GMR
The Original

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING FOR PREDICTING PHYTOCHEMICAL-INDUCED APOPTOSIS IN BREAST CANCER CELL LINES: A MOLECULAR APPROACH TO CANCER PREVENTION

Haripriya R [1], Prabu D[2*], Sindhu R[3], Gayathiri R[4], Christy Grace M A[5], Vishvesh Sanathan A[6], Naveen Karthick R[7], Banujothi A[8]

[1,8] Postgraduate, Department of Public health dentistry, SRM Dental College, Ramapuram, Bharathi Salai, Chennai, TN, India
[2*]PhD, Professor and Head , Department of Public health dentistry, SRM Dental College, Ramapuram, Bharathi Salai, Chennai, TN, India
[3]MDS, Senior lecturer, Department of Public health dentistry, SRM Dental College, Ramapuram, Bharathi Salai, Chennai, TN, India
[4] MDS, Research scholar, Department of Public health dentistry, SRM Dental College, Ramapuram, Bharathi Salai, Chennai, TN, India
[5]Assistant Professor, Department of Computer Science and Engineering (AI&ML), Easwari Engineering College, Chennai, India.
[6,7]Undergraduate, Department of Computer Science and Engineering (AI&ML), Easwari Engineering College, Chennai, India.
***Corresponding author:** Dr. Prabu D, Email id: researchphdsrm@gmail.com

ABSTRACT:
Background: Phytochemicals are increasingly being studied for their potential role in breast cancer research because many of these naturally derived compounds can influence cancer-related cellular processes. However, testing large numbers of phytochemicals experimentally is expensive, time-consuming, and difficult to scale. This study therefore explored whether machine learning could be used to predict phytochemical-induced apoptosis and help identify compounds that may deserve further experimental investigation.
Methods: A dataset containing 1,347 experimental observations collected from 54 peer-reviewed studies was prepared using information from breast cancer cell-line experiments. The variables included phytochemical type, concentration, treatment duration, breast cancer cell line, combination treatment, cell viability, and apoptosis. Ten regression models were evaluated, including Linear Regression, Support Vector Regression, K-Nearest Neighbors, Decision Tree, Gradient Boosting, LightGBM, CatBoost, Random Forest, XGBoost, and Extra Trees. Association Rule Mining was also applied to identify recurring relationships among the experimental variables, while SHAP analysis was used to explain how individual features contributed to model predictions.
Results: The findings showed that tree-based ensemble methods performed better than the simpler statistical and conventional machine-learning approaches. Extra Trees produced the highest reported training performance with an R^2 of 0.9853, followed closely by XGBoost (R^2 = 0.9852) and Random Forest (R^2 = 0.9839). Among the evaluated phytochemicals, curcumin showed the highest average apoptosis value at 52.87%, followed by resveratrol at 46.13%. Association Rule Mining revealed recurring patterns linking higher phytochemical concentrations with increased apoptosis and reduced cell viability. SHAP analysis further showed that phytochemical type had the strongest influence on apoptosis prediction, followed by breast cancer cell line, cell viability, and concentration.
Conclusion: The study demonstrates that combining machine learning with explainable AI and association rule mining can provide a practical approach for analysing phytochemical responses in breast cancer cell-line experiments. The framework may help researchers narrow down promising phytochemicals before conducting further laboratory testing. Nevertheless, the findings should be validated using independent datasets and additional experimental studies before the approach is considered for translational or clinical use.

KEYWORDS: Breast cancer; phytochemicals; machine learning; apoptosis; explainable AI; SHAP; association rule mining; Extra Trees.

INTRODUCTION:

Breast cancer remains the most commonly diagnosed cancer among women worldwide and is one of the leading causes of cancer-related deaths despite significant advances in early detection, molecular diagnostics, targeted therapies, and immunotherapy [1–6]. Although these developments have considerably improved patient survival, the global incidence of breast cancer continues to rise, particularly among younger women, emphasizing the need for effective preventive strategies alongside improved treatment options [1–6]. Current therapeutic approaches, including surgery, chemotherapy, radiotherapy, endocrine therapy, and targeted biological agents, have transformed clinical management; however, they are often associated with severe adverse effects, drug resistance, high treatment costs, and disease recurrence [4–6]. These limitations have prompted growing interest in identifying safer and more effective preventive agents, particularly those derived from natural sources.
Among the most promising candidates are phytochemicals, naturally occurring bioactive compounds found in medicinal plants. These compounds exhibit a wide range of biological activities, including antioxidant, anti-

inflammatory, antiproliferative, antiangiogenic, pro-apoptotic, and immunomodulatory effects, making them attractive for breast cancer prevention [7–10]. Several phytochemicals, such as curcumin, resveratrol, quercetin, epigallocatechin gallate, genistein, berberine, and withanolides, have demonstrated significant anticancer activity against different breast cancer cell lines, including estrogen receptor-positive (MCF-7), triple-negative (MDA-MB-231), and HER2-positive models [8–10]. Their anticancer effects are mediated through the regulation of multiple signaling pathways, including PI3K/Akt/mTOR, MAPK/ERK, NF-κB, Wnt/β-catenin, JAK/STAT, and p53-dependent apoptotic pathways, thereby inhibiting cell proliferation, metastasis, and tumor progression [8–10]. Experimental studies using breast cancer cell lines remain essential for evaluating cytotoxicity, determining half-maximal inhibitory concentration (IC_{50}), understanding molecular mechanisms, and validating the biological activity of phytochemicals before advancing to animal studies or clinical trials [9,10].

