Invasive Ductal Carcinoma (IDC) Classification with Machine Learning Methods in Breast Cancer Images

Baykal Yılmaz. Electrical &Electronics Engineering, Abdullah Gül University. Kayseri, Turkey.
baykal.yilmaz@agu.edu.tr

Mücahit Demirci. Electrical & Electronics Engineering, Abdullah Gül University. Kayseri, Turkey
mucahit.demirci@agu.edu.tr

***Abstract*—** This paper presents a machine learning (ML) method for detection and visual analysis of invasive ductal carcinoma (IDC) locations in whole slide images (WSI) of breast cancer. Machine learning is an artificial intelligence approach that learns from the experience consisting of computational methods and statistics to learn information directly from the dataset for modeling the relationships in data. It is a similar approach to how the human brain works by interpreting features such as representative layers. Invasive ductal carcinoma detection requires a huge amount of time and expertise since it consists of many scanning operations for the identification of benign and malignant areas. Moreover, IDC delineation in WSI is vital for estimating the level of the tumor in a precise way. Machine learning and deep learning (DL) methods are precisely suitable for figuring these challenges out with even a small number of samples from the corresponding dataset for training to explore features and act accordingly. Our goal is to develop a machine learning application that can identify IDC positive and IDC negative tissues and the probability of the cancerous level with training and validation data sets in separate folders which is an obligatory procedure in machine learning.

Conclusions: Obtained results indicate that machine learning algorithms are ideally suitable for improving and contributing to this field of a cancer diagnosis.

***Keywords*—** *Breast cancer, artificial neural networks, machine learning, whole-slide imaging, invasive ductal carcinoma (IDC).*

I. INTRODUCTION

Today, invasive ductal carcinoma is the most seen type of breast cancer in around 80 percent of patients [1,2]. Two main factors make it hazardous: it is malignant and can spread from part to part inside the breast. Tissue samples can be obtained by biopsy and pathology. To decide if a patient contains IDC or not and estimate the level of the disease, an expert pathologist is required which the process is generally done in conventional ways. Therefore, distinguishing the tissue locations in terms of invasive and non-invasive categories are the first steps of this project from histopathological images in breast cancer. In recent years, with

the help of convolutional neural networks (CNN), machine learning has gained great success in classification. In this paper, as a pre-processing step, image processing methods are implemented and extracted features (image density, spatial positions, mean, median, variance, and standard deviation) for classification that forming the malignant and benign tissues by differentiating foreground and background images. Second, after pre-processing the data and extracting features, predictive models developed by splitting the training and validation data. Then, selecting classification algorithms such as Decision Tree and Support Vector Machine (SVM) implemented. Afterward, trained iteratively and evaluate the selected classification algorithms. Overall, in developing this project, steps followed: First, accessing and exploration of the dataset. Second, pre-processing and feature extraction from the dataset. Next, the development of predictive models. Finally, model optimization procedures. Results & Discussion.

II. LITERATURE REVIEW

A huge amount of the literature is reviewed for investigating relevant studies and results refer to the analysis of invasive ductal carcinoma breast cancer detection with various datasets from different facilities in different countries. There are major methods such as Naïve Bayes, Support Vector Machine (SVM) Classifiers, and Ada Boost Methods, CNN Classifiers. Moreover, SVM and Random Forest classifiers are producing better results [3-6]. The classifier results are shown in Table 1 for further interpretation.

Methodology	**Accuracy**	**Precision**	**Recall**
Naive Bayes Classifier	95.61	95.65	93.61
Support Vector Machine Classifier	95.61	95.65	93.61
AdaBoost Method	95.75	95.72	96.26
Recurrent Neural Network (RNN)	91.3	91.3	89.3

Table 1 Performance Analysis of Breast Cancer IDC Detection

AdaBoost is a precision method for enhancement of the classification accuracy by combining several weak classifiers. RNNs are one of the sub-groups of Neural Network (NN). In contrast with the conventional NN, RNNs are accurate in processing the points in sequential data where the orders matter a lot. It is also ideally suitable for machine learning challenges even though the computation is slow and cannot consider any future input for its current state. Computational

Neural Networks uses spatial data among the image pixels and hence, discrete convolution is used for grayscale image processing.

