JPPIPA 12(1) (2026)

Jurnal Penelitian Pendidikan IPA

http://jppipa.unram.ac.id/index.php/jppipa/index

Multimethodology Analysis of Determinants of Breast Cancer Diagnosis Machine Learning

Dita Anggriani Lubis[1], Yuli Irnawati[2], Ayu Trisni Pamilih[2], Ria Fazelita Br Gultom[1]

[1] DIII Kebidanan Fakultas Kesehatan, Universitas Satya Terra Bhinneka, Indonesia
[2] STIkes Bakti Utama Pati, Indonesia, Indonesia

Received: August 12, 2025
Revised: December 29, 2025
Accepted: January 25, 2026
Published: January 31, 2026

Corresponding Author:
Dita Anggriani Lubis
ditalubis@satyaterrabhinneka.ac.id

DOI: 10.29303/jppipa.v12i1.12497

Abstract: Breast cancer remains one of the most prevalent and life-threatening diseases worldwide, highlighting the urgent need for accurate and interpretable diagnostic models. While machine learning has shown promise in classification tasks, many existing models lack transparency and overlook the individual contribution of cellular features essential for clinical decision-making.This study proposes an integrative and explainable framework to identify the most influential cellular-level features in distinguishing between benign and malignant breast tumors. Using a publicly available dataset comprising 569 observations and 32 numerical features, we conducted a multi-step analysis. Feature relevance was first evaluated using Pearson correlation. Random Forest and Recursive Feature Elimination (RFE) were employed to rank and refine the feature subset, followed by Principal Component Analysis (PCA) for dimensionality reduction and pattern visualization. SHapley Additive exPlanations (SHAP) were utilized to interpret individual predictions. Complementary statistical tests, including t-tests and chi-square analyses, assessed associations between tumor characteristics and diagnosis. A logistic regression model was developed to evaluate predictive performance.Key cellular features—such as mean radius, texture, and concavity—were consistently identified as highly predictive of diagnosis. RFE demonstrated that fewer than 10 features were sufficient for optimal classification. The logistic regression model achieved high accuracy, offering a simpler yet effective alternative for prediction.By combining statistical methods with interpretable machine learning, this study presents a transparent and clinically relevant approach to breast cancer diagnosis. The integration of SHAP values bridges the gap between model performance and interpretability, supporting more informed and personalized clinical decisions. Future work should consider external validation, image-based features, and patient demographic variables to enhance generalizability.

Keywords: Breast cancer; Feature selection; Interpretable machine learning; SHAP

Introduction

The research was motivated by the need to better understand the characteristics of cells that contribute to the diagnosis of breast cancer, both benign and malignant. In this endeavor, an initial search was conducted through Pearson correlation analysis to determine which cell numerical features had a significant relationship with the diagnosis results (Yang et al., 2022). In order to determine which features are most influential in distinguishing between benign and malignant cancers, a Random Forest approach was used to assess the importance of each feature in improving classification accuracy. However, given the large number of features, it is necessary to determine the optimal number of features required to achieve the best performance in detecting cancer types. Therefore, the Recursive Feature Elimination (RFE) method with cross-validation is applied to obtain the most efficient combination of features. In order to observe the possibility of visual separation between the two diagnosis groups in a lower dimension, Principal Component Analysis (PCA) was conducted to map the data in a two-dimensional. To extend the understanding

How to Cite:
Lubis, D. A., Irnawati, Y., Pamilih, A. T., & Gultom, R. F. B. (2026). Multimethodology Analysis of Determinants of Breast Cancer Diagnosis Machine Learning. *Jurnal Penelitian Pendidikan IPA*, *12*(1), 949–969. https://doi.org/10.29303/jppipa.v12i1.12497

of the contribution of each feature in diagnosis prediction at the individual level, the SHAP (SHapley Additive exPlanations) method was used, which allows local interpretation of the predictive model results (Casas et al., 2022). On the other hand, to evaluate whether there is a significant difference between the average tumor radius size based on the diagnosis classification, a t-test was conducted and visualized in the form of a density plot. Moreover, this study also evaluated the performance of the logistic regression model in predicting the probability of malignant cancer based on key features, including accuracy and prediction distribution analysis. As for the relationship between the tumor area size category and the type of patient diagnosis, a chi-square test was conducted, Cramer's V was calculated as a measure of the strength of association, and visualized using a mosaic diagram.

The purpose of the research is to identify cellular characteristics or features that have a high correlation with the diagnosis of breast cancer, both benign and malignant, through the Pearson correlation analysis approach. The research also determines the features that are most influential in distinguishing these cancer types using the Random Forest method, and identifies the optimal number of features needed to achieve the best classification accuracy through the Recursive Feature Elimination (RFE) technique. The research analyzed the possibility of visually separating diagnoses in a low-dimensional space using Principal Component Analysis (PCA), while mapping the distribution pattern of diagnoses in two-dimensional form. In addition, the research attempts to explain the contribution of each feature to the diagnosis prediction results at the individual level using the SHAP (SHapley Additive exPlanations) method, which allows for a more transparent interpretation of the model predictions. The study also statistically tested whether there is a significant difference between the mean tumor radius size based on the diagnosis classification through t-test and distribution visualization. An evaluation of the performance of the logistic regression model in predicting the probability of malignant cancer diagnosis was also conducted to assess the effectiveness of a simpler predictive approach. The study examined the association between tumor area size categories and patient diagnosis outcomes using the chi-square test and strength of association analysis through Cramer's V calculation, and visualized it in the form of a mosaic plot.

The benefit of the research is to provide a deeper understanding of cellular features relevant in breast cancer diagnosis, so as to assist in the process of identifying the main characteristics that distinguish between benign and malignant cancer types. The research is expected to provide a strong basis for optimal feature selection to improve the efficiency and accuracy of cancer classification systems. This research will utilize a machine learning approach, the results of this research have the potential to support the development of a more reliable and data-driven cancer early detection system. In addition, through the application of SHAP analysis, this research offers a transparent interpretation of the prediction model's decisions, which is particularly important in the context of clinical practice. The resulting understanding of the statistical analysis of tumor patterns and characteristics can also support more informed medical decision-making. On the other hand, the study can serve as a scientific reference in the development of predictive and diagnostic models for other types of cancer or diseases with similar approaches. More broadly, this research also encourages the integration of clinical data with modern analytical techniques as part of efforts to improve the quality of services and health systems (Soong et al., 2018).

This research has several limitations that need to be considered. The data used is limited to a dataset with 569 observations and 32 numerical features, so the analysis results are highly dependent on the scope and quality of the data. The research focuses only on numerical features without involving categorical variables or medical images such as mammograms. The diagnosis analyzed is binary, which only includes benign and malignant cancers, without considering further cancer subtypes or stages. This study did not address clinical aspects such as treatment decisions or prognosis in depth. The methods used are limited to basic statistical approaches and conventional machine learning, such as correlation analysis, Random Forest, RFE, PCA, SHAP, logistic regression, and simple statistical tests, without covering advanced techniques such as deep learning. Model validation is done internally through cross-validation and does not use external datasets for further testing. This study also did not consider patient sociodemographic factors such as age or family history, as they were not available in the dataset (Nougaret et al., 2020).

In recent years, the application of machine learning in breast cancer diagnosis and prediction has made significant progress. Studies such as and have demonstrated how a multimodal machine learning-based approach can improve the accuracy of early detection and the effectiveness of clinical decision-making. Methods such as deep learning, ensemble models, and integration of clinical and radiological data further strengthen the role of artificial intelligence in delivering precise and individualized prognosis predictions (Soto et al., 2023). Research even designed a multiprocess classification system to automatically detect breast lesions with a multi-input learning approach, while focused on optimizing long-term metastasis prediction through EHR-based deep learning. However, most of these studies focus on long-term survival prediction, neoadjuvant therapy effectiveness, or large-scale data fusion from hospitals and medical imaging. An important research gap is the lack of emphasis on understanding the individual contribution of cell microscopic features to cancer diagnosis (benign vs. malignant) in an interpretive and computational way in explainable machine learning. Many previous studies have used complex black-box models, making direct interpretation of the influence of single features at the individual patient level difficult (Dykens et al., 2023).

This research fills this gap by taking a quantitative approach that combines statistical correlation, machine learning, and modern interpretability techniques such as SHAP (SHapley Additive exPlanations). This research not only aims to identify the numerical features of cells that have a high correlation to cancer diagnosis, but also evaluates the importance of each feature through Random Forest and determines the optimal number of features needed in diagnosis classification efficiently using Recursive Feature Elimination (RFE). Furthermore, by utilizing Principal Component Analysis (PCA), this research explores whether there is a pattern of separation of diagnoses in a low-dimensional space, and examines the statistical and probabilistic relationship between features such as tumor size and diagnosis using t-test and chi-square tes (Pieters et al., 2021). The novelty of this study lies in the application of an integrated multimethod approach in a layered and interpretive manner for breast cancer diagnosis classification, which is rarely done simultaneously in a single study. In particular, the integration of the SHAP method as a means of visualizing the local contribution of features to patient diagnosis allows clinicians and researchers to gain in-depth insights that are not only predictive but also explainable, thereby supporting more transparent and targeted clinical decisions (Pacal, 2024).

Method

Dataset

Breast cancer is the most common type of cancer found in women worldwide. It is estimated that breast cancer accounts for about 25% of all cancer cases, and in 2015, more than 2.1 million people were affected globally. The disease begins when cells in the breast tissue begin to grow uncontrollably, forming a tumor that can be detected through X-ray imaging or felt as a lump in the breast area. An accurate and prompt diagnosis is crucial in determining the next medical step, be it surgery, chemotherapy, or hormonal therapy (Davidović et al., 2024).

The dataset used in the development of a machine learning-based medical classification system for detecting breast cancer is Breast Cancer Wisconsin (Diagnostic) (Gravitt et al., 2021). The dataset was contributed on October 31, 1995 and is publicly available through the UCI Machine Learning Repository. The dataset is derived from digitized images of fine needle aspiration (FNA) of breast tissue masses, where the collected features describe the characteristics of the cell nuclei contained in the image (Ampofo et al., 2023).

The dataset consists of 569 observations and 30 numerical features that are all real-valued or continuous. The variables include information such as the average size of the cell core radius, texture, perimeter, area, surface roughness, density, basin severity, number of basin points, symmetry, as well as fractal dimension. Each feature was calculated based on three statistical measurements, namely the mean (average), standard error, and worst (maximum or extreme value) values for each feature, resulting in a total of 30 features calculated from the 10 main cell nucleus characteristics (Obol et al., 2021).

Computation Models

Pearson Correlation

In statistics, Pearson correlation is one of the most common methods used to measure the extent to which two numerical variables have a linear relationship. Correlation is denoted by the symbol (r), and is mathematically defined as (Al Mudawi & Alazeb, 2022):

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \tag{1}$$

In this formula, (X_i) and (Y_i) represent the i-th value of variable (X) and (Y), respectively, while ($\overline{X}$) and ($\overline{Y}$) is the average value of each of these variables. The index (i) runs from 1 to (n), which is the total number of observations in the dataset. The numerator of the formula represents the covariance between (X) and (Y), while the denominator is the product of the standard deviation of each variable.

The Pearson correlation value is always in the range between (-1) and (+1). If the (r) value is close to +1, then this implies a strong positive linear relationship between the two variables, meaning that when the value of (X) increases, the value of (Y) tends to also increase proportionally. Conversely, if the (r) value is close to -1, then the relationship detected is perfectly negative, i.e. an increase in the value of (X) is followed by a decrease in the value of (Y). If the correlation value is close to 0, then no linear relationship pattern can be detected between the two, although there could be a non-linear relationship (Lilhore et al., 2022).

