MJFAS
Malaysian Journal of Fundamental and Applied Sciences

RESEARCH ARTICLE

Discovery of Interpretable Patterns of Breast Cancer Diagnosis via Class Association Rule Mining (CARM) With SHAP-Based Explainable AI (XAI)

Shahiratul, A. Karim[a]*, Ummul, H. Mohamad[a,b], Puteri, N. E. Nohuddin[a,b]

[a]Institute of Visual Informatics, Bangunan Akademia Siber Teknopolis, Universiti Kebangsaan Malaysia, 43600 Bangi, Malaysia; [b]iAI-UKM Research Group, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia

Abstract Breast cancer remains the most common cancer among women globally highlighting the importance of early and reliable diagnostic methods. While previous studies have applied association rules mining (ARM) to explore factors contributing to breast cancer, many lacked robust validation of the extracted rules. To address this gap and deepen our understanding of the key biological markers linked to the disease, this study proposes a hybrid framework that integrates Class Association Rule Mining (CARM) with SHapley Additive exPlanations (SHAP) values based on Random Forest (RF) and Gradient Boost (GB) models to uncover and validate meaningful diagnostic patterns. Using the Breast Cancer Coimbra (BCC) dataset comprising 116 patient samples and nine biological markers, a total of 723,938 association rules (AR) were generated with 17,720 significant class association rules (CAR) were extracted. These rules were pruned using lift, leverage and conviction to retain the most relevant ones. Among the healthy group, combinations involving low glucose, low insulin, low resistin and low Homeostatic Model Assessment (HOMA) were consistently observed, while high BMI appeared particularly among younger individuals. These features were associated with negative SHAP values validating their contribution to healthy classifications. In contrast, common patterns such as high glucose, medium resistin and medium Monocyte Chemoattractant Protein-1 (MCP.1) among middle aged individuals highlighting their influence in predicting patient classification. These features consistently showed strong positive SHAP values across both classifiers highlighting their influence in predicting patient outcomes. By combining rule extraction of CARM with feature contribution using SHAP, this study provides a validated and interpretable approach for breast cancer diagnosis. The findings highlight the importance of feature interactions and offer promising directions for personalized risk assessment and early detection.

Keywords: Breast cancer, class association rule, pattern discovery, SHAP, XAI model.

***For correspondence:**
p105766@siswa.ukm.edu.my

Received: 20 Aug. 2024
Accepted: 19 April 2025

Introduction

Breast cancer continues to be a global health issue that significantly affects women around the world. The World Health Organization (WHO) reported that there were 2.3 million women diagnosed with breast cancer with 670,000 deaths recorded in 2022 [1]. The prediction for 2024 is concerning as there are an estimation of 310,720 new breast cancer cases among women with 42,250 fatalities [2]. With the advances in digital technology in recent years, the application of data mining techniques has significantly fast-forwarded the field of cancer research to a point that they can analyze and interpret complex medical data [3].

Data mining is a step in the process of knowledge discovery in databases [4]. This involves a process of uncovering hidden information which is previously unknown to potentially valuable information from large datasets [5]. The extracted information is known as knowledge and can be presented as rules, constraints, and patterns. ARM is a well-established method used to identify interesting relationships between variables in a dataset [6]. The concept of ARM originated from market basket analysis where the goal is to determine the likelihood of a customer purchasing additional products based on their current purchases [7]. The strength of such a rule is measured by several metrics such as support and confidence. To add, CAR is an extension of traditional AR that incorporate class labels into the rule-mining process [8]. While traditional AR focus on identifying relationships between items in a dataset, CAR specifically aims to discover patterns that relate features to class labels. In breast cancer classification tasks, CAR can help identify patterns associated with diseases based on patient attributes while helping towards a more accurate risk assessment and informed decision-making.

As the data began to accumulate and become more complex, machine learning models became the highlight to address this limitation. Black box AI model are highly accurate but lack transparency, making it difficult to understand how they make decisions. Their opacity often lead to trust issues, difficulty in diagnosing errors and challenges in meeting regulatory requirements in sensitive applications like healthcare. Hence, the presence of Explainable Artificial Intelligence (XAI) models aims to provide insight and make decision and working of AI systems more transparent and understandable to humans by showing hidden details such as which features are important and how they are related [9], [10]. It explains the how, why and when [11]. SHAP is one of the XAI techniques that effectively explain more complex dependencies and relationships. SHAP can aggregate feature importance over the entire dataset and provide a global view of which features are most influential in the dataset [12]. Even though CARM identifies important patterns and relationships within the dataset, validating these patterns requires a deep understanding of how they affect the predictions of the model. Hence, SHAP can help in this validation process by clarifying how specific patterns influence the outcome of the model.

The primary objective of this study is to explore interpretable and clinically relevant patterns for breast cancer diagnosis by integrating CARM with SHAP-based explainability. By combining rule-based learning with model interpretation, this study proposes to identify robust combinations of biological features that effectively differentiate between healthy individuals and patients. Unlike previous research that primarily relied on CARM alone, our proposed method provides a dual-layered framework for uncovering complex patterns and validating them rigorously. This integration also aims to improve the reliability of extracted CAR through SHAP validation.

The paper is organized as follows: Section 2 reviews related works on ARM in breast cancer diagnosis. Section 3 covers the background theory including fuzzy sets and ARM concepts. Section 4 details the methodology including the data descriptions and the proposed research framework. Subsequently, Section 5 presents the results of the experiments. Finally, Section 6 summarizes the findings and suggests potential future research.

Related Works

ARM has been widely explored in breast cancer diagnosis and risk factors due to its ability to uncover significant patterns and relationships within datasets. ARM has proven valuable in identifying risk factors and predictive patterns for the disease. A notable study by [13] applied ARM using Apriori and FP-Growth algorithms to NED-breast cancer datasets to discover the patterns through the relationship among features began from 1-dimensional, 2-dimensional, 3-dimensional, and n-dimensional. The association result of both algorithms was almost similar with the 10-highest confidence value representing 100% confidence with support value up to 50%. Similarly, Kabir *et al*. [14] used ARM and CAR to analyze Breast Cancer Surveillance Consortium (BCSC) data to discover rule patterns of the risk factors of this disease. They first utilized the logit model to identify appropriate factors that may affect the likelihood of breast cancer. Subsequently, the significant rules of both non-breast cancer and breast cancer patients were obtained. Besides, Oladipupo *et al*. [15] employed an Interval Type-2 fuzzy ARM approach to explore the pattern discovery in the Wisconsin Original Breast Cancer (WOBC) dataset. The fuzzification of the dataset was carried out using the Hao and Mendel approach. The study obtained the associative rules with a minimum number of symptoms at confidence values as high as 91%. They also identified *High Bare Nuclei* and *High Uniformity of Cell Shape* as strong determinant factors for diagnosing breast cancer.

Integrating XAI techniques such as SHAP can enhance the interpretability of machine learning models. For example, Khater *et al*. [16] employed different ML algorithms such as Support Vector Machine (SVM), K-Nearest Neighbors (KNN), RF and Extreme Gradient Boost (XGB) on breast cancer dataset. The study explained the model behavior using three model-agnostic techniques including permutation importance, partial dependence plots and SHAP. The results had shown that the most important features were the *bare_nuclei* feature in the WOBC and the *area_worst* feature for Wisconsin Diagnostic Breast Cancer (WDBC) dataset. Further research by [17] emphasized how XAI can identify and explore the key factors that influence breast cancer recurrence. They utilized RF and TreeSHAP to identify the top five significant features which were tumor size, clinical stage III, total metastatic lymph nodes, risk of recurrence and age. Moreover, Suresh *et al*. [18] investigated the SHAP of the XGB model for the WDBC dataset. The SHAP explanation indicated that *perimeter* and *concave_points* had the highest impact on breast cancer diagnosis with an accuracy of 98.42%. A notable study by [19] had utilized XGB and SHAP to investigate the survival analysis in breast cancer. The study found that the XGB model was capable of capturing interaction effects between the features. For instance, a notable interaction was between age and pathological tumor stage (PTS). The study revealed that patients aged 20 to 60 with lower PTS (I and IIA) have a lower mortality risk than those with higher PTS (IIB, IIIA, IIIB, and IIIC). However, this difference decreases significantly for patients older than 60. This approach shows that the XAI model can generate explicit knowledge of how models make their predictions which is very important in increasing the trust in oncology and healthcare.

Previous studies have also shown that the ARM efficiently identified valuable patterns for breast cancer diagnosis. However, there is a lack of research on understanding how individual features contribute to predictions to ensure the discovered patterns by ARM are accurate and meaningful. Since the interpretability of the patterns obtained is crucial, this study proposed the integration of CAR and SHAP methods to ensure the identified patterns give clear insights into the end user such as health professionals to make better decision-making in breast cancer diagnosis.

Background Theory

This section discussed the definition of fuzzy sets and triangular membership functions. Besides, it also reviews the basic concepts of ARM and CARM with several interestingness measures employed to prune the CAR to obtain insightful patterns.

Fuzzy Sets

ARM often faces issues with numeric data since it depends on categorical features. To handle numeric data, data discretization is used to split continuous data into intervals and assign each an integer label. Traditional discretization methods encounter a "sharp boundary problem," where values close to interval edges are treated the same as those in the center which leads to information loss and reduced accuracy. Hence, this study used fuzzy sets theory where the fuzzification process converts numeric values into fuzzy values. This concept was introduced by [20] and allows the representation of uncertainty and vagueness in data. In classical set theory, an element is either strictly a member or not a member of set A, denoted as is $x \in A$ or $x \notin A$. This type of set is known as a crisp set. In contrast, a fuzzy s*et al*lows each element to have a degree of membership, which can be expressed as:

$$A = \{(x, \mu_A(x))| \; x \in X\} \tag{1}$$

where X is the universe set and $\mu_A(x): X \rightarrow [0,1]$ is the membership function that assigns a degree of membership to each element x of X ranging from 0 to 1. This degree indicates the strength of the association of the element to the fuzzy set. If $\mu_A(x) = 0$, x is not a member of A while if $\mu_A(x) = 1$, then x is a fully member of A. Values between 0 and 1 show partial membership in A.

Membership Functions

The membership function indicates the degree to which the elements belongs to a particular fuzzy set.. It includes several types such as triangular, Gaussian, trapezoidal, bell-shaped, and sigmoidal functions. This study used the triangular membership function to define the fuzzy sets that can expressed as:

$$\mu_A(X;\; a, b, c) = \begin{cases} 0 & x \leq a \\ \dfrac{x-a}{b-a} & a \leq x \leq b \\ \dfrac{c-x}{c-b} & b \leq x \leq c \\ 0 & x \geq c \end{cases} \tag{2}$$

where x is an element of a fuzzy set X, a and c denote the lower and upper boundary of fuzzy set X respectively and b represents the lower and the center of the fuzzy set X.

Association Rule Mining

Association analysis is used to find interesting relationships in large datasets. Let D represent a dataset, I is the set of items, and T is the transaction database where t denotes a transaction. A collection of k items is called a k-itemset. The support count σ of an itemset indicates how many times it appears in the dataset which can be expressed as:

$$\sigma(X) = |\{t_i | X \subseteq t_i, t_i \in T\}| \tag{4}$$

where the symbol $|\cdot|$ denotes the number of elements in a set. An AR is an implication expression of the form $X \Rightarrow Y$ where X and Y are disjoint itemset. The strength of the rule effectiveness is evaluated based on its support and confidence which can be calculated as:

$$supp(X \Rightarrow Y) = \frac{\sigma(X \cup Y)}{N} \tag{5}$$

$$conf(X \Rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} \tag{6}$$

An Apriori algorithm was introduced by [[6] for ARM that effectively handles the exponential growth of candidate itemsets by pruning based on support. It operates in two main steps:
(i) iteratively identifying frequent itemsets that meet a specified support threshold
(ii) generating AR that meet a minimum confidence level using these frequent itemsets.

This approach follows a systematic progression where it moves from frequent 1-itemsets to larger frequent itemsets using a generate-and-test strategy. AR are generated iteratively until no more antecedents can be added.

CAR is a specialized type of AR that signifies relationships between items and a specific class label. Let X be the set of items, C denote the set of class labels, and X_f denote the itemset representing feature values. CAR can be defined as the form of

$$X_f \Rightarrow \{y\} \tag{7}$$

where $\{y\} \subseteq Y$. . There are various measures to evaluate the interestingness of AR [21] such as lift, leverage and conviction are utilized which can be described as follows:

Lift

Lift measures how much more frequently the antecedent X and the consequent Y occur together than if they were statistically independent. Lift is defined as:

$$\mathrm{lift}(\mathrm{X}{\Rightarrow}Y) = \frac{conf(X \Rightarrow Y)}{supp(Y)} \tag{8}$$

A lift value greater than 1 indicates a positive correlation between X and Y while a lift value less than 1 shows a negative correlation. If the lift is 1, X and Y are uncorrelated.