Limitations of the existing methods

However, their certain limitations with the existing methods. For instance, Naïve Bayes Classifier produces distorted results when training data is not pre-processed. Support Vector Machine classifier is not suitable for large datasets and therefore not effective unless the dataset is cleaned. What is worse, if there is an imbalance problem in the dataset, AdaBoost increases error terms, and losses.

III.METHODOLOGY

Original Dataset

IDC image histopathology dataset is available in Kaggle [7] for studying this project. This dataset is referred from novel research published in Ref. [8]. There are 277,524 images, 50 x 50 pixels images extracted from thousands of IDC slide images. The images are labelled as positive or negative in terms of IDC. The dataset is downloaded from Kaggle which consists of 198,738 negative and 78,786 positive images. Due to the image file size concerns, the resolution of these images is down sampled by 16 times to be around 2.5x from 40x. Therefore, the patches are treated by 2.5x scaling down process. Figures 1 and 2 show randomly selected samples of IDC positive and negative from the dataset.

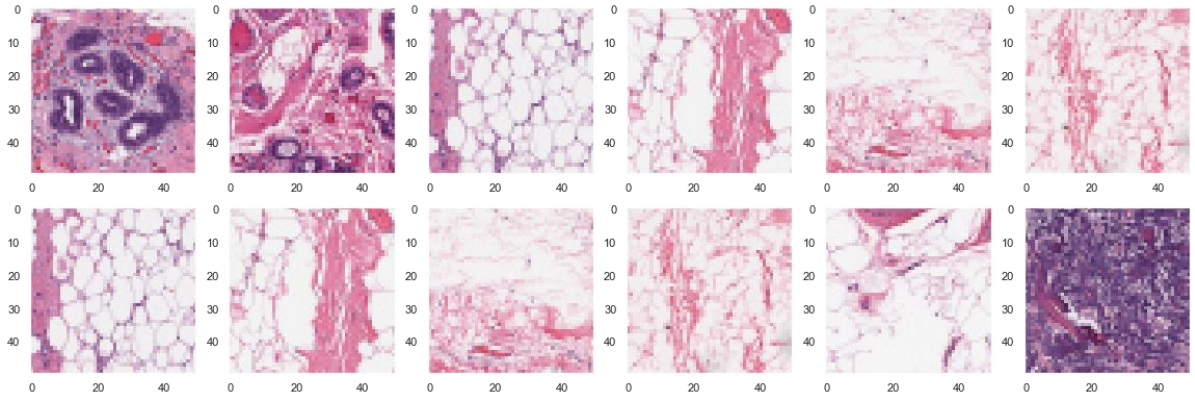

Figure 1 IDC Negative Images patches 50 x 50 (Bening)

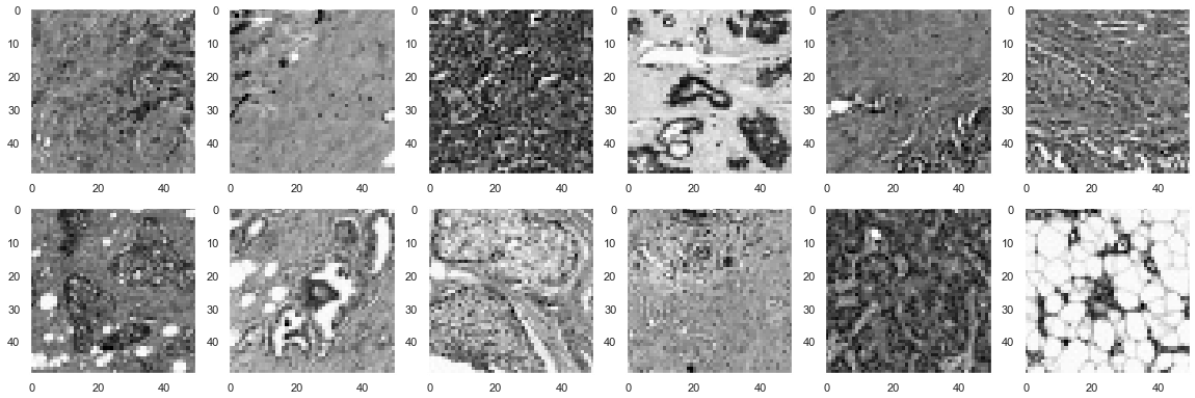

Figure 2 IDC Positive Images patches 50 x 50 (Malignant)

In Figure 2, malicious breast cancer tissues are highlighted with grayscale implementation as image segmentation.

