Available online at
https://jcrinn.com/
https://crinn.conferencehunter.com/

Journal of Computing Research and Innovation 11(2) 2026

Journal of Computing Research and Innovation

Application of Machine Learning Techniques for Predicting Breast Cancer

Wan Nashua Amira Ayob[1], Nor Hayati Shafii[2*], Nur Fatihah Fauzi[3], Diana Sirmayunie Mohd Nasir[4], Nor Azriani Mohamad Nor[5]

[1,2,3,4,5] *Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM) Perlis Branch, Arau Campus, 02600 Arau, Perlis, Malaysia.*

ARTICLE INFO

Article history:
Received 30 June 2025
Revised 5 March 2026
Accepted 6 March 2026
Online first
Published 1 September 2026

Keywords:
Prediction
Machine Learning
Random Forest
Logistic Regression
Decision Tree
Breast Cancer

DOI:
10.24191/jcrinn.v11i2.564

ABSTRACT

Breast cancer is the most prevalent invasive cancer in women and the second leading cause of cancer-related mortality among women. Interest in breast cancer research and prevention has surged recently. With the advent of data mining techniques, researchers can now efficiently extract valuable information from large databases, facilitating prediction, classification, and clustering. In this study, three classification models, namely Decision Tree, Random Forest, and Logistic Regression, were used to classify datasets related to breast cancer. The goal was to develop an accurate model to predict breast cancer and reduce the risk of death from the disease. The performance of these models was evaluated using three metrics: Precision, Recall, and F1 Score. Prediction accuracy was also measured. Comparative experiments in this study revealed that the Random Forest model outperformed the other two techniques in terms of performance and accuracy. Consequently, the study's model demonstrates significant clinical and referential value in real-world applications.

1. INTRODUCTION

Breast cancer is one of the most common cancers globally, including in Malaysia, and remains the leading cause of cancer-related death in women. Treatment for breast cancer often involves surgical resection, chemotherapy, radiotherapy, and medication to target microscopic disease that has spread from the primary tumor. In 2020, 2.3 million women were diagnosed with breast cancer, resulting in 685,000 deaths worldwide. As of the end of 2020, 7.8 million women had been diagnosed with breast cancer in the preceding five years, making it the world's most common cancer. Breast cancer causes more disability-adjusted life years loss in women than any other cancer globally. It affects women of all ages after puberty, with incidence rates increasing with age (World Health Organization, 2021).

In Malaysia, breast cancer is the most common cancer among women, with approximately one in every 19 women at risk. The number of cases is steadily rising, particularly with the growing population.

[2*] Corresponding author. *E-mail address*: norhayatishafii@uitm.edu.my
https://doi.org/10.24191/jcrinn.v11i2.564

Improvements in mammography screening are needed to enhance early detection and efficiency in medical image analysis. Early diagnosis can significantly extend the lives of cancer patients (Ismail & Sovuthy, 2019). The incidence rate in Malaysia is 38.7 per 100,000 women per year, and the total number of breast cancer patients is expected to increase further. Notably, Malaysian breast cancer patients have one of the lowest survival rates in the Asia-Pacific region, with a five-year survival rate of only 49%, compared to up to 90% in the United States (Mujar et al., 2018).

Breast imaging, including ultrasound, is commonly used for screening but has limitations such as operator dependency and potential false negatives. Traditionally, diagnostic accuracy relies heavily on a doctor's experience, which can be fallible. Recent advancements in machine learning (ML) have significantly enhanced breast cancer diagnosis and prognosis by differentiating between benign and malignant tumours and predicting patient outcomes.

The development of computing technology and patient databases has facilitated the analysis of large datasets, improved diagnostic accuracy and aiding in future medical planning. ML techniques can automate manual tasks and improve diagnostic precision. For instance, text and speech analysis can classify patient emotions, which can influence health outcomes. Despite the success of ML models in other regions, factors influencing breast cancer vary by location, necessitating models tailored to the Malaysian context.

This study aims to integrate ML techniques to analyze breast cancer data using Python and developing a mathematical model with high prediction accuracy. A dataset from the UCI Machine Learning Repository is utilized to develop a breast cancer prediction model using Python. The dataset contains 699 instances and 11 attributes, including clump thickness, uniformity of cell size, and bare nuclei.

