



Article Title: **Breast Cancer Detection Using Machine Learning Classifier**

Breast Cancer Detection Using Machine Learning Classifier

Tharun Kumar M D¹, DR. G. Soniya Priyatharsini², DR. S. Geetha³

¹ Student, Department of Computer Science and Engineering, Dr. M. G. R Educational and Research Institute of Technology, Madhuravoyal, Chennai-95, TamilNadu, India

^{2,3} Associate Professor, Department of Computer Science and Engineering, Dr. M. G. R Educational and Research Institute of Technology, Madhuravoyal, Chennai-95, TamilNadu, India

Corresponding Author, Mail ID: *soniya.cse@drmgrdu.ac.in

ABSTRACT

Breast cancer is one of the main causes of cancer death worldwide. Early diagnostics significantly increases the chances of correct treatment and survival, but this process is tedious and often leads to a disagreement between pathologists. The diagnosis is based on the qualification of histopathologist, who will look for abnormal cells. However, if the histopathologist is not well-trained, this may lead to wrong diagnosis. Computer-aided diagnosis systems showed potential for improving the diagnostic accuracy. In this work, we develop the computational approach based on deep convolution neural networks for breast cancer histology image classification. Hematoxylin and eosin stained breast histology microscopy image dataset is provided as a part of the ICIAR 2018 Grand Challenge on Breast Cancer Histology Images. Our approach utilizes several deep neural network architectures. Convolutional Neural Networks for Binary class classification and multiclass classification. The Binary class classification is used to classify the cancer cells to malignant and benign. And the Multiclass classification these classes into different subclasses like adenosis, fibroadenoma, phyllodes tumour, tabular adenoma for benign class and ductal carcinoma, lobular carcinoma, mucinous carcinoma, papillary carcinoma for malignant class. The result will show Convolutional Neural Networks outperformed the handcrafted feature based classification with high accuracy in both binary and multiclass classification.

Keywords: Machine Learning, Logistic Regression, Naïve Bayes, Support Vector Machine, Random Forest, Convolutional Neural Network, Image Processing.

1 Introduction

Image processing is a method to convert an image into digital form and perform operations on it, in order to get an enhanced image or extract useful information from it. It is a type of signal dispensation in which input is image and output is image or characteristics associated with that image. Image processing involves three steps: importing the image to the system, analyzing and manipulating the image, and output is the result of an altered image or report based on image analysis. It is among the fastest growing technologies today, with its applications in various aspects of a business.



Article Title: Breast Cancer Detection Using Machine Learning Classifier

Convolutional Neural Networks (CNNs) are a variation of multilayer perceptrons designed to require minimal preprocessing. They are inspired by biological processes in that the connectivity pattern between neurons resembles the organization of the animal visual cortex. CNNs use relatively little pre-processing compared to other image classification algorithms, allowing them to learn filters that were hand-engineered. This independence from prior knowledge and human effort is a major advantage.