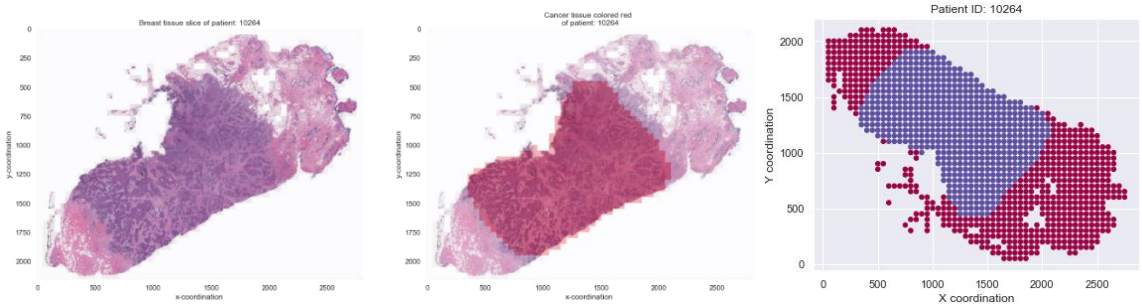

Figure 3 Instance of Whole Slide Images (WSI), IDC Positive masked in red and blue.

In figure 3, malignant places prone to appear in groups rather than being spread out all over the tissue.

Exploratory Data Analysis

The first step is to review what kind of data we are working with. Therefore, some visualizations are created as an initial exploration. In the original dataset, the major part of it was imbalanced. 72% of the image patches negative while 28%, around 78,786 images, positive cases. This imbalanced dataset would cause bias in the learning of our machine learning model. In order not to cause bias in the ML model, after the feature extraction process, as a validation strategy target information is added for generating some balance.

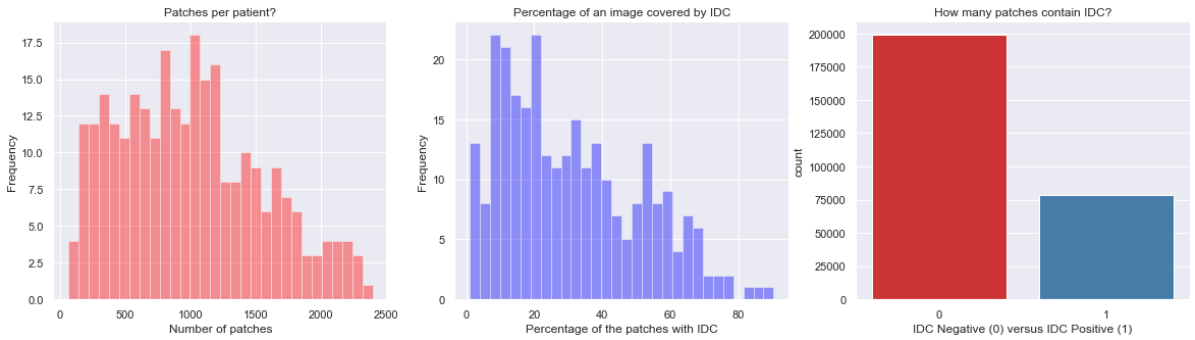

Figure 4 Exploratory Data Analysis for interpreting the dataset.