Leverage

Leverage calculates the difference between the actual frequency of X and Y occurring together and the frequency expected if X and Y were independent. A leverage value of 0 indicates no association. Leverage is calculated as:

$$\mathrm{conviction}(\mathrm{X}{\Rightarrow}Y) = \frac{1 - supp(Y)}{1 - conf(X \Rightarrow Y)} \tag{9}$$

Conviction

Conviction compares how often X would be expected to occur without Y (if X and Y were independent) to how often X actually occurs without Y. If it is greater than 1, the rule is more likely a genuine relationship rather than a random chance. Conviction can be defined as:

$$\mathrm{conviction}(\mathrm{X}{\Rightarrow}Y) = \frac{1 - supp(Y)}{1 - conf(X \Rightarrow Y)} \tag{10}$$

Methodology

This section outlines the proposed research methodology which comprises six subsections which are dataset description, proposed framework, data pre-processing, fuzzy discretization and the process of CARM. In addition, the use of SHAP plots to interpret the results of the RF and GB models is also highlighted.

Dataset Description

The BCC dataset obtained from the UC Irvine (UCI) Machine Learning Repository [22] is employed for this study. This dataset was selected for this study due to its inclusion of metabolic and biochemical features that have been associated with breast cancer risk in prior literature. Specifically, biomarkers such as glucose, insulin, BMI**,** HOMA, leptin, adiponectin and resistin have been identified as potential contributors to increased breast cancer risk [23], [24], [25], [26], [27]. These findings support the relevance and suitability of the dataset for investigating interpretable patterns related to breast cancer classification. The dataset includes 9 continuous features and a binary target indicating the presence or absence of breast cancer based on 116 samples from both 52 healthy individuals and 64 patients Table 1 listed all the features with brief descriptions and their range values (rounded to 2 decimal places).

Table 1. Description of features from BCC dataset

Feature Name	Feature Description (unit)	Range [Min-Max]
Age	Age of the patient (years)	24.00-89.00
BMI	Body Mass Index (kg/m²)	18.37-38.58
Glucose	Glucose level (mg/dL)	60.00-201.00
Insulin	Insulin level (µU/mL)	2.43-58.46
HOMA	Homeostatic Model Assessment	0.47-25.05
Leptin	Leptin level (ng/mL)	4.31-90.28
Adiponectin	Adiponectin level (µg/mL)	1.66-38.04
Resistin	Resistin level (ng/mL)	3.21-82.1
MCP.1	Monocyte Chemoattractant Protein-1 (pg/dL)	45.84-1698.44
Classification	Labels: 1 = Healthy controls, 2 = Patients	1-2

Previous research has shown its effectiveness in exploratory modeling and interpretable analysis using methods like fuzzy rule systems, SHAP-based explanations and metaheuristic feature selection in BCC dataset [28], [29], [30]. The inclusion criteria required samples to have complete values for all features while records were excluded if they contained any missing data. The distribution of these features for both classes was shown in Figure 1 using swarm plots.

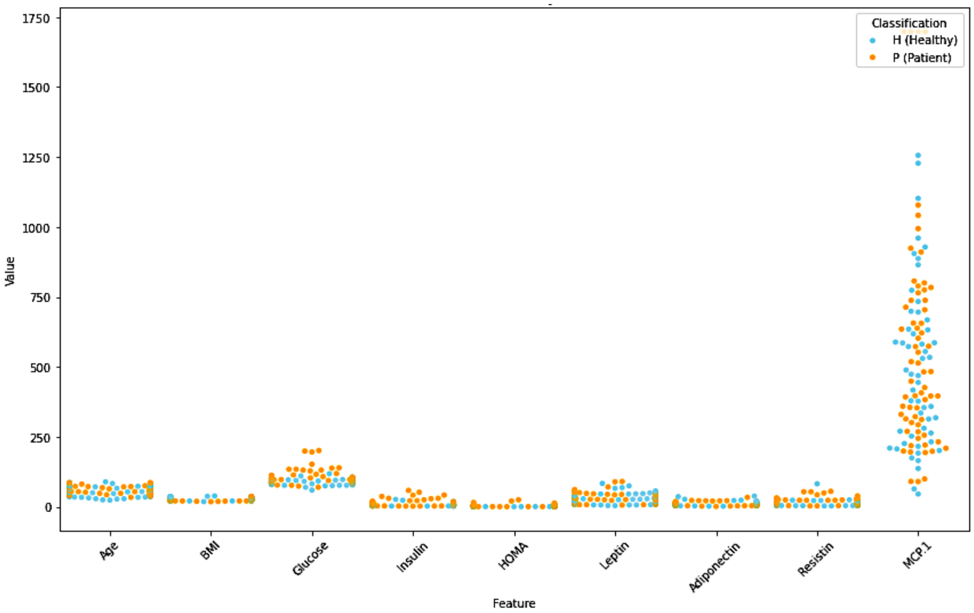


Figure 1. Distribution of BCC features

Figure 2 displays a heatmap that illustrated correlations between the features and provided insights into the relationships and patterns among variables within the dataset. The heatmap visualizes the linear relationships between features in the dataset with color coding ranging from blue to orange. Blue hues denote negative correlations while orange hues indicate positive correlations with the intensity of the color reflecting the strength of the correlation.

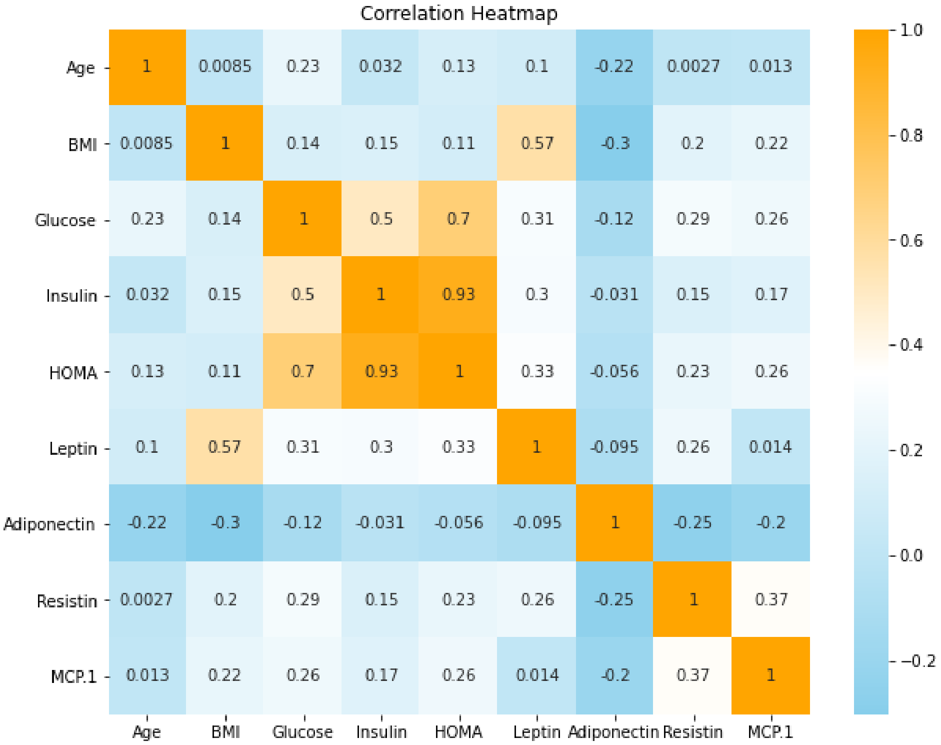


Figure 2. Correlation heatmap showing positive (orange) and negative (blue) relationships among biomarkers where deeper shades indicate stronger associations

From Figure 2, it can be observed that insulin and HOMA exhibited a very strong positive correlation (0.93). This can be translated as increased insulin levels are significantly associated with higher HOMA values. Similarly, Glucose and HOMA also showed a strong positive correlation (0.7) denoting that higher glucose levels responded to increased HOMA values. A moderate positive correlation is observed between BMI and leptin (0.57) suggesting an association between higher BMI and increased leptin levels. In contrast, adiponectin and age demonstrate a moderate negative correlation (-0.22) indicating that increased age is associated with lower adiponectin levels. Other variable pairs generally exhibit low to moderate correlations showing weaker or no significant linear relationships.

As the heatmap was used as an initial exploratory step to understand general relationships across the full dataset, hence it does not distinguish between healthy individuals and cancer patients. For group-specific insights were obtained later through SHAP-based interpretation to assess feature contributions to cancer prediction.

Proposed Framework

The proposed research framework begins with analyzing breast cancer data to ensure data quality by identifying and managing missing values. Fuzzy discretization techniques are then applied to transform the continuous variables into categorical intervals suitable for ARM. The Apriori algorithm with metrics such as lift, leverage, and conviction is employed to mine CAR in order to uncovering significant relationships between features and breast cancer diagnosis. The discovered patterns are validated using SHAP plots with RF and GB models to assess feature importance and confirm insights. Finally, critical knowledge from these analyses is obtained to enhance breast cancer diagnosis. The proposed research framework is illustrated in Figure 3.

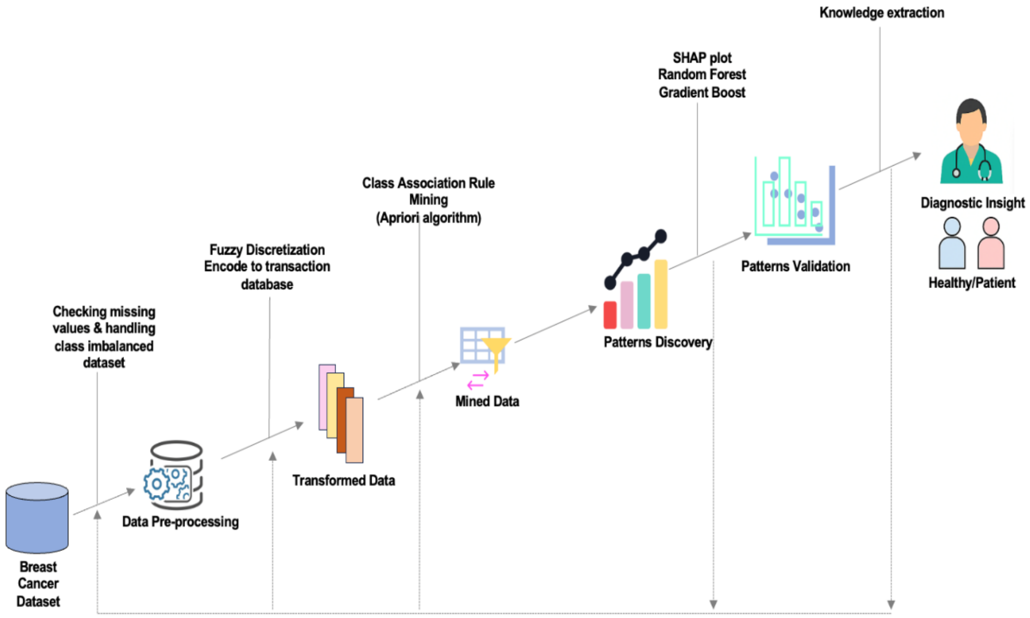


Figure 3. The proposed framework

Data Pre-Processing

Data preprocessing is a critical step to ensure the quality and relevance of the data used for analysis. The preprocessing method involves several key steps.

a. Missing values checking: This step involves verifying for the absence of missing values to ensure data cleanliness and completeness, which may arise due to human error or unavailable information.

b. Encode the categorical data: This step converts data into categorical variables to which the machine learning models can processed. In this study, the original categorical labels 'H' for Healthy and 'P' for Patients were encoded into numerical values (P=1 and H=0) to ensure compatibility with machine learning algorithms.

c. Outlier detection. Outliers are data points that significantly differ from the majority of observations in a dataset, either being unusually high or low. The Z-Score method is a common approach for detecting and removing outliers. It quantifies how far an observation deviates from the mean of the dataset which can be expressed as

$$z = \frac{(X - \mu)}{\sigma} \tag{11}$$

where X is the data point, μ is the mean of the dataset, and σ is the standard deviation. Data points with Z-Score exceeding a threshold of $z < -3, z > 3$ are typically considered outliers [31] and may be eliminated from the dataset.

d. Data transformation. Transforming raw data into a transactional format is the first step for ARM. This method treats patient records as 'transactions' which is similar to retail transactions. Each feature such as Age, BMI, and Glucose becomes an 'item' in the analysis.

e. Class imbalance handling. The dataset exhibited a class imbalance between healthy individuals and breast cancer patients. To handle this issue, the Synthetic Minority Oversampling Technique (SMOTE) technique [32] was applied to generate synthetic examples of the minority class. This approach helps to balance the class distribution and improving the classification performance.

Fuzzy Discretization

This study employs fuzzy discretization which can be carried out in five steps:

a. Identify each continuous feature and partition its range into three intervals: 'Low/Young', 'Medium/Middle', and 'High/Old' based on value thresholds depending on the suitability of the feature characteristics.

b. Use triangular membership functions to create a fuzzy set after outlier detection using the Z-Score method.

c. Assign membership values to data within each interval to define discrete boundaries.
d. Derive discrete attribute values by selecting the interval with the highest membership value for each data point.
e. Create a finalized dataset with fuzzy discretized attribute values.

