Breast cancer prediction using machine learning classification algorithms

**1.Introduction**

Base on the pdf Breast cancer is the second most diagnosed cancer among women and a leading cause of cancer-related deaths. Base on the studies Early detection improves survival rates, The study carried out two methods of diagnosing cancer patients are traditional and machine learning methods but current diagnostic methods, such as mammography (70% accuracy) and biopsies, have limitations, including human error and the need for multiple evaluations. Machine learning (ML) presents an opportunity to enhance accuracy and speed in breast cancer diagnosis by analyzing medical datasets with minimal human intervention

**2. Material and Methodology**

The study utilizes a publicly available breast cancer dataset with 11 features, including age,tumor size, metastasis, and involved lymph nodes.

Various machine learning are tested to determine their effectiveness.

* Logistic Regression (LR)
* Random Forest (RF)
* Extra Trees (ET)
* Extreme Gradient Boosting (XGB)
* Light Gradient Boosting Machine (LGBM)
* Cat boost
* Support Vector Classifier (SVC)
* Gaussian Naïve Bayes (GNB)

The dataset undergoes minimal preprocessing, with only a single missing data point removed. The classifiers are evaluated using accuracy, precision, recall, and F1-score.

3. Results

Each classifier's performance is assessed through confusion matrices, precision-recall curves, and AUC-ROC curves. Key findings include:

* Logistic Regression (91.67% accuracy) performed the best among all models.
* Random Forest and Cat boost (90.3% accuracy) were also highly reliable.
* LGBM (90.74% accuracy after feature selection) showed improvements.
* Other classifiers, such as Extra Trees, XGB, SVC, and Gaussian Naïve Bayes, showed moderate performances, with accuracies ranging from 87.1% to 88.7%.
* The most influential features for prediction were tumor size, metastasis status, and involved lymph nodes.

**4.Future importance and selection**

Feature selection is conducted to identify the most relevant patient characteristics. The top contributing factors across models were:

1. Tumor Size
2. Age
3. Metastasis
4. Involved Lymph Nodes

The less significant feature were removed, such as menopausal status and year of diagnosis, some models, like LGBM, showed notable performance improvements.

**5.Discussion**

* The study confirms ML's potential to outperform traditional diagnostic methods in accuracy and efficiency.
* Machine learning reduces human error and can automate parts of the diagnostic process.
* Feature selection plays a crucial role in optimizing performance for some classifiers.
* Despite high accuracy rates, real-world implementation would require further validationand integration into clinical workflows.

**6.Future works and improvement**

* Integration with larger and more diverse datasets to improve model generalization.
* Development of an automated tool for medical professionals based on the best-performing ML model.
* Further exploration of deep learning techniques, such as convolutional neural networks (CNNs), for breast cancer diagnosis.
* Addressing potential bias and fairness issues in AI-based medical diagnostics.

7.**Conclusion**

The study demonstrates that **ML models, particularly Logistic Regression, Random Forest, and LGBM, can significantly improve breast cancer diagnosis accuracy**. By refining **feature selection** and expanding dataset size, ML can revolutionize **early cancer detection and prognosis**. The application of ML in other diseases like **heart disease and diabetes** suggests a promising future for AI-driven healthcare.