Despite encouraging findings, the discovery and validation of phytochemicals with potential anticancer activity remain challenging. Conventional screening approaches rely on extensive in vitro experimentation involving numerous phytochemicals, different breast cancer cell lines, varying treatment concentrations, and exposure conditions, making the process labor-intensive, time-consuming, and costly [7,13]. Furthermore, the biological activity of phytochemicals is influenced by multiple experimental factors, including phytochemical concentration (dosage), breast cancer cell type, treatment duration, and the resulting biological responses, such as cell viability and apoptosis induction. The complex interactions among these variables are difficult to characterize using conventional statistical methods alone, highlighting the need for advanced computational approaches capable of identifying meaningful patterns and accurately predicting phytochemical activity [13,24].

These complex nonlinear relationships are often difficult to capture using traditional statistical methods, limiting predictive accuracy and slowing the identification of promising chemo preventive compounds [13,24].Recent advances in Artificial Intelligence (AI) and Machine Learning (ML) have transformed biomedical research by providing powerful computational tools capable of extracting meaningful patterns from large and complex biological datasets [11–13,23,24]. AI-driven approaches can integrate diverse sources of information, including chemical descriptors, molecular fingerprints, transcriptomic and proteomic profiles, metabolomic signatures, molecular docking scores, and cell viability data, to accurately predict the biological activity of phytochemicals [12,13,21–25]. By rapidly analyzing multidimensional datasets, these computational techniques can substantially reduce experimental costs, shorten the drug discovery timeline, and improve the prioritization of promising phytochemicals for laboratory validation.

Several machine learning algorithms have demonstrated excellent performance in cancer prediction and drug discovery. Support Vector Machines (SVMs) are particularly effective for high-dimensional biomedical datasets and relatively small sample sizes because they identify optimal decision boundaries between different classes [16,24]. However, their performance depends heavily on kernel selection and parameter optimization and may become computationally intensive for large datasets. Random Forest (RF), an ensemble learning method, improves prediction by combining multiple decision trees, effectively capturing nonlinear relationships while reducing overfitting. It also provides estimates of feature importance, although interpreting complex interactions remains challenging [15,24]. Gradient boosting algorithms, including XGBoost, LightGBM, and CatBoost, have gained widespread popularity because of their high predictive accuracy, efficient handling of missing data, and ability to model complex nonlinear relationships. Nevertheless, these methods require careful hyperparameter tuning and may overfit when trained on limited datasets [14]. Artificial Neural Networks (ANNs) offer flexible nonlinear modeling capabilities and have been widely applied to predicting drug sensitivity, identifying molecular biomarkers, and classifying cancer subtypes. Their major limitation, however, is the requirement for large annotated datasets and their relatively poor interpretability due to their "black-box" nature [17].

The rapid development of deep learning has further expanded AI applications in oncology. Convolutional Neural Networks (CNNs) have achieved remarkable success in mammographic image classification, histopathological image analysis, and automated feature extraction from microscopic cell images [11,17]. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks are particularly useful for analyzing sequential genomic and transcriptomic data, whereas Graph Neural Networks (GNNs) effectively learn molecular graph representations to predict compound-target interactions and drug responses [19]. More recently, Transformer-based architectures, self-supervised learning, multimodal foundation models, and Explainable Artificial Intelligence (XAI) have enabled the integration of heterogeneous biomedical datasets while improving model transparency and interpretability [18–20]. Nevertheless, deep learning approaches generally require extensive computational resources, large annotated datasets, and rigorous validation before they can be reliably translated into biomedical applications.

Selecting an appropriate AI model depends largely on the characteristics of the available data. Ensemble learning techniques, such as Random Forest and XGBoost, are particularly well suited for structured phytochemical descriptors and experimental cell-line datasets because they provide high predictive performance while maintaining reasonable interpretability. In contrast, deep learning approaches become increasingly advantageous when analyzing high-dimensional omics datasets, molecular structures, or biomedical images [11–20].

Despite these advances, several important research gaps remain. Existing AI-assisted studies have predominantly focused on synthetic drug discovery, molecular property prediction, or therapeutic response modeling, with limited attention to the prioritization of naturally derived phytochemicals for breast cancer chemoprevention [13,21–24]. Moreover, most predictive models emphasize general cytotoxicity or molecular descriptors rather

than accurately predicting apoptosis induction, a critical biological endpoint for evaluating the anticancer potential of phytochemicals. Current approaches also lack integrated frameworks that simultaneously combine apoptosis prediction, phytochemical prioritization, and interpretable machine learning using experimentally validated breast cancer cell-line data. In addition, few studies have incorporated Association Rule Mining (ARM) to identify biologically meaningful relationships between phytochemical dosage, breast cancer cell lines, cell viability, and apoptosis, while the limited use of Explainable Artificial Intelligence (XAI) restricts model transparency and biological interpretability. These limitations highlight the need for an interpretable decision-support system capable of accurately predicting apoptosis, ranking phytochemicals according to their anticancer potential, and providing transparent explanations to support experimental validation and breast cancer chemoprevention research

To address the existing research gaps, this study aims to develop an interpretable Artificial Intelligence (AI)- and Machine Learning (ML)-based decision-support framework for analyzing and prioritizing phytochemicals that may have preventive potential against breast cancer. The framework is built using experimentally validated breast cancer cell-line data and compares a broad range of predictive models, including Linear Regression, Support Vector Regression (SVR), K-Nearest Neighbors (KNN) Regression, Decision Tree Regression, Gradient Boosting Regression, Light Gradient Boosting Machine (LightGBM), CatBoost Regression, Random Forest Regression, Extreme Gradient Boosting (XGBoost) Regression, and Extra Trees Regression. Alongside these models, Association Rule Mining (ARM) is used to uncover meaningful relationships among experimental variables, while Explainable Artificial Intelligence (XAI), particularly SHapley Additive exPlanations (SHAP), is used to understand how individual variables influence model predictions. By considering factors such as phytochemical type, concentration, treatment duration, breast cancer cell line, combination treatment, cell viability, and apoptosis, the framework allows the different models to be compared in terms of predictive performance, robustness, interpretability, and generalizability. The overall goal is to create a practical and transparent computational approach that can support the identification of promising phytochemicals for further laboratory validation and future breast cancer chemoprevention research.