Mining Class Association Rules

CAR are generated to discover meaningful patterns between features with a focus on the “Classification” label. The Apriori algorithm is applied using minimum support and confidence thresholds. To evaluate the quality and relevance of the extracted association rules, we employed three objective interestingness measures such as lift, leverage and conviction. These metrics provide a deeper statistical understanding of the relationships between features beyond basic support and confidence. Lift is chosen for its ability to identify strong and meaningful associations between symptoms and diagnosis. Meanwhile, leverage is used to quantify the strength and statistical significance of these associations beyond what support and confidence provide. Besides, conviction ensures the reliability and robustness of the rules while reducing the likelihood of misleading patterns. These metrics were selected due to their effectiveness in identifying statistically meaningful patterns as demonstrated in recent studies across medical and bioinformatics domains [21], [33], [34]. Only rules meeting all predefined thresholds for these metrics are retained for further analysis. This process is illustrated in Figure 4.

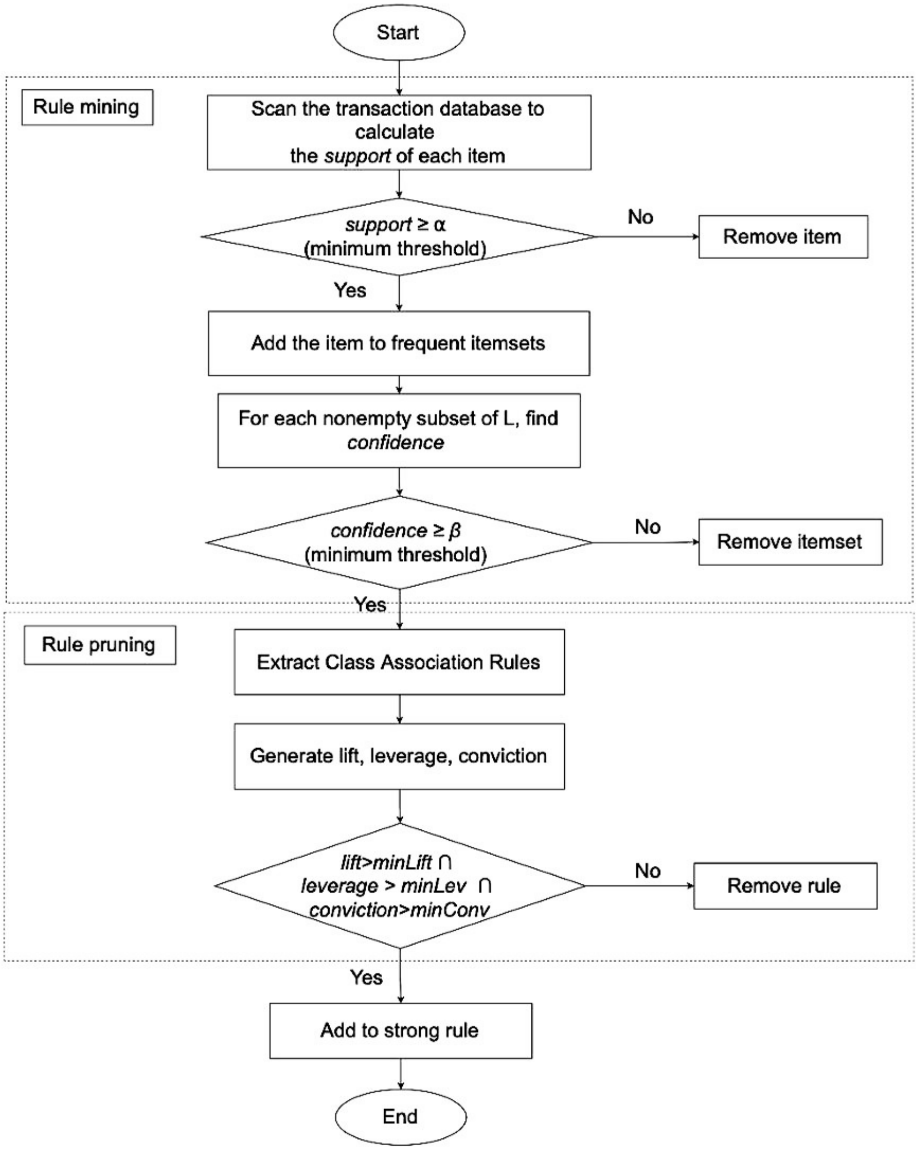

Start
Rule mining
Scan the transaction database to calculate the *support* of each item
support ≥ α (minimum threshold)
No
Remove item
Yes
Add the item to frequent itemsets
For each nonempty subset of L, find *confidence*
confidence ≥ *β* (minimum threshold)
No
Remove itemset
Yes
Rule pruning
Extract Class Association Rules
Generate lift, leverage, conviction
lift>minLift ∩ *leverage > minLev* ∩ *conviction>minConv*
No
Remove rule
Yes
Add to strong rule
End

Figure 4. Workflow of rule pruning procedure

SHAP with Random Forest and Gradient Boost

SHAP values are a comprehensive method to understand the machine learning model which is derived from principles of cooperative game theory [12]. These values measure how much each feature contributes to the output of the model. SHAP can improve the transparency of complex models and provide insights into how individual features affect predictions. It can be calculated as:

$$\varphi_i(f) = \sum_{S\subseteq N\setminus\{i\}} \frac{|S|!\,(N-|S|-1)!}{N!}[f(S\cup\{i\}) - f(S)] \tag{12}$$

where N is the set of all features and S is a subset of features that excludes the i-th feature. $f(S)$ is the prediction from the model when the only features in subset S are present while $f(S\cup\{i\})$ is the prediction from the model when the features in subset S and feature i are present. This formula measures the marginal contribution of feature i by considering all possible subsets of features. The machine learning models used with SHAP values in this study are as follows:

a. Random Forest. RF is an ensemble learning method used for classification that constructs multiple decision trees during training [35]. Each tree is built on a different random subset of the data and features. The final prediction is made by taking the majority vote on all the trees' predictions. RF ensures diversity among its trees by selecting the best feature from a random subset of features thereby reducing the risk of overfitting [36]. The prediction of RF can be represented as:

$$RF(x) = \text{mode}(\{f_1(x), f_2(x), \ldots, f_T(x)\}) \tag{13}$$

where T is the number of decision trees and $f_T(x)$ is the prediction of the T-th decision tree for the input data.

b. Gradient Boost. GB is an ensemble learning technique that sequentially builds a series of weak learners which typically decision trees where each subsequent learner corrects the errors of its predecessor. It optimizes a loss function using gradient descent to refine the predictions of the model. GB can be expressed as [37]:

$$\gamma_{jm} = \arg min_\gamma \sum_{x_i\in R_{jm}} -(y(F_{m-1}(x_i)+\gamma) - \log\left(1+e^{F_{m-1}(x_i)+\gamma}\right)) \tag{14}$$

where γ_{jm} is the value of γ that minimizes the given loss function for the j-th region R_{jm} on the m-th tree and $F_{m-1}(x)$ is the prediction of the ensemble model up to $(m-1)$-th iterations combining the initial prediction and the contributions from all previous base learners.

Results and Discussions

This section presented the result of the fuzzy discretization on the continuous features, the mining of the CAR based on various defined minimum parameters, the key patterns of the breast cancer diagnosis and the validation of the patterns using SHAP plots with RF and GB classifiers. All data pre-processing and analysis were conducted using the Python programming language in the Spyder 5.4.3 environment. The experiments were executed on a 64-bit Windows 11 Pro system with a 2.5 GHz processor and 32 GB of RAM. The experiments employed a hold-out validation strategy with 2/3 of the data allocated for training and the remaining 1/3 reserved for testing [38].

Fuzzy Discretization of the Continuous Features

In this study, fuzzy discretization transformed continuous features into a set of linguistic terms to help better interpretability of the AR. This phase involved removing outliers, generating fuzzy sets using triangular membership functions, and assigning membership values to data points within each interval to define discrete boundaries. For instance, consider the feature '*Leptin*' in the dataset which originally ranged from 4.31 to 90.28 narrowed to 4.31 to 83.48 after removing outliers. This removal ensured that the values of features were within a relevant range for analysis. The boxplots in Figure 5 (a) and (b) illustrated the distribution of '*Leptin*' before and after the removal of outliers respectively. Subsequently, the values of the membership values were converted to linguistic intervals such as 'Low', 'Medium', and 'High'. The membership function of these intervals of the 'Leptin' feature can be observed in Figure 6.

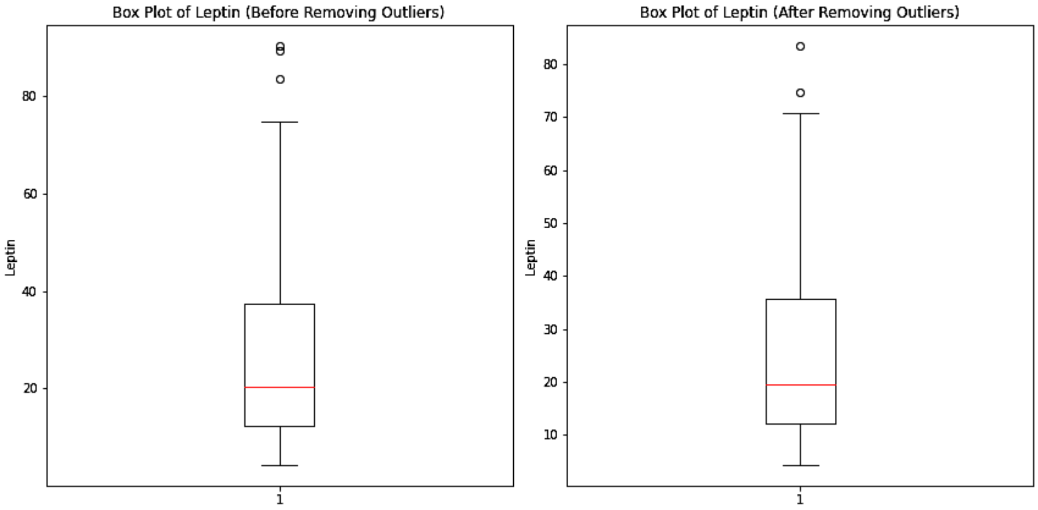


Figure 5. Box plot of *Leptin* feature (a) before and (b) after removal of outlier

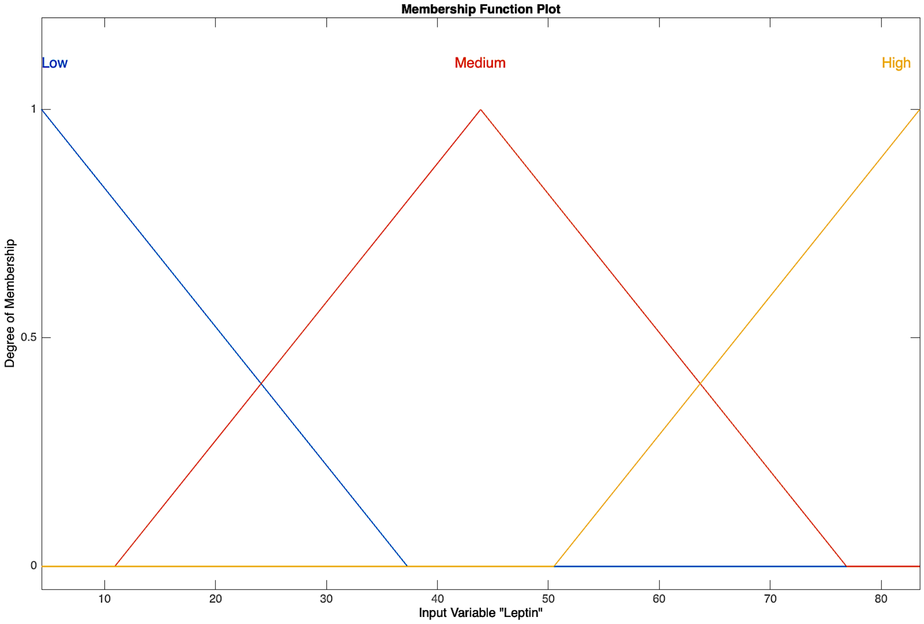


Figure 6. The triangular membership function of *Leptin* feature

Mining Insightful Rules from the Association Rule Mining

Mining insightful rules from ARM is a crucial step in understanding the underlying patterns and relationships within the dataset. Initially, the support counts were examined to identify the most common features in the dataset. Features with high support counts occur frequently and are important for generating meaningful rules. Understanding the distribution of support counts helps set appropriate minimum support thresholds for ARM. If the threshold is set too high, potentially valuable rules involving less frequent features may be ommited, whereas setting it too low may produce of numerous insignificant rules. Figure 7 depicted the support counts for each feature in the dataset.