This study could significantly benefit cancer patients by enabling early diagnosis and treatment, thereby reducing breast cancer mortality rates. Early detection is crucial as it often allows for less aggressive and more effective treatment options. Patients diagnosed early have a better chance of successful treatment and increased life expectancy, as they can receive treatment before the disease progresses to an advanced stage. Additionally, early treatment is generally less expensive than treatment for late-stage cancer. In summary, this study aims to facilitate early treatment, improving patient outcomes and reducing healthcare costs associated with late-stage breast cancer diagnosis.

2. PRELIMINARIES

This section defines and explores the theory behind the machine learning process, focusing on the three algorithms used in this study: Logistic Regression, Decision Trees, and Random Forest.

2.1 Machine Learning

Machine Learning (ML) encompasses a broad range of data analysis algorithms that build models for autonomous predictions by iteratively improving based on training data. Essentially, software program performance improves automatically over time (Jordan & Mitchell, 2015). The primary goal of an ML algorithm is to create a mathematical model that accurately fits the data. There are three types of learning: supervised, semi-supervised, and unsupervised.

Supervised learning algorithms require labeled data for training purposes. These labels, or ground truth, provide the necessary responses to specific questions. Unsupervised learning, on the other hand, clusters data with similar characteristics and generates labels that meaningfully organize the data without prior labeling. This method often requires larger training datasets compared to supervised techniques. Unsupervised learning can identify meaningful clustering labels, which can then be used in supervised training to develop effective ML techniques.

https://doi.org/10.24191/jcrinn.v11i2.564

ML algorithms evaluate datasets to extract data-driven models, prediction rules, or decision rules. To ensure that ML systems operate autonomously and effectively without human intervention, they must learn or derive knowledge from input data or experiences, such as rules or patterns. The process involves several steps: initially, the system must obtain data-driven features. Next, it analyzes these features to detect and classify any potential patterns or abnormalities. Finally, an ML algorithm is used to determine the most suitable model to represent the data's behavior or trends (Sahran et al., 2018).

In this study, the output is classified into two categories: benign and malignant. The machine learning algorithms used for this classification are Logistic Regression, Decision Tree, and Random Forest. These algorithms are widely adopted in forecasting and predictive modeling due to several key advantages as follows. First, they offer an effective balance between interpretability and predictive accuracy, allowing researchers and practitioners to obtain reliable results while maintaining a reasonable level of transparency in model interpretation. Second, they are computationally efficient and do not require excessive processing power, making them suitable for practical implementation across various settings. Third, they perform well even with moderately sized datasets, which is particularly beneficial when large-scale data is not available. Finally, these algorithms have been extensively validated across multiple industries, including healthcare and finance, demonstrating their robustness and practical relevance in diverse application domains (Hastie et al., 2009, James et al. 2013, Lessmann et al., 2015).

2.2 Logistic regression

Logistic regression is a supervised machine learning model commonly used for classification and predictive analytics. Also known as a logit model, it calculates the likelihood of an event occurring based on a set of independent variables. The model is particularly useful when the outcome is dichotomous, meaning it can take only two possible values.

In logistic regression, a logit transformation is applied to the odds (the probability of success divided by the probability of failure). This transformation, also known as the log odds or the natural logarithm of odds, helps to differentiate between classes (or categories). Unlike generative algorithms such as Naive Bayes, logistic regression does not generate new data about the class it predicts; instead, it focuses on classification.

Logistic regression is widely regarded for its simplicity and interpretability. It can be applied in various fields, such as medicine, to predict the likelihood of a disease or illness within a population. Its ability to provide probability scores for classifications makes it a valuable tool for decision-making processes in diverse applications.