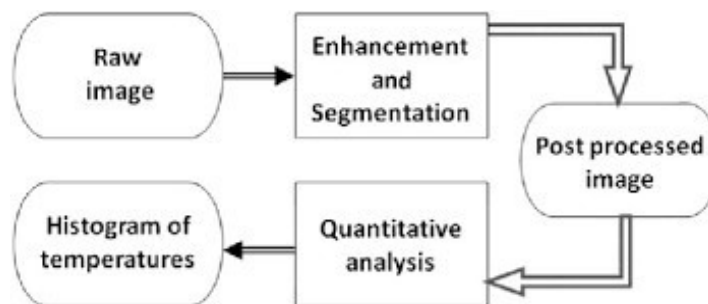


Figure 1: Basic Steps of Image Processing

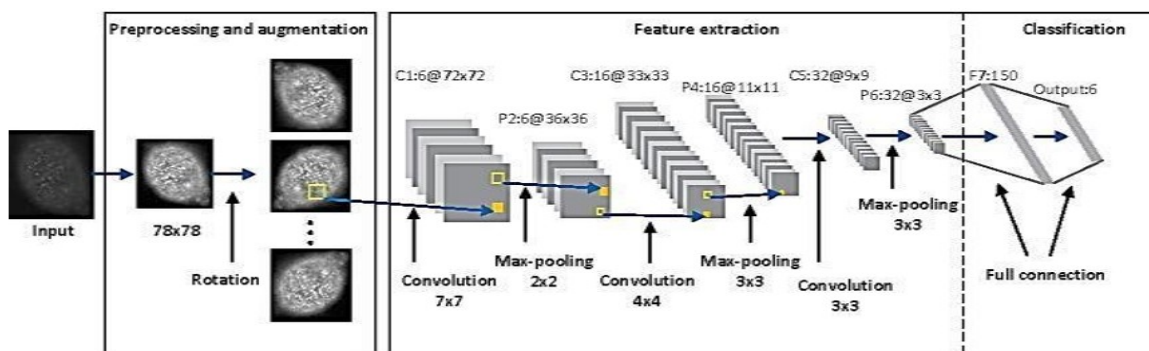


Figure 2: Image Processing Using CNN

2 Literature Review

2.1 Classification of Breast Cancer Based on Histology Images using Convolutional Neural Networks

The work is contributed by Dalal Bardou et.al [1]. It compares two machine learning approaches for the automatic classification of breast cancer histology images into benign and malignant sub-classes. The first approach is based on the extraction of handcrafted features encoded by two coding models and trained by support vector machines. The results show that convolutional neural networks outperformed the handcrafted feature based classifier, achieving accuracy between 96.15% and 98.33% for the binary classification and 83.31% and 88.23% for the multi-class classification. This paper aims to develop a reliable pattern recognition based systems to improve the quality of diagnosis.



Article Title: Breast Cancer Detection Using Machine Learning Classifier

2.2 Bacterial colony counting with Convolutional Neural Networks in Digital Microbiology Imaging

The work is contributed by Alessandro Ferrari et.al [2]. This work explores the possibility of finding effective solutions to the classification of bacterial colonies using two different machine learning approaches. The first is based on the extraction of handcrafted morphometric and radiometric features, while the second is based on a Convolutional Neural Networks deep learning architecture. To validate the proposed bacterial load estimation techniques, a large and representative database of both single and aggregated bacterial colonies was built and publicly released. Dataset enhancement approaches were also tested for performance optimization. The adopted deep learning approach outperformed the handcrafted feature based one, and also a conventional reference technique, becoming a preferable solution for the Digital Microbiology Imaging quantification task, especially in the emerging context of Full Laboratory Automation systems.

2.3 A Dataset for Breast Cancer Histopathological Image Classification

The work is contributed by Fabio A. Spanhol et.al [3]. This paper introduces a dataset of 7909 breast cancer histopathology images acquired on 82 patients, which is now publicly available from <http://web.inf.ufpr.br/vri/breast-cancer-database>. The task associated with this dataset is the automated classification of these images in two classes, which would be a valuable computer-aided diagnosis tool for the clinician. Preliminary results obtained with state-of-the-art image classification systems show that accuracy ranges from 80% to 85%, showing room for improvement. By providing this dataset and a standardized evaluation protocol to the scientific community, we hope to gather researchers in both the medical and the machine learning field to advance toward this clinical application.

2.4 Breast Cancer Multi- classification from Histopathological Images with Structured Deep Learning Model

The work is contributed by Zhongyi Han, Benzheng Wei et.al [4]. Automated breast cancer multi-classification from histopathological images is a key role in computer-aided breast cancer diagnosis or prognosis. However, it faces two main challenges: the difficulty of binary classification and the subtle differences in multiple classes due to the broad variability of high-resolution image appearances, high coherency of cancerous cells, and extensive inhomogeneity of color distribution. This study proposes a deep learning model using a newly proposed deep learning model that has achieved remarkable performance (average 93.2% accuracy) on a large-scale dataset, demonstrating the strength of the method.

2.5 Deep Convolutional Neural Networks for Breast Cancer Histology Image Analysis

The work is contributed by Zhongyi Han, Benzheng Wei et.al [4]. This work develops a computational approach based on deep convolution neural networks for breast cancer



Article Title: Breast Cancer Detection Using Machine Learning Classifier

histology image classification. It utilizes several deep neural network architectures and gradient boosted trees classifiers to report 87.2% accuracy, AUC 97.3%, and sensitivity/specificity 96.5/88.0% at the high-sensitivity operating point. This approach outperforms other common methods in automated histopathological image classification.