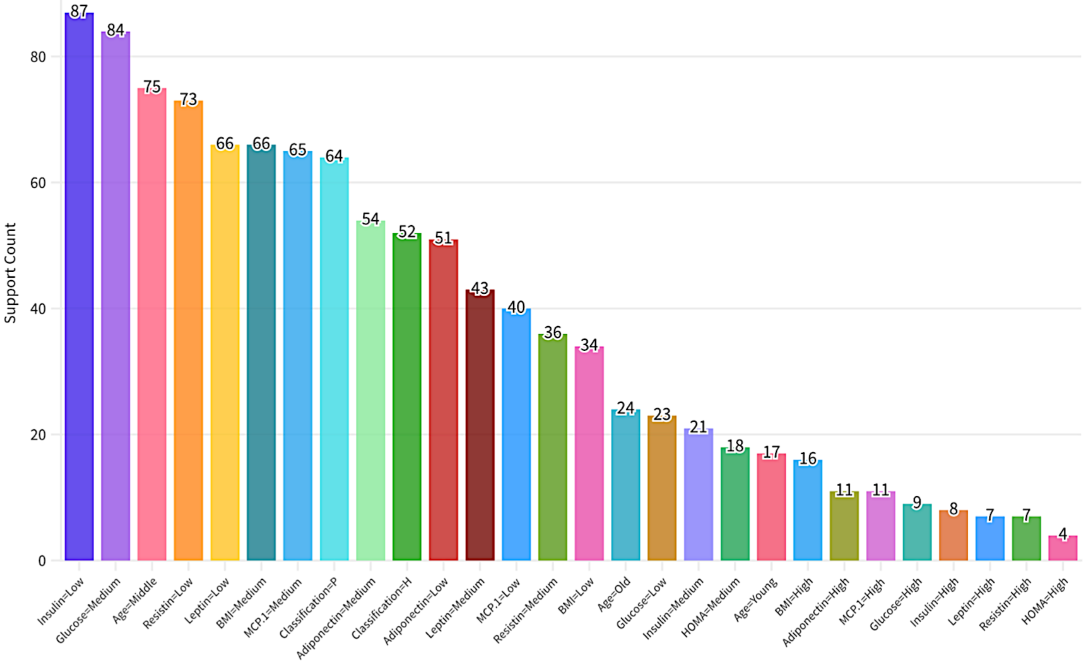


Figure 7. Support count of each feature values

As shown in Figure 7, "Insulin=Low" has the highest support count at 87, followed closely by "Glucose=Medium" at 84, indicating that these features are among the most prevalent. Features like "MCP.1=Medium", "Classification=P", and "Adiponectin=Medium" have moderate support counts ranging from 54 to 64. In contrast, features such as "Insulin=High", "Leptin-High", "Resistin=High", and "HOMA-high" have low support counts ranging from 8 to 4. This distribution helped in setting the thresholds for minimum support in ARM to obtain significant rules. High support features are likely to form more significant rules due to their frequency, while low support features, though less frequent, might still give valuable insights. Hence, in the process of performing ARM, selecting the appropriate minimum support (*minsu*p) threshold is crucial because it directly affects the number of frequent itemsets generated and the computational feasibility of the analysis. In this study, the number of AR generated was analyzed based on different *minsup* (α) and minimum confidence (*minconf*, β) thresholds. These thresholds were varied to assess their impact on the quantity and quality of the rules produced. Initially, setting the α to 0.005 resulted in the generation of 46,448 frequent itemsets. Subsequently, we tested higher α of 0.01, 0.02, 0.03, 0.04, 0.05, 0.10, 0.15, 0.20, 0.25 and 0.30. By increasing α, the number of frequent itemsets was significantly reduced to 18,612, 11,165, 7,920, 5,809, 4,443, 1,406, 587, 291, 179, and 104 respectively. For each α value, we also tested different levels of β such as 0.7, 0.8, 0.9, and 1.0. This helps in identifying strong AR to provide valuable insights while keeping the computational requirements manageable. Table 2 presented the number of AR identified at various α and β combinations.

Table 2. Number of AR based on certain α and β thresholds

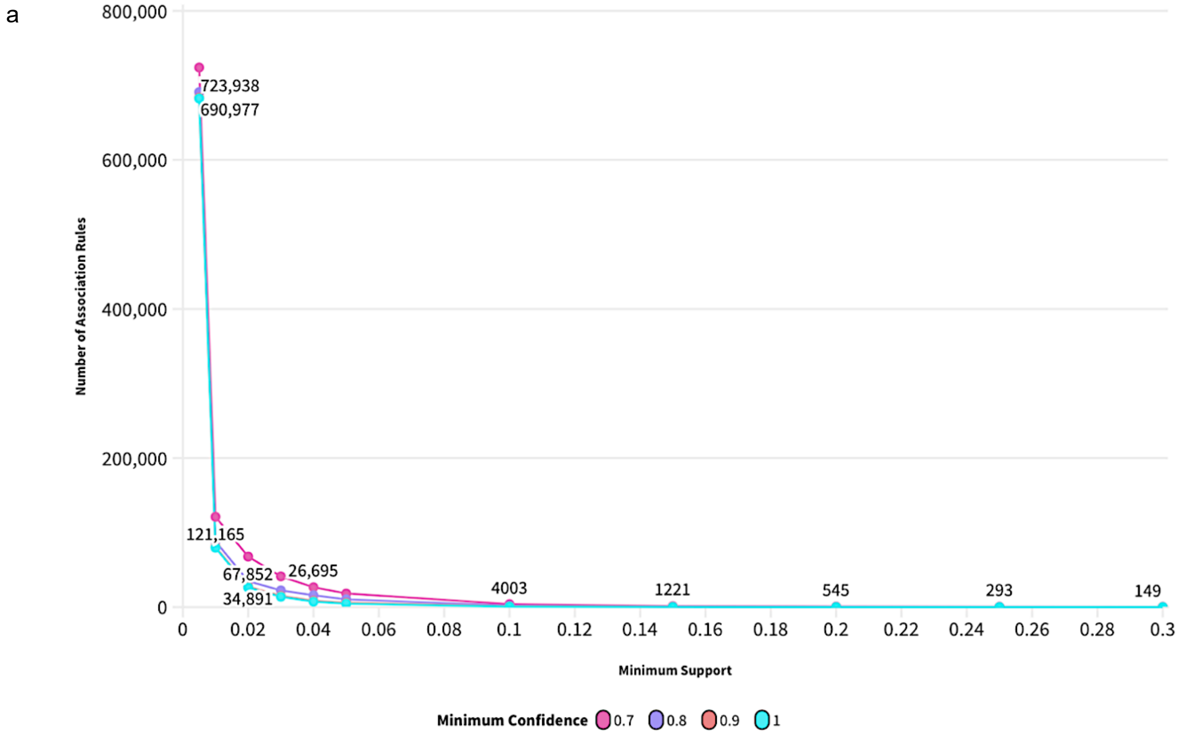
α/β	0.7	0.8	0.9	1.0
0.005	723,938	690,977	683,463	682,415
0.01	121,165	88,204	80,690	79,642
0.02	67,852	34,891	27,377	26,329
0.03	41,277	22,422	14,908	13,860
0.04	26,695	15,851	8,337	7,289
0.05	18,319	10,368	5,669	4,621
0.1	4,003	2,221	1,142	617
0.15	1,221	670	351	149
0.2	545	287	143	67
0.25	293	145	74	36
0.3	149	76	40	21

723,938 AR were identified at α of 0.2 and β of 0.7. This number drastically drops to 121,165 when the β is raised to 0.01. The reduction is even more obvious from 617 to 21 at α of 0.1 and 0.3 respectively at β of 1.0. These findings indicated that lower α and β values yielded a higher number of rules which may include many less significant ones, while higher thresholds produced fewer but potentially more meaningful and robust rules. Table 3 showed the number of CAR was identified at various α and β combinations.

Table 3. Number of CAR based on certain α and β thresholds

α/β	0.7	0.8	0.9	1.0
0.005	17,720	16,693	16,284	16,260
0.01	5,456	4,429	4,020	3,996
0.02	3,024	1,997	1,588	1,564
0.03	1,911	1,236	827	803
0.04	1,268	821	412	388
0.05	839	480	214	190
0.1	150	69	16	7
0.15	28	8	0	0
0.2	13	2	0	0
0.25	6	0	0	0
0.3	1	0	0	0

According to Table 3, at a β of 0.7, the number of CAR decreases from 17,720 to 1 only at a α of 0.005 and 0.3 respectively. The pattern was consistent across different β thresholds for each α. It can be observed that there was no strong CAR starting on α of 0.15 for higher thresholds of β. Figures 7(a) and 7(b) presented the number of AR and CAR respectively across the different α and β thresholds. The graphs clearly illustrate the steep decline in the number of rules as the thresholds increase.



Continue to next page

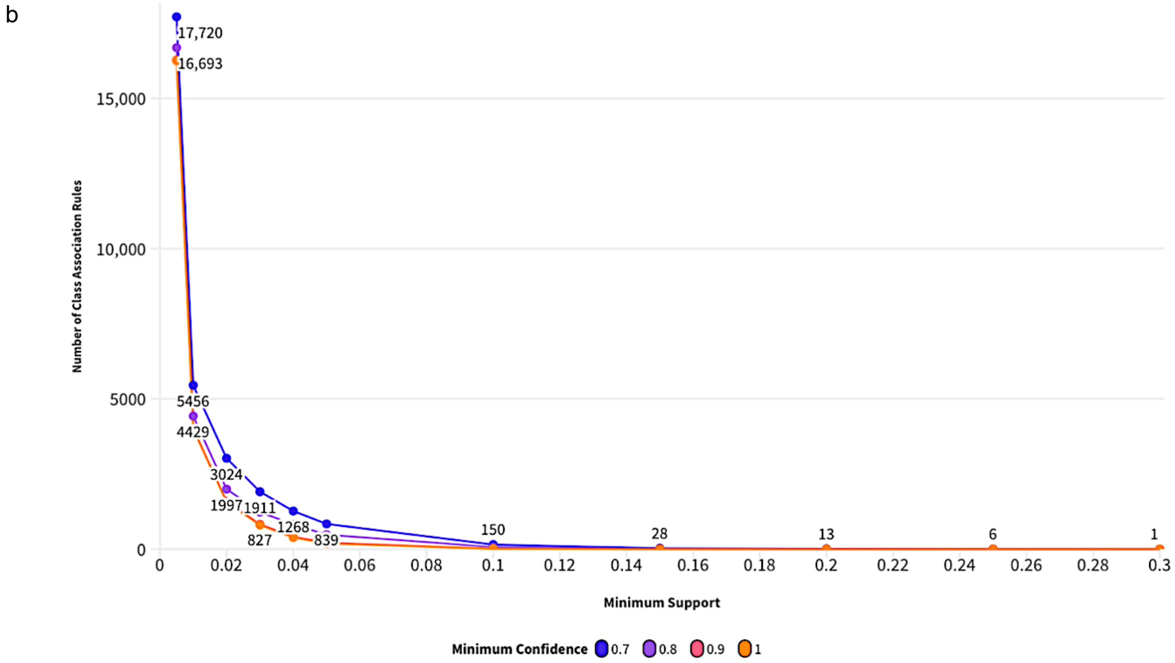


Figure 7. Number of (a) AR (b) CAR with different α and β thresholds

Comparing AR and CAR results had shown a clear trend where AR produced a large number of rules at lower thresholds, while CAR generated fewer rules that were more focused on class distinctions. For instance, with a α of 0.2 and β of 0.7, AR identified 723,938 rules, whereas CAR identified only 17,720 rules at the lowest α of 0.005. This highlights that while AR can discover broader patterns, CAR is more effective in identifying class-specific insights, crucial for breast cancer diagnosis. The results demonstrate the importance of adjusting thresholds to achieve a balance between the number and significance of the rules generated.

Key Patterns for Breast Cancer Diagnosis from Strong CAR

The process of identifying key patterns from strong CAR provides critical insights into the domain of breast cancer diagnosis. To obtain strong pruned CAR to filter out weaker associations, some pruning criteria were chosen such as *minLift* = 1.5, *minLev* = 0.01, and *minConv* = 1.4.These thresholds were determined through a trial-and-error approach during exploratory analysis by adjusting them to fit the characteristics and size of BCC dataset. Instead of relying on standard benchmarks, this dataset-specific tuning ensured selection of only the most meaningful and statistically relevant rules for further validation. The statistics of average support, confidence, lift, leverage, and conviction values for the generated pruned CAR based on different itemset sizes are evaluated based on two different classes of the target variable including “Classification=H” and “Classification=P”. The statistics of two classes of target variables are presented in Table 4 and Table 5 respectively and are visualized in Figure 8 and Figure 9.

