Machine Learning for BioRobotics

Diagnosis of Breast Cancer Using Random Forests

Team 5

Name	Section	Bench Number
Mariem Mounier	2	35
Osama Faisel	1	11
Shuaib Saleh	1	49
Mina Fakhry Azer	2	43
Yehia Said	2	51
Abdelrahman Sameh	1	53

Submitted To:

Dr. Hisham Abdeltawab

TA. Mutair & Amira

Systems and Biomedical Engineering Dept.

Cairo University

- **Abstract:**

In 2020, breast cancer was the most prevalent cancer type. Early detection significantly improves long-term survival, but current diagnostic methods are costly and time-consuming. This study utilizes supervised machine learning to accurately detect breast cancer, focusing on the Wisconsin Breast Cancer Database with 569 instances. The Random Forest algorithm, trained on subsets with 16 and 8 features, outperforms other models, achieving accuracies of 100% and 99.30%. Comparative analysis with four other algorithms confirms Random Forest as the superior method for breast cancer diagnosis.

1. Introduction:

Breast cancer, the most common cancer in 2020, has diverse symptoms and risk factors, including genetic mutations and lifestyle habits. Limited access to qualified experts and high consultation costs hinder care, particularly in developing countries. Automated clinical decision systems could address this gap, but the lack of awareness about machine learning in medicine poses a challenge.

This report aims to promote the integration of machine learning models into clinical workflows by presenting results supporting their effectiveness. Early breast cancer diagnosis is crucial, and an automated system could reduce fatalities and complement clinician diagnoses, especially in populous regions like India. The study employs supervised machine learning, focusing on the Random Forest algorithm for its superior performance in diagnosing malignant tumors.

The report is organized as follows: Section 2 reviews related work, Section 3 outlines the methodology, Section 4 details the dataset and preprocessing, and Section 5 explains feature selection. Sections 6 and 7 elaborate on model development and evaluation, while Section 8 presents results and discussion. Section 9 outlines possible future work and concludes the report.

2. Related Work:

Several studies, including, consistently highlight the Random Forest algorithm's superior performance in diagnosing various diseases. Many researches achieved an 83.85% accuracy in diagnosing coronary heart disease and demonstrated Random Forests' effectiveness in detecting breast cancer. It is found that Random Forests outperforming other methods in classifying neuroimaging data for Alzheimer's disease.

Additionally, All experiments admitted the proficiency of various supervised machine learning techniques in detecting malignant breast tumours. Azar and El-Metwally [14] achieved over 95% accuracy with various Decision Tree classifiers. Desai and Shah [15] used a Multilayer Perceptron (MLP) classifier, reaching 91.9% accuracy on the Wisconsin breast cancer dataset. Polat and Güneş [16] employed a variation of the Support Vector Machine, LS-SVM, achieving a final accuracy of 98.53%. Sarkar and Leong [17] applied K-Nearest Neighbours (KNN), reporting a 1.17% improvement over existing models.

3. Methodology:

The experiment commences by obtaining the chosen dataset from the UCI Machine Learning repository, followed by data cleaning and preprocessing to retain only pertinent information. Initial feature selection, utilizing correlation coefficients, results in a dataset with 16 features to eliminate redundancies. Subsequently, five machine learning models are trained on this data, with special tuning of hyperparameters for the Random Forest model. Further refinement occurs through three additional feature selection methods, reducing the dataset to just eight attributes. The same training procedure is applied to this streamlined dataset. Ultimately, the models undergo testing on a holdout set, and their performances are assessed.

