Invasive Ductal Carcinoma Classifier

Suryansh Ankur
University of Alberta
Edmonton, Canada
sankur@ualberta.ca

KEYWORDS

Neural, Network, SVM, Random, Forest, Breast, Cancer, Classifier

ACM Reference Format:
Suryansh Ankur. 2024. Invasive Ductal Carcinoma Classifier. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Invasive Ductal Carcinoma (IDC) is a prevalent form of breast cancer known for its aggressive nature and tendency to metastasize, posing significant health risks. Biopsies are commonly performed to extract tissue samples for diagnosis, requiring pathologists to manually identify IDC cells, differentiate them from other cancer types, or determine if the tissue is healthy. However, this manual process is time-consuming and reliant on pathologist expertise, which may be scarce in certain regions.

To address these challenges, machine learning offers a promising solution by automating the detection and localization of tumor tissue cells, thereby expediting the diagnosis process. By developing and comparing three distinct machine learning algorithms - Random Forest, Support Vector Machine, and Neural Network - we aim to evaluate their effectiveness in predicting IDC presence. This approach can potentially reduce reliance on pathologists and enhance diagnostic efficiency, particularly in areas lacking expert medical personnel.

2 PROBLEM SETUP:

Dataset I'm using is from Kaggle which consist of 162 whole mount slide images of Breast Cancer specimens scanned at 40x. From that, 277,524 patches of size 50x50 were extracted (198,738 IDC negative and 78,786 IDC positive). Each patch's file name is of the format: u_xX_yY_classC.png where u is the patient ID, X is the x-coordinate of where this patch was cropped from, Y is the y-coordination of where this patch was cropped from, and C indicates the class where 0 is non-IDC and 1 is IDC.

URL to the dataset: Link.

2.1 Data-preprocessing:

Our model goes to the parent folder 1 by 1 and appends the image into X and its label in y, non-cancerous cell has label 0 and cancerous cell has label 1, we resize the image to 50x50 and the convert it to grayscale from rgb. Since, we had a data imbalance, we only pick 78,786 out of 198,738 IDC negative images. Then we randomly shuffle the dataset to reduce biasness. We split the data between 80% training data and 20% testing data.

2.2 Baseline:

My baseline was to randomly guess on test dataset which resulted on a baseline of 50.546% accuracy. Our model needs to do at least better than randomly guessing.

3 APPROACH:

Model Selection:

Random Forest: You used the Random Forest classifier from the scikit-learn library with specific hyperparameters (number of estimators = 100 and random state = 42).

Support Vector Machine (SVM): You employed the SVM classifier with a linear kernel, suitable for linearly separable data.

Neural Network (NN): You built a neural network from scratch using Python, implementing two dense layers with Rectified Linear Unit (ReLU) activation functions followed by a softmax activation function in the output layer. You used the categorical cross-entropy loss function, Adam optimizer, and categorical accuracy for evaluating the model.

Training and Evaluation: You trained each model on the preprocessed training data and evaluated their performance using appropriate evaluation metrics (e.g., accuracy for Random Forest and SVM, categorical accuracy for Neural Network).

Comparison: After training and evaluating all models, you compared their performance based on the achieved accuracy. This comparison provides insights into the relative strengths and weaknesses of each model in solving the classification task.

Model Interpretation: By understanding the inner workings of each model (e.g., decision boundaries for Random Forest and SVM, learned weights and activations for Neural Network), you can gain insights into how they make predictions and interpret their results.

Overall, your approach demonstrates a good understanding of machine learning principles and a systematic methodology for solving classification problems. By leveraging different types of models and evaluating their performance, you can make informed decisions about which model(s) to use for your specific task. Additionally, building a neural network from scratch showcases your ability to implement deep learning models and customize them according to your requirements.

4 EVALUATION:

We use accuracy to evaluate our model, this seemed as a correct approach as in medical diagnosis tasks, such as breast cancer detection,

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/XXXXXXX.XXXXXXX

false positives and false negatives have different consequences. Misclassifying a cancerous tumor as benign (false negative) could lead to delayed treatment and potentially worsen the patient's condition. On the other hand, misclassifying a benign tumor as cancerous (false positive) could lead to unnecessary medical procedures and psychological distress for the patient.