Table 4. Number of pruned CAR with average metrics values for *Classification=H*

Size of subsets (n)	Number of CAR	Number of pruned CAR	Average support values	Average confidence values	Average lift values	Average leverage values	Average conviction values
2	51	43	0.0678	0.8384	1.8703	0.0306	2.9601
3	365	187	0.0509	0.8781	1.9589	0.0243	2.9185
4	1187	390	0.0415	0.9063	2.0218	0.0205	2.8107
5	2044	430	0.0358	0.9289	2.0722	0.0182	2.6994
6	2002	258	0.0326	0.9478	2.1142	0.0169	2.5732
7	1118	87	0.0306	0.9654	2.1536	0.0162	2.4191
8	332	18	0.0283	0.9861	2.1998	0.0154	2.2069
9	41	2	0.0259	1.0000	2.2308	0.0143	inf

Based on Table 4, as the itemset size increases, the number of generated pruned CAR also increases until n=4 and reduces significantly until n=9. For itemsets of size n=2, 51 CAR were generated with 43 pruned CAR met the pruned defined criteria. For itemsets of size n=3, 365 CAR rules were generated, with 187 showing significant interdependence between antecedents and target variables of healthy samples. The maximum itemset size observed was n=9 generating a total of 41 CAR with only 2 strong CAR obtained.

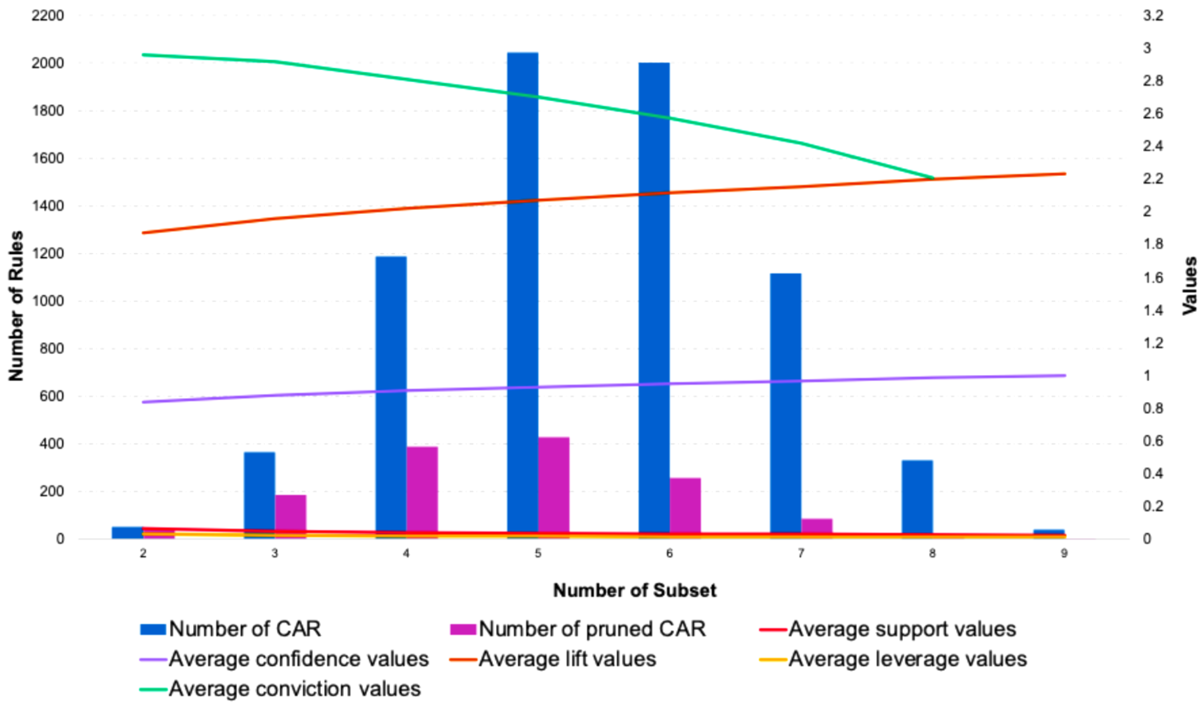


Figure 8. Graphs of the number of CAR with average metrics values for *Classification=H*

Table 5. Number of pruned CAR with average metrics values for *Classification=P*

Size of subsets (n)	Number of CAR	Number of pruned CAR	Average support values	Average confidence values	Average lift values	Average leverage values	Average conviction values
2	124	44	0.0458	0.9664	1.7515	0.0190	3.1080
3	720	157	0.0434	0.9731	1.7638	0.0180	3.3563
4	1994	253	0.0432	0.9594	1.7389	0.0176	3.2421
5	3068	231	0.0430	0.9614	1.7425	0.0177	3.3831
6	2750	154	0.0426	0.9559	1.7326	0.0175	3.2238
7	1449	63	0.0404	0.9622	1.7441	0.0168	3.0819
8	416	16	0.0356	0.9818	1.7795	0.0153	3.1379
9	50	2	0.0302	1.0000	1.8125	0.0135	inf

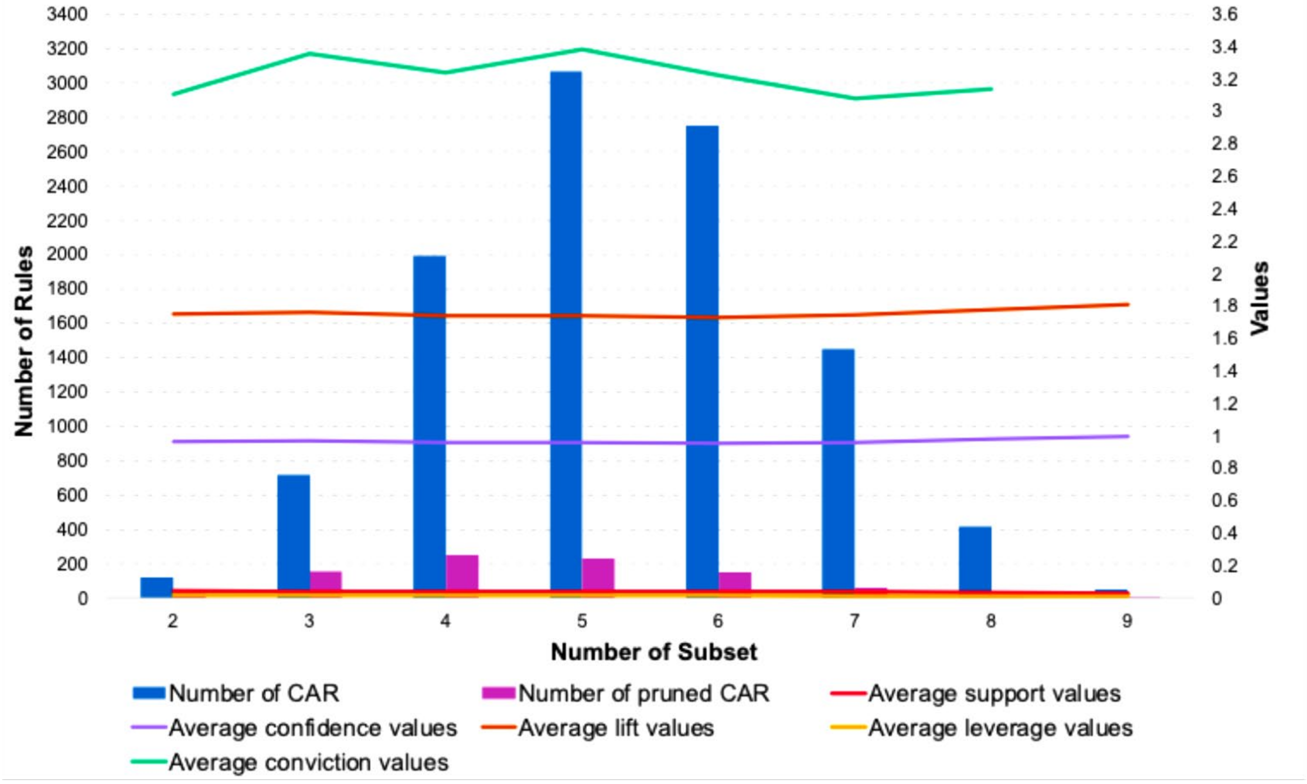


Figure 9. Graphs of the number of CAR with average metrics values for *Classification=P*

Table 6 lists the top pruned CAR for identifying the "H" (Healthy) classification with their corresponding number of subset size, confidence, lift, leverage and conviction values.

Table 6. Top pruned CAR for *Classification=H* with their confidence, lift, leverage and conviction values

Antecedent	Consequent	Number of Itemset	α	β	Lift	Leverage	Conviction
{Leptin=Low, Age=Old}	{Classification=H}	2	0.1121	1.000	2.2308	0.0618	inf
{Age=Old, BMI=Low}			0.0517	1.000	2.2308	0.0285	inf
{Age=Young, Glucose=Low}			0.0431	1.000	2.2308	0.0238	inf
{Adiponectin=Medium, Age=Young}			0.0345	1.000	2.2308	0.0190	inf
{BMI=High, Age=Young}			0.0345	1.000	2.2308	0.0190	inf
{BMI=Low, Age=Young, Adiponectin=High}	{Classification=H}	3	0.0259	1.0000	2.2308	0.0143	inf
{HOMA=Low, Age=Young, Adiponectin=High}			0.0259	1.0000	2.2308	0.0143	inf
{Age=Young, Insulin=Low, Adiponectin=High}			0.0259	1.0000	2.2308	0.0143	inf
{BMI=Low, Insulin=Low, Adiponectin=High}			0.0431	1.0000	2.2308	0.0238	inf
{BMI=Low, MCP.1=Low, Adiponectin=High}			0.0259	1.0000	2.2308	0.0143	inf
{BMI=Low, HOMA=Low, Age=Young, Adiponectin=High}	{Classification=H}	4	0.0259	1.0000	2.2308	0.0143	inf
{BMI=Low, Age=Young, Insulin=Low, Adiponectin=High}			0.0259	1.0000	2.2308	0.0143	inf
{Resistin=Low, BMI=Low, Age=Young, Adiponectin=High}			0.0259	1.0000	2.2308	0.0143	inf
{HOMA=Low, Age=Young, Insulin=Low, Adiponectin=High}			0.0259	1.0000	2.2308	0.0143	inf
{Resistin=Low, HOMA=Low, Age=Young, Adiponectin=High}			0.0259	1.0000	2.2308	0.0143	inf

Antecedent	Consequent	Number of Itemset	α	β	Lift	Leverage	Conviction
{BMI=Low, Insulin=Low, Adiponectin=High, HOMA=Low, Age=Young}	{Classification=H}	5	0.0259	1.0000	2.2308	0.0143	inf
{Resistin=Low, BMI=Low, Adiponectin=High, HOMA=Low, Age=Young}			0.0259	1.0000	2.2308	0.0143	inf
{Resistin=Low, BMI=Low, Insulin=Low, Adiponectin=High, Age=Young})			0.0259	1.0000	2.2308	0.0143	inf
{Resistin=Low, Insulin=Low, Adiponectin=High, HOMA=Low, Age=Young}			0.0259	1.0000	2.2308	0.0143	inf
{BMI=Low, Glucose=Medium, Insulin=Low, Adiponectin=High, HOMA=Low}			0.0259	1.0000	2.2308	0.0143	inf
{Resistin=Low, BMI=Low, Insulin=Low, Adiponectin=High, HOMA=Low, Age=Young}	{Classification=H}	6	0.0259	1.0000	2.2308	0.0143	inf
{Resistin=Low, BMI=Low, Glucose=Medium, Insulin=Low, Adiponectin=High, HOMA=Low}			0.0259	1.0000	2.2308	0.0143	inf
{Resistin=Low, BMI=Low, Insulin=Low, Adiponectin=High, Leptin=Low, HOMA=Low}			0.0259	1.0000	2.2308	0.0143	inf
{Resistin=Low, BMI=Low, HOMA=Low, Insulin=Low, Adiponectin=High, MCP.1=Low}			0.0259	1.0000	2.2308	0.0143	inf
{Glucose=Medium, Insulin=Low, Age=Middle, BMI=High, HOMA=Low, Adiponectin=Low}			0.0259	1.0000	2.2308	0.0143	inf
{BMI=Medium, Insulin=Low, Glucose=Low, Age=Middle, Leptin=Low, HOMA=Low, Adiponectin=Low}	{Classification=H}	7	0.0259	1.0000	2.2308	0.0143	inf
{BMI=Medium, Insulin=Low, MCP.1=Medium, Glucose=Low, Age=Middle, HOMA=Low, Adiponectin=Low}			0.0259	1.0000	2.2308	0.0143	inf
{Resistin=Low, BMI=Medium, Insulin=Low, Glucose=Low, Age=Middle, HOMA=Low, Adiponectin=Low}			0.0259	1.0000	2.2308	0.0143	inf
{BMI=Medium, MCP.1=Medium, Glucose=Low, Age=Middle, Leptin=Low, HOMA=Low, Adiponectin=Low}			0.0259	1.0000	2.2308	0.0143	inf
{Resistin=Low, BMI=Medium, Glucose=Low, Age=Middle, Leptin=Low, HOMA=Low, Adiponectin=Low}			0.0259	1.0000	2.2308	0.0143	inf
{BMI=Medium, Insulin=Low, MCP.1=Medium, Glucose=Low, Age=Middle, Leptin=Low, HOMA=Low, Adiponectin=Low}	{Classification=H}	8	0.0259	1.0000	2.2308	0.0143	inf
{Resistin=Low, BMI=Medium, Insulin=Low, Glucose=Low, Age=Middle, Leptin=Low, HOMA=Low, Adiponectin=Low}			0.0259	1.0000	2.2308	0.0143	inf
{Resistin=Low, BMI=Medium, Insulin=Low, MCP.1=Medium, Glucose=Low, Age=Middle, HOMA=Low, Adiponectin=Low}			0.0259	1.0000	2.2308	0.0143	inf
{Resistin=Low, BMI=Medium, MCP.1=Medium, Glucose=Low, Age=Middle, Leptin=Low, HOMA=Low, Adiponectin=Low}			0.0259	1.0000	2.2308	0.0143	inf
{Resistin=Low, BMI=Medium, Insulin=Low, MCP.1=Medium, Glucose=Low, Age=Middle, Leptin=Low, Adiponectin=Low}			0.0259	1.0000	2.2308	0.0143	inf
{Resistin=Low, BMI=Medium, Insulin=Low, MCP.1=Medium, Glucose=Low, Age=Middle, Leptin=Low, HOMA=Low, Adiponectin=Low}	{Classification=H}	9	0.0259	1.0000	2.2308	0.0143	inf
{Resistin=Low, Adiponectin=Medium, Age=Old, BMI=Medium, Glucose=Medium, HOMA=Low, Insulin=Low, Leptin=Low, MCP.1=Low}			0.0259	1.0000	2.2308	0.0143	inf