2.6 Gland segmentation in colon histology images using hand-crafted features and convolutional neural networks

The work is contributed by Stephen J. McKenna et.al [5]. We investigate glandular structure segmentation in colon histology images as a window-based classification problem. We compare and combine methods based on fine-tuned convolutional neural networks (CNN) and hand-crafted features with support vector machines (HC-SVM). We find that CNN outperforms HC-SVM in gland segmentation measured by pixel-wise Jaccard and Dice indices. We also observe that training a second-level window classifier on the posterior probabilities can substantially improve the segmentation performance. Finally, we show that by combining and refining the posterior probability outputs, a further performance boost is obtained.

2.7 Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images

The work is contributed by Babak Ehteshami Bejnordi et.al [6]. This text presents a context-aware stacked convolutional neural networks (CNN) for classification of breast WSIs into normal/benign, DCIS, and invasive ductal carcinoma (IDC). The system is trained and evaluated on 221 WSIs of hematoxylin and eosin stained breast tissue specimens and achieves an AUC of 0.962 for binary classification of nonmalignant and malignant slides and a three-class accuracy of 81.3%. It has potential for routine diagnostics.

To develop a reliable pattern recognition to classify breast cancer based on histological dataset using Convolutional Neural Networks (CNN) using handcrafted feature encoding models, binary and multiclass classifiers to improve the quality of diagnosis

3 Existing System

Existing System are the works that are already implemented successfully. Techniques used in the existing system are described below,

- With the recent advances in image processing and machine learning, there is an interest in attempting to develop a reliable pattern recognition based systems to improve the quality of diagnosis. In this paper, we compare two machine learning approaches for the automatic classification of breast cancer histology images into benign and malignant and into benign and malignant sub-classes.
- The first approach is based on the extraction of a set of handcrafted features encoded by two coding models (bag of words and locality constrained linear coding)



Article Title: Breast Cancer Detection Using Machine Learning Classifier

and trained by support vector machines, while the second approach is based on the design of convolutional neural networks.

- We have also experimentally tested dataset augmentation techniques to enhance the accuracy of the convolutional neural network as well as "handcrafted features + convolutional neural network" and "convolutional neural network features + classifier" configurations.
- The results show convolutional neural networks outperformed the handcrafted feature based classifier, where we achieved accuracy between 96.15% and 98.33% for the binary classification and 83.31% and 88.23% for the multi-class classification.

4 Proposed System

The proposed system is the system which is developed to enhance the shortcomings of existing system by using advanced techniques.

4.1 Advantage

Our Proposing work using CNN topology has to improve the accuracy of the previous ones for the multiclass classification task, where they reached a performance of between 83.31% and 88.23%

4.2 System Requirement

The project will require some software and hardware to develop the system and these resources are mandate to develop the project with utmost accuracy. The hardware and software required to develop the proposed system are listed below:

4.3 Hardware Requirements

Hardware specifications are technical description of the computer's components and capabilities. Processor speed, model and manufacturer, etc. So the hardware components required for the proposed system are:

Processor: Intel Core i5.

Hard disk: 1TB.

Speed: 1.80GHz

Memory: 4GB.

4.4 Software Requirements

A software requirement specification is a description of a software system to be developed. It lays out functional and non- functional requirements and it also describes the operating system and tool used in the system and they are:

Operating system: Ubuntu 16.4 LTS.

Front end: Keras.



Article Title: Breast Cancer Detection Using Machine Learning Classifier

Back end: Tensorflow, Open CV package.