In the analysis of data that has many numerical variables-in the breast cancer dataset consisting of 30 numerical features-the use of Pearson correlation is extended in the form of a correlation matrix (Liu et al., 2022). This matrix looks simultaneously at the relationship between all pairs of features. In a dataset (D) consisting of (p) numerical variables, the Pearson correlation matrix that can be formed from this data is a symmetric matrix of size $(p \times p)$, whose elements consist of the correlation values between each pair of variables (X_i) and (X_j). The matrix is written as:

$$\mathrm{R} = \begin{bmatrix} r_{X_1X_1} & r_{X_1X_2} & \cdots & r_{X_1X_p} \\ r_{X_2X_1} & r_{X_2X_2} & \cdots & r_{X_2X_p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{X_pX_1} & r_{X_pX_2} & \cdots & r_{X_pX_p} \end{bmatrix} \quad (2)$$

Each value (r_{ij}) in the above matrix are calculated using the Pearson correlation formula as described earlier, where (i) and (j) represent the variable indices. Due to the symmetrical nature of the correlation ($r_{ij} = r_{ji}$), simply calculate the value in the lower or upper triangle of the matrix.

Before calculating the correlation matrix, first separate the non-numeric variables from the dataset (patient ID column or diagnosis label), as they are not relevant for numerical correlation analysis. The process continues by calculating the Pearson correlation value for each pair of numerical variables, and the results are then placed in the appropriate position in the matrix (R).

Random Forest as a Classification Model

To create a Random Forest with (B) trees involves creating (B) bootstrap samples from dataset (D). Bootstrap sample ($D^{(b)}$) randomly sampled with replacement, of size (n) (the same number of observations as the original dataset)

$$D^{(b)} = \left(x_1^{(b)}, y_1^{(b)}\right), \left(x_2^{(b)}, y_2^{(b)}\right), \dots, \left(x_n^{(b)}, y_n^{(b)}\right), where\ b = 1,2, \dots, B \quad (3)$$

Due to the presence of replacements, some observations may appear more than once in a single observation ($D^{(b)}$), while some data may not appear at all (out-of-bag / OOB samples). At each node, a randomized number of features ($m \leq p$) of the total (p) features. The goal is to reduce the correlation between trees, making the model more stable. From the selected feature subset ($X_{j_1}, X_{j_2}, \dots, X_{j_m}$), Features that minimize impurity after splitting are selected. As for a node (t), with the proportion of class data ($c \in \{0,1\}$) is (p_c), so:

$$G(t) = 1 - \sum_{c=0}^{1} p_c^2 = 1 - (p_0^2 + p_1^2) \quad (4)$$

The split objective minimizes the weighted average Gini of the left and right nodes after splitting:

$$G_{\text{split}} = \frac{n_L}{n_t} G(t_L) + \frac{n_R}{n_t} G(t_R) \quad (5)$$

Where (n_L), (n_R) is the sum of the data on the left and right branches, and ($n_t = n_L + n_R$). Features and split points that minimize (G_{split}). The tree grows until All data in the nodes have the same label, or The number of data in the nodes < minimum samples split, or The depth of the tree reaches a certain limit (for pruning or overfitting restrictions) (Ji et al., 2023). As for After all the trees ($T^{(b)}$) is completed, predictions can be made on the new data (x). Each tree gives a prediction label:

$$h_b(x) \in \{0,1\} \quad (6)$$

The final Random Forest prediction is the majority vote of all trees:

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), \dots, h_B(x) \quad (7)$$

Explicitly:

$$\hat{y} = \begin{cases} 1 \text{ if } \sum_{b=1}^{B} h_b(x) > \frac{B}{2} \\ 0 \text{ if } \sum_{b=1}^{B} h_b(x) \leq \frac{B}{2} \end{cases} \quad (8)$$

If the prediction is probabilistic, the average can be taken as the probability of class 1:

$$P(y = 1|x) = \frac{1}{B} \sum_{b=1}^{B} h_b(x) \quad (9)$$

Mean Decrease Accuracy (MDA) is used to measure the importance of each feature in the Random Forest model, by looking at the decrease in model accuracy when the feature value is permuted. If the randomness of a feature causes the accuracy to drop drastically, then it is considered important (Ford et al., 2021). Dataset consisting of (n) observations:

$$D = \{(x_i, y_i)\}_{i=1}^{n}, x_i = (x_{i1}, x_{i2}, \dots, x_{ip}) \in \mathbb{R}^p, y_i \in \{1,2, \dots, K\} \quad (10)$$

Random Forest can be built with (B) trees:

$$\text{RF} = T^{(1)}, T^{(2)}, \dots, T^{(B)} \quad (11)$$

For each tree ($T^{(b)}$), there is an Out-Of-Bag (OOB) data set, which is data that is not used in bootstrap training. For each tree ($T^{(b)}$), can evaluate the prediction on the OOB data. The (O_b) is the index of the OOB data for the (b)th tree. Then, the OOB accuracy before permutation:

$$\text{Acc}_{\text{original}}^{(b)} = \frac{1}{|\mathcal{O}_b|}\sum_{i\in\mathcal{O}_b} \mathbf{1}\left[T^{(b)}(x_i) = y_i\right] \tag{12}$$

Where ($\mathbf{1}[\cdot]$) is an indicator function: 1 if true, 0 if false, and $T^{(b)}(x_i)$ is the prediction of the (b)-th tree against the input (x_i). As for each feature (X$_j$), random permutation of feature value (j) in OOB data can be done:

$$x_i^{\text{perm}(j)} = (x_{i1},\ldots,x_{i,j-1},x_{j,\pi(i)},x_{i,j+1},\ldots,x_{ip}) \tag{13}$$

Where $\pi(i)$ is a random permutation function of the index ($\mathcal{O}_b$). Then, an accuracy evaluation is performed on the permutation data:

$$\text{Acc}_{j,\text{perm}}^{(b)} = \frac{1}{|\mathcal{O}_b|}\sum_{i\in\mathcal{O}_b} \mathbf{1}\left[T^{(b)}(x_i^{\text{perm}(j)}) = y_i\right] \tag{14}$$

For feature (X$_j$), the decrease in accuracy in tree-(b):

$$\Delta\text{Acc}_j^{(b)} = \text{Acc}_{\text{original}}^{(b)} - \text{Acc}_{j,\text{perm}}^{(b)} \tag{15}$$

Then, take the average of all trees:

$$\text{MDA}(X_j) = \frac{1}{B}\sum_{b=1}^{B}(\text{Acc}_{\text{original}}^{(b)} - \text{Acc}_{j,\text{perm}}^{(b)}) \tag{16}$$

As for if ($\text{MDA}(X_j) \gg 0$), then the feature (X$_j$) is very important, as messing with it makes the prediction drop sharply. If ($\text{MDA}(X_j) \approx 0$), then the feature (X$_j$) did not contribute much. As for if ($\text{MDA}(X_j) < 0$), means the feature may be misleading, and the model works better without it (Spencer et al., 2021).

Feature Selection with Recursive Feature Elimination (RFE)

RFE aims to select the best subset of features that provide the highest model accuracy, by iteratively removing less important features. It can be used to avoid overfitting, improve generalization, and simplify the model. Recursive Feature Elimination (RFE) works by gradually eliminating features based on their importance. The process starts by using all available features, then a Random Forest-like model is trained to calculate the importance score of each feature. Features with the lowest scores are removed, and this step is repeated until the desired number of features remain. At each stage, the performance of the model is evaluated, so that we can determine the combination of features that provides the best accuracy. The dataset (D) is divided into 10 subsets (called folds) ($D_1, D_2, \ldots, D_{10}$), such that:

$$D = \bigcup_{k=1}^{10} D_k \text{ and } D_i \cap D_j = \emptyset \text{ for } i \neq j \tag{17}$$

Each fold (D$_k$) is approximately the size of ($n/10$). As for each iteration to ($i \in \{1,2,\ldots,10\}$), can be done training data (training set):

$$D_{\text{train}}^{(i)} = D \setminus D_i = \bigcup_{\substack{k=1\\k\neq i}}^{10} D_k \tag{18}$$

The amount of training data is by:

$$n_{\text{train}} = n - |D_i| \tag{19}$$

On validation data (test set) with ($D_{\text{val}}^{(i)} = D_i$). The amount of validation data is ($n_{\text{val}} = |D_i|$). Furthermore, it can train the model ($M^{(i)}$) at ($D_{\text{train}}^{(i)}$), and evaluation on ($D_{\text{val}}^{(i)}$). The accuracy can be obtained:

$$A_k^{(i)} = \frac{1}{|D_i|}\sum_{(x_j,y_j)\in D_i} \mathbb{1}\{\hat{y}_j^{(i)} = y_j\} \tag{20}$$

Where ($\hat{y}_j^{(i)} = M^{(i)}(x_j)$) is the model prediction, ($\mathbb{1}\{\cdot\}$) is the indicator function (1 if true, 0 if false), and (k) is the number of features used in the model (e.g. from the RFE of the kth stage). After completing the 10 iterations, the average accuracy of the model with the number of features (k) can be calculated (Pramanik et al., 2022):

$$\overline{A}_k = \frac{1}{10}\sum_{i=1}^{10} A_k^{(i)} \tag{21}$$

From the equation above, ($\overline{A}_k$) is the average accuracy of the model when using (k) features, and $A_k^{(i)}$ is at the (i)th iteration (fold) when using (k) features. As for the above for various number of features ($k \in \{1,2,\ldots,p\}$), then the optimal number of features (k) is chosen based on:

$$k^* = \arg \max_k \overline{A}_k \tag{22}$$

As for how stable the model is at each (k), it can be calculated using the equation:

$$\text{Var}(A_k) = \frac{1}{9}\sum_{i=1}^{10}(A_k^{(i)} - \overline{A}_k)^2 \tag{23}$$

As before Random Forest (RF) is a bagging-based ensembling algorithm that combines (T) independent decision trees:

$$\mathcal{F} = \{h_1, h_2, \ldots, h_T\} \tag{24}$$

Note that for each tree (h_t) is trained on a random subset of training data (with replacement) and a random subset of features. It is understood that Importance Score with Permutation aims to measure how important a feature (xj) is to the model prediction. One important metric is Mean Decrease Accuracy (MDA). For each feature (x_j) (Heidari Sarvestani et al., 2021), can be done in stages; take a validation dataset (usually Out-of-Bag samples, OOB), measure the accuracy of the original ($\text{Accuracy}_{\text{original},t}$) from a tree ($h_t$), randomizes the values of the features (x_j) which is the shape of the new dataset ($X^{(j)}_{\text{shuffled}}$) , and calculate the new accuracy ($\text{Accuracy}^{(j)}_{\text{shuffled},t}$). The difference will show how much accuracy is lost when information from the (x_j) removed which means the bigger, the more important. Mean Decrease Accuracy for the jth feature:

$$\text{MDA}_j = \frac{1}{T}\sum_{t=1}^{T}(\text{Accuracy}_{\text{original},t} - \text{Accuracy}^{(j)}_{\text{shuffled},t}) \tag{25}$$

As for the above equation, it is known that (T) is the number of trees in the Random Forest, (j) is the index of the jth feature, and ($\text{Accuracy}^{(j)}_{\text{shuffled},t}$) is the accuracy after feature (j) is randomized for the (t)-th tree. If (MDA_j) close to zero or negative which are features that do not contribute much and can be eliminated by RFE. After performing Recursive Feature Elimination (RFE) of all features (p), the average accuracy for each feature subset (k) can be measured.

