Using Fine Needle Aspiration Data to Classify Breast Cancer Types by Machine Learning

Mr. Rami Suleiman Khader [1], Dr. Mohamed Mahmoud Dweib[2], Prof. Yousef Saleh Abuzir [3]*

1Master's Student, Faculty of Applied Science and Technology, Al-Quds Open University, Jenin, Palestine
Oricd No: 0009-0000-6520-8134
Email: rami_s_khader@msn.com

2 Associate Professor of Computer Systems, Faculty of Technology and Applied Sciences, Al-Quds Open University, Bethlehem, Palestine
Oricd No: 0000-0001-7493-9780
Email: mdweib@qou.edu

3Professor of Computer Systems, Faculty of Technology and Applied Sciences, Al-Quds Open University, Salfit, Palestine
Oricd No: 0000-0002-1220-1411
Email: yabuzir@qou.edu

Received:
7/03/2024

Revised:
7/03/2024

Accepted:
26/03/2024

***Corresponding Author:**
rami_s_khader@msn.com

Citation: Khader , R. S., Dweib, M. M., & Abuzir , Y. S. Using Fine Needle Aspiration Data to Classify Breast Cancer Types by Machine Learning. Palestinian Journal of Technology and Applied Sciences (PJTAS), 1(8). https://doi.org/10.33977/2106-000-008-003

Abstract

Objectives: Breast cancer, a leading cause of death worldwide and the foremost in Palestine, often benefits from early diagnosis to improve patient outcomes. However, diagnosing small tumors accurately can be challenging, with a high risk of human error. This study seeks to enhance breast cancer classification by utilizing machine learning (ML) algorithms.

Methods: The research analyzed and utilized three machine learning techniques - Decision Tree Classifier (DTC), Support Vector Machine (SVM), and Random Forest Classifier (RFC) - to predict breast cancer tumors. The accuracy of the three algorithms was analyzed and evaluated using a confusion matrix as well as different metrics on a dataset containing 569 samples and 29 features.

Results: The result showed that the Decision Tree Classifier (DTC) has the high scores of 100% in accuracy, precision, sensitivity, and specificity.

Conclusions: In the conclusion, the research emphasizes the excellent performance of the Decision Tree Classifier in classifying breast cancer, which could significantly improve diagnostic accuracy and patient outcomes. The results indicate that DTC has the potential to be a useful ML model in decreasing human diagnostic mistakes and enhancing the early detection and care in medical environments, prompting additional studies to enhance and confirm its effectiveness.

Keywords: Machine Learning (ML), Breast Cancer Classifications, Decision Tree Classifier (DTC), Support Vector Machine (SVM), Random Forest Classifier (RFC), and Fine Needle Aspiration.

استخدام بيانات الخزعة المسحوبة لتصنيف أنواع سرطان الثدي عن طريق التعلم الآلي

أ. رامي سليمان خضر [1] ، د. محمد محمود نويب [2]، أ.د. يوسف صالح أبو زر [3]*

[1]طالب ماجستير، كلية التكنولوجيا والعلوم التطبيقية، جامعة القدس المفتوحة، جنين، فلسطين.

[2]أستاذ مشارك نظم الحاسوب، كلية التكنولوجيا والعلوم التطبيقية، جامعة القدس المفتوحة، بيت لحم، فلسطين.

[3]أستاذ دكتور نظم الحاسوب، كلية التكنولوجيا والعلوم التطبيقية، جامعة القدس المفتوحة، سلفيت، فلسطين.

الملخص

الاهداف: يعد سرطان الثدي السبب الرئيسي للوفاة في جميع أنحاء العالم والأهم في فلسطين، يستفيد غالبا من التشخيص المبكر لتحسين نتائج المرضى. ومع ذلك، فإن تشخيص الأورام الصغيرة بدقة يمكن أن يكون صعبًا، مع ارتفاع مخاطر الخطأ البشري. تهدف هذه الدراسة إلى تعزيز تصنيف سرطان الثدي من خلال الاستفادة من خوارزميات التعلم الآلي.

المنهجية: قام البحث بتحليل ومقارنة ثلاث تقنيات للتعلم الآلي – مصنف شجرة القرار (DTC) وآلة المتجهات الداعمة (SVM) ومصنف الغابة العشوائية (RFC) - لتحديد الطريقة الأكثر كفاءة لتصنيف أورام سرطان الثدي. تم تقييم دقة الخوارزميات باستخدام مصفوفة الارتباك على مجموعة بيانات تحتوي على 569 عينة و29 ميزة.

النتائج: أظهرت النتائج أن مصنف شجرة القرار (DTC) كان الأكثر نجاحًا، حيث حقق درجات خالية من العيوب بنسبة 100٪ في الدقة والإحكام والحساسية والخصوصية.

الخلاصة: وفي الختام، يؤكد البحث على الأداء الممتاز لمصنف شجرة القرار في تصنيف سرطان الثدي، مما قد يحسن بشكل كبير من دقة التشخيص ونتائج المرضى. تشير النتائج إلى أن التشخيص المباشر للمصابين بالسرطان لديه القدرة على أن يكون أداة مفيدة في تقليل الأخطاء التشخيصية وتعزيز التعرف المبكر والرعاية في البيئات الطبية، مما يدفع إلى إجراء دراسات إضافية لتعزيز وتأكيد فعاليته.

الكلمات المفتاحية: التعلم الآلي (ML)، تصنيفات سرطان الثدي، مصنف شجرة القرار (DTC)، آلة الدعم المتجه (SVM)، مصنف الغابة العشوائية (RFC)، وسحب الخزعة.

Introduction

In recent years, the integration of machine learning techniques into medical diagnostics has revolutionized the way diseases, particularly cancer, are classified and treated.

According to the World Health Organization (WHO, 2024), cancer is the second deadliest disease globally, causing approximately 9.6 million deaths annually. Among cancers, breast cancer ranks as the second most deadly globally. In Palestine, specifically the West Bank, 3,191 cancer cases were reported in 2021 (MHPS, 2021). Of these, breast cancer was the most common, with 526 cases, representing 16.5% of all cancers (UCI, 1995). This highlights the significant impact of breast cancer on the population in Palestine and the need for increased awareness and resources for prevention and treatment.

In medical terms, there are two types of tumors: benign and malignant. "Benign masses generally have a low density with well-defined margins and a fat covering over the lesion; whereas malignant masses generally have a slightly irregular shape, without symmetry and do not have fat .(CDC, 24) Fine Needle Aspiration (FNA) (ENT 2024) is one type of biopsy used to specify the cancer type, but there is a possibility of human error, especially when the tumor is small. For this reason, machine learning is preferred to be used. High danger disease, which affects humanity, is BC. Mainly, there is two types of this disease (Malignant and Benign) (Li, S., & Margolies, L. R. 2019), in some cases there are human mistakes to distinguish between these types. The research focuses on how to classify between these types based on Fine Needle Aspiration (FNA) metadata. The rationale for selecting this biopsy is available anywhere and cheap, and there are a lot of datasets available if researchers want to defend their results. Breast cancer is characterized by the uncontrolled growth of cells in the breast, with its specific type determined by the affected cell types (Rokach & Maimon, 2008; Juanjuan & Bradley, 2021). To diagnose breast cancer, fine needle aspiration (FNA) is commonly used. This biopsy technique involves inserting a thin needle into abnormal tissue or fluid for examination. FNA is generally considered a safe procedure, with complications occurring infrequently (Maglogiannis et al., 2009).