The model applies to a logit transformation to convert the linear combination of input features into a probability. The logit function (log-odds) is defined as the natural logarithm of the odds of the event.

$$logit(p) = ln\ ln\ \left(\frac{p}{1-p}\right) \tag{1}$$

where p is the probability of the positive class. The logistic function, or sigmoid function, is used to map any real-valued number into the range [0, 1]

$$\sigma(z) = \frac{1}{1+e^{-z}} \tag{2}$$

where z is linear combination of input features and their corresponding coefficients. The logistic regression model can be expressed as

$$p(x) = \sigma(w^{\top}x + b) \tag{3}$$

where $p(x)$ is the probability of the positive class given the input features x, w is the vector of coefficients (weights), x is the vector of input features and b is the bias term. The model parameters

https://doi.org/10.24191/jcrinn.v11i2.564

$(w\ and\ b)$ are estimated using a method called maximum likelihood estimation (MLE). The goal is to find the parameter values that maximize the likelihood of observing the given training data.

2.3 Decision Trees

The Decision Tree algorithm is one of the simplest and most widely used classification algorithms, belonging to the supervised learning algorithm family. Unlike other supervised learning algorithms, Decision Trees can be used to solve both regression and classification problems. The primary objective of using a Decision Tree is to build a training model that can predict the class or value of an attribute by learning simple decision rules from the training data.

The Decision Tree algorithm begins at the root node, which encompasses the entire dataset. It selects the most optimal feature to split the data, using criteria such as Gini impurity or information gain to determine the split that best separates the data into homogeneous subsets. This splitting process continues recursively, with each node creating further branches based on the values of the selected features, thereby forming a decision tree. At each node, the algorithm evaluates potential splits and chooses the one that maximizes homogeneity within the resulting subsets. This process stops when further splitting does not significantly improve the model or when all data points in a node are homogeneous. The final tree, consisting of decision nodes and leaf nodes, represents the decision-making process, which can then be used for classification or regression tasks. The process is shown in Fig. 1.

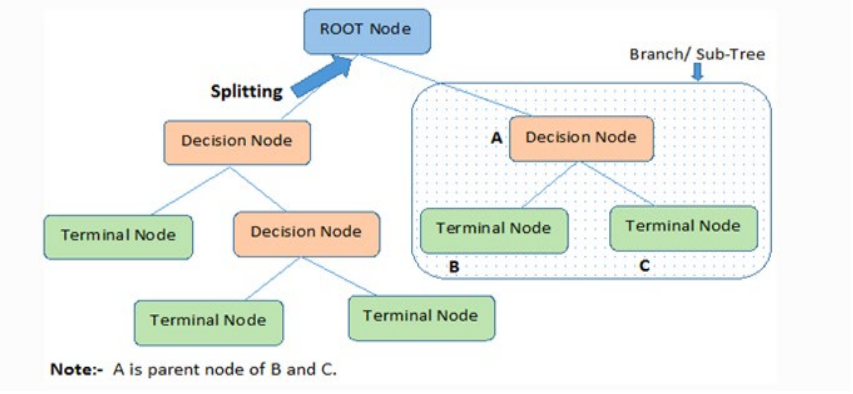

Fig. 1. Decision Tree's illustration

2.4 Random Forest

A Random Forest is a machine learning technique used for regression and classification problems, leveraging the power of ensemble learning, which combines multiple classifiers to tackle complex issues. The algorithm consists of numerous decision trees, collectively referred to as a 'forest.' The forest is trained using bagging, or bootstrap aggregation, a meta-algorithm that enhances the accuracy of machine learning models through an ensemble approach. In Random Forest, each tree is constructed from a random subset of the data, and the final prediction is determined by aggregating the outputs of all the individual trees, either by averaging (for regression) or majority voting (for classification). The accuracy of the model typically increases with the number of trees, making Random Forest a robust and reliable method for various predictive tasks. The process illustrated in Fig. 2.