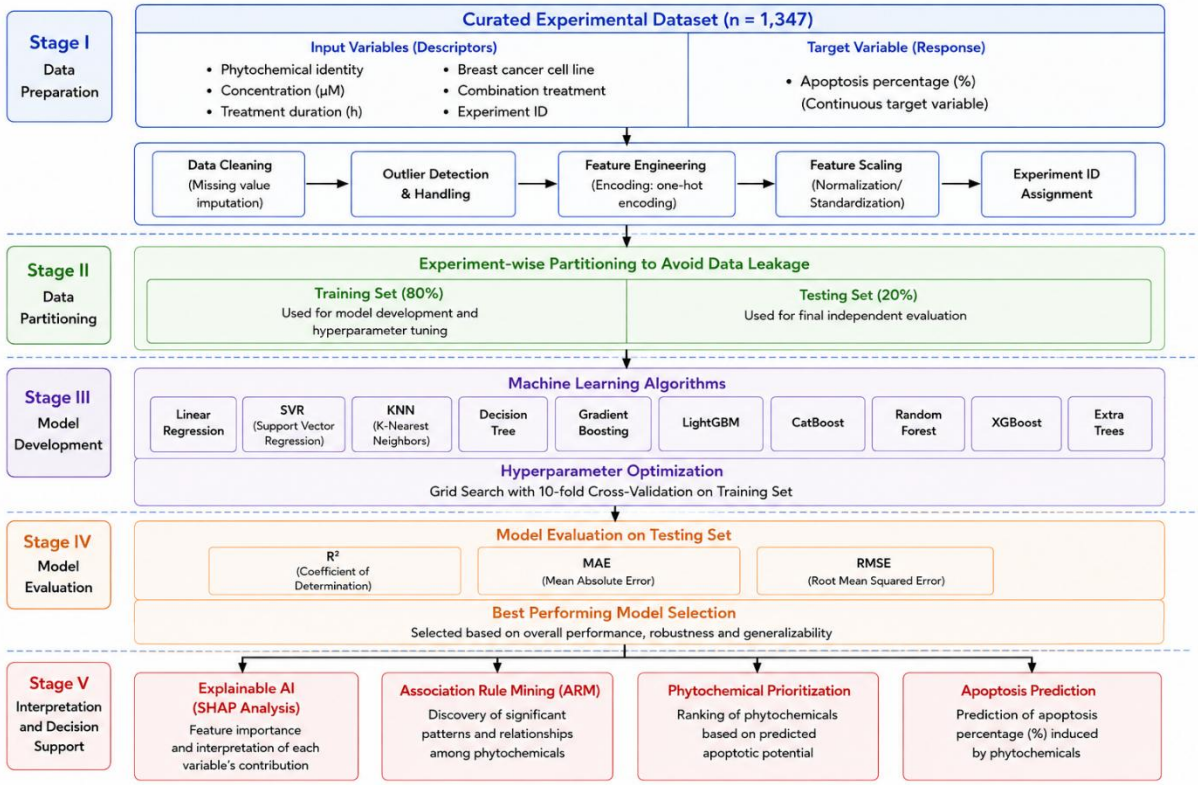
The main contributions of this study can be summarized in five key areas. First, the study develops a structured dataset by bringing together experimentally validated breast cancer cell-line data reported in peer-reviewed studies, including information on phytochemicals, concentration, treatment duration, cell lines, combination treatments, cell viability, and apoptosis. Second, Association Rule Mining is used to identify important patterns and relationships among variables such as concentration, dosage, cell viability, and apoptosis, helping to reveal trends that may not be immediately visible through conventional analysis. Third, the study introduces a phytochemical prioritization approach to rank compounds based on their apoptosis-inducing potential and support the selection of promising candidates for future experimental investigation. Fourth, multiple statistical, machine learning, and ensemble learning models are developed and compared for apoptosis prediction, allowing the most accurate, reliable, and generalizable model to be identified. Finally, Explainable AI is incorporated to make the predictions easier to understand, with SHAP and feature-importance analysis used to show which experimental factors contribute most strongly to the predicted apoptosis response. Together, these contributions provide a more complete framework that combines prediction, pattern discovery, prioritization, and explainability within a single AI-driven approach.

METHODOLOGY:

STUDY DESIGN:

This study proposes an Artificial Intelligence (AI)- and Machine Learning (ML)-driven predictive framework for identifying phytochemicals with potential preventive activity against breast cancer using experimentally validated breast cancer cell-line data. The proposed framework integrates systematic literature mining, dataset construction, data preprocessing, feature engineering, machine learning model development, explainable artificial intelligence (XAI), and model evaluation into a unified computational pipeline. The overall workflow of the proposed framework is illustrated in Figure 1.

Figure 1: Workflow of the Interpretable Machine Learning Framework for Breast Cancer Phytochemical Analysis



DATASET CONSTRUCTION :

The dataset was constructed through a comprehensive literature search of Google Scholar, PubMed, and Web of Science. Publications between 2005 and 2025 were searched using combinations of the keywords *breast cancer*, *phytochemicals*, *cell viability*, *cell apoptosis*, *combined treatment*, *natural compounds*, and *potential drug*. Duplicate publications were removed before screening.