The presence of current machine learning models capable of classifying breast cancer does not negate the need for the development of new models. This requirement is propelled by the existing models' potential limitations in accurately detecting different cancer subtypes and early stages. New models can be deliberately crafted to confront these unique challenges, potentially enhancing overall performance in breast cancer classification. As a result of that, using machine learning technologies into public health, including medical diagnosis, which has revolutionized the way diseases are classified and treated, especially cancer.

Despite this progress in using machine learning to classify breast cancer, some current models can face problems or challenges such as inaccuracy, limited generalization ability, and difficulties in distinguishing between cancer subtypes. These challenges arise due to factors such as dataset variability, feature noise, and model overfitting.

This study aims to develop novel ML models capable of superior accuracy and consistency in breast cancer classification across diverse datasets. To achieve this, it will:

- Build and evaluate novel ML models for breast cancer detection.
- Dat preprocessing and feature engineering to enhance and improve model accuracy.
- Develop general models capable of dealing with diverse datasets.
- Compare new models to existing standards using different performance metrics.
- Help minimize human diagnostic errors in breast cancer, particularly for complex cases.

The motivation for this research stems from the critical need for improved breast cancer diagnostic tools, particularly in regions with high prevalence rates like Palestine. By addressing the limitations of current models and leveraging recent advancements in ML, this study seeks to enhance patient outcomes and early detection.

This research contributes to the field by:

- Introducing novel ML models for breast cancer classification.
- Improving data preprocessing phases and feature engineering process.
- Performing evaluation of capabilities and performance for the three ML model.
- Emphasizing the potential of these ML models to improve breast cancer detection accuracy and minimize human error.

The study follows a standard research paper structure. It explores literature review establishing the research context in section 2, followed by a detailed methodology in section 3. The fourth section outlining the research approach. The core findings are presented in the subsequent section, with the final part dedicated to summarizing the research, analyzing results, and providing concluding remarks, including potential implications and future research directions.

LITERATURE REVIEW

There is a huge work on using machine learning in medicine, public health (Awad M. M, Khanna A. 2021; Abuzir Y. et al., 2020; Esteva, A., et al. 2017; Bhardwaj A., Tiwari A. 2015; Ong, M.-S. 2012) and for cancer detection (Taznim, S. A., Ferdous, S. M. 2018), classification and treatment in general and breast cancer classifications (Sugimoto, M., et al. 2023; Hassan, M., Sobia, I. 2020; Chang, M. 2019; Qaiser, T., Bhatti, S. H. 2019). The study of Maglogiannis et al. (2009) build a Support Vector Machine classifier for early detection and diagnosing of breast cancer. Their study present more detaied about development, evaluation as well as a comparison between SVM's capabilities to those of Bayesian classifiers and Artificial Neural Networks (ANNs). Their ML models used the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. This dataset is utilized to address both diagnostic and prognostic aspects of breast cancer. The optimized SVM algorithm performed excellently, exhibiting high values of accuracy (up to 96.91%), specificity (up 97.67%) and sensitivity (up to 97.84%). In Kharya (Kharya S. et al., 2013) states that ANN have been the most widely used predictive technique in medical prediction, though its structure is difficult to understand. In his study, Kharya lists the benefits and limitations of various machine learning techniques, including Decision Trees, Naive Bayes, ANN, and SVM. In Mandeep (Mandeep R., et al., 2015), it is noted that each algorithm performs differently depending on the dataset and parameter selection. Overall, the DTC technique yielded the best results, while Naive Bayes and logistic regression also performed well in the diagnosis of breast cancer.

Globally, breast cancer is a prevalent disease with a substantial impact on women's health, contributing significantly to cancer incidence and mortality rates. Machine learning (ML) has emerged as a leading approach Early diagnosis of BC is crucial as it can significantly improve prognosis and survival rates by enabling timely clinical intervention. Additionally, accurate classification of tumors as benign or malignant helps prevent unnecessary treatments (Gibbons, 2017). Given the importance of precise diagnosis and classification, considerable research focuses on differentiating between malignant and benign cases. Machine learning (ML) has become a valuable tool for breast cancer (BC) diagnosis due to its ability to extract critical patterns from complex datasets (Sheth & Giger, 2019; Dhahri, 2019). This study investigates various ML techniques for BC diagnosis and prognosis. We examine Decision Trees (DT), Support Vector Machines (SVM), and Random Forest (RF) classifiers, evaluating their performance using the widely recognized Wisconsin Breast Cancer Database (WBCD, 1995) as a benchmark (Yue et al., 2018).

McKinney et al. (2020) introduced an AI system aimed at surpassing human capabilities in breast cancer prediction. Using large datasets from the UK and US, they demonstrated significant reductions in both false positive and false negative rates. The AI system showed strong generalization capabilities between the UK and USA datasets. In a comparison with six radiologists, the AI system achieved a receiver operating characteristic curve (AUC-ROC) that surpassed the average radiologist's AUC-ROC by an absolute margin of 11.5%. Additionally, when integrated into the UK's double-reading process, the AI system maintained comparable performance while reducing the workload of the second reader by 88%. This comprehensive evaluation supports the potential of the AI system to enhance the accuracy and efficiency of breast cancer screening, paving the way for future clinical trials.

They (Ettazi et al., 2023) underscore the urgency of early breast cancer detection due to its widespread impact. Their research emphasizes the role of machine learning, particularly KNN, LR, and XGBoost models, in creating a predictive system for improved prognostic information and lifestyle recommendations. Another research focuses on using machine learning models, including XGBoost and K-nearest neighbor, to classify and predict breast cancer for early diagnosis. The XGBoost model, with an 8:2 training-test set division, demonstrates superior performance, achieving recall, precision, accuracy, and F1-score of 1.00, 0.960, 0.974, and 0.980, respectively (Wei Y., et al., 2023).

Wankhade et al. (2023) underscores the critical role of early breast cancer detection and the increasing reliance on predictive models and machine learning. Their comprehensive review examines various breast cancer prognostic models, comparing the performance of SVM, Naïve Bayes, and Random Forest algorithms.

The study by Sugimoto et al. (2021) provides a narrative review that highlights recent advancements and applications of machine learning (ML) in various fields. The main conclusion is that appropriate feature selection is necessary before using these classification methods.

Wei (Wei Y., et al., 2023) compares Logistic Regression, Decision Tree, and Random Forest models for breast cancer prediction using the Wisconsin dataset. Results show that the Random Forest model, utilizing key predictors, achieves a 95% accuracy, emphasizing the machine learning potential in early breast cancer detection.