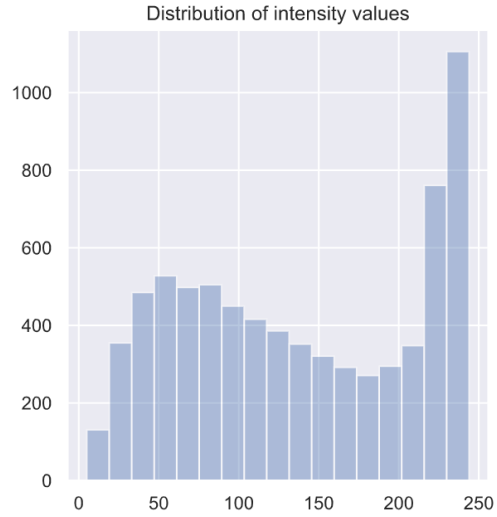

Figure 5 Intensity Value Distribution of the Malicious Images

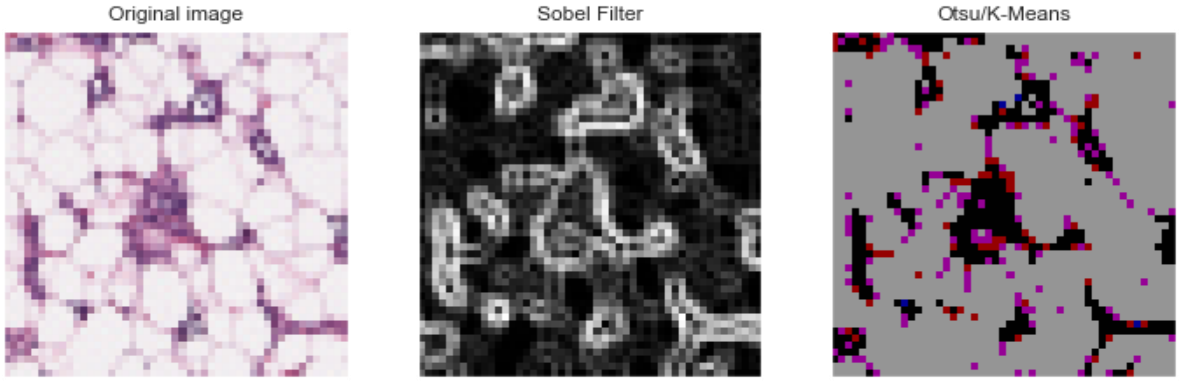

Figure 6 Image Segmentation: Sobel Edge Detection, Otsu's Method for Masking on malicious tissues.

In Figure 6, it can be interpreted those malicious tissues are located and masked with the help of foreground and background features extracted from the images. Otsu's method works at high accuracy levels. In the same figure, the Sobel filter detected the edges of the malicious tissues for further implementations.

Feature Extraction

Feature extraction is a major step of machine learning workflow since provides the information from images which is vital for ML algorithms to produce the most accurate results. *The scikit-image* library is used for feature extraction with Grayscale Pixel Values since our images are represented by pixels and it is the simplest way to create image features. There is a library named *io.imread(image path)* that converts an image into its numerical form. After that, using with NumPy arrays (n x m x 3) for Red Blue Green RGB converted into grayscale shapes as (n x m x 1). After that, we obtained image dimensions as input. Then, another method named thresholding was applied for separating the foreground and background of the image that is a simple image segmentation

applied on image pixel parameters for segmenting the background. Then Otsu's method was applied for finding optimal threshold values. At the end of this process, some statistical features and image-related features are obtained such as mean, standard deviation, variance, corners, image pixel intensities, sizes, and patterns of the cells. It also created binary images which were a pre-step for clustering and vector quantization. Afterward, Sobel filter is implemented for edge detection by calculating the gradient of image intensity with the values at each pixel in an image and provided the rate of change from intensity distribution shown in Figure 5.

Principal Component Analysis (PCA)

PCA is also a famous feature extraction algorithm in ML. It finds the eigenvectors of a variance matrix with the highest eigenvalues and then uses them to form the data into a new and dimensionally reduced space [9]. It reveals how a variable is associated with another variable (named as co-variance matrix). It reduces feature numbers by reconstructing the same dataset with a smaller number of the original features. Before implementing PCA, feature standardization methods are applied to normalize the breast cancer dataset. The role of using PCA is that there are numerous numbers of features, and it makes predictions less accurate. After applying PCA, the number of features is reduced to 150 which is shown in Figure 7. What is more, it also contributed to this project by reducing the over-fitting of the breast cancer data.