Principal Component Analysis (PCA)

PCA aims to find a linear combination of the original features so that the maximum variance of the data is explained by the new components (principal components), reduce the dimensionality while retaining as much information as possible, and transform the database towards the eigenvectors of the standardized feature covariance matrix. As can be memorized by having data ($X \in \mathbb{R}^{n \times p}$), where (n) is the number of observations (rows), and (p) is the number of features (columns) (Gravitt et al., 2021). The Standardization is as follows:

$$Z = \frac{X - \overline{X}}{s} \tag{26}$$

where from the above equation, ($\overline{X}$) is the column mean (feature), and (s) is the standard deviation of each column. The covariance matrix of the normalized data can also be calculated:

$$S = \frac{1}{n-1} Z^T Z \tag{27}$$

The Matrix of ($S \in \mathbb{R}^{p \times p}$) stores the covariance between features. In Eigen Decomposition, the eigenvalue can be found (λ) than eigenvector ($\mathbf{v}$) of the covariance matrix:

$$S\mathbf{v}_i = \lambda_i \mathbf{v}_i \tag{28}$$

From the equation above, ($\mathbf{v}_i$) is the (i)th eigenvector, referred to as the (i)th principal component, (λ_i) is the eigenvalue of (i), expressing the amount of variance explained by component (i). This can be sorted:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \tag{29}$$

From this, we can reduce the dimension, we can only take (k) main components (k = 2), namely PC1 and PC2). Project the data to the new space:

$$Z_{\text{PCA}} = Z \cdot V_k \tag{30}$$

From ($V_k \in \mathbb{R}^{p \times k}$) is a matrix consisting of eigenvectors ($\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$) , dan ($Z_{\text{PCA}} \in \mathbb{R}^{n \times k}$) is the result of data projection.

1.1.1 SHAP (Shapley Additive Explanations)

SHAP expresses the local prediction of an ML model for an observation ($x \in \mathbb{R}^M$) in additive form:

$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i \tag{31}$$

Based on the equation above, f(x) is the model output for input (x). (ϕ_0) is the global expectation value of the model that is the baseline prediction. (ϕ_i) is the Contribution of the i-th feature, calculated by the Shapley Value method. The Model Prediction f(x) is the result of model prediction for an observation ($x = (x_1, x_2, \ldots, x_M)$. Expected value of the model output when there is no feature information (Zhuang & Guan, 2022):

$$\phi_0 = \mathbb{E}[f(x)] \tag{32}$$

Usually calculated as the average of the model output over all data:

$$\phi_0 = \frac{1}{N}\sum_{j=1}^{N} f(x^{(j)}) \tag{33}$$

Based on the equation above which is a prediction if we know nothing about the input. The average marginal contribution of the i-th feature to the prediction f(x), calculated as:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(M - |S| - 1)!}{M!} \cdot [f_{S \cup \{i\}}(x) - f_S(x)] \quad (34)$$

In SHAP, suppose ($F = \{1,2,\dots,M\}$) is the set of all features in the model. To calculate the contribution of the i-th feature, we consider all the subsets of ($S \subseteq F \setminus \{i\}$), i.e. all combinations of features that do not include feature (i). The model prediction used on subset (S), written as $f_S(x)$, represents the model output when only the information of the features in (S) is known, while other features (including (i)) are considered unknown and represented by their expected values or marginalization results. The Weight equation:

$$w_S = \frac{|S|!\,(M - |S| - 1)!}{M!} \quad (35)$$

Based on the above equation, which is the Shapley coefficient, guarantees fairness for all sequences. The Shapley value feature satisfies four key axioms of game theory. If two features contribute equally in all subsets, then ($\phi_i = \phi_j$) . If feature (i) does not change the prediction output in any subset, then ($\phi_i = 0$). As for the two models (f) and (g), SHAP for (f + g) is:

$$\phi_i^{f+g} = \phi_i^f + \phi_i^g \quad (36)$$

Parametric Statistical Tests

In the analysis, parametric statistical testing was performed using the independent two-sample t-test to evaluate whether there was a statistically significant difference in the mean values between the two breast cancer diagnosis groups, namely the Benign and Malignant groups.

The main purpose of the test is to determine whether the differences in mean values are merely coincidental or do indeed represent real differences in the wider population. Mathematically, the null hypothesis (H_0) being tested states that the mean radius mean of the two groups is the same, ($\mu_1 = \mu_2$). In other words, there was no mean difference between the groups with benign and malignant diagnoses. Conversely, the alternative hypothesis (H_1) states that there is a difference between the two means, ($\mu_1 \neq \mu_2$) , which indicates that the mean radius value can be used to significantly distinguish between benign and malignant tumors (Pacal, 2024).

In the analysis, the main objective was to test whether there was a significant difference in the mean radius values between the two breast cancer diagnosis groups, namely the benign-diagnosed group and the malignant-diagnosed group. This test is important because the size of the cancer cell radius can interpret the morphological characteristics that statistically distinguish the two types of diagnosis.

The first step was to calculate the descriptive statistics of each group. Suppose ($\overline{x}_1$) is the average score for the Benign group, and ($\overline{x}_2$) is the average for the Malignant group. Suppose also that (s_1^2) as the variance of the Benign group and (s_2^2) as the variance of the Malignant group, with respective sample sizes of (n_1) and (n_2).

To test the hypothesis of the difference between two means in samples that are independent and have variances that are not assumed to be equal, Welch's t-test method is used (Kumawat et al., 2023). The test statistic (t) of the method is defined by the formula:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad (37)$$

Mathematically, the numerator of the formula represents the difference between the means of the two groups, while the denominator is the root of the sum of the relative variances (i.e. variance divided by sample size) of each group (Chisale Mabotja et al., 2021). This takes into account the inequality in the distribution of data between groups. However, due to the assumption that the variances of the two groups are not equal, the degrees of freedom (df) in the t-distribution are not calculated using the classic formula ($n_1 + n_2 - 2$), but with the Welch-Satterthwaite formula, as follows

$$df = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{(\frac{s_1^2}{n_1})^2}{n_1 - 1} + \frac{(\frac{s_2^2}{n_2})^2}{n_2 - 1}} \quad (38)$$

The above formula is used to calculate the degrees of freedom more accurately under different variance conditions, with the numerator being the square of the denominator in the t statistical formula, and the denominator being the sum of the squared errors of each group that have been normalized to the sample size minus one.

Once the (t) and (df) values are obtained, the test statistic can be compared with the t-distribution for those degrees of freedom to obtain the p-value. This p-value is the basis for making statistical decisions. If the value of ($p < \alpha$) (As for the significance level ($\alpha = 0{,}05$)), then it can be interpreted that the difference in mean radius between the Benign and Malignant groups is statistically significant, so the null hypothesis (which states that there is no mean difference) is rejected. Conversely, if ($p \geq \alpha$), then there is not enough evidence to conclude a difference, so we fail to reject the null hypothesis (Yu et al., 2022).

Logistic Regression

The model is used to predict the probability of a binary dependent variable ($Y \in \{0,1\}$). The predictor variables used are:

$$X = (X_1, X_2, \dots, X_k)^T \tag{39}$$

Modelable:

$$P(Y = 1 \mid X) = p \ and \ P(Y = 0 \mid X) = 1 - p \tag{40}$$

Instead of directly modeling the probability (p), the log-odds can be modeled:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) \tag{41}$$

The logistic model assumes that the logit is linearly related to the predictor variables:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \tag{42}$$

$$\log\left(\frac{p}{1-p}\right) = X^T \beta \tag{43}$$

As with $\boldsymbol{X}^T = (1, X_1, X_2, \dots, X_k)$, and $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \dots, \beta_k)$. Can be done inverting the Logit Function. The goal is to express (p) explicitly:

$$\log\left(\frac{p}{1-p}\right) = \eta \Rightarrow \frac{p}{1-p} = e^{\eta} \tag{44}$$

From this can be multiplied both segments by (1 - p):

$$p = e^{\eta}(1-p) \Rightarrow p + pe^{\eta} = e^{\eta} \Rightarrow p(1+e^{\eta}) = e^{\eta} \Rightarrow p = \frac{e^{\eta}}{1+e^{\eta}} \tag{45}$$

$$p = \frac{1}{1+e^{-\eta}} = \frac{1}{1+e^{-(\beta_0+\beta_1X_1+\cdots+\beta_kX_k)}} \tag{46}$$

The above function is called a sigmoid (logistic) function, because the Value of ($p \in (0,1)$) for ($\eta \to \infty$), where ($p \to 1$) for ($\eta \to -\infty$), and ($p \to 0$). This function is smooth, differentiable, and suitable for binary probability models. The sigmoid function (Johnson et al., 2025):

$$f(\eta) = \frac{1}{1+e^{-\eta}} \tag{47}$$

Where the above equation forms an S-curve symmetrical to the vertical axis at ($\eta = 0$), if f(0) = 0.5, given $f(\eta) \in (0,1)$. As for observation (i), the variable ($Y_i \in \{0,1\}$) follows the Bernoulli distribution:

$$P(Y_i = y_i \mid X_i) = p_i^{y_i}(1-p_i)^{1-y_i} \tag{48}$$

So the likelihood of all data (independent):

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \tag{49}$$

Take log:

$$\ell(\beta) = \sum_{i=1}^{n} [y_i \log(p_i) + (1-y_i)\log(1 - p_i)] \tag{50}$$

As for the parameter estimation (β) obtained with this log-likelihood maximization. The logistic regression model models:

$$P(Y = 1 \mid X) = \frac{1}{1+e^{-(\beta_0+\beta_1X_1+\cdots+\beta_kX_k)}} \tag{51}$$

Based on the above equation using the probabilistic approach of the logit function:

$$\log\left(\frac{p}{1-p}\right) = \boldsymbol{X}^T\boldsymbol{\beta} \tag{52}$$

This function ensures that the probability value stays between 0 and 1, and the coefficient of the (β_j) shows the direction and strength of the relationship (X_j) against the log-odds of the event (Y=1).

Chi-Square Test and Association Analysis

In this study, one of the numerical variables analyzed is the mean area, which is the average size of the tumor area detected in breast tissue. This variable is continuous, so it needs to be categorized first in order to analyze its association with the categorical variable of diagnosis (with Malignant and Benign categories. If there is a set of values of the variable `area mean` denoted as:

$$X = \{x_1, x_2, x_3, \dots, x_n\} \tag{53}$$

As for ($x_i \in \mathbb{R}$), is the average size value of the tumor area of the i-th individual, and (n) represents the total number of individuals or observations in the dataset. The median value of the set (X), which is denoted as ($\tilde{X}$), is the center value when all elements ($x_i \in X$) sorted from smallest to largest. Mathematically:

$$\tilde{X} = \text{Median}(X) \tag{54}$$

The median value is used as a dividing limit to group the size of the tumor area into two categories. Each value ($x_i \in X$) will be categorized into one of two classes, based on its comparison with the median value ($\tilde{X}$). The categorization rule is expressed as follows:

$$\text{Category}(x_i) = \begin{cases} \text{"Small", If } x_i \leq \tilde{X} \\ \text{"Big", If } x_i > \tilde{X} \end{cases} \qquad (55)$$

From the above, all mean area values in the dataset will be divided into two groups, namely Small for values smaller than or equal to the median, and Large for values larger than the median.

The purpose of clustering is to convert the numerical variable \`area mean\` into a two-class categorical variable (\`Small\` and \`Large\`), so that it can be compared directly with the categorical variable \`diagnosis\`. This allows the use of non-parametric statistical tests such as the Chi-Square Test and categorical association measures such as Cramér's V, which require the data to be in categorical form.

The Chi-Square test is a statistical method used to test whether there is dependence between two categorical variables. In the analysis, the main objective was to determine whether breast cancer diagnosis-which consists of two categories, malignant (M) and benign (B)-had a significant association with tumor area size category. Tumor area size was categorized into two groups, large and small, based on the median division of the numerical value of the mean area variable. In other words, the test aims to examine whether the proportion of diagnoses is significantly different between the large and small tumor groups, which may indicate an association between the degree of cancer malignancy and the size of the patient's tumor area. The frequency table of the number of cases based on the combination of the two categories (Chisale Mabotja et al., 2021).

Table 1. Number of cases based on the combination of the two categories

	Small (C_1)	Big (C_2)	Total
Benign (B)	O_{11}	O_{12}	R_1
Malignant (M)	O_{21}	O_{22}	R_2
Total	C_1	C_2	N

In analyzing the relationship between two categorical variables, namely diagnosis (with Malignant and Benign categories) and paint area (Large and Small categories based on the median of the area mean), the Chi-Square test is the appropriate statistical tool to use. The aim is to determine whether the two variables are interdependent or not, i.e. whether there is a statistically significant relationship between the type of diagnosis and the size of the tumor area.