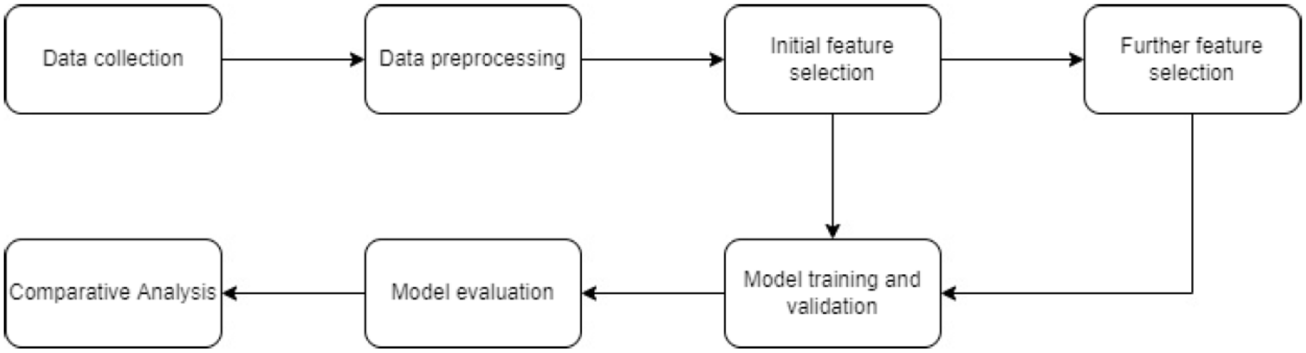

Data collection
Data preprocessing
Initial feature selection
Further feature selection
Comparative Analysis
Model evaluation
Model training and validation

Fig. 1. Machine learning process

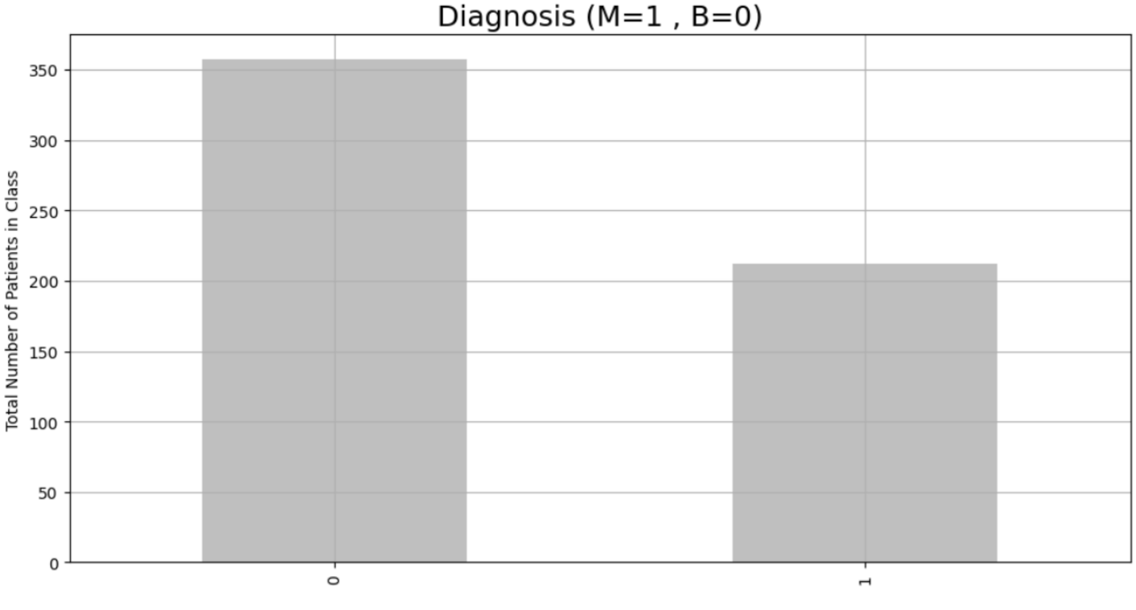
4. Dataset Description and Preprocessing:

The experiment employs the Wisconsin Breast Cancer Diagnostic dataset from the UC Irvine Machine Learning Repository [3]. This dataset comprises 569 tumours instances, with 212 being malignant and 357 benign. The dataset's imbalance, with less than 40% in the positive class, requires upscaling. Table 1 outlines ten attributes, describing cell nucleus measurements from a tumour image, with mean, standard error (se), and worst values, resulting in 30 features. The last attribute indicates the diagnosis as benign or malignant. Features are identified by name and value type (mean/se/worst).

The dataset, comprising 32 columns, excludes an irrelevant arbitrary ID column. The diagnosis column serves as the target variable, and all 596 entries, free of errors or missing data, are included in the training set. One-hot encoding transforms the Diagnosis feature into binary values, and the remaining features consist of decimal values with four significant digits.

Addressing the class imbalance issue, the minority class is up sampled, resulting in a final dataset of 714 instances to prevent biased predictions. To mitigate the impact of widely varying feature ranges on certain supervised learning methods like K-Nearest Neighbours, a Min-Max scaler normalizes the features to a uniform range between 0 and 1.