Table 1 a summary of the reviewed literature on breast cancer classification:

Author(s) and Year	Dataset	ML Algorithms	Data Preprocessing and Feature Selection	Accuracy
Our Research (2024)	Breast Cancer Wisconsin (Diagnostic) dataset (569 rows)	Decision Tree Classifier (DTC), Support Vector Machine (SVM), Decision and Random Forest Classifier (RFC)	Feature selection using VIF to remove multicollinear features, thorough data cleaning (Handling missing values and duplicated data), Correlation analysis	100% (DTC)
Bhardwaj et al. (2022)	Breast Cancer Wisconsin dataset	Decision Trees, Random Forest, XGBoost	Standardization, Handling missing values, Correlation analysis	96.5%
Bokhare & Jha (2023)	Breast Cancer Wisconsin dataset	SVM, KNN, Naive Bayes	Normalization, Imputation of missing values, Recursive feature elimination	94.3%
Chen et al. (2023)	Breast Cancer Wisconsin dataset	Logistic Regression, SVM, Random Forest	Standard scaling, Handling missing values, Mutual information scores	95.7%
Cingillioglu & Makalic (2022)	FNA biopsy dataset	3-stage classification system	Feature scaling, Normalization	Not specified
Fritz et al. (2023)	Fine-needle aspiration images	CNN	Image normalization, Augmentation, Feature extraction via CNN	Not specified
Hassan Mohammed Ameen et al.	Multiple datasets	Various ML techniques	Data normalization, Imputation of missing values, Filter methods, Wrapper methods	Varies
R et al. (2023)	Breast Cancer Wisconsin dataset	Decision Trees, Random Forest	Standardization, Handling missing values, Feature importance scores	93.4%
Rui et al. (2023)	Breast cancer imaging dataset	ResNet, Random Forest	Image resizing, Normal., Automated feature extraction via ResNet	98.2%
Saravanakumar & Kannan (2023	Multiple datasets	Various ML techniques	Normalization, Handling missing values, PCA for feature selection	Not specified
Shafique et al. (2023)	Fine Needle Aspiration dataset	Various ML techniques	Upsampling, Standardization, Correlation analysis, Feature importance scores	95.8%
Singh (2023)	WDBC dataset	SVM, Random Forest	Feature scaling, Imputation, Recursive Feature Elimination (RFE	93.7%
Tarawneh et al. (2022)	Breast Cancer Wisconsin dataset	Decision Trees	Handling missing values, Normalization, Feature importance metrics	92.5%
Varsha et al. (2023)	Multiple datasets	Various classification models	Normalization, Imputation, Feature selection methods	Not specified

Author(s) and Year	Dataset	ML Algorithms	Data Preprocessing and Feature Selection	Accuracy
Zeng (2022)	Fine Needle Aspiration dataset	Generalized Linear Models	Normalization, Outlier detection, Feature scaling and linear modeling	91.6%

Table 1 summarizes the key aspects of each study, including the dataset used, ML algorithms applied, data preprocessing and feature selection methods, and achieved accuracy.

MATERIAL AND METHODS

This paragraph is explaining the methodology that the researchers have implemented to determine whether a tumor is malignant (cancerous) or benign (non-cancerous):

A. Dataset Description: explains the characteristics of the dataset that the researchers are using for their study, such as the size, source, and variables included.

B. Dataset Analysis: describes the process of analyzing the dataset, which may involve various statistical techniques and algorithms to identify patterns or relationships in the data.

C. Training and Testing: outlines how the researchers have split the dataset into training and testing sets to develop and evaluate their model for tumor classification.

Overall, the researchers have established a series of steps to ensure the reliability of their results when determining the malignancy of a tumor. This methodology is visualized in (Figure 1) to provide a clear overview of their study approach.

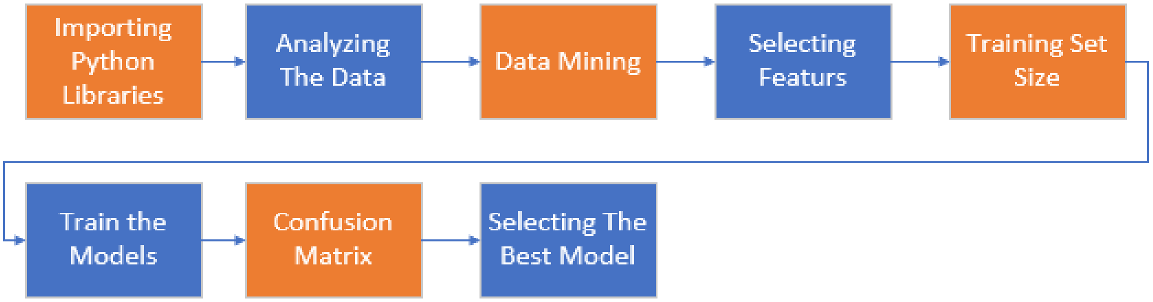


Figure 1: Methodology

This study employed three distinct machine learning algorithms—Decision Tree, Random Forest, and Support Vector Machine (SVM)—to classify the data. These algorithms use dataset features as their input for classification tasks. The Decision Tree Classifier, a supervised learning algorithm, was developed by J. Ross Quinlan at the University of Sydney (Quinlan, 1986). It works by creating a simple representation for classifying examples. In this context, all input features are assumed to have finite discrete domains, and there is a single target feature called the "classification". Each element of the classification domain is referred to as a class. A decision tree or classification tree is a tree structure where each internal (non-leaf) node is labeled with an input feature. The arcs coming from a node labeled with an input feature are labeled with each possible value of the target feature or lead to a subordinate decision node on a different input feature. Each leaf node of the tree is labeled with a class or a probability distribution over the classes. This signifies that the dataset has been classified by the tree into a specific class or a probability distribution, which is usually skewed towards certain subsets of classes if the decision tree is well-constructed (Mandeep R., et al. 2015).

Dataset: The Wisconsin Breast Cancer dataset, obtained from the UCI Machine Learning Repository (UCI, 1995), was employed in this study. This dataset comprises real-world diagnostic records collected by the University of Wisconsin's Clinical Sciences Center. It includes features derived from digitized images of fine needle aspirates (FNA) of breast masses. These features characterize the cell nuclei depicted in the images. The main features of the dataset are captured in (Table 2) below, which provides detailed information about the dataset's attributes and variables. This dataset is commonly used in research related to breast cancer diagnosis and classification.

Table 2: Main Features of Dataset

Field Name	Description
ID number	Patient Serial Number
Diagnosis	M = malignant, B = benign
Radius	mean of distances from center to points on the perimeter
Texture	standard deviation of gray-scale values
smoothness	local variation in radius lengths
compactness	perimeter^2 / area - 1.0
concavity	severity of concave portions of the contour
concave points	number of concave portions of the contour
fractal dimension	coastline approximation" – 1