Note: inf = infinite

Table 6 presents the top pruned CAR associated with the healthy classification where all rules demonstrate a confidence value of 1.0 and an infinite conviction score. The infinite conviction values observed in these rules reflect perfect confidence which means that the presence of the antecedent features consistently predicts a healthy outcome within the dataset. In order to avoid overfitting often associated with such high-confidence rules, additional rule pruning based on multiple metrics such as lift and leverage was applied to ensure the significant and reliable patterns obtained. Among the 2-itemset rules, notable combinations include {Leptin=Low, Age=Old}, {BMI=Low, Age=Old}, and {Glucose=Low, Age=Young}. These rules suggest that lower levels of leptin, BMI or glucose in conjunction with specific age groups are highly indicative of a healthy classification. Interestingly, {BMI=High, Age=Young} also appears as a valid rule indicating that higher BMI may not always signify risk particularly among younger individuals.

As the number of features increases, patterns involving Insulin=Low, HOMA=Low, Resistin=Low and Glucose=Low or Medium become more significant. These are often combined with Age=Young, BMI=Low and MCP.1=Low. In more complex rules including 7 to 9 features, healthy classifications are still observed even when some moderate-risk values such as BMI=Medium or MCP.1=Medium are present. For example, the rule {Resistin=Low, BMI=Medium, Insulin=Low, MCP.1=Medium, Glucose=Low, Age=Middle, Leptin=Low, HOMA=Low, Adiponectin=Low} continues to predict the healthy class despite the presence of medium BMI or MCP.1 levels. This implies that the cumulative influence of low glucose, insulin, resistin, and HOMA levels may outweigh the moderate risk associated with other features leading to healthy classification These pruned CAR successfully highlight the patterns of healthy people using various biological markers in the dataset. The patterns of “Classification=H” are visualized in the network graph in Figure 10.

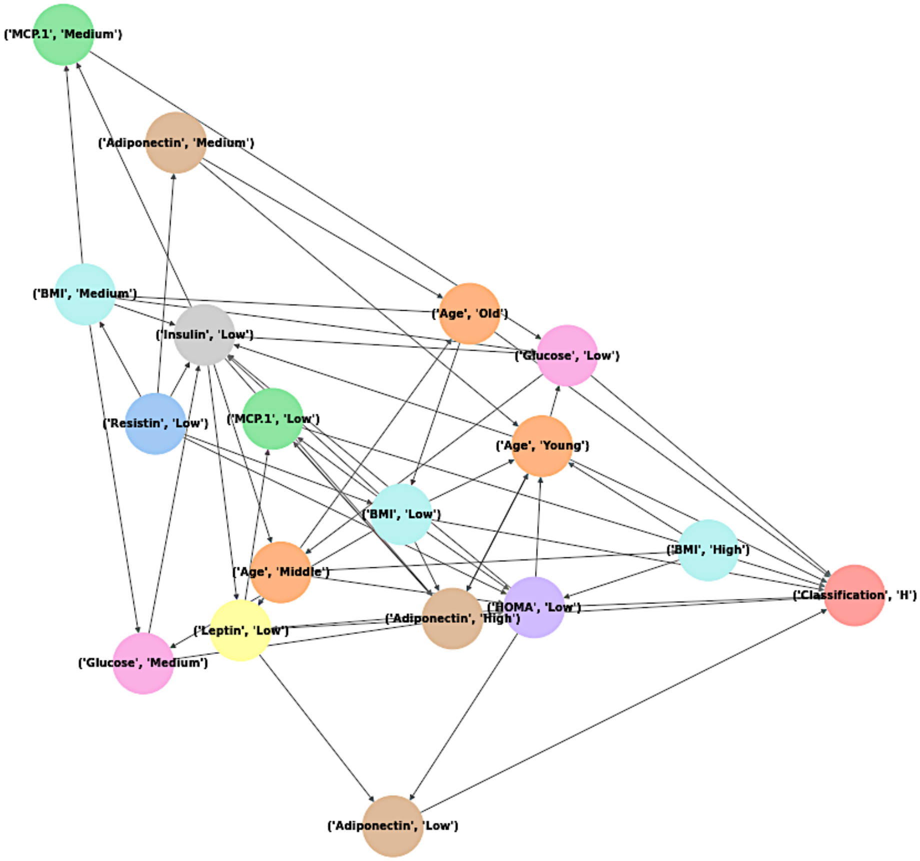


Figure 10. Network graph for classification=H

Table 7 lists the top pruned CAR for identifying the "P" (Patient) classification along with their subset sizes, confidence, lift, leverage and conviction values.

Table 7. Top pruned CAR for *Classification=P* with their confidence, lift, leverage, and conviction values

Antecedent	Consequent	Number of Itemset	α	β	Lift	Leverage	Conviction
{Glucose=High, Adiponectin=Medium}	{Classification=P}	2	0.0517	1.0000	1.8125	0.0232	inf
{BMI=Medium, Adiponectin=High}			0.0431	1.0000	1.8125	0.0193	inf
{Insulin=High, Adiponectin=Low}			0.0345	1.0000	1.8125	0.0155	inf
{Glucose=High, Adiponectin=Low}			0.0259	1.0000	1.8125	0.0116	inf
{HOMA=High, Adiponectin=Low}			0.0259	1.0000	1.8125	0.0116	inf
{Age=Middle, MCP.1=Medium, Adiponectin=High}	{Classification=P}	3	0.0259	1.0000	1.8125	0.0116	inf
{HOMA=Low, BMI=Medium, Adiponectin=High}			0.0259	1.0000	1.8125	0.0116	inf
{BMI=Medium, Insulin=Low, Adiponectin=High}			0.0259	1.0000	1.8125	0.0116	inf
{BMI=Medium, Leptin=Medium, Adiponectin=High}			0.0259	1.0000	1.8125	0.0116	inf
{BMI=Medium, MCP.1=Medium, Adiponectin=High}			0.0259	1.0000	1.8125	0.0116	inf
{Glucose=Medium, BMI=Medium, Age=Middle, Adiponectin=High}	{Classification=P}	4	0.0259	1.0000	1.8125	0.0116	inf
{HOMA=Low, BMI=Medium, Age=Middle, Adiponectin=High}			0.0259	1.0000	1.8125	0.0116	inf
{BMI=Medium, Age=Middle, Insulin=Low, Adiponectin=High}			0.0259	1.0000	1.8125	0.0116	inf
{BMI=Medium, Age=Middle, Leptin=Medium, Adiponectin=High}			0.0259	1.0000	1.8125	0.0116	inf
{Resistin=Low, BMI=Medium, Age=Middle, Adiponectin=High}			0.0345	1.0000	1.8125	0.0155	inf
{Resistin=Low, Glucose=Medium, BMI=Medium, Adiponectin=High, Age=Middle}	{Classification=P}	5	0.0259	1.0000	1.8125	0.0116	inf
{BMI=Medium, Insulin=Low, Adiponectin=High, Age=Middle, HOMA=Low}			0.0259	1.0000	1.8125	0.0116	inf
{Resistin=Low, BMI=Medium, Adiponectin=High, Age=Middle, HOMA=Low}			0.0259	1.0000	1.8125	0.0116	inf
{Resistin=Low, BMI=Medium, Insulin=Low, Adiponectin=High, Age=Middle}			0.0259	1.0000	1.8125	0.0116	inf
{Resistin=Low, BMI=Medium, Leptin=Medium, Adiponectin=High, Age=Middle}			0.0259	1.0000	1.8125	0.0116	inf
{BMI=Low, Glucose=Medium, Insulin=Low, Age=Middle, HOMA=Low, Adiponectin=Low}	{Classification=P}	6	0.0431	1.0000	1.8125	0.0193	inf
{BMI=Low, Glucose=Medium, Age=Middle, Leptin=Low, HOMA=Low, Adiponectin=Low}			0.0345	1.0000	1.8125	0.0155	inf
{BMI=Low, Glucose=Medium, MCP.1=Medium, Age=Middle, HOMA=Low, Adiponectin=Low}			0.0345	1.0000	1.8125	0.0155	inf
{Resistin=Low, BMI=Low, Glucose=Medium, Age=Middle, HOMA=Low, Adiponectin=Low}			0.0345	1.0000	1.8125	0.0155	inf
{Resistin=Low, BMI=Medium, Insulin=Low, Adiponectin=High, Age=Middle, HOMA=Low}			0.0259	1.0000	1.8125	0.0116	inf
{BMI=Low, Glucose=Medium, Insulin=Low, Age=Middle, Leptin=Low, HOMA=Low, Adiponectin=Low}	{Classification=P}	7	0.0345	1.0000	1.8125	0.0155	inf
{BMI=Low, Glucose=Medium, Insulin=Low, MCP.1=Medium, Age=Middle, HOMA=Low, Adiponectin=Low}			0.0345	1.0000	1.8125	0.0155	inf
{Resistin=Low, BMI=Low, Glucose=Medium, Insulin=Low, Age=Middle, HOMA=Low, Adiponectin=Low}			0.0345	1.0000	1.8125	0.0155	inf
{Resistin=Low, BMI=Low, Glucose=Medium, Age=Middle, Leptin=Low, HOMA=Low, Adiponectin=Low}			0.0345	1.0000	1.8125	0.0155	inf
{BMI=Low, Glucose=Medium, MCP.1=Medium, Age=Middle, Leptin=Low, HOMA=Low, Adiponectin=Low}			0.0259	1.0000	1.8125	0.0116	inf

Antecedent	Consequent	Number of Itemset	α	β	Lift	Leverage	Conviction
{BMI=Low, Glucose=Medium, Insulin=Low, MCP.1=Medium, Age=Middle, Leptin=Low, HOMA=Low, Adiponectin=Low}	{Classification=P}	8	0.0345	1.0000	1.8125	0.0155	inf
{Resistin=Low, BMI=Low, Glucose=Medium, Insulin=Low, Age=Middle, Leptin=Low, HOMA=Low, Adiponectin=Low}			0.0259	1.0000	1.8125	0.0116	inf
{Resistin=Low, BMI=Low, Glucose=Medium, Insulin=Low, MCP.1=Medium, Age=Middle, HOMA=Low, Adiponectin=Low}			0.0259	1.0000	1.8125	0.0116	inf
{Resistin=Low, BMI=Low, Glucose=Medium, MCP.1=Medium, Age=Middle, Leptin=Low, HOMA=Low, Adiponectin=Low}			0.0259	1.0000	1.8125	0.0116	inf
{Resistin=Low, BMI=Low, Glucose=Medium, Insulin=Low, MCP.1=Medium, Age=Middle, Leptin=Low, Adiponectin=Low}			0.0259	1.0000	1.8125	0.0116	inf
{Glucose=Medium, BMI=Medium, Insulin=Low, MCP.1=Medium, Age=Middle, Resistin=Medium, Leptin=Low, HOMA=Low, Adiponectin=Low}	{Classification=P}	9	0.0345	1.0000	1.8125	0.0155	inf
{Resistin=Low, BMI=Low, Glucose=Medium, Insulin=Low, MCP.1=Medium, Age=Middle, Leptin=Low, HOMA=Low, Adiponectin=Low}			0.0259	1.0000	1.8125	0.0116	inf

Note: inf = infinite

Table 7 presents the top pruned CAR for patient classification where all rules show a confidence value of 1.0 and an infinite conviction. Similar to the healthy classification, this indicates that the presence of the antecedent features consistently leads to a patient outcome within the dataset. Multiple pruning metrics such as lift and leverage were used to retain only the most significant and reliable patterns and reduce potential overfitting. In the 2- and 3-itemset rules, common patterns include combinations such as {Glucose=High, Adiponectin=Medium} and {Insulin=High, Adiponectin=Low}. These rules suggest that elevated glucose or insulin especially when combined with moderate or low adiponectin levels are predictive of patient classification. Additionally, rule of {Age=Middle, MCP.1=Medium, Adiponectin=High} indicating that middle-aged individuals with increased MCP.1 are frequently associated with breast cancer in this dataset.

