel)., Mar. 2019, 11(3).

[18] Cruz-Roa A et al. Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. 2014, p. 904103.

[19] Breast histopathology images. [Online]. Available: https://www.kaggle.com/paultimothymooney/breast-histopathology-images.

[20] Magoulas GD, Prentza A. Machine learning in medical applications, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), LNAI, 2001, 2049: 300–307.

[21] Davida B. Bag of visual words in a nutshell. 2018. [Online]. Available: https://towardsdatascience.com/bag-of-visual-words-in-a-nutshell-9ceea97ce0fb.

[22] Green K. Generating and applying a bag of visual words model for image classification, 2017. [Online]. Available: http://www.deepcore.io/2017/04/18/generating-and-applying-a-bag-of-visual-words-model-for-image-classification/#page-content.

[23] Al Chanti D, Caplier A. Improving bag-of-visual-words towards effective facial expressive image classification, Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, 2018, pp. 145–152.

[24] Karim AAA, Sameer RA. Image classification using bag of visual words (BoVW), Al-Nahrain J. Sci., Dec. 2018, 21(4): pp. 76–82.

[25] Tyagi D. Introduction to ORB (Oriented FAST and Rotated BRIEF). 2019. [Online]. Available: https://medium.com/analytics-vidhya/introduction-to-orb-oriented-fast-and-rotated-brief-4220e8ec40cf.

[26] Rosten E, Drummond T. Fusing points and lines for high performance tracking. Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, 2005, 2:1508-1515.

[27] Viswanathan DG. Features from Accelerated Segment Test (FAST) Deepak Geetha Viswanathan 1., 2011.

[28] Rublee E, Rabaud V, Konolige K, Bradski G. ORB: An efficient alternative to SIFT or SURF, 2011 International Conference on Computer Vision, 2011, pp. 2564–2571.

[29] Rosten E, Porter R, Drummond T. Faster and better: a machine learning approach to corner detection, IEEE Trans. Pattern Anal. Mach. Intell., Jan. 2010, 32(1): 105–119.

[30] Calonder M, Lepetit V, Strecha C, Fua P. BRIEF: binary robust independent elementary features, Springer Berlin Heidelberg, 2010, pp. 778–792.

[31] KumarSingh B, Verma K, Thoke AS. Investigations on impact of feature normalization techniques on classifier performance in breast tumor classification, Int. J. Comput. Appl., Apr. 2015, 116(19): 11–15.

[32] Kallipolitis A, Maglogiannis I. Creating visual vocabularies for the retrieval and classification of histopathology images. in 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2019, pp. 7036–7039.

[33] Evgeniou T, Pontil M. Support vector machines: theory and applications, 2001, pp. 249–257.

[34] Gandhi R. Support vector machine — introduction to machine learning algorithms, 2018. [Online]. Available: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47.

[35] Schölkopf B. The kernel trick for distances. Adv. Neural Information Processing Systems. 2001.

[36] Wang LP ed. Support vector machines: theory and applications. Springer Science & Business Media, 177: 29-45, 2005.

[37] Joyce J. Bayes' theorem. 2003.

[38] Karim M, Rashedur MR. Decision tree and naive bayes algorithm for classification and generation of actionable knowledge for direct marketing. 2013.

[39] Peterson LF. K-nearest neighbor. Scholarpedia 4.2 (2009): 1883.

[40] Magboo MSA, Coronel, A. 30-day hospital readmission prediction model for diabetic patients within the 30-70 age group, Proceedings of Academics World 130th International Conference, Madrid, Spain, 10th -11th June, 2019.

[41] Chicco D. Ten quick tips for machine learning in computational biology. BioData Min., Dec. 2017, 10(1): 35.

[42] Introduction to the precision-recall plot, 2017 [Online] Available: https://classeval.wordpress.com/introduction/introduction-to-the-precision-recall-plot/

[43] Liu P, Choo KKR, Wang L, Huang F. SVM or deep learning? A comparative study on remote sensing image classification. Soft Comput., Dec. 2017, 21(23): 7053–7065.

[44] Huynh BQ, Li H, Giger ML. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. J. Med. Imaging, Aug. 2016, 3(3): 034501.

[45] Borges Sampaio W, Moraes Diniz E, Corrêa Silva A, Cardoso de Paiva A, Gattass M. Detection of masses in mammogram images using CNN, geostatistic functions and SVM. Comput. Biol. Med., Aug. 2011, 41(8): 653–664.

[46] Izonin I, Trostianchyn A, Duriagina Z, Tkachenko R, Tepla T, Lotoshynska N. The combined use of the wiener polynomial and SVM for material classification task in medical implants production, Int. J. Intell. Syst. Appl. 2018, 10(9): 40-47. DOI: 10.5815/ijisa.2018.09.05

[47] Bodyanskiy YE, Perova I, Vynokurova O, Izonin I. Adaptive wavelet diagnostic neuro-fuzzy system for biomedical tasks, *14th International Conference on Advanced Trends in Radioelectronics Telecommunications and Computer Engineering (TCSET)*, February 20–24, 2018, pp. 299-303.