4.5 System Architecture

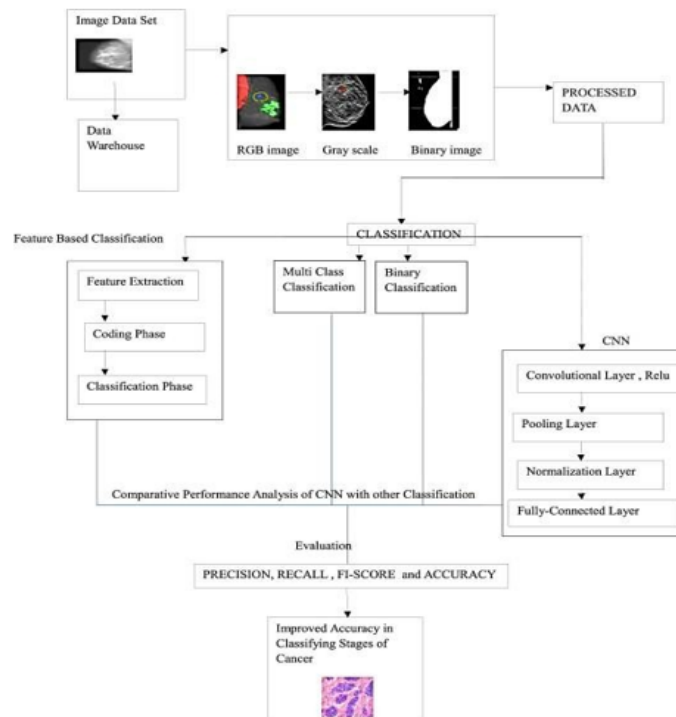


Figure 3: *Architecture of Breast Cancer Detection Using ML*

4.6 Problem Formulation

Right now, breast cancer is affecting a lot of people. This disease's cause is dependent on human factors and is difficult to pinpoint. Also, it takes a great deal of work from doctors and physicians to identify the type of cancer and determine if it is benign or malignant. Even for doctors, determining the final result may be challenging when several tests, such as those for clump thickness, uniformity of cell size, uniformity of cell form, etc., are involved in the detection of breast cancer. Due to this, the use of machine learning and computing in general as diagnostic tools has increased over the last few years. Computer-based solutions for diagnostics are widely employed for the diseases that claim many lives. In operations rooms, robotics plays a very important and required function. The rigorous treatment rooms also provide the skilled systems. It is also acceptable to use a different aspect of artificial intelligence to identify breast cancer. According to reports, breast cancer is the second most frequent type of cancer that affects women, and by the year 2002, it was the most common disease worldwide.

This cancer may be a rather prevalent type in young women, making it the second leading cause of cancer death. In the US, about one in eight females have an increased risk of



Article Title: Breast Cancer Detection Using Machine Learning Classifier

developing breast cancer over their lifetime. One breast cell dividing uncontrollably causes the start of breast cancer, which manifests as a visible growth known as a tumour. The tumour may be cancerous or benign. The right classification in identifying whether the tumour is benign or malignant may save lives. Thus, the need for accurate classification inside the clinic may be the cause of the doctors' and specialists' great anxiety. When scientists first realised the value of making constrained decisions to cure particular diseases, the importance of artificial intelligence has grown significantly during the previous 25 years. One of the most important diseases in medicine where the classification problem plays a really important part is that of breast cancer diagnosis, where the use of machine learning and data processing as tools for diagnosis becomes incredibly effective.

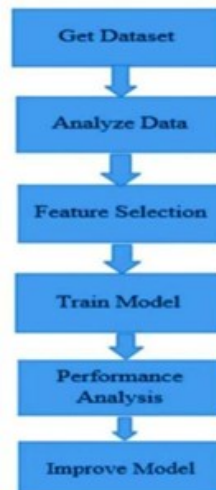


Figure 4: *Sequential Flow of Dataset and Analysis*

Hence, using machine learning techniques will make it easier for clinicians to identify breast cancer and classify tumours correctly as benign or malignant. In order to study various machine learning methods, we'll employ a number of algorithms and apply them to a dataset of breast cancer cases. Machine learning methods will be our main focus, including Naive Bayes, K-nearest Neighbor, logistic regression, reinforcement learning, and support vector machines. We'll mostly research these different algorithms and evaluate their output.

5 Working Principle

The fundamental goal of principle component analysis (PCA) is to reduce the number of variables in a data collection that are strongly or weakly associated to one another while maintaining the maximum amount of variance within the data set. The same is completed by redesigning the variables into a substitute set of variables known as the principal elements (or simply, the PCs), which are orthogonal and arranged such that the retention of variation contained within the original variables reduces as we tend to move down within the order. As



Article Title: Breast Cancer Detection Using Machine Learning Classifier

a result, the first main element in this method keeps the majority of the variation that was given within the original elements. As the Manfred Eigen vectors of a covariance matrix are the main components, they are orthogonal. It's important to note that the dataset for the PCA technique needs to be scaled. The relative scale affects the outcomes.