The first step of this test is to form a 2×2 two-way contingency table. In the table above, each cell is filled by the actual observation frequency which is called as (O_{ij}), i.e. the number of cases belonging to row (i) and column (j). (O_{11}) is the number of patients with Benign diagnosis and Small area size. The total number in each row is denoted by (R_i), while the total number in each column is called (C_j). The total number of observations in this table is expressed as (N), and the relation that:

$$N = R_1 + R_2 = C_1 + C_2 \qquad (56)$$

After knowing the actual observation values, the next step is to calculate the expected frequencies, which are denoted by (E_{ij}). This value represents the number of cases that should appear in each cell if there is no relationship between diagnosis and area size, or in other words, if the two variables are independent. The formula used to calculate the expectation value is:

$$E_{ij} = \frac{R_i \cdot C_j}{N}, \qquad (57)$$

which is the product of the row total and column total, divided by the total number of observations. This formula ensures that expectations are built proportionally to the size of the data, without any assumption of relationships. If the value of (O_{ij}) and (E_{ij}), then the Chi-Square test statistic can be calculated, which is denoted by the (χ^2) symbol. The formula is:

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \qquad (58)$$

The above formula calculates the difference between the observation and expectation in each cell, squares it, and divides it back by the expectation value. All these results are then summed up to obtain one final Chi-Square value. The larger this value, the larger the difference between what is observed and what is expected-and the more likely it is that the diagnosis and area size variables are not independent (Campos et al., 2021).

However, the value of (χ^2) tidak dapat langsung diinterpretasikan tanpa mengacu pada distribusi Chi-Square theory. Therefore, it is necessary to determine the degrees of freedom, which are calculated by the formula:

$$df = (r-1)(c-1) \qquad (59)$$

Based on the above equation (r) is the number of rows and (c) is the number of columns of the contingency table. In this case, since the table consists of 2 rows (Malignant, Benign) and 2 columns (Small, Large), then:

$$df = (2-1)(2-1) = 1 \qquad (60)$$

After that, the value of (χ^2) obtained is compared with the critical value of the Chi-Square distribution with 1 degree of freedom at the level of significance ($\alpha = 0.05$). If the value of (χ^2) is greater than the critical value of the distribution, the null hypothesis (H0)-that there is no relationship between diagnosis and area size-is rejected. This rejection means that there is sufficient statistical evidence to conclude that there is a significant relationship between the type of cancer diagnosis and tumor area size (Zhang et al., 2022).

In inferential statistical analysis, especially when the Chi-Square Test is used to test the association between two categorical variables, it is important to not only know whether the association is statistically significant (through the p-value), but also how strong the association is. For this purpose, one of the most commonly used measures of association strength is Cramér's V. Cramér's V is a symmetric association measure used for two categorical variables, and has a value between 0 and 1, where a value of 0 indicates no association between two variables, and a value close to 1 means a very strong association. While Cramér's V is referred to as the Phi Coefficient (φ) in the case of 2 × 2 tables, Cramér's V applies generally to all sizes of contingency tables (including >2 categories). Mathematically, Cramér's V is defined as:

$$V = \sqrt{\frac{\chi^2}{N(k-1)}} \tag{61}$$

With the above information, (χ^2) is the statistical value of the Chi-Square test that has been calculated based on the contingency table, (N) is the total number of observations in the dataset, and (k) is the minimum number of categories of the two variables, namely:

$$k = min(r, c) \tag{62}$$

where (r) is the number of rows (categories of the first variable), (c) is the number of columns (categories of the second variable). In a 2 × 2 contingency table as is common in the case of cancer diagnosis (Benign vs Malignant; and Small vs Large tumor size), then ($k = \min(2,2) = 2$)

Result and Discussion

Based on the research that has been carried out, the data consists of 569 observations and 32 variables, preliminary analysis of the data structure and summary statistics shows that there are significant differences in characteristics between benign and malignant cancer cases based on morphological features of breast tissue cells. The diagnosis variable classifies each sample into "M" (malignant) and "B" (benign) categories, while the other variables represent the size, shape, texture, and surface roughness of the cell nucleus measured numerically (Young & Argáez, 2020).

The mean values of features such as mean radius, mean perimeter and mean area show larger sizes than the minimum and first quartile values, indicating that many samples have relatively large cell nuclei, which are usually associated with malignant cancer cases. The maximum value for the mean radius reached more than 28 mm, while the minimum value was only about 7 mm. Similarly, the mean area variable had a very wide range, from about 144 to 2501, meaning that there was a high variation between samples (Shoghi et al., 2019).

Characteristics such as concavity mean and compactness mean, which measure the complexity and density of the cell nucleus contour, have distributions that indicate that most cells with a malignant diagnosis tend to have higher values, implying a more irregular and compact cell shape. The worst values of each feature-such as radius worst, area worst, and concavity worst-which represent the most extreme sizes and shapes of the observed cells, are also very high, supporting the hypothesis that malignancy correlates with the extreme values of these features.

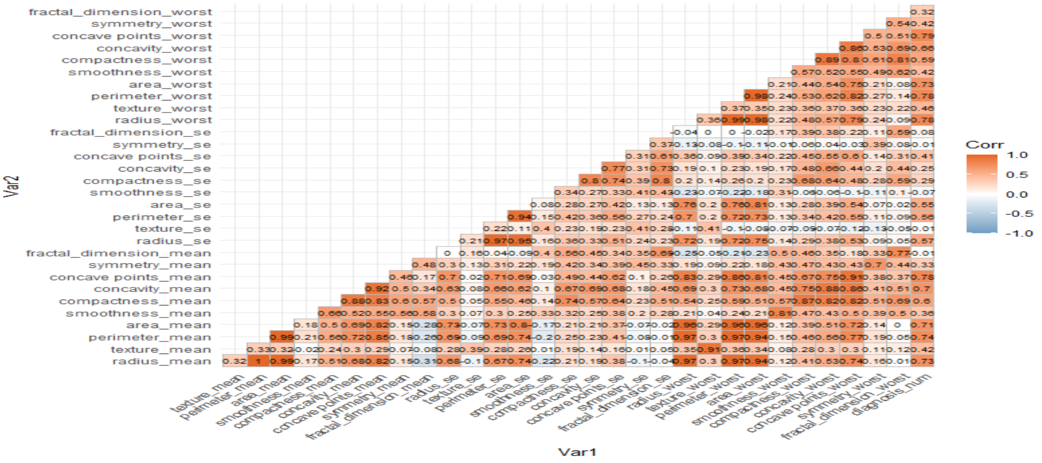

Figure 1. Correlation relationship between variables in breast cancer dataset based on Pearson coefficient

It can be seen in Figure 1 which interprets the correlation relationship between variables in the breast cancer dataset based on the Pearson coefficient. The colors in the matrix are graded from blue (negative correlation), white (no correlation), to orange (positive correlation),. High and positive correlation values, as for the radius mean, perimeter mean, and area mean variables, indicate that the larger the size of tumor cells, the greater the likelihood of other characteristics increasing simultaneously. A negative correlation between certain features could mean that there are variables that are in opposite directions in their contribution to the diagnosis (Dieli-Conwright et al., 2018).

Based on the Random Forest model built to predict breast cancer diagnosis, some of the most influential features in determining whether a case is benign or malignant are the shape and size characteristics of the tumor mass. In particular, the highest values of "concave points worst", "area worst", and "perimeter worst" show that masses with conspicuous concave edges and large size and perimeter tend to be more strongly associated with malignant diagnoses. This implies that medically speaking, tumors with unsmooth surfaces (many concave indentations) and large sizes are more likely to be cancerous (Triberti et al., 2019). Meanwhile, features such as "radius worst" and "texture worst" also play an important role, showing that the maximum radius size and surface texture of a tumor measured at its worst can be critical indicators for classification.

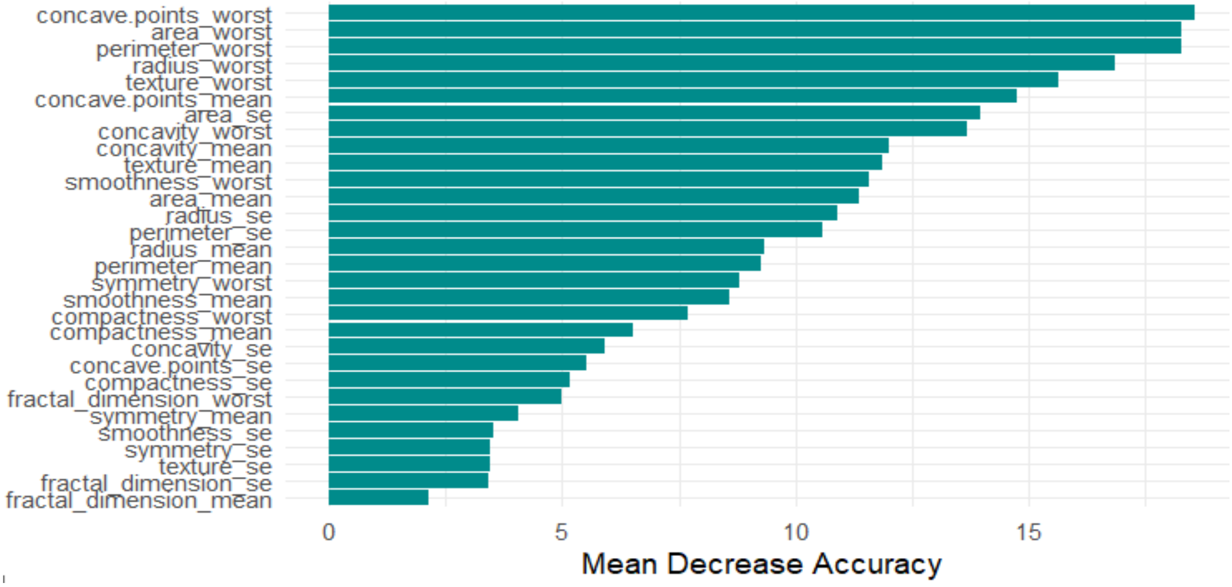

Figure 2. Random Forest model built to predict breast cancer diagnosis

Figure 2 shows the features of the Random Forest model used to predict breast cancer diagnosis. Each horizontal bar represents how much the accuracy of the model drops if the feature is removed - the longer the bar, the more important the feature. Features such as concave points worst, area worst, and perimeter worst are at the top, indicating that they contribute the most to the model's accuracy (Islami et al., 2022). As based on the research that has been conducted, the table containing the 10 most important features of the Random Forest model in predicting breast cancer diagnosis along with their Mean Decrease Accuracy values (Alcaraz et al., 2020).

Table 2 shows the most influential features in model accuracy based on the decrease in accuracy when the feature is removed (Mean Decrease Accuracy). A higher value means that the feature is more important to the model in differentiating between benign and malignant diagnoses.

Table 2. The 10 most important features of the Random Forest model in predicting breast cancer diagnosis along with the Mean Decrease Accuracy value

Feature	Mean Decrease Accuracy
Concave points worst	18.56885
area worst	18.29562
perimeter worst	18.28076
radius worst	16.86471
texture worst	15.63577
Concave points mean	14.74637
area se	13.95269
concavity worst	13.68329
concavity mean	12.0022
texture mean	11.8535

Figure 3 shows an upward trend in accuracy with respect to the number of features used. The curve rises sharply from 1 to about 6 features, then gradually slopes to peak at 9 features. The best accuracy point is specially marked in bright pink with the annotation "Best", confirming that the selection of 9 features is the optimum point-offering the best balance between model complexity and prediction performance. After this point, adding more features does not significantly improve accuracy, suggesting that additional features are redundant or may even add noise (Rock et al., 2020).