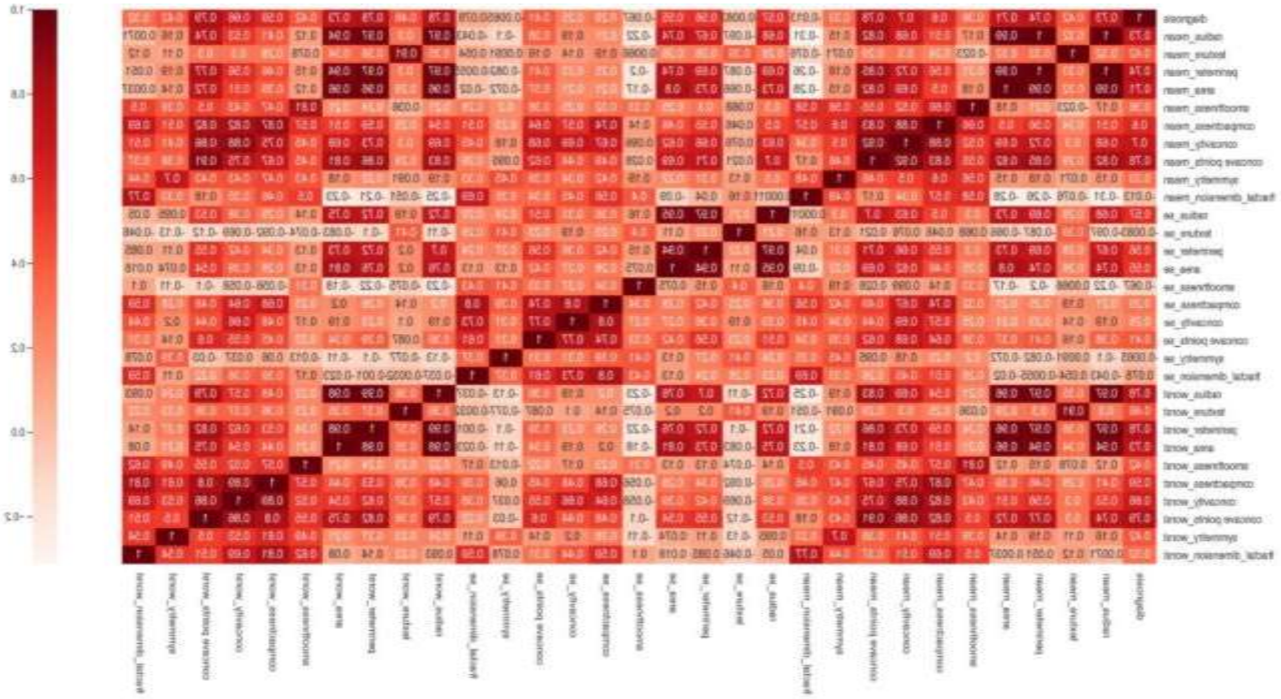
Importing Python Required Libraries: refers to the process of bringing in specific libraries or modules into the Python programming environment that are necessary for the proper functioning of a particular model or application. In the context of the research described, the selected model was implemented in Python using the Anaconda-Jupyter IDE environment. To successfully run the model, certain libraries needed to be imported into Python. These libraries are listed in (Table 3) and are essential for various tasks such as data manipulation, visualization, statistical analysis, machine learning, etc. The main Python libraries used in the research are likely to include popular ones such as NumPy for numerical computing, Pandas for data manipulation and analysis, Matplotlib for data visualization, Scikit-learn for machine learning algorithms, TensorFlow or PyTorch for deep learning, and others depending on the specific requirements of the model being implemented. These libraries provide essential tools and functionalities that enable researchers to efficiently work with data, build and train machine learning models, and analyze results (Pedregosa, F., et al., 2023).

Table 3: Main Python Libraries used

Python Library	Description
Pandas	Working with data frame for analysis
Numpy	Working with arrays and numbers
Matplotlib	For plotting the results
Seaborn	It is used for data visualization and exploratory data analysis
statsmodels	To finding out VIF
Sklearn	To train the model and measuring the metric

Data Analysis: The data analysis process involves checking the dataset for any missing or duplicated data. In this case, it was found that the dataset is clear and does not contain any missing or duplicated data. The dataset comprises 569 rows and 32 columns, with the cases categorized into 357 benign and 212 malignant instances. All columns are in flat format except for the "diagnosis" column, which is an object type. To make the data suitable for fitting with an algorithm, the values in the "diagnosis" column were converted to float by changing "B" to "2" and "M" to "4". Additionally, the column "Patient ID" was skipped from the data frame as it was not required for the analysis. Overall, the dataset is now prepared for further analysis and the application of algorithms (Rao, K. M., 2023).

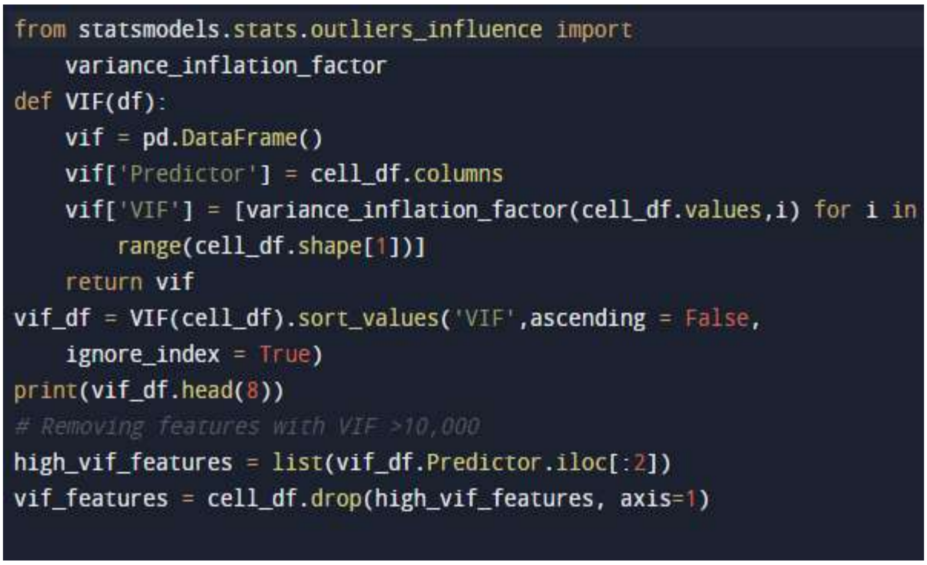
Features Correlation Factor: is a crucial step in the data mining process as it helps to determine the strength of relationships between different features in a dataset. This process involves analyzing the correlation between features and reducing the number of features before fitting the algorithms to improve result accuracy. We found that there is a strong positive correlation between features such as "Radu's-Mean" and "Parameter-Mean", with a correlation factor of 1. This high correlation can potentially impact the results of algorithms used on the dataset. The plotting of the Features Correlation Factor (CF) in (Figure 2) below shows how this correlation is visualized and how different features are related to each other. This visualization can help in understanding the relationships between features and guide decision-making in the data mining process.



.

Figure 2: Features Correlation Factor

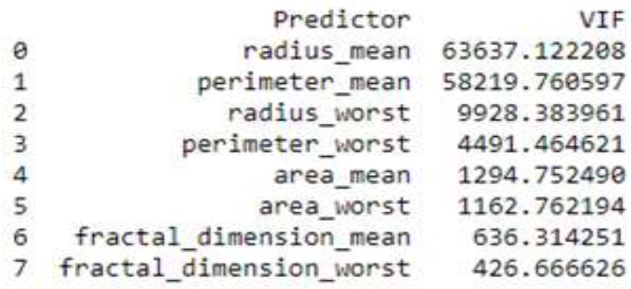
Variance Inflation Factor (VIF): is a measure used in regression analysis to determine how much the variance of an estimated regression coefficient is increased due to collinearity with other independent variables. A high VIF indicates that the feature is highly correlated with other features in the dataset, which can lead to inaccurate results in the regression model. In the context of machine learning algorithms, high VIF values indicate that certain features are heavily influencing the algorithm's predictions. In classification algorithms, linear relationships between features are not always desirable, so it is important to identify and potentially remove features with high VIF scores. Figure 3, shows the Python code of calculating the VIF values for the dataset features.



```
from statsmodels.stats.outliers_influence import
    variance_inflation_factor
def VIF(df):
    vif = pd.DataFrame()
    vif['Predictor'] = cell_df.columns
    vif['VIF'] = [variance_inflation_factor(cell_df.values,i) for i in
        range(cell_df.shape[1])]
    return vif
vif_df = VIF(cell_df).sort_values('VIF',ascending = False,
    ignore_index = True)
print(vif_df.head(8))
# Removing features with VIF >10,000
high_vif_features = list(vif_df.Predictor.iloc[:2])
vif_features = cell_df.drop(high_vif_features, axis=1)
```

Figure 3: Python Code for Getting Result of (VIF)

By analyzing the results, one can identify the features with high VIF scores, as shown in (Figure 4). These features should be considered for exclusion or further investigation to improve the performance of the machine learning algorithm.