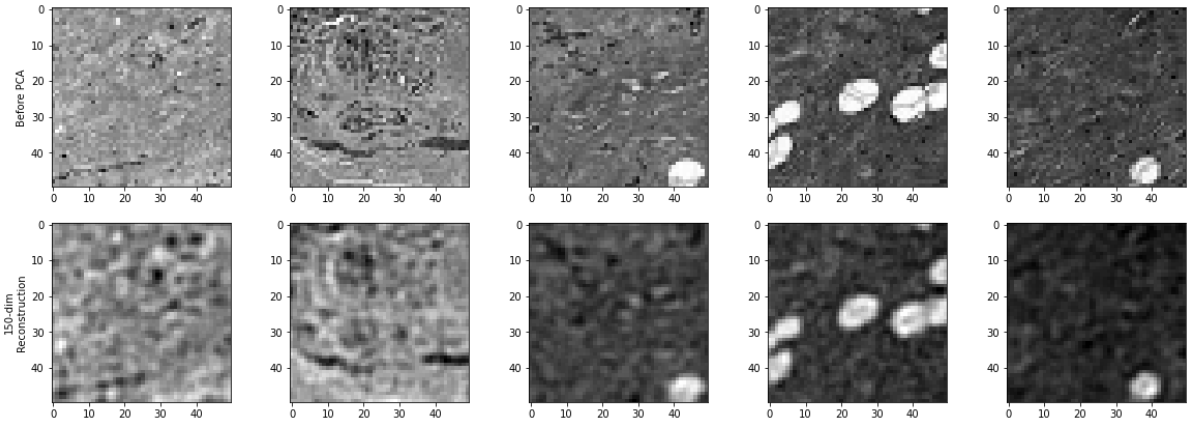

Figure 7 Before and after PCA Implementation on the dataset

PCA graph helps us to underline that the first 150 components include 80 percent of the variance in the breast cancer data. The Variance Ratio after the PCA graph shows that with 150 components in Figure 8.

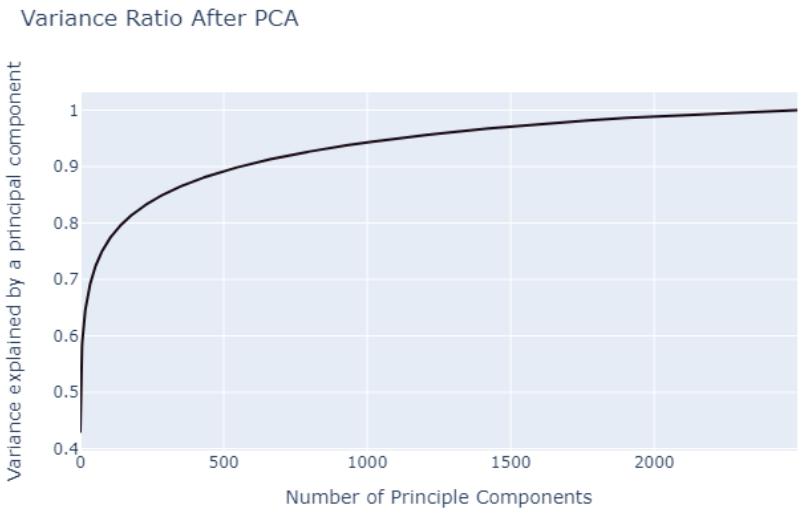

Figure 8 Variance Ratio after PCA

The graph also shows that we reduced the array of images to 150 from 2500 but able to sustain 80% of the valuable information. Overall, we reduced the dimension numbers by 96 percent. A **high variance** states that the data points are spread out from the mean, and one another [10].

Data Split

In our machine learning model, the train-test split method is used for splitting the dataset into train and test sets. After that, Principal Components Analysis (PCA) and Support Vector Machine (SVM) algorithms are implemented for reducing the size of the dataset but not losing critical information and originality. The ratio of malignant to benign images is not balanced. The original dataset is not uniformly distributed. Since it would cause bias on the machine learning model a Python library named *train-test split function from sklearn model selection* is used. After this process, we obtained a uniformly distributed dataset for train and test classes.

Training Dataset

The samples from the breast cancer dataset were used to fit the model. Simply, the model sees and learns from this dataset.