https://doi.org/10.24191/jcrinn.v11i2.564

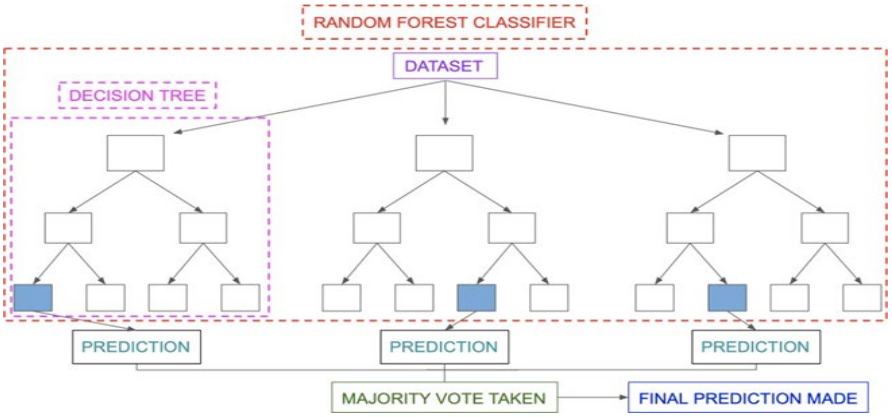

Fig. 2. Random Forest's illustration

2.5 Model evaluation

Model evaluation in machine learning is a critical process that involves assessing the performance of a trained model to ensure it generalizes well to unseen data. This process typically involves splitting the dataset into training and testing subsets, where the training set is used to build the model, and the testing set is used to evaluate its performance. Effective model evaluation ensures that the model not only performs well on the training data but also delivers accurate and reliable predictions on new, unseen data, thereby increasing its practical utility and reliability. In this study, four evaluation metrics are employed to measure the effectiveness of the model, including accuracy, precision, recall and F1 score.

(i) Accuracy

The accuracy of a classifier measures how effectively it can predict instances into the correct classification. It is calculated by dividing the number of correct predictions by the total number of instances in the dataset. It's important to note that accuracy is highly dependent on the threshold chosen by the classifier and can vary across different testing sets. Therefore, while accuracy provides a general overview of the classification performance, it may not be the best metric for comparing different classifiers. As a result, accuracy is often calculated using the following Eq. (4)

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \tag{4}$$

(ii) Precision

Precision in machine learning is a metric that measures the accuracy of positive predictions made by the classifier. It focuses on the proportion of true positive predictions (correctly predicted positives) out of all instances predicted as positive, including both true positives and false positives. Precision is particularly valuable in scenarios where minimizing false positives is crucial. Mathematically, precision is calculated as

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{5}$$

Where True Positives (TP) are instances correctly predicted as positive and False Positives (FP) are instances incorrectly predicted as positive. A high precision score indicates that the classifier is making

https://doi.org/10.24191/jcrinn.v11i2.564

accurate positive predictions, minimizing false positives. It is a critical metric in applications where the cost of false positives is high, such as medical diagnostics or fraud detection.

(iii) Recall

Recall, which is also commonly known as sensitivity or true positive rate, is a fundamental metric in machine learning that measures the proportion of actual positive instances that are correctly identified by the classifier. It focuses on how well the classifier identifies all positive instances, including those that are missed (false negatives). Mathematically, recall is calculated as

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{6}$$

True Positives (TP) are instances correctly predicted as positive and False Negatives (FN) are instances incorrectly predicted as negative. A high recall score indicates that the classifier is effectively capturing a large portion of positive instances from the dataset. It is particularly important in applications where it's crucial to avoid false negatives, such as medical diagnostics (where missing a positive diagnosis can be critical) or search and rescue operations.

(iv) F1 Score

The F1 score is a metric in machine learning that combines both precision and recall into a single measure. It provides a balance between these two metrics, making it useful when both precision and recall are equally important. It is calculated as follows.

$$F1\ Score\ = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{7}$$

The F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0. It is particularly useful in binary classification settings where there is an imbalance between the classes, and both false positives and false negatives need to be minimized. By considering both precision and recall, the F1 score provides a comprehensive assessment of a classifier's performance.