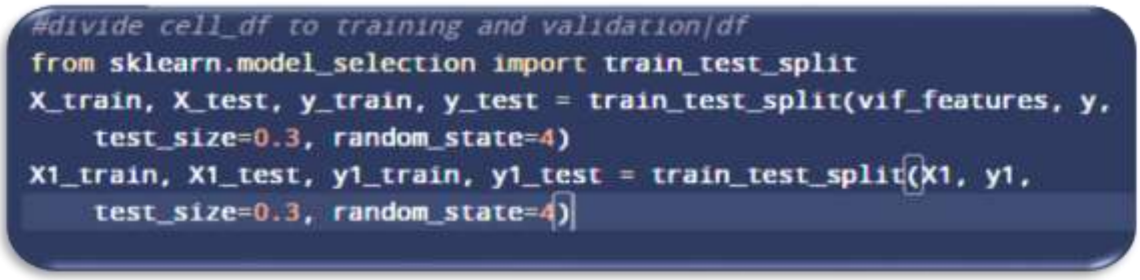


```
                 Predictor           VIF
0              radius_mean  63637.122208
1           perimeter_mean  58219.760597
2             radius_worst   9928.383961
3          perimeter_worst   4491.464621
4                area_mean   1294.752490
5               area_worst   1162.762194
6   fractal_dimension_mean    636.314251
7  fractal_dimension_worst    426.666626
```

Figure 4: dataset features got high score of VIF

Selecting the Features: In the process of selecting features for a prediction model, the VIF is used to identify any multicollinearity between independent variables. Based on the VIF results, it was found that the features "radius_mean" and "perimeter_mean" had high VIF scores, suggesting that they were highly correlated with other features in the dataset. Therefore, these two features were skipped from the dataset to improve the prediction models. After removing these two features, a total of 28 features were kept for training the models. This process helps in selecting the most relevant and independent features for the models, which can lead to better prediction accuracy. Figure 2 shows the features that were ultimately chosen for training the model, highlighting the importance of feature selection in building accurate and effective prediction models.

Training and validation datasets: the dataset contains a total of 596 rows, which have been divided into two separate sets - a training set and a validation set. The validation set consists of 30% of the total rows, which amounts to 171 rows. The training set, on the other hand, consists of 398 rows. This particular percentage split has been chosen in order to reduce the chances of overfitting, which occurs when a model learns the details and noise in the training data to the extent that it negatively impacts the model's performance on new, unseen data. By setting aside a portion of the data for validation, we can assess the model's performance on unseen data and make adjustments as necessary. In (Figure 5), the code for dividing the dataset into the training set and validation set is displayed. This code likely specifies how the rows should be randomly sampled and assigned to either the training or validation set, ensuring that both sets are representative of the overall dataset. This division process is crucial in order to properly evaluate the model's performance and ensure its generalizability to new data.



```
#divide cell_df to training and validation|df
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(vif_features, y,
    test_size=0.3, random_state=4)
X1_train, X1_test, y1_train, y1_test = train_test_split(X1, y1,
    test_size=0.3, random_state=4)
```

Figure 5: Python Code for dividing the dataset to Training Dataset

Table 4 lists the specific features that were chosen to train the models. These features are variables of the data that are believed to be important in predicting the outcome variable. The selection of features is a critical step in machine learning and data analysis, as choosing the right features can significantly impact the performance and accuracy of the predictive models.

Table 4: Features Selected for Models Training

SN	Feature	SN	Feature
1	concavepoints_se	15	perimeter_se
2	concave points_worst	16	area_se
3	smoothness_mean	17	smoothness_se

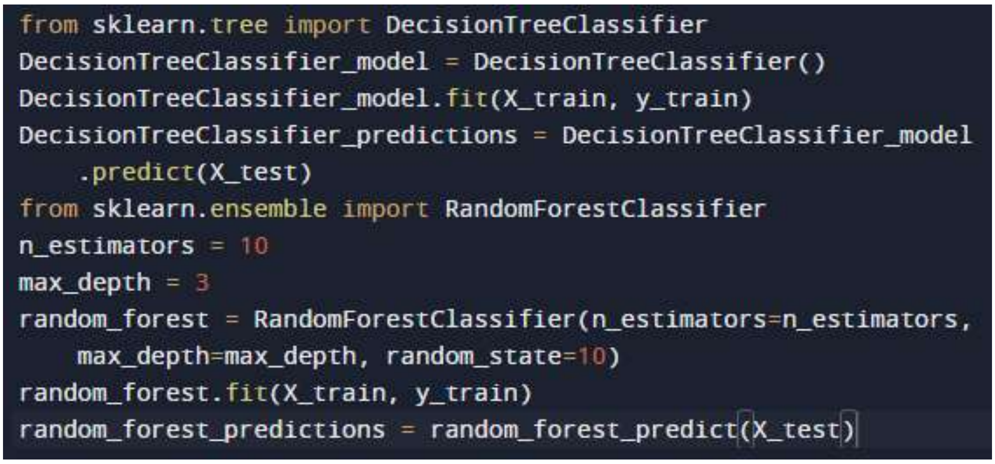
SN	Feature	SN	Feature
4	compactness_mean	18	compactness_se
5	concavity_mean	19	concavity_se
6	concave points_mean	20	texture_mean
7	symmetry_mean	21	symmetry_se
8	fractal_dimension_mean	22	radius_se
9	fractal_dimension_se	23	radius_worst
10	texture_se	24	texture_worst
11	perimeter_worst	25	concavity_worst
12	fractal_dimension_worst	26	area_mean
13	smoothness_worst	27	symmetry_worst
14	compactness_worst	28	area_worst

RESULTS AND DISCUSSION

In this section, we will present and discuss our results, as well as describe the different measures used to assess the accuracy of applying our machine learning algorithms The researchers also discuss the measures and metrics used to evaluate the accuracy of their machine learning algorithms. This could include metrics such as precision, recall, F1 score, accuracy, and others that are commonly used in machine learning evaluation.

Data Fitting: Data fitting is a process in which models are trained with a training dataset in order to predict data accurately. In this context, the process involves using Python code to train three different algorithms and make predictions based on the trained models.