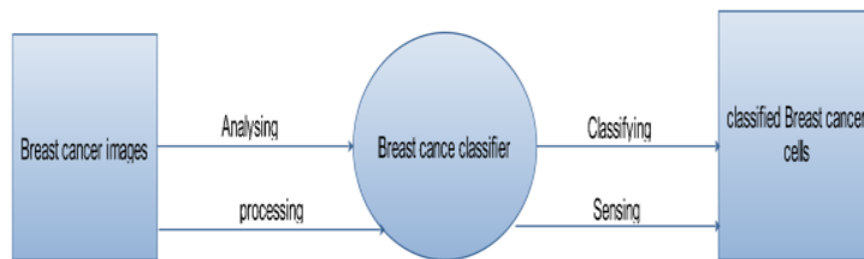


Figure 5: Data Flow Diagram on Breast Cancer Dataset

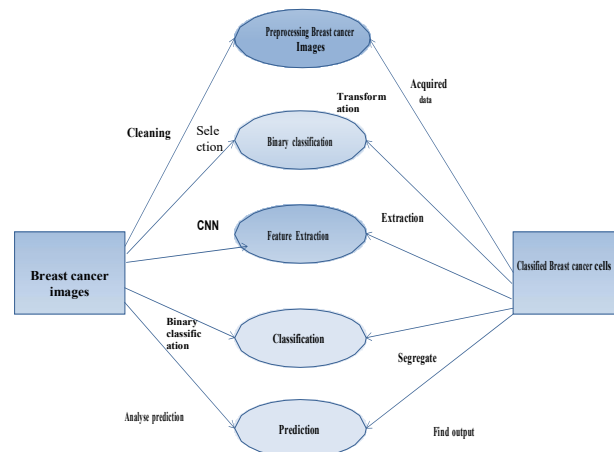


Figure 6: Data Flow Diagram of Breast Cancer Classifier

In simple terms, it's a method of summarising data. Consider some wine bottles perched on a board. Each wine is distinguished by characteristics including colour, body, age, and others. Yet, redundancy may occur since numerous of them may reside in connected properties. In order to balance this situation, PCA might summarise each wine in the stock using fewer features. Intuitively, Principal Part Analysis will give the user a lower- dimensional representation of this object, a projection or "shadow," once seen from its most informative angle.

6 Result Analysis

Following the application and implementation of machine learning models, the next stage is to determine the model's effectiveness, or how well it performed on the datasets. This is done by applying the models to the previously created test dataset. 30% of the dataset for the prediction of breast cancer was included in the test dataset. Breast cancer prediction also



Article Title: Breast Cancer Detection Using Machine Learning Classifier

underwent 10-fold cross-validation. Several measures have been applied in order to evaluate and compare the performances of the various algorithms.

Performance Metrics

The effectiveness of the Machine Learning algorithms in this thesis has been assessed using a variety of performance criteria. Performance measures related to classifications are presented in this study because it genuinely deals with classification problems. When predicting Breast Cancer, a positive instance indicates the patient has Breast Cancer if the target variable is 1 (malignant). It is also a negative instance if the target variable is 0 (benign), indicating that the patient does not have cancer.

Confusion Matrix

The effectiveness of a classification algorithm is summarised by a method known as a confusion matrix. By comparing the proportion of positive occurrences that are correctly or mistakenly classified with that of negative instances, it is conceivably the simplest technique to control how well a classification model performs. The actual labels are represented by the rows in a confusion matrix, while the anticipated labels are represented by the columns.

True Positives

When a patient has complications (in this case, breast cancer) and is also identified by the model as having complications, this is referred to as an instance where both the predicted and actual classes are true (1).

True Negatives

The occurrences where the predicted class and actual class are both False (0), i.e., when a patient does not have complications but is also identified by the model as not having complications, are known as true negatives.

False Negatives

These are instances where the expected class is False (0) but the actual class is True (1), i.e., when the model predicts that a patient doesn't have complications but they actually do.

False Positives

False positives arise when a patient is classified by the model as having complications even when they do not, for example, when the predicted class is True (1) and the actual class is False (0).

Table 1: Observation of Classification

Actual Class	Predicted class	
	Class = Yes	Class = No
	Class = Yes	Class = No
Class = Yes	True Positive	False Negative
Class = No	False Positive	True Negative



Article Title: Breast Cancer Detection Using Machine Learning Classifier

Normalized Matrix

Results are more effectively represented using the Normalized Confusion Matrix. The outcomes are comparable to the confusion matrix's. The values are spread out between 0 and 1. Prediction is made easier when data are evenly distributed.