3. METHODOLOGY

Fig. 3 depicts the general process of developing the machine learning model and describes the steps of the machine learning process used in this study.

https://doi.org/10.24191/jcrinn.v11i2.564

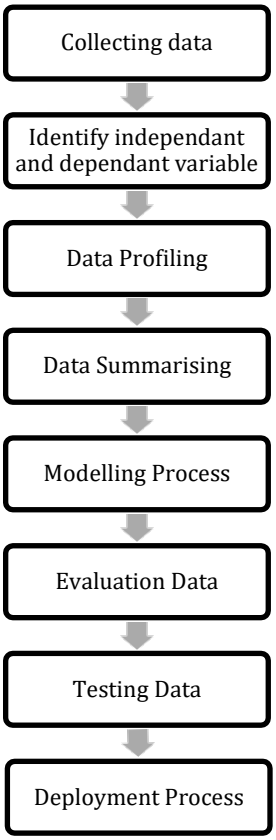

Fig. 3. Research framework

The sample dataset for this study was obtained from the Kaggle website and is classified as secondary data. There are 699 breast cancer patient records with 10 attributes. The independent variables in this study consist of Clump_Thickness, Uniformity_of_Cell_Size, Uniformity_of_Cell_Shape, Marginal_Adhesion, Single_Epithelial_Cell_Size, Bare_Nuclei, Bland_Chromatin, Normal_Nucleoi, Mitoses and the dependent variable is Class. All attributes are derived from microscopic examination of fine needle aspirate (FNA) samples of breast tissue. Each variable reflects specific cellular characteristics that help distinguish between benign and malignant tumors. Clump_Thickness refers to the thickness of cell clusters observed in the sample, Uniformity_of_Cell_Size measures consistency in cell size, Uniformity_of_Cell_Shape evaluates consistency in cell shape, Marginal_Adhesion indicates how strongly cells adhere to one another, Single_Epithelial_Cell_Size assesses the size of individual epithelial cells, Bare_Nuclei refers to nuclei that appear without surrounding cytoplasm in the sample, Bland_Chromatin describes the texture and uniformity of chromatin (genetic material) inside the nucleus, Normal_Nucleoi measures prominence of nucleoli inside the nucleus and lastly Mitoses represents the number of cells undergoing division.

The following Python coding shows the process of data analysis step by step.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
```

https://doi.org/10.24191/jcrinn.v11i2.564

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report

# ==========================================
# 1. HELPER FUNCTIONS
# ==========================================

def plot_confusion_matrix(cm, target_names, title='Confusion matrix',
cmap=plt.cm.summer):
    """
    Renders a visual heatmap of the confusion matrix.
    """
    plt.figure(figsize=(8, 6))
    plt.imshow(cm, interpolation='nearest', cmap=cmap)
    plt.title(title)
    plt.colorbar()

    tick_marks = np.arange(len(target_names))
    plt.xticks(tick_marks, target_names, rotation=45)
    plt.yticks(tick_marks, target_names)

    # Annotate each cell with the integer count
    width, height = cm.shape
    for x in range(width):
        for y in range(height):
            plt.annotate(str(cm[x][y]), xy=(y, x),
                         horizontalalignment='center',
                         verticalalignment='center', color='black', fontsize=22)

    plt.ylabel('True label')
    plt.xlabel('Predicted label')
    plt.tight_layout()
    plt.show()

# ==========================================
# 2. DATA LOADING & CLEANING
# ==========================================

# Load dataset from UCI repository
url = 'https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-
wisconsin/breast-cancer-wisconsin.data'
columns = [
    'Sample code', 'Clump Thickness', 'Uniformity of Cell Size',
    'Uniformity of Cell Shape', 'Marginal Adhesion', 'Single Epithelial Cell Size',
    'Bare Nuclei', 'Bland Chromatin', 'Normal Nucleoli', 'Mitoses', 'Class'
]
data = pd.read_csv(url, names=columns)

# Drop non-predictive ID column
data = data.drop(['Sample code'], axis=1)

# Handle missing values: replace '?' with NaN and convert to float
data = data.replace('?', np.nan)
data['Bare Nuclei'] = pd.to_numeric(data['Bare Nuclei'])

# Impute missing values in 'Bare Nuclei' with the median
data['Bare Nuclei'] = data['Bare Nuclei'].fillna(data['Bare Nuclei'].median())

print(f"Dataset Dimensions: {data.shape[0]} instances, {data.shape[1]} attributes")

# ==========================================
# 3. DATA PREPROCESSING
# ==========================================
```