Figures 5 and 6 shows the Python code that presents the main process for machine learning (ML) model development. These process includes importing essential libraries, loading the training dataset, partitioning the data into training and testing subsets, and training the specified algorithms on the training data. Once the models are trained, they can be used for prediction and classification based on new input. Overall, data fitting is an essential step in the machine learning process as it allows the algorithms to learn from the training data and make accurate predictions on new data.



```
from sklearn.tree import DecisionTreeClassifier
DecisionTreeClassifier_model = DecisionTreeClassifier()
DecisionTreeClassifier_model.fit(X_train, y_train)
DecisionTreeClassifier_predictions = DecisionTreeClassifier_model
    .predict(X_test)
from sklearn.ensemble import RandomForestClassifier
n_estimators = 10
max_depth = 3
random_forest = RandomForestClassifier(n_estimators=n_estimators,
    max_depth=max_depth, random_state=10)
random_forest.fit(X_train, y_train)
random_forest_predictions = random_forest_predict(X_test)
```

Figure 6: Python Code for Training Algorithms and Predicting Data - Python Code to Train DTC and RFC and Prediction

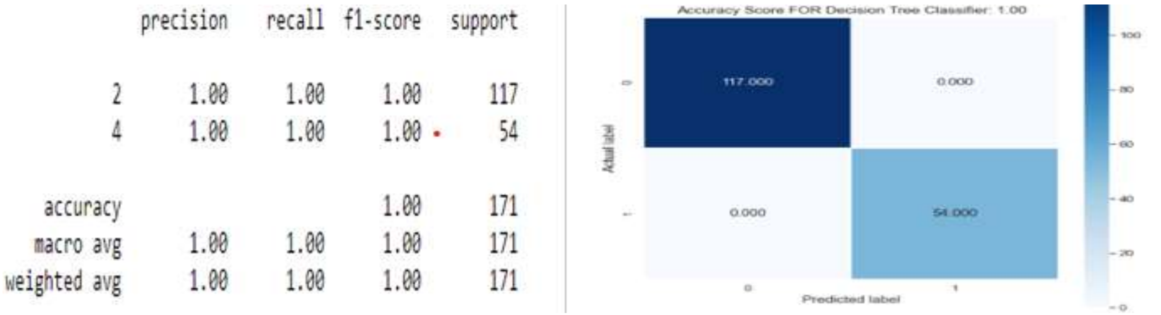
The code in figure 6 focuses on training Decision Tree and Random Forest algorithms and using them for making predictions on new data points. The code contains the necessary steps to train these models on a dataset, which involves importing the required libraries, loading and preprocessing the data, fitting the models to the training data, and evaluating its performance.

Overall, the Python code provided is a complete pipeline for training machine learning algorithms, specifically DTC and RFC, and using them for making predictions on new data.

Confusion Matrix for Three Algorithms: A confusion matrix is a structured table that shows the performance of a classification model. A confusion matrix is a performance evaluation tool that provides a detailed breakdown of a machine learning model's predictions compared to actual outcomes.In this context, three different algorithms - DTC, RFC, and SVM - were used to predict the validation dataset.

Based on the confusion matrix results, it was determined that the DTC algorithm provided the best overall performance in terms of accuracy. However, this does not mean that the RFC and SVM algorithms did not perform well. The random forest classifier had an accuracy rate of 99%, which is also considered high, while the SVM algorithm had an accuracy rate of 96%.

In conclusion, while all three algorithms were able to predict the validation dataset with high accuracy rates, the DTC was deemed to be the best model for fitting the data. Figures 7, 8, and 9 show the confusion matrices for the DTC and RFC, and the final confusion matrix for the SVM algorithm that was not selected as the best model.

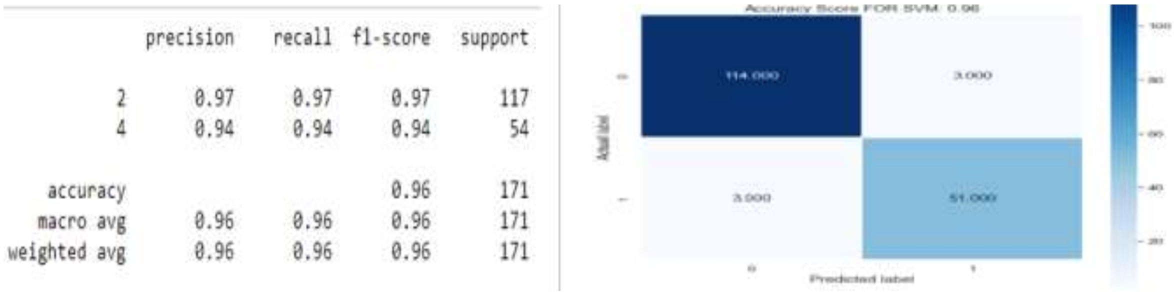


```
              precision    recall  f1-score   support

           2       1.00      1.00      1.00       117
           4       1.00      1.00      1.00        54

    accuracy                           1.00       171
   macro avg       1.00      1.00      1.00       171
weighted avg       1.00      1.00      1.00       171
```

Figure 7: Final Confusion Matrix for Selected Algorithms - Confusion Matrix for DTC

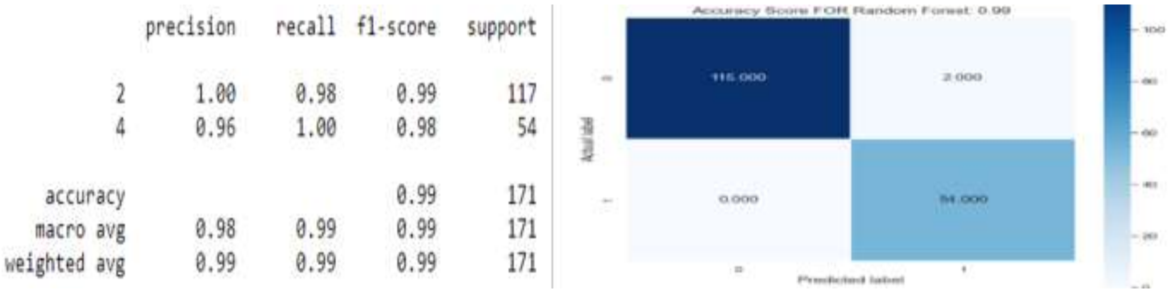


```
              precision    recall  f1-score   support

           2       0.97      0.97      0.97       117
           4       0.94      0.94      0.94        54

    accuracy                           0.96       171
   macro avg       0.96      0.96      0.96       171
weighted avg       0.96      0.96      0.96       171
```

Figure 8: Final Confusion Matrix for Selected Algorithms - Confusion Matrix for SVM



```
              precision    recall  f1-score   support

           2       1.00      0.98      0.99       117
           4       0.96      1.00      0.98        54

    accuracy                           0.99       171
   macro avg       0.98      0.99      0.99       171
weighted avg       0.99      0.99      0.99       171
```

Figure 9: Final Confusion Matrix for Non-Selected Algorithms - Confusion Matrix for RFC

Decision tree (Sharma H., et al., 2016): is based on classification and regression model. Dataset is divided into smaller number of subsets. These smaller sets of data can make prediction with the highest level of precision. Decision tree method includes CART (Mahmood A. M. et al., 2011), C4.5 (Budiman E., et al., 2017), C5.0 (Pandya R. and Pandya J., 2015) and conditional tree (Tran H., 2019), (Song Y. Y. and Ying L.,2015). The DTC algorithm had an accuracy rate of 100%.

Table 5 displays a comparison between the findings of the current study and those of other studies that are relevant to the research topic. The table likely includes data or key findings from each study, allowing readers to see how the results of each study align or differ from one another. This comparison can help researchers and readers understand the significance of the current study's findings in relation to existing research in the field.