Validation Dataset

To evaluate the model without unbiasing, this validation dataset is used on the training dataset during the tuning model hyperparameters. The validation dataset is mainly used for setting results and updating higher-level hyperparameters. It can also be named as Development set since this dataset assists during the development process of the model.

Test Dataset

This sample of the data is mainly used to provide an unbiased evaluation of the final model fit on the training dataset. It is only used once a model is completely trained. It provides the gold standard used to evaluate the model. Machine learning models are generally trained using these observations. Then, the learning algorithm is fed by these observations to update its parameters during the learning phase. After the training phase, we test the machine learning model by using observations from the test set. Hence, how the model reacts to new observations can be measured properly.

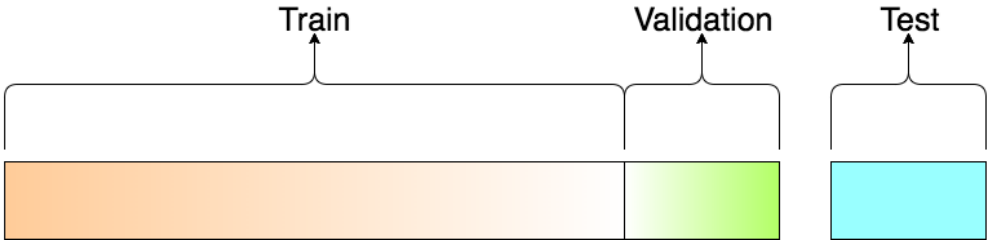

Figure 9 Visualization of the splits. Training 70%, Validation 15%, Testing 15% in the ML project. Adapted from: towards data science.

Split Ratio

To interpret machine learning model behavior, some observations are required that are not used in the training process. Otherwise, the model would be biased shown in Figure 9.

IV. MODEL EVALUATION

In this phase, how accurate the model train is evaluated. Firstly, the scores were checked out on the trained model and obtained accuracy levels of each model. Evaluation metrics' selection is directly related to the machine learning task, for instance, classification. Classification accuracy is also the most preferred metric for classification problems.

Classification Algorithms	Optimized Accuracy	Accuracy Levels
Random Forest Classifier	77.02%	68.25%
K Neighbors Classifier	70.80%	67.53%
Stochastic Gradient Descent Classifier	73.71%	63.11%
Decision Tree Classifier	68.14%	67.13%
Linear Support Vector Classifier	69.45%	66.81%
Support Vector Machine Classifier	75.20%	74.40%
SVM-RBF Kernel Classifier	78.90 %	73.78%