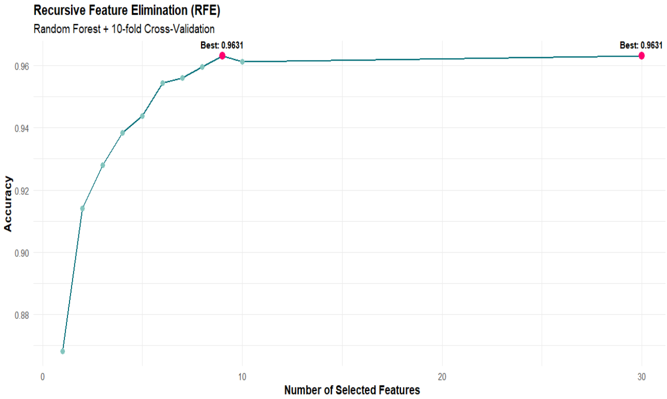

Figure 3. Recursive Feature Elimination (RFE) with Random Forest + 10 Fold Cross-Validation

The research has been conducted from recursive feature elimination (RFE) with random forest shows that of all the variables in this breast cancer dataset, there are nine top variables that are able to provide the highest diagnosis prediction accuracy, which is about 96.31% with a Kappa value of 0.9205. This means that the model is quite reliable in distinguishing between malignant and benign cancer cases. The most medically important variables that determine the diagnosis are area worst, concave points worst, perimeter worst, radius worst, and concave points mean. All of these features are closely related to the size and contour of the tumor cells as seen from medical imaging, where the larger and more irregular the shape of the cells, the higher the probability that the tumor is malignant. Therefore, these features have strong clinical value in aiding early and accurate diagnosis of breast cancer (Tanaka et al., 2022).

Table 3 is informative about the effectiveness of the number of features in predicting breast cancer diagnosis using the Random Forest algorithm with 10-fold cross validation. The table shows that the accuracy of the model increases as the number of features used increases, until it reaches its peak when using nine selected features. The highest accuracy was recorded at 0.9631, which was accompanied by a Kappa value of 0.9205-a strong indicator that the model was not only accurate, but also consistent in distinguishing between diagnosis categories (benign and malignant).

Medically, the features selected by RFE mean that the morphological characteristics of tumor cells, especially at the worst or extreme parameters (such as area worst, concave points worst, and perimeter worst), have enormous predictive power in distinguishing malignant from benign cancers. This is consistent with the clinical literature, which emphasizes that irregularities in cell shape and size are key indicators of malignancy (Li et al., 2022).

Table 3. Informative about the effectiveness of the number of features in predicting breast cancer diagnosis using Random Forest algorithm with 10-fold cross validation

Number of Variables	Accuracy	Kappa	SD Accuracy	SD Kappa
1	0.8681	0.7193	0.04717	0.09786
2	0.914	0.8137	0.04534	0.09991
3	0.9279	0.8444	0.03341	0.07293
4	0.9385	0.8673	0.03417	0.07491
5	0.9438	0.8796	0.02714	0.0583
6	0.9543	0.9014	0.02625	0.05685
7	0.9561	0.9055	0.02341	0.05056
8	0.9596	0.9127	0.02027	0.04499
9	0.9631	0.9205	0.01931	0.04203
10	0.9613	0.9168	0.01814	0.0395
30	0.9631	0.9203	0.01931	0.04225

Principal Component Analysis (PCA) of the normalized breast cancer data shows that most of the variation in the data can be explained by the first few principal components. The first principal component (PC1) alone accounts for about 44.27% of the total variation, while the second component (PC2) adds 18.97%, so these two components cumulatively explain about 63.24% of the total variation in the data. By adding up to the third component (PC3), the explanatory coverage increases to 72.64%.

After the fifth component, the value of the proportion of variance starts to decrease dramatically, indicating that most of the important information in the data has been successfully summarized in just the first few dimensions (McGuire et al., 2015). The first 10 principal components are already able to explain more than 95% of the overall variation in the data. This means that the original dimension of the data consisting of 30 variables can be significantly reduced without losing much important information, which is very beneficial for visualization, interpretation, and efficiency in the application of predictive algorithms or classification of cancer diagnosis (Sukma et al., 2022). The table below shows the PCA for breast cancer data, specifically displaying the standard deviation, proportion of variance, and cumulative proportion values of the first to tenth principal components:

Table 4. PCA for breast cancer data

Key Components	Standard Deviation	Proportion of Variance	Cumulative Proportion
PC1	3.6444	0.4427	0.4427
PC2	2.3857	0.1897	0.6324
PC3	1.6787	0.0939	0.7264
PC4	1.4074	0.066	0.7924
PC5	1.284	0.055	0.8473
PC6	1.0988	0.0403	0.8876
PC7	0.8217	0.0225	0.9101
PC8	0.6904	0.0159	0.926
PC9	0.6457	0.0139	0.9399
PC10	0.5922	0.0117	0.9516

It can be seen in Figure 4 that the Principal Component Analysis (PCA) interpretation of the breast cancer data shows a fairly clear separation between the two diagnosis groups, malignant and benign cancer, based on the first two principal components (PC1 and PC2) which explain about 63.2% of the overall variation in the data. The data points on the graph represent the representation of each patient in the reduced two-dimensional space, where red represents malignant cases and blue for benign cases (Malik et al., 2021).

The distribution in Figure 4 shows that patients with different diagnoses tend to cluster in different feature spaces, indicating that the statistical characteristics of the original data (such as the texture, size, and shape of the tumor cells) can indeed distinguish the two types of diagnoses significantly. The additional elliptical lines surrounding each group show a normal distribution around the group centers, reinforcing the separation pattern between cancer categories (Poltavets et al., 2018).

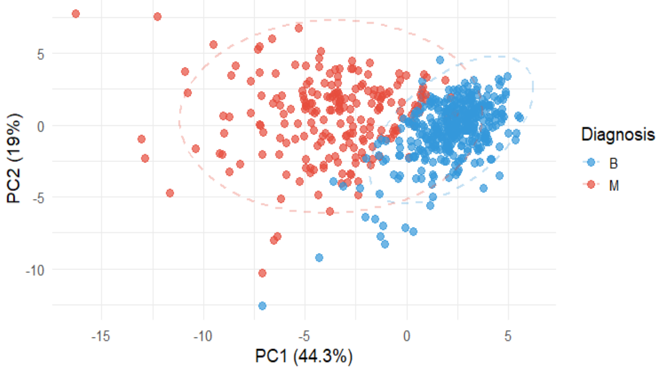

Figure 4. Malignant and benign cancers, based on the first two principal components (PC1 and PC2) which explain about 63.2% of the overall variation in the data

Based on the analysis using Shapley values in the random forest model for cancer diagnosis, it is found that the probability of the tested individual to belong to the malignant cancer class is 0.076, much lower than the overall average probability of 0.373. This means that the model predicts the individual is less likely to have malignant cancer, or in other words, more likely to belong to the benign category (Takahashi et al., 2021).

The contribution of each feature to the prediction means that some features have a significant negative influence on increasing the probability of malignant cancer. Among these features, perimeter worst, area worst, and radius worst are the most dominant factors in decreasing the probability of malignant cancer prediction, contributing -0.066, -0.064, and -0.060 negatively to the probability score, respectively. This implies that the relatively low values of these measures (e.g. perimeter worst = 76.51, area worst = 351.9) help the model infer that the tumor is benign.

Based on SHAP interpretation, the predicted probability of malignant cancer of 0.076 is mainly influenced by features that indicate relatively small tumor size and shape and do not characterize malignancy. Features such as worst perimeter, worst area, and worst radius contribute the most negatively, lowering the probability of malignant cancer. In contrast, some features such as concavity mean and concave.points worst did slightly increase the probability, but not significantly enough to change the final result. Overall, the model predicts that patients are likely to have benign tumors (Kobryn et al., 2023).

Meanwhile, features such as concave.points worst, concavity mean, and concavity worst make a positive contribution to the probability of malignant classification, although the contribution is not large enough to reverse the final prediction. A value of concave.points worst = 0.15, for example, increases the probability of malignant cancer by +0.033, which implies a trend towards sharper and inward tumor shapes, a common feature of malignant tumors.

However, since the other values of the key features favor benign classification, the final result still shows a low probability of malignant cancer

Table 5. Contribution of each feature to the prediction

Feature	Feature Value	Contribution (phi)	Contribution Direction	Interpreting
perimeter worst	76.51	-0.06604	Negative	The size of the worst tumor perimeter is relatively small, lowering the chance of malignancy
area worst	351.9	-0.06384	Negative	Small maximum area, favoring benign classification
radius worst	10.85	-0.06016	Negative	Small maximum radius, lowering the chance of malignant tumors
area se	7.23	-0.02182	Negative	Small area deviation, strengthening the suspicion of benign tumor
texture worst	22.82	-0.0213	Negative	Maximum texture is inconspicuous, reducing the suspicion of malignancy
fractal dimension se	0.01	-0.02082	Negative	Low-boundary irregularity, strengthening benign prediction
concavity se	0.15	-0.02032	Negative	Small localized incidence, lower probability of malignancy
perimeter mean	70.15	-0.0172	Negative	Small average perimeter, favoring benign classification
smoothness worst	0.11	-0.01396	Negative	Smooth surface, not characteristic of malignant tumors
radius mean	10.57	-0.01228	Negative	Small average radius size, reinforcing the benign classification
...	...	...	...	...
concavity worst	0.6	0.02732	Positive	Maximum significant sharp basin, slightly raising suspicion of malignancy
concave.points mean	0.06	0.02798	Positive	The concave dots are quite high, pushing the prediction towards malignancy
concavity mean	0.23	0.03106	Positive	Average likelihood is high, increasing the probability of malignant cancer
concave.points worst	0.15	0.03254	Positive	High maximum basin point, pushing towards malignant prediction

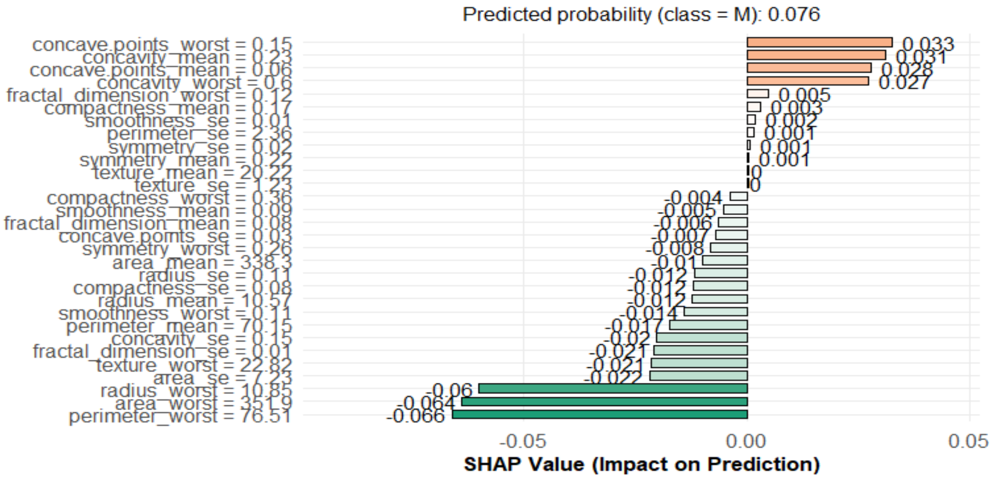

Figure 5. SHAP (Shapley Additive Explanations) which defines the contribution of each feature to the prediction of the probability of malignant cancer in a patient observation

Figure 5 is the SHAP (Shapley Additive Explanations) which defines the contribution of each feature to the prediction of the probability of malignant cancer in a patient observation (Cameron et al., 2020). Each horizontal bar represents one feature, sorted by its influence on the model (Miake-Lye et al., 2019). The color of the bar defines the direction of the contribution: greenish color indicates a negative effect (decreasing the probability of malignant cancer), while yellowish to orange color indicates a positive effect (increasing the probability). The length of the bar defines the magnitude of the feature's impact on the final prediction. In this case, most of the features contribute negatively to the prediction, leading to a low probability that the patient has malignant cancer (predicted value: 0.076). This means that the patient's overall tumor characteristics more closely resemble a benign tumor pattern (Rompis et al., 2019).