Studies were included if they reported experimental investigations using human breast cancer cell lines; evaluated phytochemicals either individually or in combination with therapeutic agents; reported quantitative measurements of cell viability and/or apoptosis; and provided sufficient experimental information on phytochemical concentration and treatment duration. Studies involving animal models, clinical trials, review articles, editorials, conference abstracts lacking experimental data, or datasets with insufficient experimental information were excluded.

Following the screening process, 54 peer-reviewed publications satisfied the eligibility criteria and were included in the final dataset. Experimental values were manually extracted from tables and text. When numerical values were presented only in graphical form, digital data extraction was performed using Plotly Express to obtain quantitative measurements. The final dataset consisted of 1,347 independent experimental observations obtained from multiple breast cancer cell lines and included more than 16 phytochemical–drug combinations evaluated under different concentrations and exposure durations.

The primary input variables (descriptors) included phytochemical identity; phytochemical concentration (μM); treatment duration (h); breast cancer cell line; combination treatment information. The output variables consisted of experimentally measured cell viability (%) and cell apoptosis (%).

DATA PREPROCESSING:

The collected data originated from multiple independent studies, extensive preprocessing was performed to ensure consistency. Missing numerical values were imputed using mean imputation, whereas missing categorical variables were replaced using the mode. Experimental concentrations and dosage values were standardized into common SI units wherever necessary.

Treatment duration was converted into a standardized numerical variable by extracting the reported exposure time and converting it into hours.Categorical variables, including phytochemical names, breast cancer cell lines, and treatment combinations, were transformed into numerical representations using one-hot encoding.To avoid information leakage, observations originating from identical experiments but performed using different concentrations of the same phytochemical were assigned a common experiment identifier. During model development, all observations belonging to a particular experiment were allocated exclusively to either the training or testing dataset, thereby preventing highly correlated samples from appearing in both subsets. All preprocessing techniques were applied for training and test dataset.

Feature engineering was performed to improve the predictive capability of the developed machine learning models.Continuous variables, including phytochemical concentration, dosage, and exposure duration, were

normalized and transformed into machine-readable numerical features.Additional interaction features were generated to represent combination therapies involving phytochemicals and conventional drugs.
For Association Rule Mining (ARM), continuous concentration values were discretized into four categories:None (0 μM);Low (1–10 μM);Medium (11–100 μM);High (>100 μM).Similarly, cell viability values were categorized according to ISO 10993-5 as Non-cytotoxic (>70%) and Cytotoxic (≤70%)To evaluate different toxicity levels, cell viability was further classified into four categories Strongly cytotoxic (≤40%);Moderately cytotoxic (41–60%) ;Weakly cytotoxic (61–80%) and Non-cytotoxic (81–100%).These categorical variables were subsequently used during association rule mining.

DATA PARTITIONING

To avoid information leakage, observations originating from identical experiments but performed using different concentrations of the same phytochemical were assigned a common experiment identifier. The complete dataset was partitioned into 80% training and 20% testing subsets based on experiment identifiers rather than individual observations. All observations belonging to a particular experiment were allocated exclusively to either the training or testing dataset, thereby preventing highly correlated samples from appearing in both subsets, minimizing experimental dependency, and reducing the risk of data leakage. The training set was used for model development and hyperparameter tuning, while the testing set was reserved for independent evaluation.

MACHINE LEARNING AND MODEL DEVELOPMENT

All predictive models were developed using Python (version 3.12) with the Scikit-learn machine learning framework and other relevant libraries. Multiple supervised regression algorithms were implemented and comparatively evaluated to predict apoptosis percentage. Linear Regression was used as a baseline statistical model to estimate the linear relationship between the input variables and apoptosis percentage. Support Vector Regression (SVR) was employed to model complex relationships by identifying an optimal regression function within a defined error margin and, when required, using kernel functions to capture nonlinear patterns. K-Nearest Neighbors (KNN) Regression predicted apoptosis based on the average response of the most similar neighboring observations in the feature space. Decision Tree Regression generated predictions by recursively partitioning the data according to feature-based decision rules, thereby capturing nonlinear relationships and interactions among variables. Gradient Boosting Regression constructed an ensemble of sequential decision trees, with each successive tree attempting to correct the prediction errors of the preceding trees. Light Gradient Boosting Machine (LightGBM) employed an efficient gradient-boosting framework with leaf-wise tree growth, enabling rapid training and effective modeling of complex nonlinear relationships. CatBoost Regression is a gradient-boosting algorithm specifically designed to efficiently handle categorical variables while reducing prediction bias and overfitting through ordered boosting. Random Forest Regression combined predictions from multiple independently constructed decision trees generated from bootstrap samples and random subsets of features, thereby improving predictive stability and reducing overfitting. XGBoost Regression implemented an optimized gradient-boosting approach that incorporates regularization, efficient tree construction, and iterative error correction to achieve high predictive accuracy. Finally, Extra Trees Regression (Extremely Randomized Trees) generated an ensemble of highly randomized decision trees by introducing additional randomness in both feature and split-point selection, which can reduce variance and improve generalizability. The predictive performance of all models was subsequently compared to identify the most accurate, robust, and generalizable algorithm for apoptosis percentage prediction.
Random Forest was chosen as the primary baseline model because it is reliable, handles nonlinear relationships well, and is less prone to overfitting, making it suitable for structured biological data. The model was optimized using Grid Search with 10-fold cross-validation. Key parameters such as the number of trees, maximum tree depth, minimum samples required to split a node, minimum samples required at leaf nodes, and the number of features considered at each split were tuned to identify the best-performing configuration.Random Forest showed strong predictive performance with an R^2 value of 0.9839. However, when all models were compared, Extra Trees achieved the highest overall performance with an R^2 of 0.9853, followed closely by XGBoost with an R^2 of 0.9852. Therefore, Extra Trees was identified as the best-performing model, while Random Forest remained one of the strongest and most reliable ensemble models evaluated in the study.