https://doi.org/10.24191/jcrinn.v11i2.564

```
# Split features (X) and target (y)
# Class 2 = Benign, Class 4 = Malignant
X = data.iloc[:, :-1].values
y = data.iloc[:, -1].values

# Split into 80% training and 20% testing
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=0)

# Feature Scaling: Essential for Logistic Regression convergence
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)

# ========================================
# 4. MODEL TRAINING & EVALUATION
# ========================================

# --- Model A: Logistic Regression ---
print("\n--- Logistic Regression ---")
lr_classifier = LogisticRegression(random_state=0)
lr_classifier.fit(X_train, y_train)
y_pred_lr = lr_classifier.predict(X_test)

# Evaluate
cm_lr = confusion_matrix(y_test, y_pred_lr)
print(f"Accuracy: {accuracy_score(y_test, y_pred_lr) * 100:.2f}%")
plot_confusion_matrix(cm_lr, np.unique(y_test), title='Logistic Regression CM')



# --- Model B: Decision Tree (Entropy) ---
print("\n--- Decision Tree (Entropy) ---")
dt_classifier = DecisionTreeClassifier(criterion='entropy', random_state=0)
dt_classifier.fit(X_train, y_train)
y_pred_dt = dt_classifier.predict(X_test)

# Evaluate using manual accuracy calculation
cm_dt = confusion_matrix(y_test, y_pred_dt)
accuracy_dt = np.diag(cm_dt).sum() / cm_dt.sum()
print(f"Accuracy: {accuracy_dt * 100:.2f}%")


# --- Model C: Random Forest ---
print("\n--- Random Forest ---")
rf_classifier = RandomForestClassifier(n_estimators=10, criterion='entropy',
random_state=0)
rf_classifier.fit(X_train, y_train)
y_pred_rf = rf_classifier.predict(X_test)

# Evaluate
cm_rf = confusion_matrix(y_test, y_pred_rf)
accuracy_rf = np.diag(cm_rf).sum() / cm_rf.sum()
print(f"Accuracy: {accuracy_rf * 100:.2f}%")
```

Referring to the Python code above, a helper function is a small function designed to support a main function in performing specific tasks within a program. Its primary purpose is to break complex problems into smaller, more manageable components. Instead of writing one large function that performs multiple operations, the program can be organized into several smaller functions, each handling a specific task. This approach improves code modularity, readability, reusability, and maintainability.

https://doi.org/10.24191/jcrinn.v11i2.564

The next stage involves data cleaning and data preprocessing, which are essential steps in data analysis and machine learning. Raw data is often incomplete, inconsistent, or contains errors. Therefore, these processes aim to improve data quality by correcting inaccuracies and preparing the dataset so that it becomes accurate, consistent, and suitable for analysis or modelling.

Finally, model training and model evaluation are key stages in machine learning used to develop and assess predictive models. Model training allows the algorithm to learn patterns from the data and build a predictive model, while model evaluation assesses the model's performance to ensure that its predictions are accurate and reliable before it is applied to real-world situations.

4. RESULT AND DISCUSSION

The experimental results conclusively demonstrated that the random forest model outperformed other models as a forecasting tool, boasting the highest accuracy score. Specifically, the random forest achieved an accurate score of 97.81%, surpassing both the decision tree model (96.35%) and the logistic regression model (97.08%). This suggests that the predictions made by the random forest model are the most accurate among the evaluated models.

Table. 1. Model evaluation metrics

	Accuracy (%)	**Precision**	**recall**	**F1 Score**
Logistic Regression	97.08	0.96	0.98	0.97
Decision Tree	96.35	0.96	0.96	0.96
Random Forest	97.81	0.97	0.98	0.98

The result in Table 1 proved that Random Forest is a powerful and flexible machine learning algorithm that improves predictive accuracy by combining multiple decision trees. It is particularly effective in handling large datasets with higher dimensions and provides reliable predictions while mitigating the risk of overfitting. Its ability to estimate feature importance also makes it a valuable tool for understanding complex datasets.