The data undergoes an 80/20 training-test split to maximize training data, with a sufficient 20% test set due to the absence of many outliers in the data, ensuring an objective evaluation of final models.



5. Feature Selection:

5.1. Initial feature selection:

The initial stage of feature selection is based on the Pearson correlation coefficient. For pairs of features having a correlation coefficient higher than 0.8, one of the features is dropped to avoid multicollinearity. This process results in 14 features being removed and 16 features remaining.

In our project we used the preprocessed data initial selection because it is the best initial selection made. And this dataset is used to train and evaluate the first set of models. It will be referred to as the initial dataset from now on.

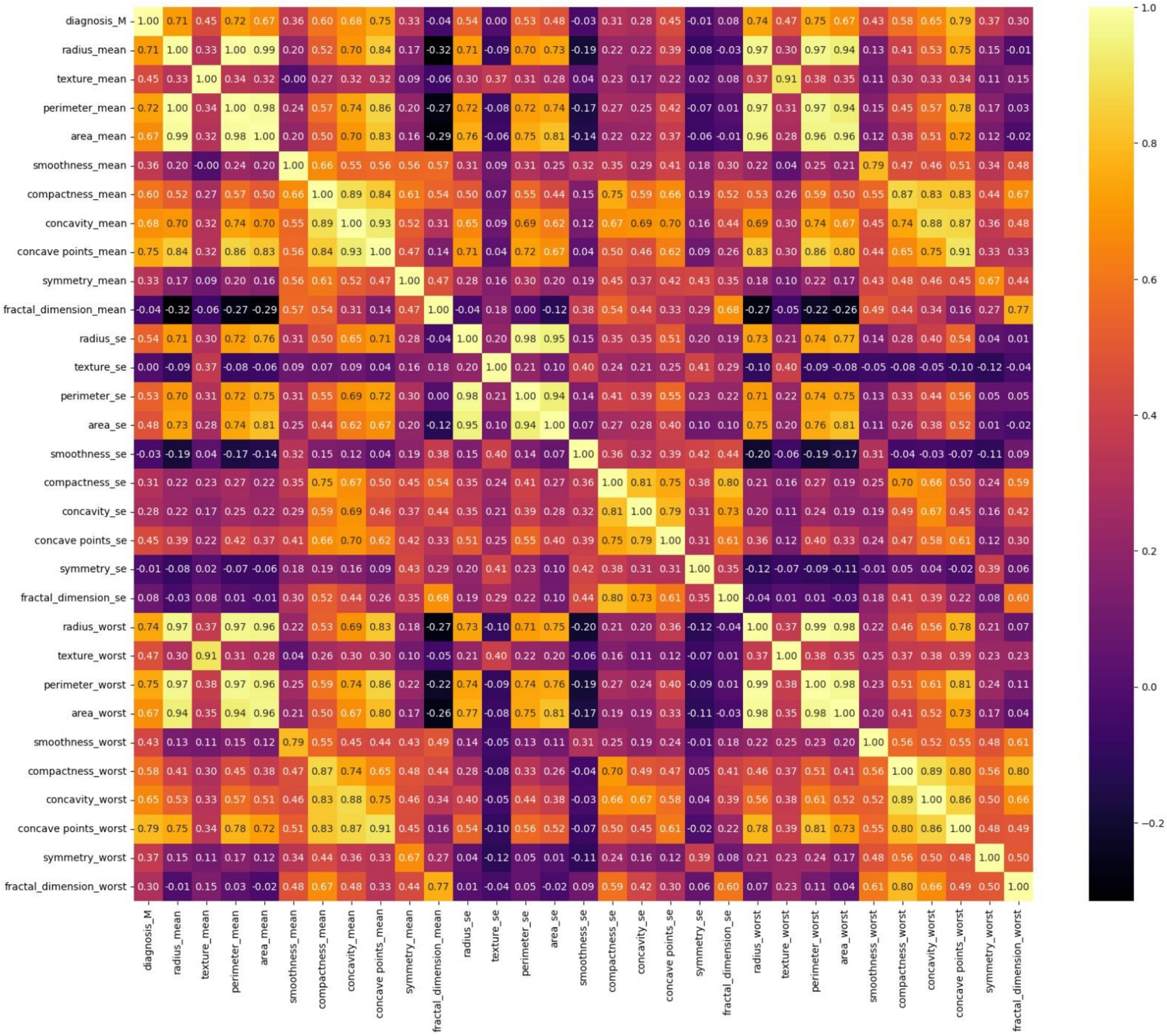


Fig. 3. Feature correlation map

The set of features contained in the initial dataset is detailed in set F_1 as follows:

F_1 = {diagnosis_M, radius_mean, texture_mean, smoothness_mean, compactness_mean, symmetry_mean, fractial_dimension_mean, radius_se, texture_se, compactness_se, smoothness_se, concave_point_se, symmetry_se, fractial_dimension_se, smoothness_worst, symmetry_worst, fractial_dimension_worst}

5.2. Further feature selection:

The second stage of feature selection is based on two different methods which are Logistic Regression and Univariate Selection.

5.2.1 Logistic Regression:

In this method, feature importance are determined using a Logistic Regression model. The seven features with the highest importance are then selected, which are:

{ radius_mean, texture_mean, compactness_mean, fractial_dimension_mean, radius_se, smoothness_worst, symmetry_worst }

5.2.2. Univariate Selection:

Univariate selection selects the required features by performing univariate statistical tests on the data. In this case, the ANOVA F-value is used to determine the eight best features, which are:

{ radius_mean, texture_mean, smoothness_mean, compactness_mean, radius_se, concave_point_se, smoothness_worst, symmetry_worst}

5.2.3. Minimal Dataset:

It is the common features between the two methods to make sure that these features are not affected by biases in the individual methods. The set of features contained in the minimal dataset is detailed in set F_2 as follows:

F_2 = { radius_mean, texture_mean, compactness_mean, radius_se, smoothness_worst, symmetry_worst}

6. Model Development:

Five models are selected for a preliminary comparison- Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT), Multilayer Perceptron (MLP), and K-Nearest Neighbours (KNN). They are trained on the initial dataset with default parameters, and their results are compared. The models are 5-fold cross-validated using recall value as the scoring function. The recall value is used as it is more acceptable for a medical model to return a false positive than a false negative. The Random Forest model performs the best with a cross-validation recall score of **0.9838**.

Thus, the Random Forest model is selected as our primary model. The four remaining models are trained using their default parameters on the initial and the minimal dataset. Random Forest's hyper-parameters are precisely tuned to maximize the primary model's performance.

7. Model Evaluation:

The five models are trained and evaluated on both the initial and minimal datasets. As accuracy is not a sufficient metric in the medical field, the models are compared based on a multitude of measures. The particular metrics used are detailed below.

7.1. Basic definitions:

- True Positive (TP): The count of malignant tumours classified as malignant by the model.
- True Negative (TN): The count of benign tumours classified as benign by the model.
- False Positive (FP): The count of benign tumours classified as malignant by the model.
- False Negative (FN): The count of malignant tumours classified as benign by the model.

7.2. Metrics used:

- Accuracy: The percentage of tumours whose malignancy was correctly predicted.
- Precision: TP / (TP + FP)
- Recall: TP / (TP + FN)
- F1 Score: (Precision * Recall) / (Precision + Recall)
- ROC-AUC Score: Area under the Receiver Operating Characteristic Curve.

All five models are trained and evaluated using these metrics on both the initial and minimal datasets. Their performances are collated and analyzed in the Results section.

8. Results and Discussion:

8.1. Initial Dataset:

The Random Forest model outperforms all other models by achieving the highest score on all the measured metrics. The Support Vector Machine model performs the second best, supporting the claims made in the related works previously analysed in the paper. The K-Nearest Neighbours performs the worst, and we conclude that clustering methods may not be ideal for malignant tumour detection.

Model	Accuracy (%)	Precision	Recall	F1-Score	ROC-AUC
Random Forest	97.3%	0.953	0.983	0.968	0.973
SVM	97.2%	0.967	0.967	0.967	0.971
Decision Tree	93.7%	0.884	0.983	0.931	0.943
MLP	95.8%	0.938	0.967	0.952	0.959
KNN	92.3%	0.932	0.887	0.909	0.918

Table 1. Model comparison on the initial dataset

8.2. Minimal Dataset:

Model	Accuracy (%)	Precision	Recall	F1-Score	ROC-AUC
Random Forest	97.9%	0.953	1	0.976	0.981
SVM	97.2%	0.967	0.967	0.967	0.971
Decision Tree	95.8%	0.924	0.983	0.953	0.961
MLP	95.1%	0.923	0.967	0.944	0.953
KNN	92.3%	0.932	0.887	0.909	0.918

Table 2. Model comparison on the Minimal dataset

The Random Forest model continues to achieve the best metrics, with an accuracy of 97.9% and a perfect recall score. This is ideal, as the model does not return any false negatives. Furthermore, the model performs at par with the model trained on the initial dataset, having used just half the number of features. A smaller dataset will allow clinicians to work more efficiently in the field. This will also reduce the computational power necessary to train and run the model. Fig. 4 and Fig. 5 depict the ROC curves for the RF model on the initial and minimal datasets.