The Welch Two Sample t-test of the mean radius variable based on cancer diagnosis shows that there is a highly statistically significant difference between the mean radius values in the benign cancer group (B) and the malignant cancer group (M). The average mean radius value in group B was about 12.15, while in group M it was much higher, about 17.46. The obtained t-statistic value of -22.209 with degrees of freedom of about 289.71, yields a very small p-value (p < 2.2e-16), which means that this difference is highly significant and unlikely to have occurred by chance. The 95% confidence interval for the mean difference is in the range of -5.79 to -4.85, which is entirely below zero, indicating that the mean radius is consistently lower in patients with benign diagnoses compared to malignant ones. This finding means that mean radius can be an important indicator in distinguishing between benign and malignant cancer types.

Figure 6 interprets the distribution of mean radius values for each cancer diagnosis group, namely benign (B) and malignant (M), in the form of density curves. It can be seen that the distribution curve for malignant cancer (colored pink) is significantly shifted to the right compared to the curve for benign cancer (colored blue), which implies that patients with a malignant cancer diagnosis tend to have higher mean radius values. This pattern reinforces the previous t-test results, where there was a statistically significant difference between the two groups, with the mean radius being greater in the malignant group than the benign group. This implies that mean radius is potentially a relevant variable in differentiating cancer types (Osaili et al., 2023).

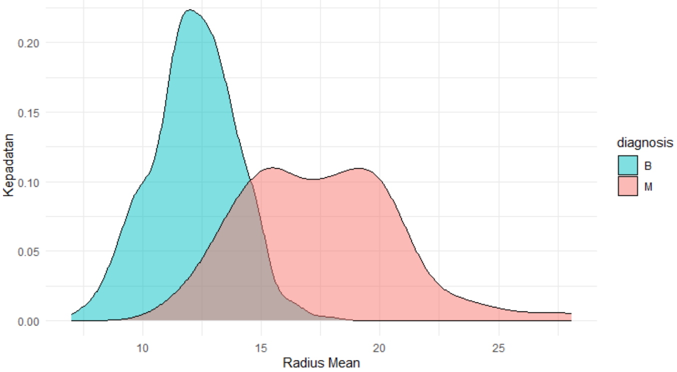

Figure 6. Distribution of mean radius values for each cancer diagnosis group, namely benign (B) and malignant (M), in the form of density curves

Based on the logistic regression analysis conducted to model breast cancer diagnosis (expressed as a categorical variable between benign and malignant), it was found that the variables radius mean, texture mean, perimeter mean, and area mean significantly affect the likelihood of a person being diagnosed with malignant cancer. This model implies that as the radius mean increases, the likelihood of malignant cancer decreases significantly, indicated by the negative coefficient of -9.428 and a very small p value (p < 0.001). In contrast, texture mean, perimeter mean, and area mean had a positive relationship with malignant cancer diagnosis, meaning that an increase in these characteristics correlated with an increased risk of malignancy. In particular, the perimeter mean showed the statistically strongest influence on diagnosis with a positive coefficient value of 1.150 and very high significance (p < 0.001). However, the model raises a warning that there are fitted probabilities that are numerically close to 0 or 1, indicating possible overfitting or perfect separability issues in the data (De Falco, 2012). This suggests that the model is able to separate benign and malignant data very clearly based on the combination of predictor variables, but this also needs to be further reviewed to ensure the generalizability of the model to new data. Shown in the table below are the logistic regression results in tabular form summarizing the coefficient estimates, standard errors, z-values, and significance values (p-values) for each of the predictor variables on cancer diagnosis:

Table 6. Logistic regression summarizing coefficient estimates, standard errors, z-values, and significance values (p-values) for each predictor variable on cancer diagnosis

Variable	Coefficient (Estimate)	Std. Error	Z value	P-value	Significance
(Intercept)	1.77291	6.87011	0.258	0.79636	Not significant
Radius mean	-9.42874	1.63958	-5.751	8.89E-09	*** (highly significant)
Texture mean	0.23761	0.04603	5.162	2.44E-07	*** (highly significant)
Perimeter mean	1.15066	0.16436	7.001	2.54E-12	*** (highly significant)
Area mean	0.03277	0.01182	2.771	0.00558	** (significant)

Table 6 shows that the radius mean variable has a negative and highly significant effect on cancer diagnosis, while the texture mean, perimeter mean, and area mean have a significant positive effect, with the perimeter mean implying the strongest influence on the likelihood of malignant cancer.

The breast cancer diagnosis prediction probability distribution histogram image shown below represents a clear separation between the two diagnosis groups, Benign and Malignant. Medically, the graph illustrates that the logistic regression model is able to distinguish the majority of malignant and benign cancer cases based on the physical features of the cells (such as radius, texture, perimeter, and area).

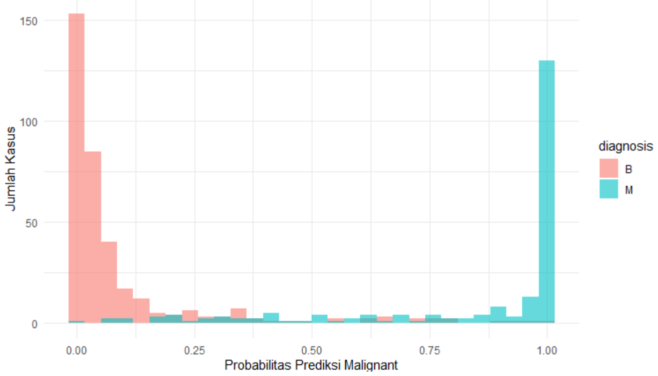

Figure 7. Prediction probability distribution of breast cancer diagnosis

Cases with malignant diagnoses tend to have higher prediction probabilities (closer to 1), while benign cases are more on the low probability side (closer to 0). This suggests that clinically, the model can be used as an early screening tool to estimate the risk of malignancy in patients based on tumor cell measurement results. Evaluation of the model's accuracy showed excellent performance, with a prediction accuracy rate of 91.92%. From the classification table, the model successfully classified 340 benign cases and 183 malignant cases correctly, and only made mistakes in 46 cases (17 false positives and 29 false negatives) (Shields et al., 2021).

Based on the research conducted, there is a highly statistically significant relationship between the average size of breast cancer cell area and the type of diagnosis (benign or malignant). In the analyzed data, most cases with small cell areas were categorized as benign tumors (268 benign cases vs 17 malignant cases), while cases with large cell areas tended to be classified as malignant tumors (195 malignant cases vs 89 benign cases). The chi-square test yielded a statistical value of \mathrm{(}\chi^2=236.53\mathrm{)} with a very small p value ($p < 2.2e-16$), meaning that the association between the two variables did not occur by chance. The strength of the association between area size and diagnosis was confirmed by the Cramer's V value of 0.648, indicating a strong association between cell area size and the likelihood of breast cancer diagnosis as malignant or benign. This finding reinforces the notion that average cell area size is an important indicator in the early classification of breast cancer.

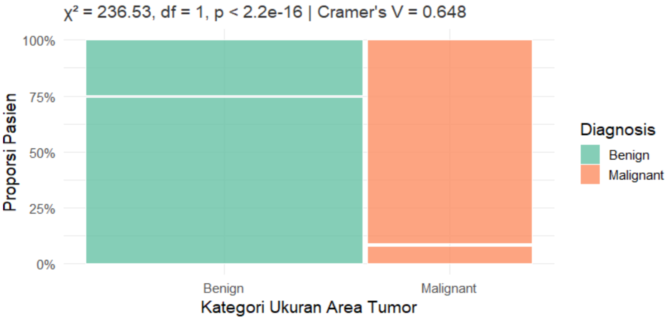

Figure 8. Prediction probability distribution of breast cancer diagnosis

Medically, the visualization shown in Figure 8 indicates that the average size of the breast tumor cell area has a strong association with the malignancy of the tumor. The proportion of patients with malignant tumors is much higher in the large area size group compared to the small size group. In contrast, benign tumors were more common in the group with smaller cell area size. The findings confirmed that the larger the area size of the cells detected in the tumor tissue, the more likely the tumor was malignant. This has important implications for clinical practice, especially in the process of early detection and diagnosis of breast cancer, as the morphological size of the cells can be used as an early indicator in distinguishing the potential malignancy of the tumor.

Discussion

The worst concave points feature has a large contribution in the Random Forest and SHAP models, showing a strong association with cell malignancy. In histopathology, the irregular contours of the nucleus and cell membrane are classic indicators of dysplasia and malignant transformation. Literature that has been done confirms the shape abnormality is the result of cytoskeleton reorganization that is common in cancer. The modern Random Forest and SHAP models that have been conducted in the study verify quantitatively that cells with irregular basin contours are highly likely to be malignant. This means that the integration of AI feature importance and classical morphological indicators can strengthen the diagnosis.

High radius mean area mean and perimeter mean features are associated with malignant cancer. Biologically, cancer cells have larger nuclei (hyperchromatic and pleomorphic). According to research conducted, Increased nucleus volume and diameter were positively correlated with cancer aggressiveness. The Random Forest model automatically identifies that nucleus size-related features are strong predictors, which is consistent with this study. This confirms the classical theory with modern machine learning approaches.

Features such as `texture worst` and `compactness mean` show significantly different distributions between benign and malignant tumors. Study conducted by shows that malignant tumors tend to have higher textural heterogeneity, which is also reflected at the microscopic level. Textural feature extraction in radiomics has become a key approach in precision oncology. Features such as `texture worst` automatically capture microscopic heterogeneity in histopathology images (Braun et al., 2020)

The features `area worst`, `perimeter worst`, and `radius worst` rank highest in prediction contribution. In clinical practice, the maximum (limb) (Zahid Iqbal & Campbell, 2023) value is used to determine severity (in tumor grading) (Shtar et al., 2021). The "worst" feature implies that the most abnormal condition is often more relevant for prediction than the average value. The ML model in this study confirms that the "worst" feature is statistically more informative than the mean, supporting the clinical direction of using extreme features as the gold standard diagnostic threshold.

PCA shows that >95% of the data variance can be summarized in just 10 components. In pattern recognition, PCA is a standard tool for reducing information redundancy without sacrificing accuracy. Used in biomedical signal/image processing to speed up classification. In the era of medical big data, PCA's effectiveness in reducing model complexity while maintaining interpretability is a key cornerstone of AI-assisted diagnosis, especially in embedded or edge-computing-based systems.

In the study, the model only needed 9 features to achieve high accuracy (96.31%). RFE is a common method in feature selection, where only a subset of truly relevant features are retained, improving model generalization and reducing overfitting (Kumawat et al., 2023). In explainable AI (XAI) for healthcare, simplifying models without losing performance is essential for clinical adoption (Kabassi & Alepis, 2020). It also reduces the burden of interpretation for doctors and increases transparency.

Some features such as low-value `area worst' actually lowered malignancy prediction. Benign tumors tend to be small, compact and non-spreading (Zhu et al., 2018). Small values in some morphological features are strong indicators of benignity. With the SHAP interpretation performed by this study, negative values can be explained as a protective factor against cancer prediction. This suggests that not only are large values important, but the context of the features also determines.

There is a high correlation between radius, perimeter and area. The law of cell geometry explains that abnormal growth in one aspect of size usually affects other aspects (tumors enlarge isotropically). This corresponds to the volume-surface theory in pathology (Uddin et al., 2022). The application of correlation will enable indirect estimation and strengthen predictive features with a multivariate approach, which is particularly effective in models such as Random Forest (Sandra et al., 2022).

The Random Forest Model provides high accuracy, and SHAP shows the transparency of the model locally and globally. Study on SHAP explains how important interpretability is in machine learning-based diagnosis. Based on the research conducted, Random Forest excels in non-linear detection, is robust to noise, and is well suited for complex medical datasets. SHAP becomes a bridge between AI and human decision-making in clinical diagnosis that is interpretable and actionable, just like the research that has also been done (Oršolić et al., 2021).

