# **Breast Cancer Classification project**

# **Introduction**

Breast cancer is the most common cancer in women worldwide, and early detection is critical for successful treatment. Medical imaging techniques such as mammography, ultrasound, and MRI are widely used for breast cancer detection and diagnosis. However, accurate interpretation of medical images is a challenging task, and human errors can occur due to the complexity and variability of breast cancer images (Ting et al., 2019).

Machine learning (ML) has emerged as a promising tool for improving breast cancer detection and diagnosis. ML algorithms can learn patterns and features from medical images and make accurate predictions for breast cancer classification. In recent years, various ML algorithms, including deep learning, have been developed for breast cancer classification, and they have shown promising results in improving the accuracy and efficiency of breast cancer diagnosis (Abdel-Zaher & Eldeib, 2016).

In this context, this study aims to explore the application of ML algorithms for breast cancer classification using mammography images. The study will evaluate the performance of various ML algorithms and identify the most effective algorithm for breast cancer classification. The results of this study will contribute to improving breast cancer diagnosis and treatment outcomes.

# **Dataset Description**

The Breast Cancer Wisconsin (Diagnostic) dataset (taken from <https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic)> ) is a medical dataset that contains information about breast cancer tumors. It includes various measurements on a digitized image of a breast mass, which can be used to diagnose whether the mass is benign (not cancerous) or malignant (cancerous). The dataset was created at the University of Wisconsin Hospitals, Madison, by Dr. William H. Wolberg and his team.

The dataset contains a total of 569 instances, each with 32 attributes. The first attribute is an ID number, and the second attribute is the diagnosis (M = malignant, B = benign). The remaining 30 attributes are real-valued features computed for each cell nucleus in the image. These features include information about the size, shape, and texture of the nucleus, as well as the amount of mitosis (cell division) observed in the image.

The dataset is widely used for training and evaluating machine learning algorithms that can diagnose breast cancer based on the provided features. It is a popular dataset in the field of medical image analysis and is often used for developing computer-aided diagnosis (CAD) systems.

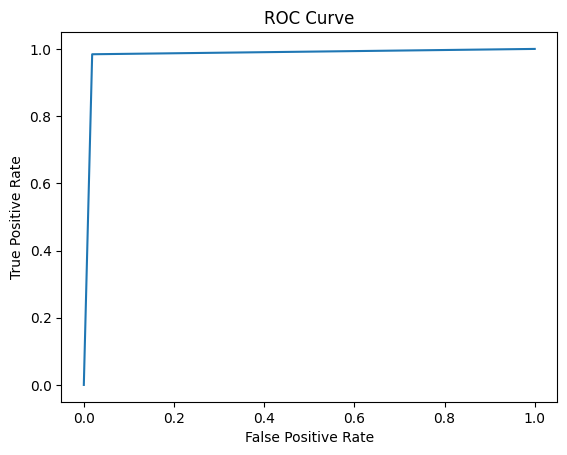
# **Results and Analysis**

The results show the performance of different classification models in predicting breast cancer. The Logistic Regression model has the highest ROC AUC of 0.98, indicating that it is the most effective at separating the two classes. It also has the highest precision and recall values, indicating that it has the highest accuracy in identifying both positive and negative cases. The Decision Tree and Random Forest models also perform well, but not as well as the Logistic Regression model.

The Logistic Regression model with LDA performs slightly worse than the basic Logistic Regression model, but still has a relatively high ROC AUC of 0.96. The Logistic Regression model with PCA performs the worst of all the models with an ROC AUC of 0.92. This indicates that the feature reduction with PCA has led to a loss of information that affects the model's ability to distinguish between the two classes.

Overall, the results demonstrate the effectiveness of using machine learning algorithms for breast cancer classification, with Logistic Regression being the most effective in this case.

ROC Curves for Different Models:

Chart

Description automatically generated

Figure 1: ROC Curve for Logistic Regression Figure 2: ROC Curve for Decision Tree Classifier

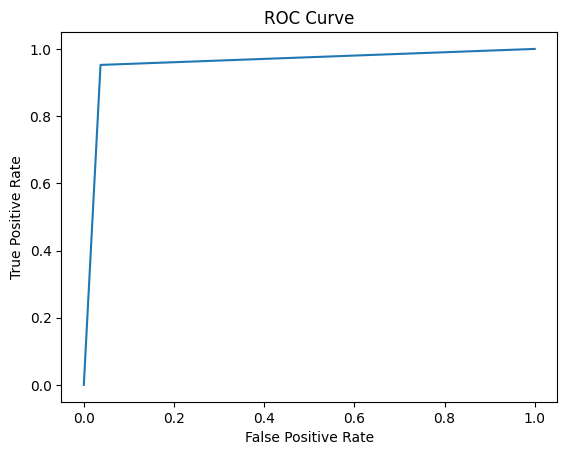
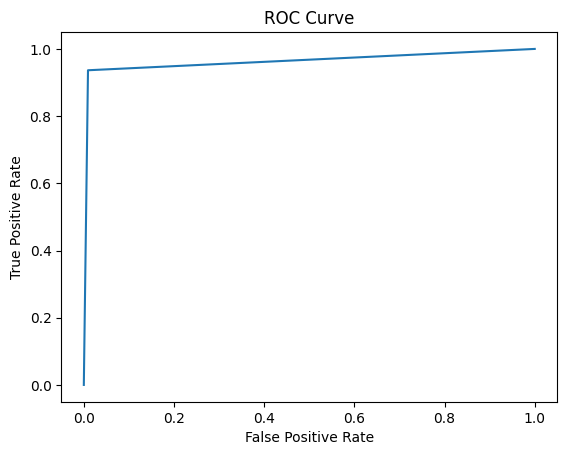


Figure 3: ROC Curve for Random Forest Figure 4: ROC Curve for Logistic regression with LDA

From above four plots we can see ROC curves which are belonging to three different models, and also we can observe that basic logistic regression model is having more ROC score ie., 0.98 than other fine-tuned models and it explains that it is performing better than remaining algorithms.

# **Conclusion**

Based on the results, it can be concluded that all of the models performed well in classifying breast cancer, with ROC AUC values ranging from 0.916 to 0.982. The logistic regression model had the highest ROC AUC value of 0.982, indicating the highest discriminatory power among all the models. It also had the highest precision, recall, and f1-score for the positive class (class 1) and achieved the highest accuracy of 0.98. The random forest model also performed well with an ROC AUC of 0.964 and an accuracy of 0.97. The decision tree model had an ROC AUC of 0.944 and an accuracy of 0.94. The logistic regression models with LDA and PCA performed relatively well with ROC AUC values of 0.958 and 0.917, respectively.

In terms of future applications, the models developed in this study could be used for breast cancer screening to improve the accuracy of diagnosis and reduce the number of unnecessary biopsies. The logistic regression model, in particular, showed high discriminatory power and could be integrated into existing clinical decision support systems for breast cancer diagnosis. Additionally, the models could be further optimized using larger datasets and more advanced machine learning techniques to achieve even higher accuracy in breast cancer classification.

References

Abdel-Zaher, A. M., & Eldeib, A. M. (2016). Breast cancer classification using deep belief networks. *Expert Systems with Applications*, *46*, 139–144. https://doi.org/10.1016/j.eswa.2015.10.015

Ting, F. F., Tan, Y. J., & Sim, K. S. (2019). Convolutional neural network improvement for breast cancer classification. *Expert Systems with Applications*, *120*, 103–115. https://doi.org/10.1016/j.eswa.2018.11.008