MODEL EVALUATION

The predictive performance of the developed regression models was evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2). These metrics quantify the differences between the observed and predicted values and provide a comprehensive assessment of model accuracy and goodness-of-fit. The evaluation metrics are defined as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where $n$denotes the total number of observations, $y_i$represents the observed value, $\hat{y}_i$denotes the predicted value, and $\underline{y}$is the mean of the observed values. Lower values of MAE, MSE, and RMSE indicate better predictive performance, whereas an R^2value closer to 1 indicates a better fit between the predicted and observed values.

Explainable Artificial Intelligence

To enhance model transparency and biological interpretability, Explainable Artificial Intelligence (XAI) techniques were incorporated into the predictive framework.SHapley Additive exPlanations (SHAP) were employed to quantify the contribution of each feature toward individual model predictions and overall model performance.Feature importance analysis enabled identification of the most influential phytochemical descriptors associated with breast cancer cell apoptosis and viability, thereby facilitating biological interpretation of the computational predictions.

Association Rule Mining (ARM)

Association Rule Mining (ARM) was employed to identify meaningful relationships among experimental variables, including phytochemical concentration, breast cancer cell line, cell viability, and apoptosis induction. An association rule is represented as:

X→ Y

where:X can contain experimental characteristics such as phytochemical concentration. Y represents the biological response of interest, such as apoptosis or cell viability.The strength and significance of the generated rules were evaluated using three standard metrics: support, confidence, and lift, defined as:

$$\text{Support}(X \rightarrow Y) = \frac{n(X \cap Y)}{N}$$

$n(X \cap Y)$ = number of observations/transactions containing both X and Y

N = total number of observations/transactions in the dataset.

where $\sigma(X \cup Y)$denotes the number of transactions containing both $X$and Y, $\sigma(X)$represents the number of transactions containing the antecedent X, and $N$is the total number of transactions. Support measures the frequency of occurrence of a rule, confidence quantifies the conditional probability of the consequent given the antecedent, and lift evaluates the strength of the association relative to random co-occurrence, with values greater than one indicating a positive association. To ensure the discovery of biologically meaningful and reliable patterns, rule filtering was performed by retaining only those rules that satisfied predefined minimum thresholds for support, confidence, and lift, while redundant and statistically insignificant rules were discarded. The filtered association rules were subsequently used to prioritize phytochemicals exhibiting strong associations with apoptosis induction in breast cancer cell lines.

Association Rule Mining (ARM) was performed to identify hidden relationships between phytochemical characteristics and biological responses.ARM analysis requires categorical variables; therefore, continuous variables were discretized before analysis.The generated rules were evaluated using three standard metrics Support, representing the frequency with which a rule appears within the dataset;Confidence, indicating the conditional probability that the consequent occurs when the antecedent is present; and Lift, measuring the strength of association between variables. A lift value greater than one indicates a positive association between predictor variables and biological outcomes.Association rules with high support, confidence, and lift values were considered biologically meaningful.

Prioritization and Apoptosis Prediction

The final framework ranks phytochemicals according to their predicted anticancer efficacy and identifies key molecular and experimental factors influencing breast cancer cell responses. The integration of machine learning,

association rule mining, and explainable AI provides a robust decision-support system for prioritizing phytochemicals for future experimental validation and translational breast cancer prevention research.

RESULTS:

The proposed AI-based model was developed through a stepwise process that included data preprocessing, feature engineering, machine learning model development, performance evaluation, explainable AI, and association rule mining. The main aim of the model was to predict phytochemical-induced apoptosis in breast cancer cell lines and to identify the factors that had the greatest influence on the predicted response.

Ten regression algorithms were developed and compared, including Linear Regression, Support Vector Regression, K-Nearest Neighbors, Decision Tree, Gradient Boosting, LightGBM, CatBoost, Random Forest, XGBoost, and Extra Trees. Their performance was evaluated using R^2, RMSE, MAE, and training time in order to identify the model that provided the most accurate and reliable prediction of apoptosis.

Overall, the tree-based ensemble models performed better than the simpler statistical and conventional machine learning approaches. Extra Trees showed the best overall performance with an R^2 value of 0.9853, followed closely by XGBoost with an R^2 of 0.9852 and Random Forest with an R^2 of 0.9839. These results suggest that ensemble models were better suited to capture the complex and nonlinear relationships between phytochemical type, treatment conditions, breast cancer cell line, cell viability, and apoptosis.

Table 1: Training Performance Comparison Table

Category	Model	R^2	RMSE	MAE	Training Time(s)
Statistical Model	Linear Regression	0.500463	0.168864	0.132082	0.0005
Classical Machine Learning	SVR	0.713588	0.117526	0.004385	0.0119
	KNN	0.903454	0.914805	0.068234	0.035781
	Decision Tree	0.983744	0.026642	0.087877	0.01
Ensemble Learning	Gradient Boosting	0.877336	0.076912	0.057475	0.1386s
	Light GBM	0.949059	0.049565	0.676127	0.026639
	CatBoost	0.978017	0.032559	0.016227	0.7014
	Random Forest	0.983913	0.027853	0.009344	0.0852
	XGBoost	0.985193	0.026722	0.007684	0.0549
	Extra Trees	0.985281	0.026642	0.087877	0.005710

Table 1 presents the training performance of ten machine learning models, including one statistical model, three classical machine learning models, and six ensemble learning models. Model performance was evaluated using the coefficient of determination (R^2), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and training time.Among all the evaluated models, the Extra Trees algorithm achieved the highest predictive performance with an R^2 of 0.9853, indicating excellent goodness-of-fit while maintaining a low RMSE (0.0266) and the shortest training time (0.0057 s) among the ensemble methods. XGBoost and Random Forest also demonstrated competitive performance, achieving R^2 values of 0.9852 and 0.9839, respectively, with low prediction errors. Decision Tree performed well among the classical machine learning models, attaining an R^2 of 0.9837 with a low RMSE (0.0266). In contrast, Linear Regression exhibited the lowest predictive capability (R^2 = 0.5005), indicating that linear relationships alone were insufficient to accurately model the phytochemical dataset.