Conclusion

Research has comprehensively confirmed that the morphological characteristics of breast tumor cells, particularly the size and shape of the cell nucleus, play a central role in differentiating benign and malignant cancers. Features such as "area worst", "concave points worst", and "perimeter worst" proved to be particularly powerful in indicating malignancy, as extreme values of these features correlated closely with the irregular, large, and non-smooth-edge structures of cancer cells - features typical of malignant tumors in the clinical literature. From a medical point of view, these findings reinforce the understanding that the larger the size and morphological irregularity of the cell nucleus, the higher the likelihood of the cancer being malignant. Prediction models based on Random Forest algorithm, Recursive Feature Elimination (RFE), and interpretation of Shapley Values show that statistical features of cells such as "concavity", "compactness", and "texture" also contribute significantly to the diagnosis. Therefore, these features have high diagnostic value in clinical practice and can be utilized as a basis in the development of medical image-based cancer early detection systems. In practical terms, the study successfully identified nine key features that are optimal in predicting breast cancer diagnosis with an accuracy rate of 96.31% and a Kappa value of 0.9205, indicating the consistency and reliability of the model in distinguishing benign and malignant cases. The use of Random Forest model combined with 10-fold cross validation provides high accuracy and generalization, while avoiding overfitting. Furthermore, the Principal Component Analysis (PCA) approach was able to reduce the data dimension from 30 variables to 2 to 10 principal components without losing much important information (>95% of data variation explained). This is very useful for efficient visualization and interpretation of models in real applications, such as in the development of AI-based diagnosis aids in hospitals or clinics. Meanwhile, the application of model interpretation methods such as Shapley values provides much-needed transparency in machine learning-based systems, allowing clinicians to understand the rationale for the system's predictions and strengthening confidence in its use. Based on the results of the research that has been done, there are several suggestions that can be used as a reference for further research. First, it is recommended that the prediction model that has been built be further tested on real clinical data with a wider number and variety. This aims to test the external validity and measure the generalization ability of the model in a real clinical environment. Secondly, the integration of numerical data with medical image data such as mammograms or histopathology images can be a promising approach to improve accuracy and enrich diagnostic information. This multimodal approach has the potential to provide a more comprehensive picture of cancer characteristics. Third, future research can consider the use of deep learning algorithms such as Convolutional Neural Network (CNN) or Long Short-Term Memory (LSTM) that are able to capture non-linear and complex patterns, especially in large data. Fourth, longitudinal analysis or repeated data over time is also important to understand changes in cell morphological characteristics that can support early cancer detection. Improving the interpretability of prediction models is also an important aspect, especially by utilizing interpretation methods such as SHAP or LIME. This is so that the prediction results are not only accurate, but can also be easily understood by medical personnel in clinical decision-making.

Acknowledgment
In this section, authors acknowledge all contributions from other persons and any individual or organization who have contributed to the research that was done. Begin with "The Author(s) wish to thank."

Authors Contribution
All authors contribution together for each stage of research.

Funding
All of the sources of funding for the work described in this publication are acknowledged here.

Conflicts of Interest
All financial or non-financial competing interests must be declared in this section. If you do not have any competing interests, please state "The author(s) declare no competing interest in this study"

References

Al Mudawi, N., & Alazeb, A. (2022). A Model For Predicting Cervical Cancer Using Machine Learning Algorithms. *Sensors*, *22*(11), 4132. https://doi.org/10.3390/S22114132

Alcaraz, K. I., Wiedt, T. L., Daniels, E. C., Yabroff, K. R., Guerra, C. E., & Wender, R. C. (2020). Understanding And Addressing Social Determinants To Advance Cancer Health Equity In The United States: A Blueprint For Practice, Research, And Policy. *Ca: A Cancer Journal For Clinicians*, *70*(1), 31–46. https://doi.org/10.3322/caac.21586

Ampofo, A. G., Boyes, A. W., Asibey, S. O., Oldmeadow, C., & Mackenzie, L. J. (2023). Prevalence And Correlates Of Modifiable Risk Factors For Cervical Cancer And Hpv Infection Among Senior High School Students In Ghana: A Latent Class Analysis. *Bmc Public Health*, *23*(1), 340. https://doi.org/10.1186/s12889-022-14908-w

Braun, M., Klingelhöfer, D., Oremek, G. M., Quarcoo, D., & Groneberg, D. A. (2020). Influence Of Second-Hand Smoke And Prenatal Tobacco Smoke Exposure On Biomarkers, Genetics And Physiological Processes In Children—An Overview In Research Insights Of The Last Few Years. *International Journal Of Environmental Research And Public Health*, *17*(9), 3212. https://doi.org/10.3390/ijerph17093212

Cameron, A. R., Meyer, A., Faverjon, C., & Mackenzie, C. (2020). Quantification Of The Sensitivity Of Early Detection Surveillance. *Transboundary And Emerging Diseases*, *67*(6), 2532–2543. https://doi.org/10.1111/tbed.13598

Campos, N. G., Demarco, M., Bruni, L., Desai, K. T., Gage, J. C., Adebamowo, S. N., De Sanjose, S., Kim, J. J., & Schiffman, M. (2021). A Proposed New Generation Of Evidence-Based Microsimulation Models To Inform Global Control Of Cervical Cancer. *Preventive Medicine*, *144*, 106438. https://doi.org/10.1016/j.ypmed.2021.106438

Casas, C. P. R., Albuquerque, R. De C. R. De, Loureiro, R. B., Gollner, A. M., Freitas, M. G. De, Duque, G. P. Do N., & Viscondi, J. Y. K. (2022). Cervical Cancer Screening In Low-And Middle-Income Countries: A Systematic Review Of Economic Evaluation Studies. *Clinics*, *77*, 100080. https://doi.org/10.1016/j.clinsp.2022.100080

Chisale Mabotja, M., Levin, J., & Kawonga, M. (2021). Beliefs And Perceptions Regarding Cervical Cancer And Screening Associated With Pap Smear Uptake In Johannesburg: A Cross-Sectional Study. *Plos One*, *16*(2), E0246574. https://doi.org/10.1371/journal.pone.0246574

Davidović, M., Asangbeh, S. L., Taghavi, K., Dhokotera, T., Jaquet, A., Musick, B., Van Schalkwyk, C., Schwappach, D., Rohner, E., & Murenzi, G. (2024). Facility-Based Indicators To Manage And Scale Up Cervical Cancer Prevention And Care Services For Women Living With Hiv In Sub-Saharan Africa: A Three-Round Online Delphi Consensus Method. *Jaids Journal Of Acquired Immune Deficiency Syndromes*, *95*(2), 170–178. https://doi.org/10.1097/QAI.0000000000003343

De Falco, S. (2012). The Discovery Of Placenta Growth Factor And Its Biological Activity. *Experimental & Molecular Medicine 2012 44:1*, *44*(1), 1–9. https://doi.org/10.3858/Emm.2012.44.1.025

Dieli-Conwright, C. M., Courneya, K. S., Demark-Wahnefried, W., Sami, N., Lee, K., Sweeney, F. C., Stewart, C., Buchanan, T. A., Spicer, D., Tripathy, D., Bernstein, L., & Mortimer, J. E. (2018). Aerobic And Resistance Exercise Improves Physical Fitness, Bone Health, And Quality Of Life In Overweight And Obese Breast Cancer Survivors: A Randomized Controlled Trial 11 Medical And Health Sciences 1117 Public Health And Health Services. *Breast Cancer Research*, *20*(1), 1–10. https://doi.org/10.1186/s13058-018-1051-6

Dykens, J. A., Peterson, C. E., Holt, H. K., & Harper, D. M. (2023). Gender Neutral Hpv Vaccination Programs: Reconsidering Policies To Expand Cancer Prevention Globally. *Frontiers In Public Health*, *11*, 1067299. https://doi.org/10.3389/fpubh.2023.1067299

Ford, S., Tarraf, W., Williams, K. P., Roman, L. A., & Leach, R. (2021). Differences In Cervical Cancer Screening And Follow-Up For Black And White Women In The United States. *Gynecologic Oncology*, *160*(2), 369–374. https://doi.org/10.1016/j.ygyno.2020.11.027

Gravitt, P. E., Silver, M. I., Hussey, H. M., Arrossi, S., Huchko, M., Jeronimo, J., Kapambwe, S., Kumar, S., Meza, G., Nervi, L., Paz-Soldan, V. A., & Woo, Y. L. (2021). Achieving Equity In Cervical Cancer Screening In Low- And Middle-Income Countries (Lmics): Strengthening Health Systems Using A Systems Thinking Approach. *Preventive Medicine*, *144*, 106322. https://doi.org/10.1016/j.ypmed.2020.106322

Heidari Sarvestani, M., Khani Jeihooni, A., Moradi, Z., & Dehghan, A. (2021). Evaluating The Effect Of An Educational Program On Increasing Cervical Cancer Screening Behavior Among Women In Fasa, Iran. *Bmc Women's Health*, *21*, 1–8. https://doi.org/10.1186/s12905-021-01191-x

Islami, F., Guerra, C. E., Minihan, A., Yabroff, K. R., Fedewa, S. A., Sloan, K., Wiedt, T. L., Thomson, B., Siegel, R. L., Nargis, N., Winn, R. A., Lacasse, L., Makaroff, L., Daniels, E. C., Patel, A. V., Cance, W. G., & Jemal, A. (2022). American Cancer Society's Report On The Status Of Cancer Disparities In The United States, 2021. *Ca: A Cancer Journal For Clinicians*, *72*(2), 112–143. https://doi.org/10.3322/caac.21703

Ji, L., Chen, M., & Yao, L. (2023). Strategies To Eliminate Cervical Cancer In China. *Frontiers In Oncology*, *13*, 1105468. https://doi.org/10.3389/fonc.2023.1105468

Johnson, A. J., Johnson, M. J., Williams, J. B., Muscari, E., Palmo, L., Ruiz, M., Bush, B., & Campbell, L. C. (2025). Cervical Cancer Prevention Behaviors In Young Black Women. *Women's Health*, *21*, 17455057251326008. https://doi.org/10.1177/17455057251326008

Kabassi, K., & Alepis, E. (2020). Learning Analytics In Distance And Mobile Learning For Designing Personalised Software. In *Machine Learning Paradigms* (Bll 185–203). Springer. https://doi.org/10.1007/978-3-030-13743-4_10

Kobryn, A., Nian, P., Baidya, J., Li, T. L., & Maheshwari, A. V. (2023). Intramedullary Nailing With And Without The Use Of Bone Cement For Impending And Pathologic Fractures Of The Humerus In Multiple Myeloma And Metastatic Disease. *Cancers*, *15*(14), 3601. https://doi.org/10.3390/cancers15143601

Kumawat, G., Vishwakarma, S. K., Chakrabarti, P., Chittora, P., Chakrabarti, T., & Lin, J. C.-W. (2023). Prognosis Of Cervical Cancer Disease By Applying Machine Learning Techniques. *Journal Of Circuits, Systems And Computers*, *32*(01), 2350019. https://doi.org/10.1142/s0218126623500196

Li, G., Gong, S., Wang, N., & Yao, X. (2022). Toxic Epidermal Necrolysis Induced By Sintilimab In A Patient With Advanced Non-Small Cell Lung Cancer And Comorbid Pulmonary Tuberculosis: A Case Report. *Frontiers In Immunology*, *13*, 989966. https://doi.org/10.3389/fimmu.2022.989966

Lilhore, U. K., Poongodi, M., Kaur, A., Simaiya, S., Algarni, A. D., Elmannai, H., Vijayakumar, V., Tunze, G. B., & Hamdi, M. (2022). Hybrid Model For Detection Of Cervical Cancer Using Causal Analysis And Machine Learning Techniques. *Computational And Mathematical Methods In Medicine*, *2022*(1), 4688327. https://doi.org/10.1155/2022/4688327

Liu, G., Mugo, N. R., Bayer, C., Rao, D. W., Onono, M., Mgodi, N. M., Chirenje, Z. M., Njoroge, B. W., Tan, N., & Bukusi, E. A. (2022). Impact Of Catch-Up Human Papillomavirus Vaccination On Cervical Cancer Incidence In Kenya: A Mathematical Modeling Evaluation Of Hpv Vaccination Strategies In The Context Of Moderate Hiv Prevalence. *Eclinicalmedicine*, *45*. Retrieved from https://www.thelancet.com/journals/eclinm/article/PIIS2589-5370(22)00036-0/fulltext

Malik, M., Parveen Kiyani, I., Rana, S., Hussain, A., & Bin Aslam Zahid, M. (2021). *Quality Of Life And Psychological Distress During Cancer: A Prospective Observational Study Involving Liver Cancer Patients*. Retrieved from http://libraryaplos.com/xmlui/handle/123456789/6325

Mcguire, A., Brown, J. A. L., Malone, C., Mclaughlin, R., & Kerin, M. J. (2015). Effects Of Age On The

Detection And Management Of Breast Cancer. *Cancers*, *7*(2), 908–929. Https://Doi.Org/10.3390/Cancers7020815

Miake-Lye, I. M., Mak, S., Lee, J., Luger, T., Taylor, S. L., Shanman, R., Beroes-Severin, J. M., & Shekelle, P. G. (2019). Massage For Pain: An Evidence Map. In *Journal Of Alternative And Complementary Medicine*. https://doi.org/10.1089/acm.2018.0282

Nougaret, S., Addley, H., Sala, E., & Sahdev, A. (2020). Ovarian Cancer 19. *Husband & Reznek's Imaging In Oncology*, 378. CRC Press.