Accuracy

One of the criteria used to evaluate classification models is accuracy. The percentage of a prediction that is accurate. It calculates the proportion of accurate forecasts among all of the model's predictions.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN}$$

Recall

It is a measurement of the percentage of patients—among those who actually experience the complications—who were expected to have them.

$$\text{Recall} = \frac{TP}{TP+FN}$$

Precision

It is defined as a percentage of patients who are genuinely affected by problems among those who the model has identified as having difficulties.

$$\text{Precision} = \frac{TP}{TP+FP}$$

Specificity

Specificity has a direct relationship with a classifier's ability to identify undesirable outcomes. It is exactly what Recall is not. It is a measurement of the proportion of patients—among those who genuinely lacked difficulties—who are labelled as not having complications.

F1 Score

The F1 score is the precision and recall weighted average. As a result, this score takes into account false positives and false negatives. Although intuitively it is not as easy to understand as accuracy, F1 is often more useful than accuracy.

$$\text{F1 Score} = \frac{2 * (\text{Recall} * \text{Precision})}{(\text{Recall} + \text{Precision})}$$



Article Title: Breast Cancer Detection Using Machine Learning Classifier

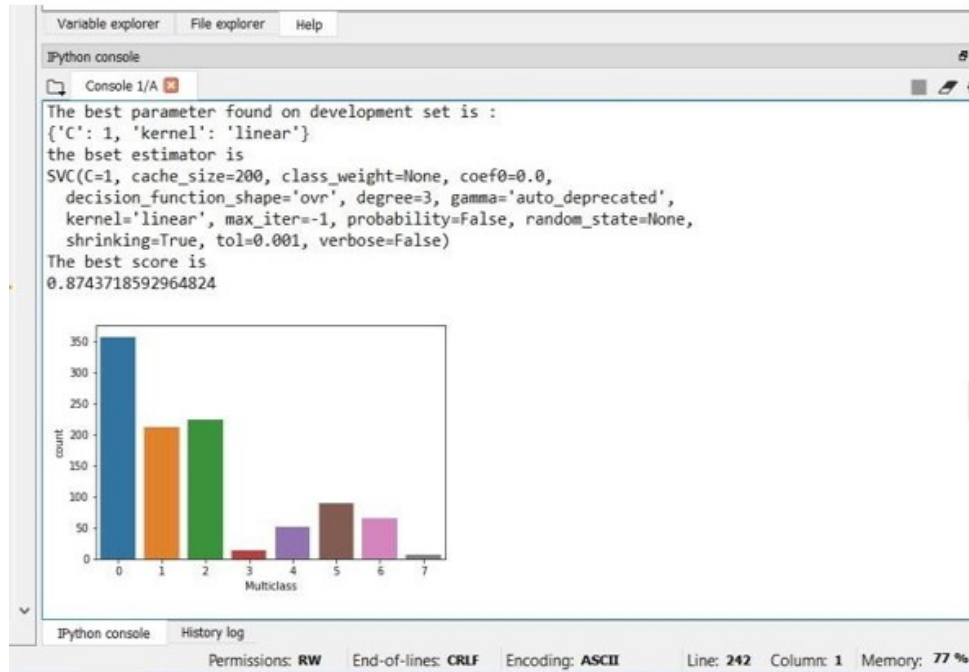


Figure 7: Histogram Representation of Dataset

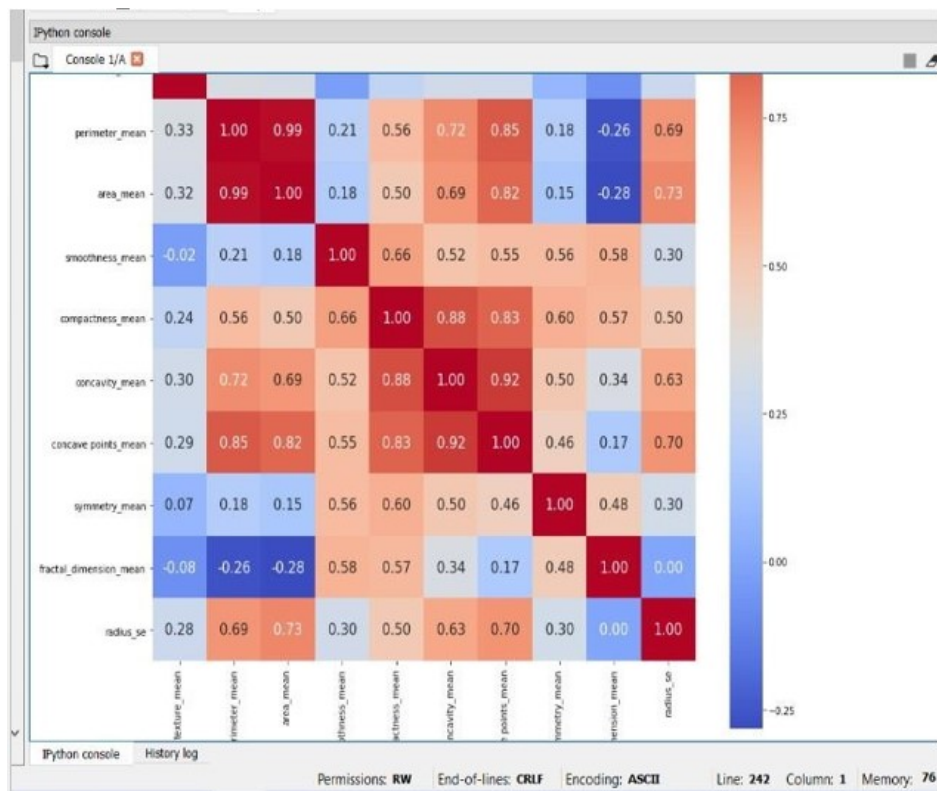


Figure 8: Graphical Representation of Dataset



Article Title: Breast Cancer Detection Using Machine Learning Classifier

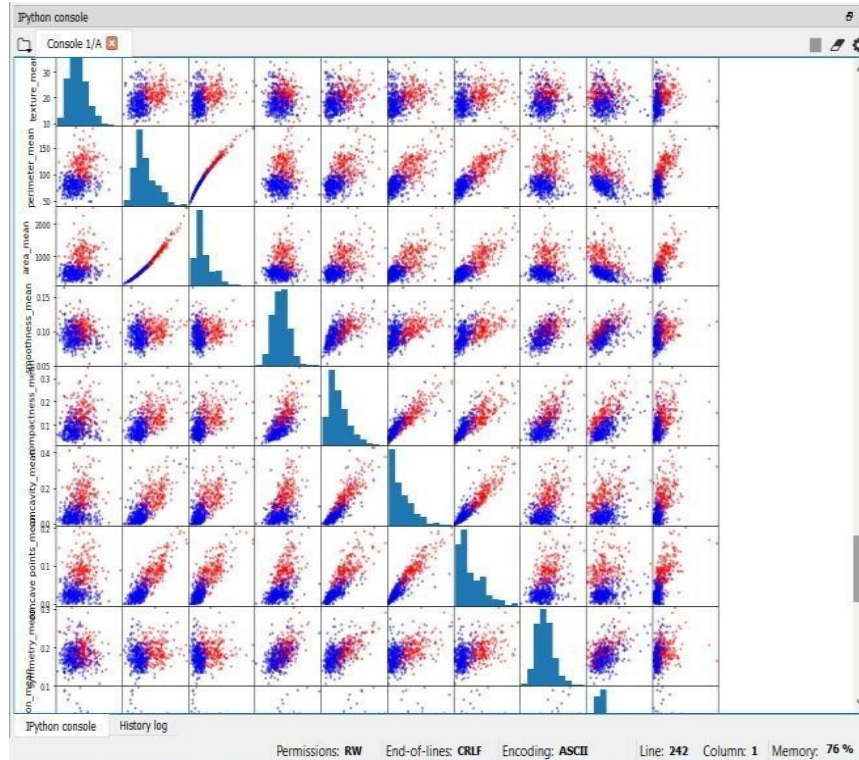


Figure 9: Final Outcome in Graphical Representation

7 Conclusion