Overall, the results demonstrate that ensemble learning algorithms consistently outperformed statistical and conventional machine learning approaches, achieving higher R^2 values and lower prediction errors. These findings suggest that ensemble-based methods are more effective for capturing the complex nonlinear relationships present in phytochemical data, with the Extra Trees model emerging as the most suitable model for subsequent analysis.

Figure 2: Predicted vs Actual apoptosis

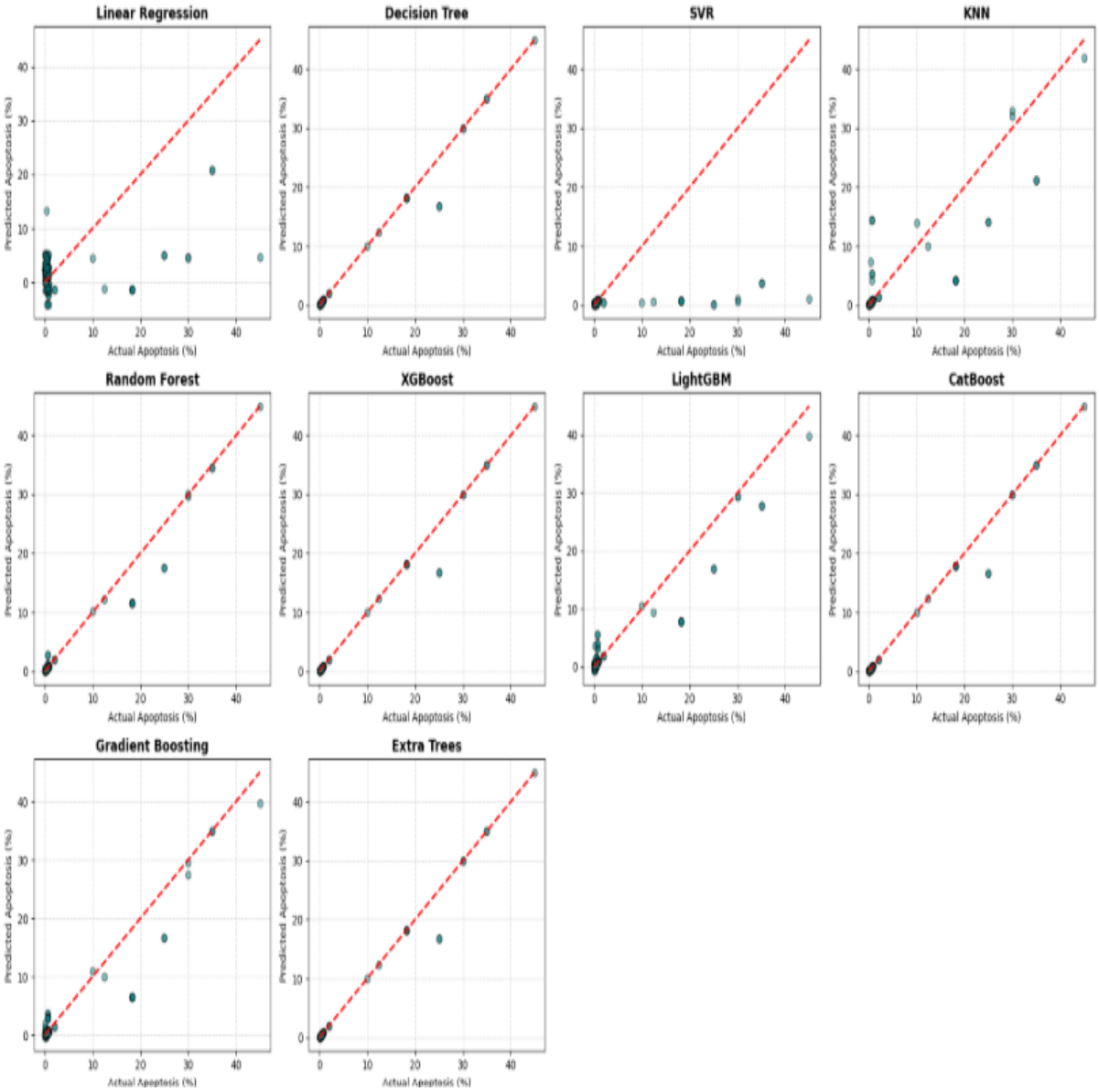


Figure 1 illustrates the relationship between the predicted and actual apoptosis rates for all evaluated machine learning models. The red dashed diagonal line (1:1 line) represents the ideal prediction, where the predicted values perfectly match the observed values. Models with data points clustered closely around this reference line exhibit higher predictive accuracy and lower prediction error.