Obol, J. H., Lin, S., Obwolo, M. J., Harrison, R., & Richmond, R. (2021). Knowledge, Attitudes, And Practice Of Cervical Cancer Prevention Among Health Workers In Rural Health Centres Of Northern Uganda. *Bmc Cancer*, *21*, 1–15. https://doi.org/10.1186/s12885-021-07847-z

Oršolić, D., Pehar, V., Šmuc, T., & Stepanić, V. (2021). Comprehensive Machine Learning Based Study Of The Chemical Space Of Herbicides. *Scientific Reports*, *11*(1), 11479. https://doi.org/10.1038/s41598-021-90690-w

Osaili, T. M., Dhanasekaran, D. K., Zeb, F., Faris, M. E., Naja, F., Radwan, H., Cheikh Ismail, L., Hasan, H., Hashim, M., & Obaid, R. S. (2023). A Status Review On Health-Promoting Properties And Global Regulation Of Essential Oils. *Molecules*, *28*(4), 1809. https://doi.org/10.3390/molecules28041809

Pacal, I. (2024). Maxcervixt: A Novel Lightweight Vision Transformer-Based Approach For Precise Cervical Cancer Detection. *Knowledge-Based Systems*, *289*, 111482. https://doi.org/10.1016/j.knosys.2024.111482

Pieters, M. M., Proeschold-Bell, R. J., Coffey, E., Huchko, M. J., & Vasudevan, L. (2021). Knowledge, Attitudes, And Practices Regarding Cervical Cancer Screening Among Women In Metropolitan Lima, Peru: A Cross-Sectional Study. *Bmc Women's Health*, *21*, 1–13. https://doi.org/10.1186/s12905-021-01431-0

Poltavets, V., Kochetkova, M., Pitson, S. M., & Samuel, M. S. (2018). The Role Of The Extracellular Matrix And Its Molecular And Cellular Regulators In Cancer Cell Plasticity. *Frontiers In Oncology*. https://doi.org/10.3389/fonc.2018.00431/bibtex

Pramanik, R., Biswas, M., Sen, S., Souza Júnior, L. A. De, Papa, J. P., & Sarkar, R. (2022). A Fuzzy Distance-Based Ensemble Of Deep Models For Cervical Cancer Detection. *Computer Methods And Programs In Biomedicine*, *219*, 106776. https://doi.org/10.1016/j.cmpb.2022.106776

Rock, C. L., Thomson, C., Gansler, T., Gapstur, S. M., Mccullough, M. L., Patel, A. V, Andrews, K. S., Bandera, E. V, Spees, C. K., Robien, K., Hartman, S., Sullivan, K., Grant, B. L., Hamilton, K. K., Kushi, L. H., Caan, B. J., Kibbe, D., Black, J. D., Wiedt, T. L., … Doyle, C. (2020). American Cancer Society Guideline For Diet And Physical Activity For Cancer Prevention. *Ca: A Cancer Journal For Clinicians*, *70*(4), 245–271. https://doi.org/10.3322/caac.21591

Rompis, K., Wowor, V. N. S., & Pangemanan, D. H. C. (2019). Tingkat Pengetahuan Bahaya Merokok Bagi Kesehatan Gigi Mulut Pada Siswa Smk Negeri 8 Manado. *E-Clinic*, *7*(2). https://doi.org/10.35790/ecl.v7i2.24023

Sandra, L., Marcel, Gunarso, G., Fredicia, & Riruma, O. W. (2022). Are University Students Independent: Twitter Sentiment Analysis Of Independent Learning In Independent Campus Using Roberta Base Indolem Sentiment Classifier Model. *2021 International Seminar On Machine Learning, Optimization, And Data Science (Ismode)*, 249–253. https://doi.org/10.1109/ismode53584.2022.9743110

Shields, H. J., Traa, A., & Van Raamsdonk, J. M. (2021). Beneficial And Detrimental Effects Of Reactive Oxygen Species On Lifespan: A Comprehensive Review Of Comparative And Experimental Studies. *Frontiers In Cell And Developmental Biology*, *9*, 628157. https://doi.org/10.3389/fcell.2021.628157

Shoghi, M., Shahbazi, B., & Seyedfatemi, N. (2019). The Effect Of The Family-Centered Empowerment Model (Fcem) On The Care Burden Of The Parents Of Children Diagnosed With Cancer. *Asian Pacific Journal Of Cancer Prevention*, *20*(6), 1757–1764. https://doi.org/10.31557/apjcp.2019.20.6.1757

Shtar, G., Rokach, L., Shapira, B., Nissan, R., & Hershkovitz, A. (2021). Using Machine Learning To Predict Rehabilitation Outcomes In Postacute Hip Fracture Patients. *Archives Of Physical Medicine And Rehabilitation*, *102*(3), 386–394. https://doi.org/10.1016/j.apmr.2020.08.011

Soong, T. R., Dinulescu, D. M., Xian, W., & Crum, C. P. (2018). Frontiers In The Pathology And Pathogenesis Of Ovarian Cancer: Cancer Precursors And" Precursor Escape". *Hematology/Oncology Clinics Of North America*, *32*(6), 915–928. Retrieved from https://www.sciencedirect.com/science/article/pii/S0889858818307639

Soto, M. L. Q., Guillén, J. C., Aguayo, J. M. B., Valdes, J. H., Ruíz, G. B., Morales, F. E., Sanchez, A. S., Campas, C. Y. Q. C., Ornelas, R. M. R., & González, M. Del R. M. (2023). Adherence Model To Cervical Cancer Treatment In The Covid-19 Era. *Baghdad Science Journal*, *20*(4 (Si)), 1559–1569. Retrieved from https://bsj.uobaghdad.edu.iq/home/vol20/iss4/26/

Spencer, J. C., Brewer, N. T., Coyne-Beasley, T., Trogdon, J. G., Weinberger, M., & Wheeler, S. B. (2021). Reducing Poverty-Related Disparities In Cervical

Cancer: The Role Of Hpv Vaccination. *Cancer Epidemiology, Biomarkers & Prevention*, *30*(10), 1895–1903. https://doi.org/10.1158/1055-9965.epi-21-0307

Sukma, D. I., Prabowo, H. A., Setiawan, I., Kurnia, H., & Fahturizal, I. M. (2022). Implementation Of Total Productive Maintenance To Improve Overall Equipment Effectiveness Of Linear Accelerator Synergy Platform Cancer Therapy. *International Journal Of Engineering*, *35*(7), 1246–1256. Retrieved from https://shorturl.asia/X289L

Takahashi, Y., Sone, K., Noda, K., Yoshida, K., Toyohara, Y., Kato, K., Inoue, F., Kukita, A., Taguchi, A., & Nishida, H. (2021). Automated System For Diagnosing Endometrial Cancer By Adopting Deep-Learning Technology In Hysteroscopy. *Plos One*, *16*(3), E0248526. https://doi.org/10.1371/journal.pone.0248526

Tanaka, T., Shindo, T., Hashimoto, K., Kobayashi, K., & Masumori, N. (2022). Management Of Hydronephrosis After Radical Cystectomy And Urinary Diversion For Bladder Cancer: A Single Tertiary Center Experience. *International Journal Of Urology*, *29*(9), 1046–1053. https://doi.org/10.1111/iju.14970

Triberti, S., Savioni, L., Sebri, V., & Pravettoni, G. (2019). Corrigendum To Ehealth For Improving Quality Of Life In Breast Cancer Patients: A Systematic Review. *Cancer Treatment Reviews*, *81*, 1–14. https://doi.org/10.1016/j.ctrv.2019.101928

Uddin, N., Jaya, S., Purwanto, E., Putra, A. A. D., Fadhilah, M. W., & Ramadhan, A. L. R. (2022). Machine-Learning Prediction Of Informatics Students Interest To The Mbkm Program: A Study Case In Universitas Pembangunan Jaya. *2021 International Seminar On Machine Learning, Optimization, And Data Science (Ismode)*, 146–151. https://doi.org/10.1109/ismode53584.2022.9743125

Yang, C., Qin, L., Xie, Y., & Liao, J. (2022). Deep Learning In Ct Image Segmentation Of Cervical Cancer: A Systematic Review And Meta-Analysis. *Radiation Oncology*, *17*(1), 175. https://doi.org/10.1186/s13014-022-02148-6

Young, C., & Argáez, C. (2020). Manual Therapy For Chronic Non-Cancer Back And Neck Pain: A Review Of Clinical Effectiveness. *Manual Therapy For Chronic Non-Cancer Back And Neck Pain: A Review Of Clinical Effectiveness*. Retrieved from https://europepmc.org/article/NBK/nbk562937

Yu, Z., Yang, X., Dang, C., Wu, S., Adekkanattu, P., Pathak, J., George, T. J., Hogan, W. R., Guo, Y., & Bian, J. (2022). A Study Of Social And Behavioral Determinants Of Health In Lung Cancer Patients Using Transformers-Based Natural Language Processing Models. *Amia Annual Symposium Proceedings*, *2021*, 1225. Retrieved from https://pmc.ncbi.nlm.nih.gov/articles/PMC8861705/

Zahid Iqbal, M., & Campbell, A. G. (2023). Agilest Approach: Using Machine Learning Agents To Facilitate Kinesthetic Learning In Stem Education Through Real-Time Touchless Hand Interaction. *Telematics And Informatics Reports*, *9*(December 2022), 100034. https://doi.org/10.1016/j.teler.2022.100034

Zhang, M., Sit, J. W. H., Chan, D. N. S., Akingbade, O., & Chan, C. W. H. (2022). Educational Interventions To Promote Cervical Cancer Screening Among Rural Populations: A Systematic Review. *International Journal Of Environmental Research And Public Health*, *19*(11), 6874. https://doi.org/10.3390/ijerph19116874

Zhu, X., Xu, Q., Tang, M., Li, H., & Liu, F. (2018). A Hybrid Machine Learning And Computing Model For Forecasting Displacement Of Multifactor-induced landslides. *Neural Computing and Applications*, *30*, 3825–3835. https://doi.org/10.1007/s00521-017-2968-x

Zhuang, J., & Guan, M. (2022). Modeling the mediating and moderating roles of risk perceptions, efficacy, desired uncertainty, and worry in information seeking-cancer screening relationship using HINTS 2017 data. *Health Communication*, *37*(7), 897–908. https://doi.org/10.1080/10410236.2021.1876324