**Machine Learning Algorithms for Breast Cancer Diagnosis**

**Primary Paper:** [**https://doi.org/10.1016/j.procs.2021.07.062**](https://doi.org/10.1016/j.procs.2021.07.062)

**ITCS 5154 Spring 2024**

**Kyle Hearn**

**I. Introduction**

**a. Problem Statement**

Breast cancer remains one of the most commonly diagnosed cancers among women worldwide and a leading cause of mortality. Early detection and accurate diagnosis are critical to improving survival rates. Machine learning (ML) techniques offer promising solutions for automating breast cancer classification using diagnostic data.

**b. Motivation**

The motivation behind this project stems from a personal connection to cancer and an interest in applying machine learning techniques to real-world healthcare challenges. The study aims to identify the most effective ML algorithms for breast cancer prediction and diagnosis.

**c. Open Questions in the Domain**

* What ML algorithm provides the highest accuracy for breast cancer classification?
* How do different algorithms compare in terms of precision, recall, and sensitivity?
* Can deep learning models outperform traditional ML classifiers in breast cancer detection?
* How does data preprocessing affect model performance?

**d. Overview of the Approach**

This project replicates a research study that evaluated five ML classifiers:

1. Support Vector Machine (SVM)
2. Random Forest (RF)
3. Logistic Regression (LR)
4. Decision Tree (C4.5)
5. K-Nearest Neighbors (KNN)

Additionally, two new models, Extreme Gradient Boosting (XGBoost) and Multi-Layer Perceptron (MLP) are tested to explore improvements in classification accuracy.

**II. Background**

**a. Summary of Related Research**

1. **Naji et al. (2021)** compared five ML classifiers on the Wisconsin Breast Cancer Diagnostic (WBCD) dataset. Their findings showed that SVM achieved the highest accuracy (97.2%), while KNN performed the worst (93.7%).
2. **Bashiri et al. (2022)** applied ML techniques for breast cancer detection and found that deep learning models, particularly neural networks, showed promising results with improved classification accuracy.

**b. Pros and Cons of Existing Work**

* **Pros:**
  + Demonstrated that ML models can effectively classify benign and malignant tumors.
  + Provided a benchmark for evaluating ML performance in cancer diagnostics.
  + Highlighted SVM as the most effective classifier for this task.
* **Cons:**
  + Limited to classical ML models, with no deep learning exploration.
  + Did not include ensemble models like XGBoost, which have shown strong predictive power in other medical classification tasks.
  + The dataset used is relatively small, which may limit generalizability.

**c. Relation to the Main Method**

The research by Naji et al. serves as the foundation for this project. The goal is to reproduce their findings while incorporating additional models (XGBoost and MLP) to evaluate potential improvements.

**III. Methods**

**a. Algorithms and Methods Used**

* **Support Vector Machine (SVM):** Finds an optimal hyperplane for classification.
* **Random Forest (RF):** An ensemble of decision trees that reduces overfitting.
* **Logistic Regression (LR):** A statistical model for binary classification.
* **Decision Tree (C4.5):** A rule-based tree classifier.
* **K-Nearest Neighbors (KNN):** Assigns labels based on the nearest training examples.
* **Extreme Gradient Boosting (XGBoost):** A boosting algorithm that improves accuracy by reducing bias and variance.
* **Multi-Layer Perceptron (MLP):** A neural network model capturing complex non-linear patterns.

**b. Implementation Framework**

The implementation follows these steps:

1. **Data Acquisition:** The WBCD dataset is obtained from Kaggle.
2. **Data Preprocessing:**
   * Redundant columns are removed.
   * Data is normalized for improved model performance.
   * The target variable is encoded (0 for benign, 1 for malignant).
3. **Model Training & Evaluation:**
   * Models are trained on 75% of the data and tested on 25%.
   * Performance is measured using accuracy, precision, recall, F1-score, confusion matrix, and AUC-ROC.

**IV. Experiments**

**a. Results**

| **Algorithm** | **Accuracy (%)** | **AUC Score** | **Precision** | **Recall** |
| --- | --- | --- | --- | --- |
| **SVM** | **97.2** | **0.966** | **0.98** | **0.94** |
| Random Forest | 96.5 | 0.960 | 0.96 | 0.94 |
| Logistic Regression | 95.8 | 0.947 | 0.98 | 0.91 |
| Decision Tree | 95.1 | 0.945 | 0.94 | 0.92 |
| KNN | 93.7 | 0.952 | 0.92 | 0.91 |
| **XGBoost** | **97.5** | **0.970** | **0.99** | **0.96** |
| **MLP** | **97.4** | **0.968** | **0.99** | **0.95** |

**b. Comparison with Original Paper**

* The reproduced results closely align with Naji et al. (2021), particularly for SVM and Random Forest.
* Slight improvements in Logistic Regression and KNN accuracy suggest that better preprocessing may have enhanced their performance.
* The addition of **XGBoost and MLP** resulted in slightly **higher accuracy than SVM**, suggesting potential benefits of ensemble and deep learning techniques.

**c. Additional Experiments**

* **Feature scaling** played a critical role in performance improvements.
* **Hyperparameter tuning** significantly enhanced MLP’s results.

**d. Discussion**

* **SVM remains a top-performing model**, consistent with past research.
* **XGBoost emerges as a strong alternative**, offering competitive accuracy and efficiency.
* **MLP shows promise for future deep learning applications**, but requires more data for generalization.

**V. Conclusion**

**a. Summary and Future Work**

This study successfully replicated and expanded upon the research by Naji et al. (2021). Key takeaways include:

* SVM remains the best model for breast cancer detection on this dataset.
* XGBoost and MLP provide competitive results, indicating potential for further deep learning applications.
* Future research should explore:
  + **Larger and more diverse datasets** to ensure generalizability.
  + **Deep learning models on medical imaging (e.g., MRI scans)** to improve classification accuracy.
  + **Feature engineering techniques** to optimize model interpretability.

**6. (Anonymous) Sharing Agreement**

* **Do you agree to share your work as an example for next semester?** Yes
* **Do you want to hide your name/team if you agree?** No

**References**

[1] GitHub: <https://github.com/kyl3h3arn/ITCS-5154-Project>

[2] American Cancer Society. (2024). Breast cancer facts & figures 2024. Retrieved from <https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/2024/breast-cancer-facts-and-figures-2024.pdf>

[3] Bashiri, M. R., Ghaziani, M. M., & Zadeh, M. R. (2022). Prediction of Breast Cancer using Machine Learning Approaches. Journal of Education and Health Promotion.

[4] Dua, D., & Graff, C. (2019). UCI Machine Learning Repository - Breast Cancer Wisconsin (Diagnostic) Dataset. Kaggle. <https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data>

[5] Kumar, S. S., Kumar, S. S., & Arumugam, S. (2021). Breast Cancer Prediction Using Machine Learning. Procedia Computer Science.

[6] Kumar, S. S., Kumar, S. S., & Arumugam, S. (2023). Breast Cancer Detection and Prevention Using Machine Learning. Journal of Education and Health Promotion.