As the rules increase in complexity, BMI=Medium appears frequently alongside other metabolic features such as Glucose=Medium, HOMA=Low, and Insulin=Low. Although some individual features like insulin or HOMA are typically associated with healthy profiles, their co-occurrence with other risk factors such as moderate glucose or adiponectin levels results in patient classification. In the larger itemsets like 5- to 9-feature rules, combinations in the 6-itemset such as {BMI=Low, Glucose=Medium, MCP.1=Medium, Age=Middle, HOMA=Low, Adiponectin=Low} highlights the presence of MCP.1=Medium as a recurring feature in patient classification particularly among middle-aged individuals. Likewise, the 9-itemset {Glucose=Medium, BMI=Medium, Insulin=Low, MCP.1=Medium, Age=Middle, Resistin=Medium, Leptin=Low, HOMA=Low, Adiponectin=Low} demonstrates the additional contribution of Resistin=Medium in identifying patients. These rules suggest that even when insulin resistance markers are low, moderate levels of MCP.1 and resistin may play a significant role in distinguishing patient profiles within the dataset. The patterns of “Classification=P” are visualized in the network graph in Figure 11.

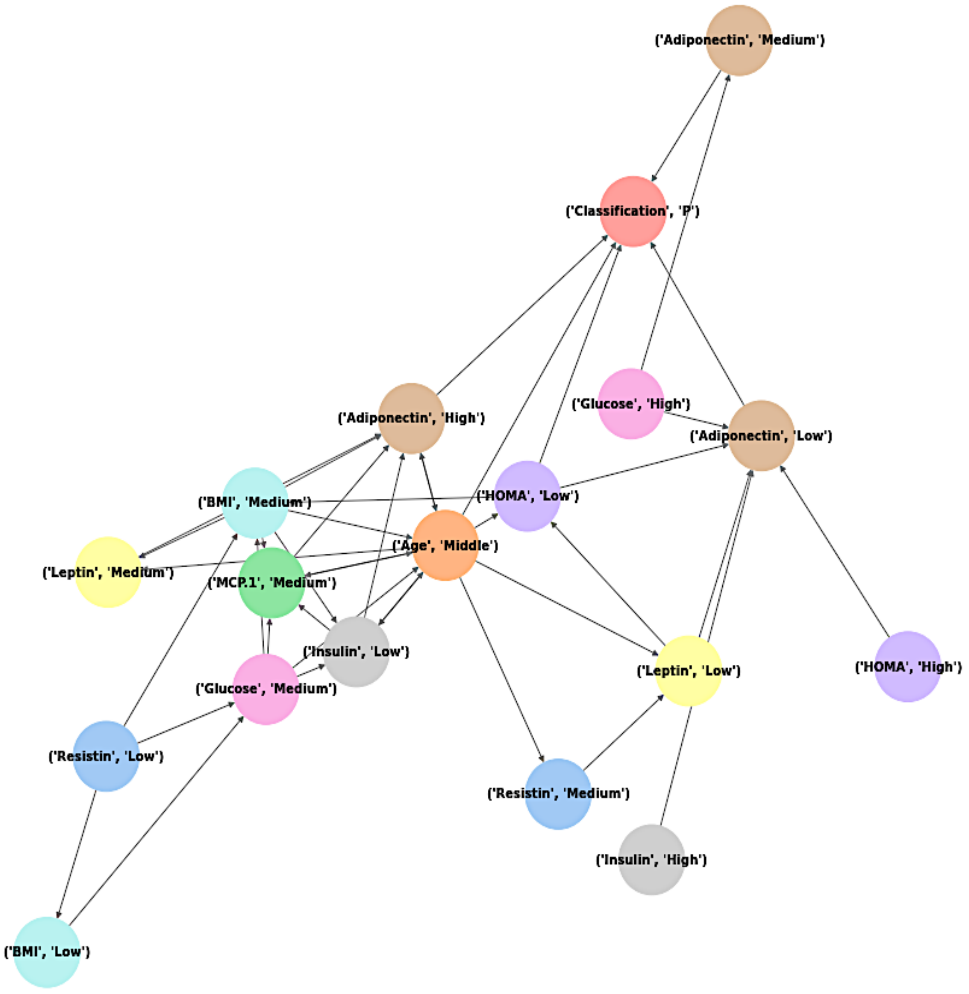


Figure 11. Network graph for classification=P

Patterns Validation Using SHAP Plots with Random Forest and Gradient Boost

After applying SMOTE to address class imbalance, the models were trained and SHAP analysis was conducted to interpret feature contributions based on the balanced dataset. In order to ensure consistency across methods, patterns identified through pruned CAR were compared and validated with SHAP values. This mutual assessment confirmed that the dominant features highlighted by SHAP values were also supported by the rules extracted from the pruned CAR analysis. In SHAP plots, red color indicates higher feature values while blue indicates lower feature values. The position of the points along the x-axis shows the SHAP value and represents the impact of that feature on the model's prediction. A positive SHAP value increases the likelihood of the "Patient" classification while a negative SHAP value indicates that the feature value increases the likelihood of the "Healthy" classification. Figure 12 (a) and (b) visualized the SHAP summary plot for “Patient” classification using RF and GB classifiers respectively.

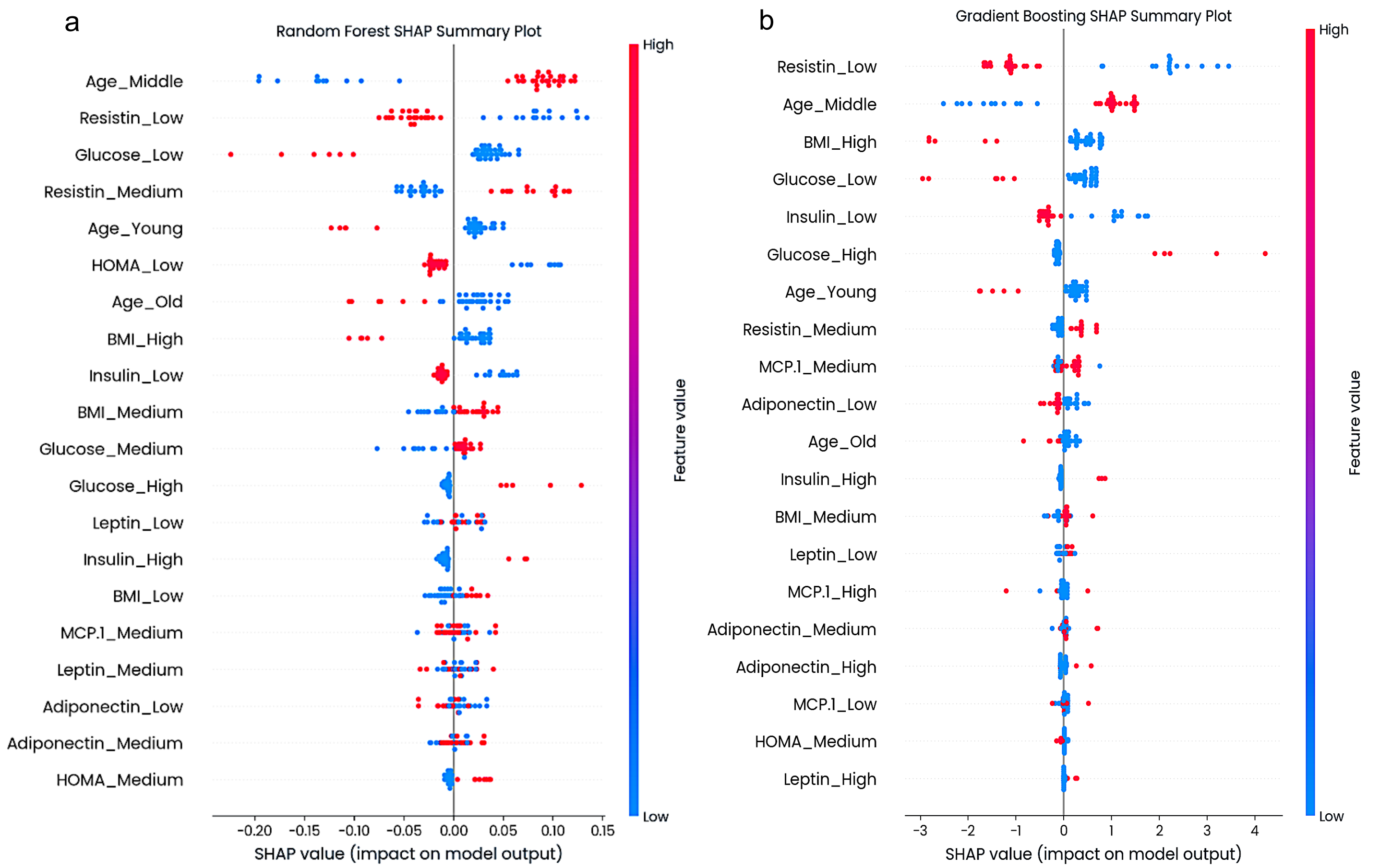


Figure 12. SHAP summary plot of (a) RF (b) GB showing top contributing features. Features toward the right (positive SHAP values) contribute to patient classification, while those on the left are associated with healthy outcomes

Figure 12 (a) presents the SHAP summary plot for the RF model. Each dot represents a single instance with red indicating the feature is present and blue indicating it is absent. SHAP values on the right push the prediction toward Patient class, while dots on the left push toward Healthy class. Notably, the presence of features like Resistin_Low, Glucose_Low, Age_Young, HOMA_Low, Age=Old and BMI_High appear in red on the left side indicating that their presence contributes to a healthy prediction. In contrast, the features such as Age_Middle, Resistin_Medium and Glucose_High appears in red on the right suggesting these characteristics are associated with increased likelihood of being classified as a patient.

Figure 12(b) shows the SHAP summary plot for the GB model. The plot indicates that the features such as Resistin_Low, Glucose_Low, Insulin_Low, Age_Young and BMI_High appear on the left side of the plot when present suggesting they are associated with healthier individuals. On the other hand, the presence of features such as Age_Middle, Glucose_High, Resistin_Medium and MCP.1_Medium contributes positively toward predicting a patient. These patterns are consistent with those observed in the RF model with additional MCP.1_Medium on GB.

The patterns extracted through CAR were validated against the SHAP summary plots of both the RF and GB models to ensure consistency between rule-based and model-based interpretations. For the healthy classification, features such as Glucose_Low, Insulin_Low, HOMA_Low, Resistin_Low, and BMI_High appeared with negative SHAP values positioned toward the left indicating their contribution to healthy predictions. These findings support several high-confidence CAR rules including {Glucose=Low, Age=Young}, {Resistin=Low, BMI=Low, Adiponectin=High, HOMA=Low, Age=Young}, and {Resistin=Low, BMI=Medium, Insulin=Low, Glucose=Low, Age=Middle, Leptin=Low, HOMA=Low, Adiponectin=Low} where the presence of these low-risk metabolic markers consistently led to healthy classification. Notably, BMI_High also showed a negative SHAP impact in the context of younger individuals which is consistent with the rule {BMI=High, Age=Young} suggesting that elevated BMI may not always indicate adverse outcomes in this subgroup. On the other hand, for the patient classification, the SHAP summary plots similarly confirmed several key features identified through CAR. Glucose_High, Age_Middle, MCP.1_Medium and Resistin_Medium were among the top contributors consistently showing positive SHAP values indicating their contribution toward the patient class. These findings validate rules such as {Glucose=High, Adiponectin=Medium}, {Age=Middle, MCP.1=Medium, Adiponectin=High} and the more complex {Glucose=Medium, BMI=Medium, Insulin=Low, MCP.1=Medium, Age=Middle, Resistin=Medium, Leptin=Low, HOMA=Low, Adiponectin=Low}.In particular, the repeated presence of MCP.1=Medium and Resistin=Medium in both CAR rules and SHAP plots suggests a strong influence of inflammatory markers in distinguishing patients. Although some features such as Insulin_Low and HOMA_Low are typically associated with healthy outcomes, their co-occurrence with moderate-risk indicators like glucose, MCP.1 and resistin appears to shift the prediction toward the patient class as reflected in both the rules and SHAP interpretations.