5 FIGURES

Confusion Matrix for the models

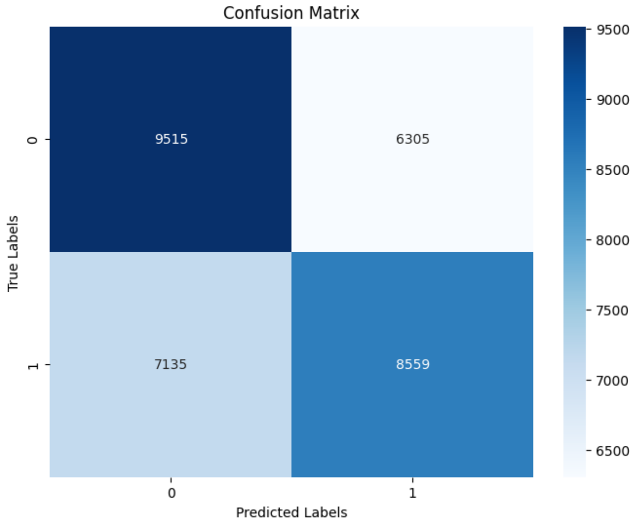

Figure 1: Support Vector Machine

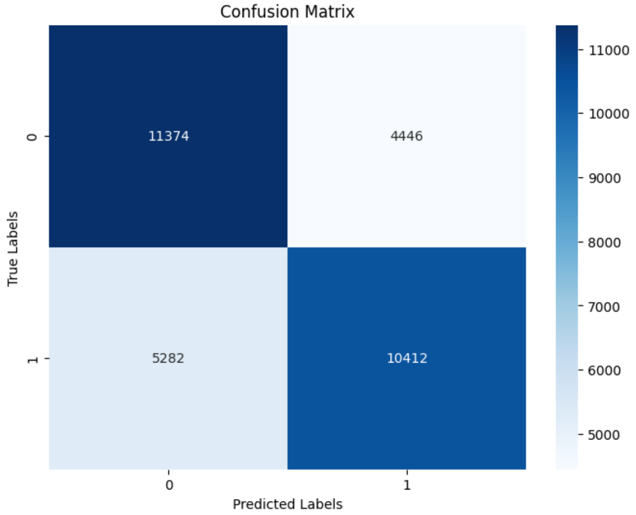

Figure 2: Neural Network

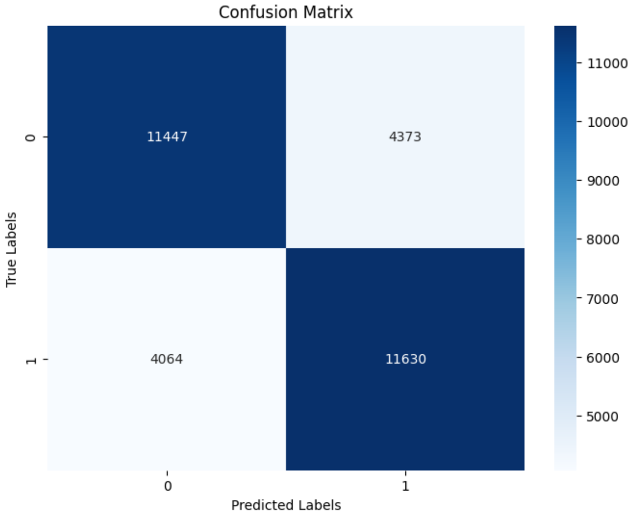

Figure 3: Random Forest

6 RESULT:

Random Baseline (50% accuracy): The random baseline serves as a point of reference, indicating the minimum level of performance that any model should surpass. Achieving an accuracy close to 50% suggests that the classification task is not trivial and requires learning meaningful patterns from the data. Support Vector Machine (57% accuracy): SVM performed better than the random baseline but still fell short of satisfactory performance. The margin of improvement over the random baseline indicates that SVM learned some discriminative features but may not have captured all the relevant patterns in the data. Further optimization or exploration of different kernel functions and hyperparameters might yield better results. Neural Network (70% accuracy): The neural network achieved a significant improvement over the random baseline and SVM. Its ability to learn complex patterns and relationships in the data contributed to the higher accuracy. However, the performance might be limited by factors such as the resolution of the images and the size of the dataset. Fine-tuning the architecture, training parameters, and preprocessing techniques could potentially yield even better results. Random Forest (74% accuracy): Surpassing all other models, Random Forest emerged as the top performer. Its ensemble approach and robustness to noisy data likely contributed to its superior performance. Random Forest's ability to handle high-dimensional data and limited training samples might have been advantageous in this scenario. The result underscores the importance of considering diverse machine learning algorithms and experimenting with different approaches.

7 IMPLICATIONS:

The results suggest that Random Forest could be a suitable choice for this classification task, given its performance and robustness. Further experimentation with ensemble methods or hybrid approaches could potentially enhance performance even more. Investigating strategies to improve the quality and resolution of the images, as well as expanding the dataset, might lead to better performance across all models. Regular evaluation and comparison of different models are essential to ensure the selection of the most effective approach for the given problem domain. Understanding the limitations and strengths of each model is crucial for making informed decisions and driving improvements in classification accuracy.

8 CHALLENGES FACED:

There were few challenges that were faced during production of code. SVM took more than 6 hours to fit the model, finding appropriate layer for the Neural Net and experimenting with different hyperparameters was an intensive task.

9 REFERENCES:

Neural Network from Scratch, Harrison Kinsley, Daniel Kukieła : https://books.google.ca/books?id=Ll1CzgEACAAJ