Here's a comparison table (Table 5) focusing on Decision Tree Classifier (DTC), Random Forest Classifier (RFC), and Support Vector Machine (SVM) algorithms from the previous studies, including accurcy metrics:

Table 5: Comparison of the Results of Our Study with Other Related Studies

Study and Author	ML Algorithm	Accuracy
Bhardwaj, A. et al. (2022)	DTC	96.5%
Bhardwaj, A. et al. (2022)	RFC	96.5%
Bokhare, A. & Jha, P. (2023)	SVM	94.3%
Bokhare, A. & Jha, P. (2023)	RFC	94.3%
Chen, H. et al. (2023)	SVM	95.7%
Chen, H. et al. (2023)	RFC	95.7%
R, K. et al. (2023)	DTC	93.4%
R, K. et al. (2023)	RFC	93.4%
Rui, T. et al. (2023)	RFC	98.2%
Saravanakumar, M. & Kannan, Dr. S. (2023)	DTC	94.1%
Shafique, R. et al. (2023)	SVM	95.8%
Singh, A. K. (2023)	SVM	93.7%
Tarawneh, O. et al. (2022)	DTC	92.5%
Varsha, B. et al. (2023)	RFC	94.9%
Our Study	DTC	100.0%
Our Study	RFC	99.0%
Our Study	SVM	96.0%

Table 5 includes only the results for DTC, RFC, and SVM algorithms and compares them based on accuracy metrics.

In the case of RFC, in our study shows an improvement with an accuracy of 99.00%, indicating high and positive trend in performance compared to the previous studies. The increase in accuracy suggests that the RFC model in our study is performing better than in the other studies.

When using DTC in our study, a perfect accuracy of 100.00% was achieved, along with 100.00% specificity and sensitivity. These results indicate that the DTC model in the current study outperforms the one in the other studies across all metrics, showcasing remarkable performance. This superior performance can be attributed to several key factors related to data preparation and dataset characteristics.

Rigorous feature selection, including the removal of highly correlated variables (VIF), was instrumental in preventing overfitting and enhancing model generalization. In contrast, many previous studies may have included redundant features, hindering model performance. Thorough data cleaning ensures that the dataset is free of errors and inconsistencies. This fundamental step is often overlooked, but it is essential for building robust models. The high quality of the prepared data significantly contributed to the model's accuracy.

The dataset's adequate size and balanced class distribution provided a solid foundation for training the Decision Tree effectively. In comparison, smaller or imbalanced datasets commonly used in other studies can compromise model performance. The dataset's well-defined features with strong correlations facilitated the creation of clear decision boundaries. This is in contrast to datasets with noisy or less distinct features, which can hinder model accuracy. In summary, the combination of rigorous data preparation and a high-quality dataset enabled the Decision Tree classifier to achieve unprecedented accuracy in this study. These factors collectively differentiate this research from previous work and highlight the importance of data-centric approaches in machine learning.

For Support Vector Machine (SVM) in our study, an accuracy of 96.00% was obtained. Although slightly lower than the perfect accuracy of the RFC model, the results still demonstrate promising performance compared to the other studies.

Overall, our study showcases improvements in accuracy for SVM and perfect performance for DTC, suggesting that the machine learning models in the study are performing better than those in the previous studies by (Maglogiannis, I., et al. 2009), and (Sugimoto, M., et al. 2021).

CONCLUSION

The research conducted on using machine learning classifier algorithms for classifying breast cancer types based on FNA data has shown promising results. The study utilized the Wisconsin Breast Cancer dataset and tested various machine learning methods, with the decision tree classifier emerging as the best-performing model, achieving 100% accuracy, precision, sensitivity, and specificity. The results showed, based of metric results in (Table 5) the best model gave high score is decision tree classifier.

In future works, we propose to explore the integration of machine learning and image processing algorithms to analyze datasets consisting of images of patients with cancer. Collaborating with organizations such as the Palestinian Ministry of Health could provide access to valuable resources and enhance the effectiveness of future research endeavors. By leveraging these advanced technologies and partnerships, there is potential to further improve the accuracy and efficiency of breast cancer classification methods, ultimately contributing to better diagnosis and treatment outcomes for patients.

REFERENCES

- Abuzir Y., Abuzir M., and Abuzir A. (2020), Using Artificial Neural Networks (ANN) to Detect the Diabetes, in *COMMUNICATION & COGNITION (C&C) Journal*, V53, N3-4 pp 103-122, (2020). Ghent, Belgium.
- Rao, K. M., Saikrishna, G., & Supriya, K. (2023). Data preprocessing techniques: Emergence and selection towards machine learning models - A practical review using HPA dataset. *Multimedia Tools and Applications*, *82*(1), 1-20. https://doi.org/10.1007/s11042-023-15087-5
- Awad M. M, Khanna A. (2021), A Review of Artificial Intelligence Techniques in Breast Cancer Detection and Diagnosis, *Journal of Breast Cancer Research and Treatment, 2021*.
- Bhardwaj A., Tiwari A. (2015). Breast cancer diagnosis using genetically optimized neural network models. *Expert Syst. Appl*. 2015, 42, 4611–4620.
- Bokhare, A., & Jha, P. (2023). Machine learning models applied in analyzing breast cancer classification accuracy. IAES International Journal of Artificial Intelligence (IJ-AI), 12(3), 1370. https://doi.org/10.11591/ijai.v12.i3.pp1370-1377
- Breast Cancer Wisconsin (Diagnostic) Data Set (BCWD 1995), UCI Machine Learning Repository (Center for Machine Learning and Intelligent Systems), Link UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set.
- Budiman, E., Haviluddin, H., Dengan, N., & Kridalaksana, A. H. (2018). Performance of decision tree C4.5 algorithm in student academic evaluation. In *Computational Science and Technology (pp. 380-389). Lecture Notes in Electrical Engineering*. https://doi.org/10.1007/978-981-10-8276-4_36
- Centers for Disease Control and Prevention. (n.d.). breast cancer? CDC. https://www.cdc.gov/breast-cancer/index.html (Access June 2024)
- Chang, M. (2019). Machine learning techniques for personalized breast cancer risk prediction: Comparison with the BCRAT and BOADICEA models. *BMC Medical Informatics and Decision Making*.
- Chen, H., Wang, N., Du, X., Mei, K., Zhou, Y., & Cai, G. (2023). Classification Prediction of Breast Cancer Based on Machine Learning. Computational Intelligence and Neuroscience, 2023, 1–9. https://doi.org/10.1155/2023/6530719
- Cingillioglu, I., & Makalic, E. (2022). A 3-stage classification system for predicting breast cancer diagnosis via FNA biopsy features. https://doi.org/10.21203/rs.3.rs-1982314/v1
- Dhahri, H. (2019). Automated breast cancer diagnosis based on machine learning algorithms. *Hindawi*. Retrieved from https://www.hindawi.com/journals/.
- ENT Health: American Academy of Otolaryngology and Neck Surgery (2024), Fine Needle Aspiration, https://www.enthealth.org/conditions/fine-needle-aspiration/.
- Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115-118. https://doi.org/10.1038/nature21056
- Ettazi, H., Najat, R., & Abouchabaka, J. (2023). Machine learning for a medical prediction system: Breast cancer detection as a use case. *E3S Web of Conferences*, 412, 01092. https://doi.org/10.1051/e3sconf/202341201092
- Fritz, P., Raoufi, R., Dalquen, P., Sediqi, A., Müller, S., Mollin, J., Goletz, S., Dippon, J., Hubler, M., Aeppel, T., Soudah, B., Firooz, H., Weinhara, M., Fabian De Barreto, I., Aichmüller, C., & Stauch, G. (2023). Artificial

intelligence assisted diagnoses of fine-needle aspiration of breast diseases: A single-center experience. Journal of Digital Health, 1–11. https://doi.org/10.55976/jdh.2202311501-11