Among the evaluated algorithms, the Extra Trees, XGBoost, Random Forest, Decision Tree, and CatBoost models demonstrated the best predictive performance, with most observations closely aligned with the 1:1 reference line, indicating excellent agreement between predicted and experimental apoptosis rates. The Extra Trees model exhibited the highest consistency between predicted and observed apoptosis rates, corroborating the quantitative evaluation presented in the performance metrics and supporting its selection as the optimal model for apoptosis prediction.

Gradient Boosting and LightGBM also showed satisfactory predictive capability, although a few samples exhibited moderate deviations from the ideal line. In contrast, Linear Regression and Support Vector Regression (SVR) displayed considerable dispersion from the reference line, suggesting limited ability to capture the nonlinear relationships within the dataset. The K-Nearest Neighbors (KNN) model produced reasonable predictions for several samples but exhibited larger deviations for high apoptosis values, indicating reduced prediction accuracy compared with the best-performing ensemble models.

Figure 3: Learning Curve (X-axis: Training samples; Y-axis: R²)

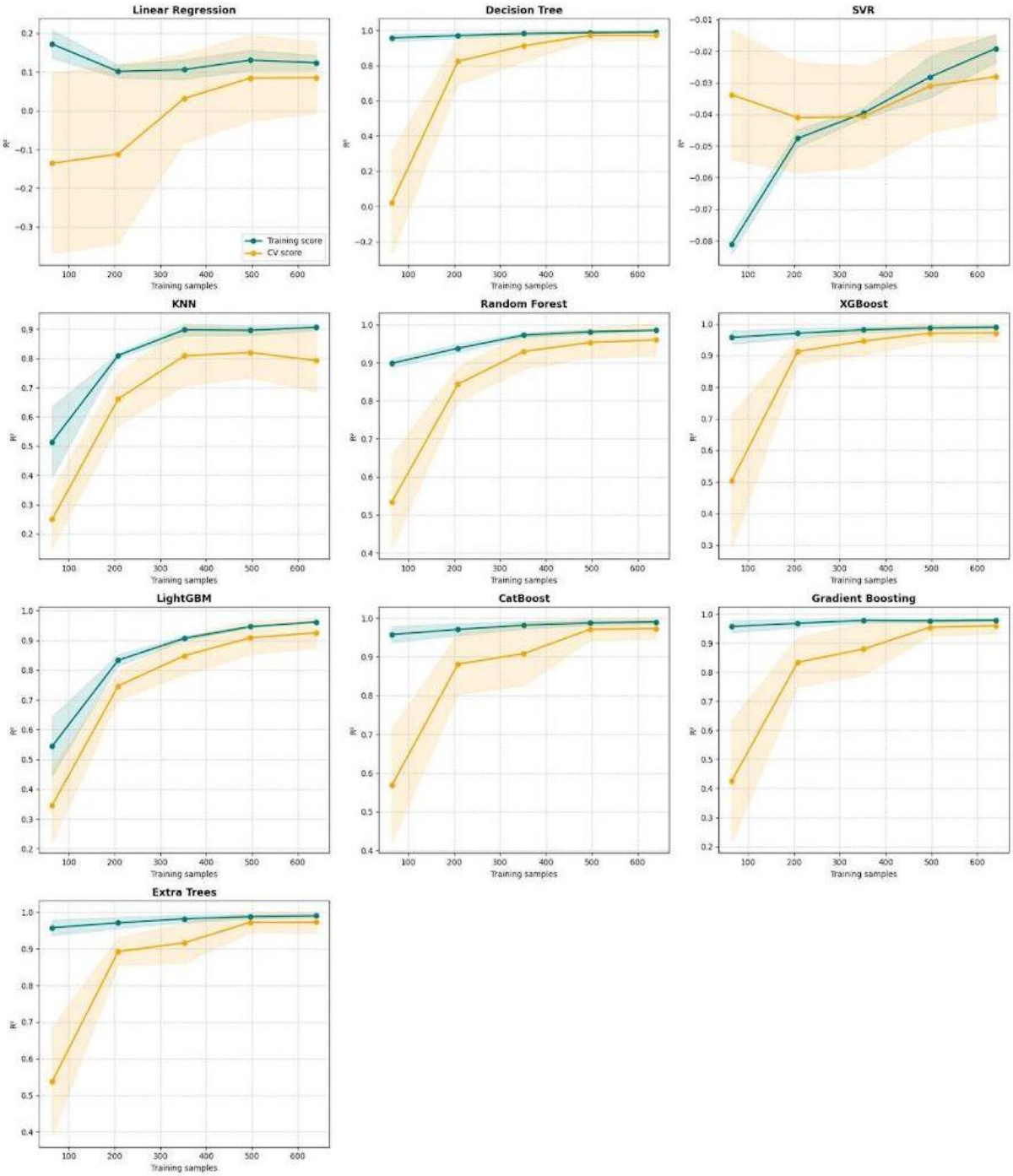


Figure 3. Learning curves of the evaluated machine learning models showing training and cross-validation (CV) R² scores as a function of training sample size. Ensemble learning algorithms including Extra Trees, CatBoost, XGBoost, Random Forest, Gradient Boosting, and Decision Tree demonstrated rapid convergence and high predictive performance, whereas Linear Regression and Support Vector Regression exhibited comparatively poor generalization, indicating limited suitability for modeling the nonlinear characteristics of the apoptosis dataset. Extra Trees demonstrated the most stable performance, characterized by minimal variance and near-complete convergence of the training and validation curves at higher sample sizes.

Table 2. One-way ANOVA results for Dosage (µg/mL) across study groups

ANOVA					
DOSAGE(µg/mL)					
	Sum of Squares	df	Mean Square	F	Sig.