Table 2 IDC (+) and IDC (-) Classification Performances of algorithms

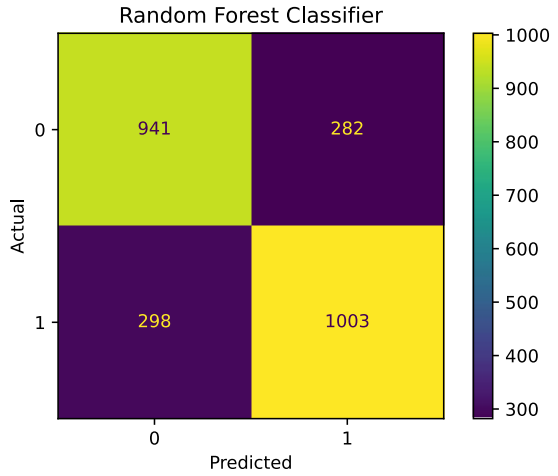

Figure 9 Random Forest Classifier IDC Prediction

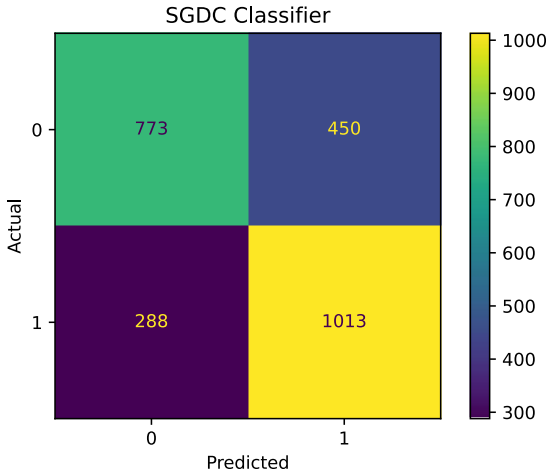

Figure 10 Stochastic Gradient Descent Classifier

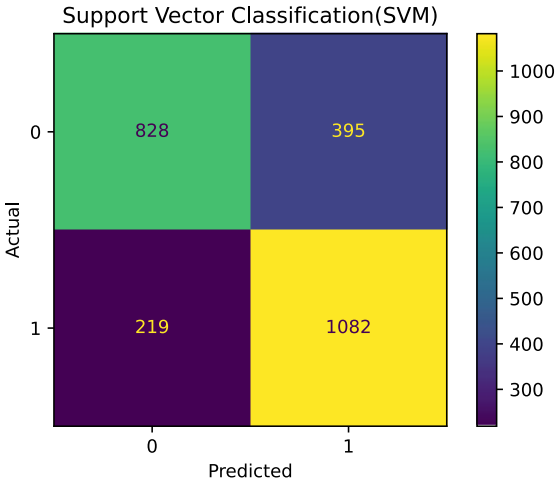

Figure 11 Support Vector Machine Classifier

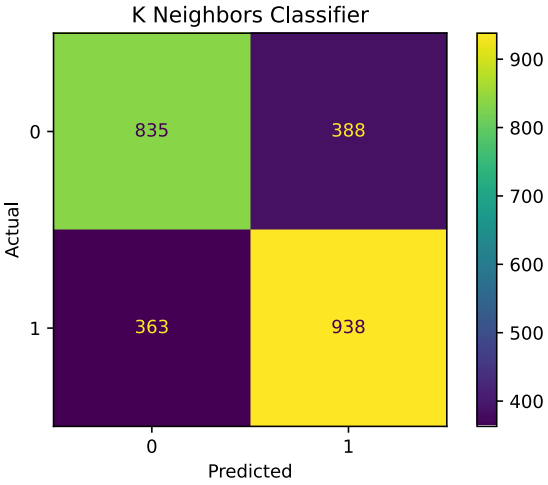

Figure 12 K-Neighbors Classifier

Results & Discussion

In Figure 11, Support Vector Machine Classifier is implemented because it is fast, but it does not produce better results in large machine learning datasets. Therefore, the accuracy is around 77%. In Figure 12, the KNN Algorithm simple to understand how it works and implement. Training duration is less. However, since this is a high-dimension dataset, the accuracy is affected negatively. In Table 2, a comparison of the different classification algorithms is listed. According to our literature review, accuracy levels of the classifications [12] are slightly higher than our results since in our case data augmentation is not implemented. In a similar study, researchers used Google Automatic Machine Learning (AutoML) models for detection of invasive ductal carcinoma [12, 13] tissues and obtained the most accurate results in this field since AutoML optimizes selection, implementation, and parameter tuning.

Random Forest Classifier

It is one of the easiest and flexible algorithms for implementation in both classification and regression. It builds decision trees on randomly selected data samples, obtains, and evaluates the best solutions in terms of evaluations for IDC positive or negative detection [11]. In this project, the main reason for selection is Random Forest Algorithm does not suffer from overfitting problems in machine learning applications shown in Figure 9 by using a confusion matrix since it provided detailed true and false classifications for each class.

RBF Support Vector Machine Classifier

Radial Basis Function kernels are the most generalized form of kernel implementation. It is widely used since it is similar to the normal distribution. It computes the similarity between two points to estimate how close they are to each other. In our implementation, RBF Support Vector Machine Classifier produced the best result (78.90% Accuracy). It is better than SVM and similar to the KNN algorithm but superior to them since reduces complexity by storing support vectors instead of the entire dataset.

V. CONCLUSION

Probability maps are obtained with the help of binary classification since it predicts negative and positive IDC groups according to the confidence threshold probability values received by the image segmentation process. The optimal threshold is found 0.30 reflecting different degrees of confidence intervals in IDCs. Based on the threshold values, intensity, and probability models produced as the intensity matrix parameters extracted from images, this step is easily conducted shown in Figure 12.