Without using PCA, Random Forest has achieved excellent accuracy scores of 0.97. Logistic regression (0.923) and K- Neighbors (0.9349) aren't far behind either. SVM has an accuracy rating of 0.917. Out of the six models, Decision Tree performs the worst, scoring 0.834. With the exception of Decision tree, all algorithms' accuracy decreases when PCA is applied. Even though accuracy decreases, PCA performs best after being applied, and accuracy values are still greater than those of Random Forest (0.917). Much may be learned about the performance of the algorithms by taking into consideration the other performance matrices. While Logistic Regression and SVM perform better when PCA is applied to the dataset, Decision Tree and K-Neighbors perform better without PCA. After PCA is used, SVM and Logistic Regression both achieve recall scores of 1.000, which is essential for disease prediction, despite reductions in all other performance measures for both the aforementioned methods. We may infer that Logistic Regression and Support Vector Analysis with PCA perform better when it comes to Breast Cancer Prediction for this dataset employed by keeping in mind that PCA reduces the run time exponential to great extends in datasets (both small and large alike).



Article Title: Breast Cancer Detection Using Machine Learning Classifier

Reference

1. Dalal Bardou; Kun Zhang; Sayed Mohammad Ahmad, Year: 2018, "Classification of Breast Cancer Based on Histology Images Using Convolutional Neural Networks," IEEE Access, Vol: 6, May.2018, pp. 24680 – 24693.
2. A. Ferrari; S. Lombardi; A. Signoroni, Year: 2017, "Bacterial colony counting with convolutional neural networks in digital microbiology imaging", Pattern Recognit., Vol: 61, pp. 629 – 640.
3. F. A. Spanhol; L. S. Oliveira; P. R. Cavalin; C. Petitjean; L. Heutte, Year: 2017, "Deep features for breast cancer histopathological image classification," in Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC), Oct. 2017, pp. 1868 – 1873.
4. F. A. Spanhol; L. S. Oliveira; C. Petitjean; L. Heutte, Year: 2016, "Breast cancer histopathological image classification using convolutional neural networks," in Proc. Int. Joint Conf. Neural Netw. (IJCNN), Jul. 2016, pp. 2560 – 2567.