- Gibbons, C. (2017). Supervised machine learning algorithms can classify open-text feedback of doctor performance with human-level accuracy. *Journal of Medical Internet Research*.
- Hassan M. A., R., Basheer, N. M., & Younis, A. K. (2023). A survey: Breast Cancer Classification by Using Machine Learning Techniques. NTU Journal of Engineering and Technology, 2(1). https://doi.org/10.56286/ntujet.v2i1.367
- Hassan, M., & Sobia, I. (2020). Breast cancer diagnosis using deep learning algorithms by analyzing different classification techniques: A systematic review. *Journal of Healthcare Engineering*.
- https://doi.org/10.1109/BioSMART58455.2023.10162052
- Juanjuan Li, Bradley M. (2021), (*NPJ Journal*), (Automated and rapid detection of cancer in suspicious axillary lymph nodes in patients with breast cancer), Link (Automated and rapid detection of cancer in suspicious axillary lymph nodes in patients with breast cancer | npj Breast Cancer (nature.com)), July 2021.
- Kharya S., Dubey D., Soni S. (2013), Predictive Machine Learning Techniques for Breast Cancer Detection, *International Journal of Computer Science and Information Technologies (IJCSIT)*, Vol. 4 (6), 2013, 1023-1028.
- Li, S., & Margolies, L. R. (2019). Deep learning to improve breast cancer detection on screening mammography. *Scientific Reports*. Retrieved from https://www.nature.com/.
- Maglogiannis, I., Zafiropoulos, E., & Anagnostopoulos (2009), An intelligent system for automated breast cancer diagnosis andprognosis using SVM based classifiers, *Applied intelligence journal, Volume 30, Issue1*, February 2009.
- Mahmood, M., Imran, M., Satuluri, N., Kuppa, M. R., & Rajesh, V. (2011). An improved CART decision tree for datasets with irrelevant features. In *Proceedings of the International Conference on Swarm, Evolutionary, and Memetic Computing* (pp. 539-549).
- Mandeep R, M., Chandorkar, P., Dsouza, A., & Kazi, N. (2015). Breast cancer diagnosis and recurrence prediction using machine learning techniques. *International Journal of Research in Engineering and Technology*.
- McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., ... Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788), 89-94. https://doi.org/10.1038/s41586-019-1799-6.
- Ministry of Health – State of Palestine MHPS. (2021). Health Annual Report Palestine.
- Ong, M.-S. (2012). Automated identification of extreme-risk events in clinical incident reports. *Journal of the American Medical Informatics Association*.
- Pandya, R., & Pandya, J. (2015). C5.0 algorithm to improve decision tree with feature selection and reduced error pruning. *International Journal of Computer Applications*, 117(16), 18-21.
- Pedregosa, F., Varoquaux, G., Gramfort, A., & others. (2023). Scikit-learn: *Machine learning in Python*. Journal of Machine Learning Research, 24, 1-9. https://doi.org/10.5555/3548367.3548368
- Qaiser, T., & Bhatti, S. H. (2019). Machine learning approaches for breast cancer classification. *Expert Systems with Applications*.
- Quinlan, J. R. (1986). *Induction of Decision Trees*. Machine Learning, 1(1), 81-106
- Krishna R, K., T M, R., Gopal M. G., N., & G, K. (2023). Breast Cancer Classification Using Machine Learning. International Research Journal on Advanced Science Hub, 5(Issue 05S), 88–93. https://doi.org/10.47392/irjash.2023.S012
- Rokach, L., & Maimon, O. (2008). Data mining with decision trees: Theory and applications. World Scientific Publishing Co.
- Rui, T., Tianyi, W., Yifan, X., Hongji, S., & Toe, T. T. (2023). Breast image classification based on ResNet and Random Forest multilayer classifier model. 2023 5th International Conference on Bio-Engineering for Smart Technologies (BioSMART), 1–6.
- Saravanakumar, M., & Kannan, Dr. S. (2023). Pattern Recognition in Breast Cancer Using Machine Learning. INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT, 07(03). https://doi.org/10.55041/IJSREM18255
- Shafique, R., Rustam, F., Choi, G. S., Díez, I. D. L. T., Mahmood, A., Lipari, V., Velasco, C. L. R., & Ashraf, I. (2023). Breast Cancer Prediction Using Fine Needle Aspiration Features and Upsampling with Supervised Machine Learning. Cancers, 15(3), 681. https://doi.org/10.3390/cancers15030681

- Sharma, H., & Kumar, S. (2016). A survey on decision tree algorithms of classification in data mining. *International Journal of Science and Research*, 5(4), 2094-2097.
- Sheth, D., & Giger, M. L. (2019). Artificial intelligence in the interpretation of breast cancer on MRI. *Journal of Magnetic Resonance Imaging*. https://doi.org/10.1002/jmri.26878
- Singh, A. K. (2023). Breast Cancer Classification Using ML on WDBC. In K. Kumar Singh, M. K. Bajpai, & A. Sheikh Akbari (Eds.), Machine Vision and Augmented Intelligence (Vol. 1007, pp. 609–619). Springer Nature Singapore. https://doi.org/10.1007/978-981-99-0189-0_48
- Song, Y. Y., & Ying, L. (2015). Decision tree methods: Applications for classification and prediction. *Shanghai Archives of Psychiatry*, 27(2), 130.
- Sugimoto, M., Hikichi, S., Takada, M., & Toi, M. (2021). Machine learning techniques for breast cancer diagnosis and treatment: A narrative review. *Annals of Breast Surgery, 7*. https://abs.amegroups.org/article/view/7085
- Tarawneh, O., Otair, M., Husni, M., Abuaddous, Hayfa. Y., Tarawneh, M., & Almomani, M. A. (2022). Breast Cancer Classification using Decision Tree Algorithms. International Journal of Advanced Computer Science and Applications, 13(4). https://doi.org/10.14569/IJACSA.2022.0130478
- Taznim, S. A., & Ferdous, S. M. (2018). Integrating big data and machine learning techniques for cancer risk prediction. *International Conference on Bangla Speech and Language Processing*.
- Tran, H. (2019). A survey of machine learning and data mining techniques used in multimedia systems.
- Varsha, B., Sneka, P., Tanuja, A., & Shana, J. (2023). Classification Models for Breast Cancer Detection. In A. Chitra, V. Indragandhi, & W. Razia Sultana (Eds.), Intelligent and Soft Computing Systems for Green Energy (1st ed., pp. 255–264). Wiley. https://doi.org/10.1002/9781394167524.ch19
- Wankhade, Y., Toutam, S., Thakre, K., Kalbande, K., & Thakre, P. (2023). Machine learning approach for breast cancer prediction: A review. In *2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)* (pp. 566-570). https://doi.org/10.1109/ICAAIC56838.2023.10141164
- Wei, Y., Zhang, D., Gao, M., Tian, Y., He, Y., Huang, B., & Zheng, C. (2023). Breast cancer prediction based on machine learning. *Journal of Software Engineering and Applications*, 16, 348-360. https://doi.org/10.4236/jsea.2023.168018
- World Health Organization. WHO (2024). Cancer. Retrieved from https://www.who.int/
- Yue, W., Wang, Z., Chen, H., & Payne, A. M. (2018). Machine learning with applications in breast cancer diagnosis and prognosis. *Designs*, 2(2), 13. https://doi.org/10.3390/designs 2020013
- Zeng, C. (2022). An Application of Generalized Linear Models to Fine Needle Aspiration in Breast Cancer. Highlights in Science, Engineering and Technology, 8, 178–184. https://doi.org/10.54097/hset.v8i.1125.