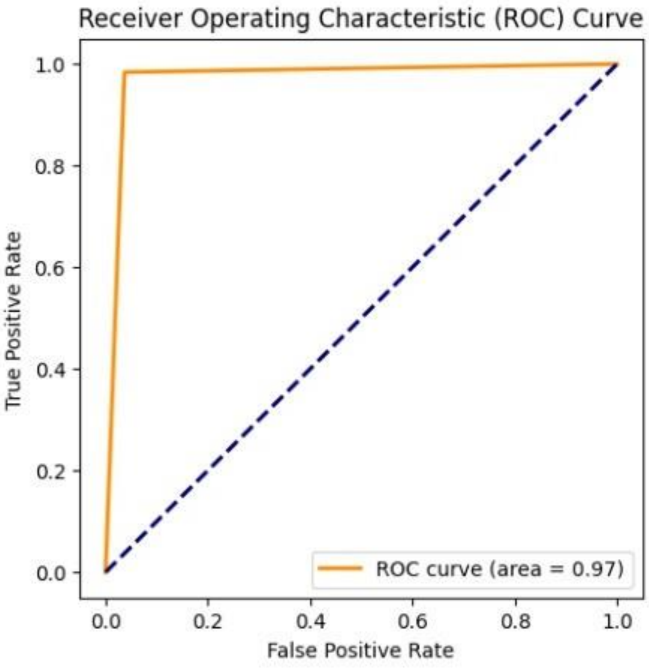


Fig. 4. ROC of RF model
on initial dataset

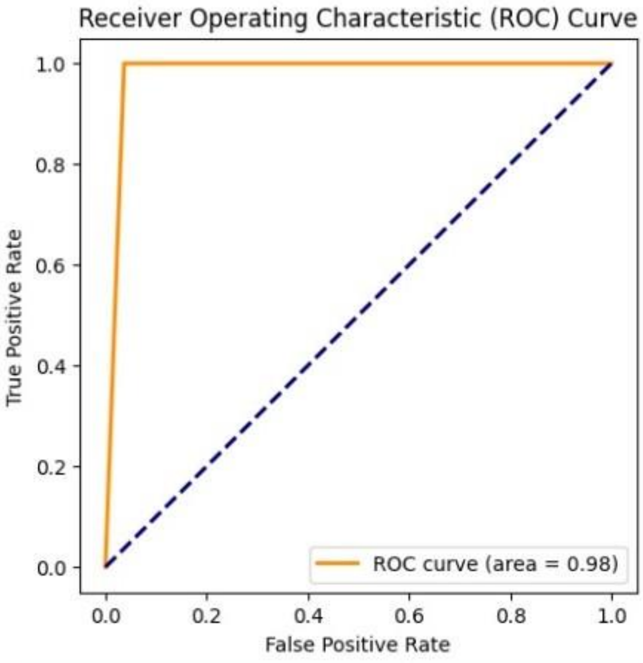


Fig. 5. ROC of RF model
on minimal dataset

9. Conclusion and Future work:

In our report, the Random Forest algorithm was trained, and its hyperparameters were tuned to diagnose breast cancer efficiently and accurately. Further, these RF models are benchmarked against four other supervised learning techniques. The RF model trained on the initial dataset achieves perfect metrics, far surpassing the performances of the other four models on the same dataset. The models trained on the minimal dataset paint a similar picture- the RF model achieves near-perfect metrics while the other models lag behind. Thus, the Random Forest model may be used by oncologists in the field to confirm their diagnoses. It may also be used in rural areas to increase the reach of breast cancer awareness and treatment.

Future work may be found in developing image processing techniques for the detection of malignant tumours. The experiment in the report utilized numerical measures extracted from images. Increases in efficiency may be found if it was not necessary to extract said measurements and directly feed the photos of the tumours into the model.