5. CONCLUSION AND RECOMMENDATIONS

This study evaluated the performance of three machine learning algorithms i.e. Logistic Regression, Decision Tree, and Random Forest in predicting breast cancer using Python. Based on accuracy score comparisons, Random Forest achieved the highest performance (97.81%), followed by Logistic Regression (97.08%) and Decision Tree (96.35%), indicating that Random Forest provides the most reliable predictions among the models tested. Its superior performance may be attributed to its ensemble structure, which reduces overfitting and improves classification accuracy when handling datasets with multiple explanatory variables. In contrast, Decision Tree models are more prone to overfitting as the number of splits increases, while Logistic Regression performs better when noise variables are limited. Therefore, Random Forest is recommended as the preferred model for breast cancer classification in this context.

For future research, it is suggested that larger and more diverse datasets be incorporated to enhance predictive robustness, including additional clinical and biological variables that may improve diagnostic and prognostic accuracy. Continuous data updates and the inclusion of broader health-related factors could further strengthen model performance and practical applicability.

https://doi.org/10.24191/jcrinn.v11i2.564

6. ACKNOWLEDGEMENTS/FUNDING

We express our sincere gratitude for the invaluable support and resources generously provided by the UiTM Perlis Branch, which have played a crucial role in facilitating the completion of this research. The encouragement, understanding, and unwavering confidence in our work from the UiTM Perlis community have been a significant source of motivation, driving our research efforts forward. We are deeply appreciative of the supportive and nurturing environment fostered by the UiTM Perlis Branch, which has enabled and inspired us to carry out this study successfully.

7. CONFLICT OF INTEREST STATEMENT

We certify that the article is the Authors' and Co-Authors' original work. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The article has not received prior publication and is not under consideration for publication elsewhere. This research/manuscript has not been submitted for publication, nor has it been published in whole or in part elsewhere. We testify to the fact that all Authors have contributed significantly to the work, validity and legitimacy of the data and its interpretation for submission to Journal of Computing Research and Innovation.

8. AUTHORS' CONTRIBUTIONS

Wan Nashua Amira Ayob and **Nor Hayati Shafii** contributed to the conceptualization of the study, development of the methodology, and performed the formal statistical and machine learning analyses. **Wan Nashua Amira Ayob** also prepared the original draft of the manuscript. **Nor Hayati Shafii** was responsible for data curation, data preprocessing, implementation of the analytical models using appropriate software, and preparation of visualizations and figures. **Nur Fatihah Fauzi** conducted the investigation, assisted in data collection, and managed the research resources and dataset organization. **Diana Sirmayunie Mohd Nasir** and **Nor Azriani Mohamad Nor** contributed to the methodological refinement, provided supervision throughout the research process, and assisted in reviewing and editing the manuscript.

9. REFERENCES

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction (2nd ed.)*. Springer. https://doi.org/10.1007/978-0-387-84858-7.

Ismail, N. S., & Sovuthy, C. (2019). Breast cancer detection based on deep learning technique. In *2019 International UNIMAS STEM 12th Engineering Conference (EnCon)* (pp. 89–92). https://doi.org/10.1109/EnCon.2019.8861256.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R.* Springer. https://doi.org/10.1007/978-1-4614-7138-7.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415.

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research, 247*(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030.

https://doi.org/10.24191/jcrinn.v11i2.564

Mujar, N. M. M., Dahlui, M., & Taib, N. A. (2018). Presentation, diagnosis, and treatment among patients with breast cancer in Malaysia. *Journal of Global Oncology, 4*(3), 25s–25s. https://doi.org/10.1200/jgo.18.10280.

Sahran, S., Qasem, A., Omar, K., Albashih, D., Adam, A., Norul Huda Sheikh Abdullah, S., Abdullah, A., Iqbal Hussain, R., Ismail, F., Abdullah, N., Hayati Md Pauzi, S., & Abd Shukor, N. (2018). Machine learning methods for breast cancer diagnostic. In *Breast Cancer and Surgery*. IntechOpen. https://doi.org/10.5772/intechopen.79446.

World Health Organization. (2021, November 19). *What is breast cancer*? https://www.who.int/news-room/fact-sheets/detail/breast-cancer.

https://doi.org/10.24191/jcrinn.v11i2.564