The observed associations in this study between low glucose, low insulin, low HOMA and high BMI with healthy individuals especially among younger age groups are supported by recent literature. For example, Augustin *et al*. [39] highlighted the role of stable glucose and insulin levels in reducing breast cancer risk. Pan *et al*. [40] also found that women with lower HOMA levels which indicating lower insulin resistance had a reduced risk of developing breast cancer. Besides, Liu *et al*. [41] reported that while overall higher BMI slightly increases the risk of breast cancer, women who had a higher BMI before menopause were actually less likely to develop the disease. This may be due to hormonal differences where obese young women tend to have irregular menstrual cycles leading to less exposure to estrogen which is a hormone linked to breast cancer [42]. Similarly, Yee *et al*. [43] found that women with more body fat during early adulthood had a lower risk of breast cancer which supporting the idea that high BMI in younger women might have a protective effect.

Conversely, features such as high glucose, medium resistin, and medium MCP.1 were frequently linked to patient classifications in CARM and SHAP results. These associations are well-supported by recent literature. For instance, Qiu *et al*. [44] demonstrated that hyperglycemia not only increases breast cancer risk but also contributes to tumor proliferation, migration and resistance to chemotherapy through abnormal glucose metabolism. In alignment, Zoroddu *et al*. [45] reported significantly higher resistin concentrations in breast cancer patients indicating its impact in cancer development. Additionally, Barulina *et al*. [46] identified MCP.1 as a key inflammatory marker in a cytokine-based predictive model for early breast cancer diagnosis further validating its relevance in classifying patients.

Conclusion and Future Works

This study presents a methodological contribution by integrating CARM with SHAP-based explainability using RF and GB classifier to investigate interpretable diagnostic patterns in breast cancer classification. In contrast to previous approaches that often lacked systematic validation, this hybrid approach improves reliability by cross-validating rule significance through SHAP values. From a total of 723,938 AR generated, 17,720 were identified as significant CAR and pruned using thresholds for lift, leverage and conviction to ensure statistical strength and interpretability. interpretability. The analysis revealed that features such as low glucose, low insulin, low HOMA and high BMI were frequently observed among healthy individuals particularly in younger age groups. These findings are supported by previous studies [39], [40], [41], [42], [43]. Middle-aged individuals with low MCP.1 and low resistin were also common among healthy classifications. In contrast, patient classifications were characterized by high glucose levels, medium MCP.1, medium resistin and middle age as indicated by recent findings [44], [45], [46]. These CAR-derived patterns were further validated using SHAP summary plots where negative SHAP values for low glucose, high BMI, and low HOMA aligned with healthy predictions while high positive SHAP values for high glucose, moderate resistin and MCP.1 supported their association with patient outcomes. Our findings indicate that the proposed approach was effective in uncovering significant patterns relevant to breast cancer diagnosis. The identified patterns reveal the relationships between biological features and may support the development of tools for breast cancer risk assessment and clinical decision-making.

While this study offers significant contributions to breast cancer diagnosis, future research can build on these findings to explore more possibilities. Utilizing on multiple datasets will allow for the exploration of even more complex patterns. By incorporating datasets that include clinical annotations such as cancer stage and treatment history further validation of risk factor patterns across different stages of disease progression and treatment conditions can be conducted. In addition, clinical validation with healthcare professionals would help assess the effectiveness of the framework. Collaboration with clinical experts is planned for future work to validate the diagnostic relevance of the extracted features and improve real-world applicability.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgement

The research is supported by Universiti Kebangsaan Malaysia through grant no. TAP-K021917. High appreciation goes to the above sponsor.

References

[1] World Health Organization. (2024, February 1). *Global cancer burden growing, amidst mounting need for services*. https://www.who.int/news/item/01-02-2024-global-cancer-burden-growing--amidst-mounting-need-for-services

[2] Siegel, R. L., Giaquinto, A. N., & Ahmedin, J. (2024). Cancer statistics, 2024. https://doi.org/10.3322/caac.21820

[3] Bellazzi, R., & Zupan, B. (2008). Predictive data mining in clinical medicine: Current issues and guidelines. *International Journal of Medical Informatics, 77*(2), 81–97. https://doi.org/10.1016/j.ijmedinf.2006.11.006

[4] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine, 17*(3).

[5] Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

[6] Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *ACM SIGMOD Record, 22*(2), 207–216. https://doi.org/10.1145/170036.170072

[7] Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. *Proceedings of 20th International Conference on Very Large Data Bases (VLDB'94)*, 487–499. https://citeseer.ist.psu.edu/agrawal94fast.html

[8] Liu, B., Hsu, W., & Ma, Y. (1998). Integrating classification and association rule mining. *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining (KDD 1998)*.

[9] Barredo Arrieta, A., *et al.* (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[10] Kok, I., Okay, F. Y., Muyanli, O., & Ozdemir, S. (2023). Explainable artificial intelligence (XAI) for Internet of Things: A survey. *IEEE Internet of Things Journal, 10*(16), 14764–14779. https://doi.org/10.1109/JIOT.2023.3287678

[11] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys, 51*(5), 1–45. https://doi.org/10.1145/3236009

[12] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 2017*(December), 4766–4775.

[13] Fahrudin, T. M., Syarif, I., & Barakbah, A. R. (2017). Discovering patterns of NED-breast cancer based on association rules using apriori and FP-growth. *Proceedings of the International Electronics Symposium on Knowledge Creation and Intelligent Computing (IES-KCIC 2017), 2017*(1), 132–139. https://doi.org/10.1109/KCIC.2017.8228576

[14] Kabir, M. F., Ludwig, S. A., & Abdullah, A. S. (2018). Rule discovery from breast cancer risk factors using association rule mining. *Proceedings of the 2018 IEEE International Conference on Big Data (Big Data 2018)*, 2433–2441. https://doi.org/10.1109/BigData.2018.8622028

[15] Oladipupo, O., Olajide, O., Adubi, S., Oyelade, J., & Omogbadegun, Z. (2021). An interval type-2 fuzzy association rule mining approach to pattern discovery in breast cancer dataset. *Journal of Computer Science, 17*(3), 330–348. https://doi.org/10.3844/JCSSP.2021.330.348

[16] Khater, T., *et al.* (2023). An explainable artificial intelligence model for the classification of breast cancer. *IEEE Access, PP*, 1. https://doi.org/10.1109/ACCESS.2023.3308446

[17] Liu, Y., Fu, Y., Peng, Y., & Ming, J. (2024). Clinical decision support tool for breast cancer recurrence prediction using SHAP value in cooperative game theory. *Heliyon, 10*(2), e24876. https://doi.org/10.1016/j.heliyon.2024.e24876

[18] Suresh, T., Assegie, T. A., Ganesan, S., Tulasi, R. L., Mothukuri, R., & Salau, A. O. (2023). Explainable extreme boosting model for breast cancer diagnosis. *International Journal of Electrical and Computer Engineering, 13*(5), 5764–5769. https://doi.org/10.11591/ijece.v13i5.pp5764-5769

[19] Moncada-Torres, A., van Maaren, M. C., Hendriks, M. P., Siesling, S., & Geleijnse, G. (2021). Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. *Scientific Reports, 11*(1), 1–14. https://doi.org/10.1038/s41598-021-86327-7

[20] Zadeh, L. A. (1965). Fuzzy sets. *Information and Control, 8*, 338–353.

[21] Akbas, K. E., *et al.* (2022). Assessment of association rule mining using interest measures on the gene data. *Medical Records, 4*(3), 286–292. https://doi.org/10.37990/medr.1088631

[22] Patrcio, M., Pereira, J., Crisstomo, J., Matafome, P., Seia, R., & Caramelo, F. (n.d.). *Breast Cancer Coimbra*. UCI Machine Learning Repository.

[23] Nam, G. E., Zhang, Z. F., Rao, J., Zhou, H., & Jung, S. Y. (2021). Interactions between adiponectin-pathway

polymorphisms and obesity on postmenopausal breast cancer risk among African American women: The WHI SHARe study. *Frontiers in Oncology, 11*(July). https://doi.org/10.3389/fonc.2021.698198

[24] Panigoro, S. S., *et al.* (2021). The association between triglyceride-glucose index as a marker of insulin resistance and the risk of breast cancer. *Frontiers in Endocrinology, 12*(October), 1–7. https://doi.org/10.3389/fendo.2021.745236

[25] Ke, J., *et al.* (2021). Glucose intolerance and cancer risk: A community-based prospective cohort study in Shanghai, China. *Frontiers in Oncology, 11*(August). https://doi.org/10.3389/fonc.2021.726672

[26] Sudan, S. K., *et al.* (2024). Obesity and early-onset breast cancer and specific molecular subtype diagnosis in Black and White women. *JAMA Network Open, 7*(7), e2421846. https://doi.org/10.1001/jamanetworkopen.2024.21846

[27] Diao, S., *et al.* (2021). Obesity-related proteins score as a potential marker of breast cancer risk. *Scientific Reports, 11*(1), 1–11. https://doi.org/10.1038/s41598-021-87583-3

[28] Sarker, P., Ksibi, A., Jamjoom, M. M., Choi, K., Al Nahid, A., & Samad, M. A. (2025). Breast cancer prediction with feature-selected XGB classifier, optimized by metaheuristic algorithms. *Journal of Big Data, 12*(1). https://doi.org/10.1186/s40537-025-01132-7

[29] Alfian, G., *et al.* (2022). Predicting breast cancer from risk factors using SVM and extra-trees-based feature selection method. *Computers, 11*(9). https://doi.org/10.3390/computers11090136

[30] Kazerani, R. (2024). Improving breast cancer diagnosis accuracy by particle swarm optimization feature selection. *International Journal of Computational Intelligence Systems, 17*(1). https://doi.org/10.1007/s44196-024-00428-5

[31] Anusha, P. V., Anuradha, C., Chandra Murty, P. S. R., & Kiran, C. S. (2019). Detecting outliers in high dimensional data sets using Z-score methodology. *International Journal of Innovative Technology and Exploring Engineering, 9*(1), 48–53. https://doi.org/10.35940/ijitee.A3910.119119

[32] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*(1), 321–357.

[33] Darrab, S., Broneske, D., & Saake, G. (2024). Exploring the predictive factors of heart disease using rare association rule mining. *Scientific Reports, 14*(1), 1–26. https://doi.org/10.1038/s41598-024-69071-6

[34] Liu, Y., *et al.* (2023). A novel FCTF evaluation and prediction model for food efficacy based on association rule mining. *Frontiers in Nutrition, 10*(August), 1–11. https://doi.org/10.3389/fnut.2023.1170084

[35] Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32.

[36] Han, S., & Kim, H. (2019). On the optimal size of candidate feature set in random forest. *Applied Sciences, 9*(5). https://doi.org/10.3390/app9050898

[37] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189–1232.

[38] Wang, L., Jiang, S., & Jiang, S. (2021). A feature selection method via analysis of relevance, redundancy, and interaction. *Expert Systems with Applications, 183*(June), 115365. https://doi.org/10.1016/j.eswa.2021.115365

[39] Augustin, L. S. A., *et al.* (2017). Low glycemic index diet, exercise and vitamin D to reduce breast cancer recurrence (DediCa): Design of a clinical trial. *BMC Cancer, 17*(1), 1–13. https://doi.org/10.1186/s12885-017-3064-4

[40] Pan, K., *et al.* (2020). Insulin resistance and breast cancer incidence and mortality in postmenopausal women in the Women's Health Initiative. *Cancer, 126*(16), 3638–3647. https://doi.org/10.1002/cncr.33002

[41] Liu, K., *et al.* (2018). Association between body mass index and breast cancer risk: Evidence based on a dose–response meta-analysis. *Cancer Management and Research, 10*, 143–151. https://doi.org/10.2147/CMAR.S144619

[42] Mohanty, S. S., & Mohanty, P. K. (2021). Obesity as potential breast cancer risk factor for postmenopausal women. *Genes & Diseases, 8*(2), 117–123. https://doi.org/10.1016/j.gendis.2019.09.006

[43] Yee, L. D., Mortimer, J. E., Natarajan, R., Dietze, E. C., & Seewaldt, V. L. (2020). Metabolic health, insulin, and breast cancer: Why oncologists should care about insulin. *Frontiers in Endocrinology, 11*(February), 1–25. https://doi.org/10.3389/fendo.2020.00058

[44] Qiu, J., Zheng, Q., & Meng, X. (2021). Hyperglycemia and chemoresistance in breast cancer: From cellular mechanisms to treatment response. *Frontiers in Oncology, 11*(February), 1–12. https://doi.org/10.3389/fonc.2021.628359

[45] Zoroddu, S., Di Lorenzo, B., Paliogiannis, P., Mangoni, A. A., Carru, C., & Zinellu, A. (2024). Resistin and omentin in breast cancer: A systematic review and meta-analysis. *Clinica Chimica Acta, 562*(July), 119838. https://doi.org/10.1016/j.cca.2024.119838

[46] Barulina, M., Gergenreter, Y., Zakharova, N., Maslyakov, V., Fedorov, V., & Ulitin, I. (2023). Predictive diagnosis of breast cancer based on cytokine profile. *Engineering Proceedings, 33*(1), 2–8. https://doi.org/10.3390/engproc2023033004