Overall, invasive ductal carcinoma detection with machine learning algorithms is a difficult and complex task for breast cancer diagnosis since high accuracy levels are essential to detect IDC for the treatment of breast cancer cases. This study is a fundamental machine learning model for this field of disease as previous researchers mainly used deep learning methods for this task requires large datasets and time for analyses and learn.

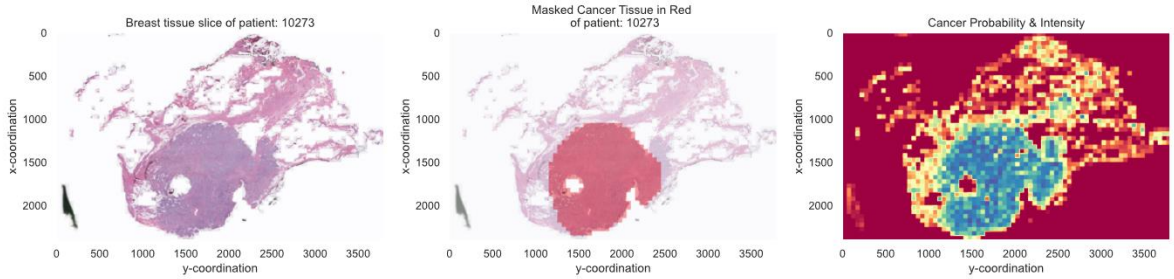

Figure 12 IDC Probability Heatmap

VI. ACKNOWLEDGEMENTS

We would like to thank Professor Bülent Yılmaz, Dr. Hakan Aksebzeci, and Dr. Burak Ünal, and Teaching Assistants: Oğuzhan Ayyıldız and Fatih Altındiş for assisting us throughout the Biomedical System Design Capsule and on this project.

REFERENCES

[1] National Breast Cancer Foundation, Invasive Ductal Carcinoma (IDC), 2020. https://www.nationalbreastcancer.org/invasive-ductal-carcinoma.

[2] Breastcancer.org, "Invasive Ductal Carcinoma (IDC)", 2020. https://www.breast cancer.org/symptoms/types/idc.

[3] Anji Reddy V., Soni Badal. Breast cancer identification and diagnosis techniques. Machine Learning for Intelligent Decision Making, Springer (2020)

[4] Pan Qiao, Zhang Yuanyuan, Chen Dehua, Xu Guangwei
Character-Based Convolutional Grid Neural Network for Breast Cancer Classification IEEE (2017)

[5] Khan SanaUllah, Islam Naveed, Jan Zahoor, Din Ikram Ud, Rodrigues Joel J.P.C.
A novel deep learning-based framework for the detection and classification of breast cancer using transfer learning Pattern Recognition Letters, Elsevier (2019)

[6] Huang Qinghua, Chen Yongdong, Liu Longzhong, Tao Dacheng, Li Xuelong
On combining biclustering mining and AdaBoost for breast tumor classification.

[7] P. Mooney, Breast Histopathology Images, Kaggle, 2018. https://www.kaggle.com /paultimothymooney/breast-histopathology-images/data.

[8] A. Cruz-Roa, A. Basavanhally, F. Gonzalez, H. Gilmore, M. Feldman, S. Ganesan, et al., Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. SPIE Medical Imaging, 2014.

[9] A. Geron, Hands-On Machine Learning with Scikit-Learn and TensorFlow, first ed., O'Reilly Book, 2017, pp. 84–92.

[10] P. Pandey, AutoML: The next wave of machine learning. Heartbeat in Medium, 2019. https://heartbeat.fritz.ai/automl-the-next-wave-of-machine-learning-549 4baac615f.

[11] M. Heller, Automated Machine Learning or AutoML Explained, InfoWorld, 2019. https://www.infoworld.com/article/3430788/automated-machine-learning-or-au toml-explained.html.

[12] A. Borkowski, C. Wilson, S. Borkowski, et al., Google Auto ML versus Apple Create ML for Histopathologic Cancer Diagnosis; Which Algorithms Are Better? 2019.

[13] Bengio, Y., "Learning deep architectures for ai," Foundations and Trends in Machine Learning 2, 1–127 (Jan. 2009).