5. Z. Han; B. Wei; Y. Zheng; Y. Yin; K. Li; S. Li, Year: 2017, "Breast cancer multiclassification from histopathological images with structured deep learning model," Sci. Rep., Vol: 7, Jun. 2017, Art. no: 4172.
6. Stephen J. McKenna; Shazia Akbar; Siyamalan Manivannan; Wenqi Li, Year: 2016, "Gland segmentation in colon histology images using hand-crafted features and convolutional neural networks," 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), June 2016, pp. 1405 – 1408.
7. Babak Ehteshami Bejnordi; Guido Zuidhof; Maschenka Balkenhol; Peter Bult; Bran van Ginneken, Year: 2017, "Journal of Medical Imaging" (Bellingham, Oct.2017).
8. Scott Mayer McKinney; Marcin Sieniek; Varun Godbole; Jonathan Godwin; Natasha Antropova; Hutan Ashrafian; Trevor Back; M. Chesus; G. S. Corrado; A. Darzi; M. Etemadi, Year: 2020, "International evaluation of an AI system for breast cancer screening", Nature 2020, Vol: 577, pp. 89 – 94.
9. S. Khan; N. Islam; Z. Jan; I. U. Din; J. J. C. Rodrigues, Year: 2019, "A novel deep learning based framework for the detection and classification of breast cancer using transfer learning", Pattern Recognit. Lett. 2019, Vol: 125, pp. 1 – 6.
10. A. Murphy, Year: 2021, "Breast Cancer Wisconsin (Diagnostic) Data Analysis Using GFS-TSK", In North American Fuzzy Information Processing Society Annual Conference; Springer: Cham, Switzerland, 2021, pp. 302 – 308.
11. Year: 2022, "Cancer Research UK. Breast Cancer Statistics", 2022. Available online:



Article Title: Breast Cancer Detection Using Machine Learning Classifier

<https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer#heading-Two> (accessed on 1 October 2022).

12. World Health Organization. Cancer:Key Facts. 2022. Available online:
<https://www.who.int/news-room/fact-sheets/detail/cancer> (accessed on 1 October 2022).

13. Fredika M. Robertson; Melissa Bondy; Wei Yang; Hideko Yamauchi; Shannon Wiggins; Samira Kamrudin; Savitri Krishnamurthy; H. Le- Petross; L. Bidaut; A. N. Player; S. H. Barsky, Year: 2010, “Inflammatory breast cancer: The disease, the biology, the treatment”, CA A Cancer J. Clin. 2010, Vol: 60, 351 – 375.