Between Groups	8647.833	13	665.218	4.950	0.001*
Within Groups	26874.840	200	134.374		
Total	35522.672	213			

Table 3. One-way ANOVA results for Cell Viability (%) across study groups

ANOVA					
Cell Viability (%)					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1710032.746	13	131540.981	1.154	0.316
Within Groups	22787729.850	200	113938.649		
Total	24497762.590	213			

Table 4. One-way ANOVA results for Apoptosis (%) across study groups

ANOVA					
Apoptosis (%)					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	14439.931	13	1110.764	5.016	0.0001
Within Groups	44284.834	200	221.424		
Total	58724.765	213			

Table 2 shows one-way ANOVA splits the total variability in dosage into variability between groups and within groups. The result shows a statistically significant difference in mean dosage across groups, $p = 0.0001$. **Table 3** tests whether mean cell viability (%) differs across the same groups. The ANOVA is not statistically significant, $p = 0.316$, which indicates there is no evidence of a meaningful difference in average cell viability between groups in this dataset. **Table 4** evaluates whether mean apoptosis (%) differs across the groups. The ANOVA shows a statistically significant group effect, $p < 0.001$, meaning apoptosis levels vary significantly among groups.

Table 5: Phytochemical Ranking of phytochemical compunds:

Rank	Compound	Average apoptosis (%)
1	curcumin	52.87%
2	Resveratrol	46.13%
3	Quercetin	36.74%
4	Genistein	35.04%
5	Fisetin	26.29%
6	Kaempferol	22.52%
7	Sulforaphane	21.57%

Table 5 presents the ranking of the selected phytochemicals based on their average apoptosis induction (%) obtained from the AI-driven analysis. Curcumin exhibited the highest apoptotic activity, with an average apoptosis rate of 52.87%, followed by Resveratrol (46.13%). Quercetin ranked third, achieving an average apoptosis rate of 36.74%, while Genistein demonstrated 35.04%. Moderate apoptotic effects were observed for Fisetin (26.29%) and Kaempferol (22.52%), whereas Sulforaphane showed the lowest average apoptosis among the evaluated phytochemicals (21.57%). Overall, the ranking indicates that Curcumin and Resveratrol exhibited substantially greater apoptosis-inducing potential than the remaining compounds, suggesting their superior anticancer efficacy within the evaluated dataset.

Table 6: Top Frequent Itemsets for Association Rule Mining :

Rank	Frequent Itemset	Support (%)	Frequency
1	Low Apoptosis, Low Concentration	38.8	388
2	High Concentration, High Apoptosis	34.0	340
3	High Concentration, Low Viability	33.8	338
4	Low Concentration, Low Dosage	33.4	334
5	High Dosage, High Concentration	31.8	318

Table 6 presents the top five frequent itemsets identified through association rule mining. The combination of low apoptosis and low concentration showed the highest support (38.8%) and frequency (388 occurrences), followed by high concentration with high apoptosis (34.0%) and high concentration with low viability (33.8%). Other notable patterns included low concentration with low dosage and high dosage with high concentration. These findings reveal recurring relationships among dosage, concentration, apoptosis, and cell viability, highlighting important patterns within the phytochemical dataset.

Figure 4: SHAP Summary Plot for Explainable AI

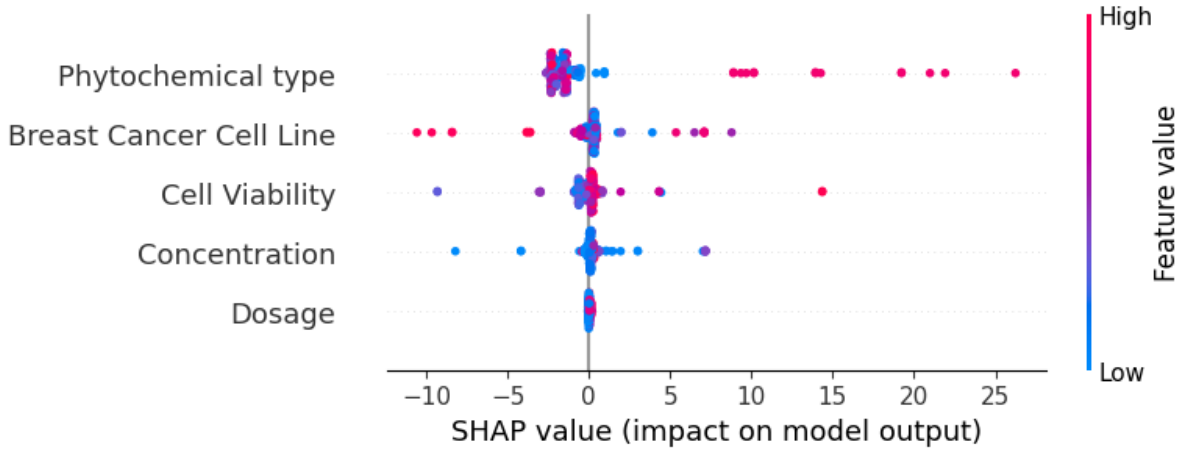


Figure 4 presents the SHAP (SHapley Additive exPlanations) summary plot, illustrating the relative importance and contribution of each feature to the model's predictions. Phytochemical type emerged as the most influential predictor, exhibiting the widest distribution of SHAP values and the greatest impact on model output. Breast cancer cell line ranked as the second most important feature, followed by cell viability and concentration, which showed moderate contributions to prediction variability. In contrast, dosage had the smallest SHAP value distribution, indicating a comparatively limited influence on the model's predictions. The color gradient demonstrates that both high and low feature values can either increase or decrease the predicted outcome depending on their interaction with other variables, highlighting the complex nonlinear relationships captured by the XGBoost model.

Figure 5:SHAP Dependence Plots Explainable AI:

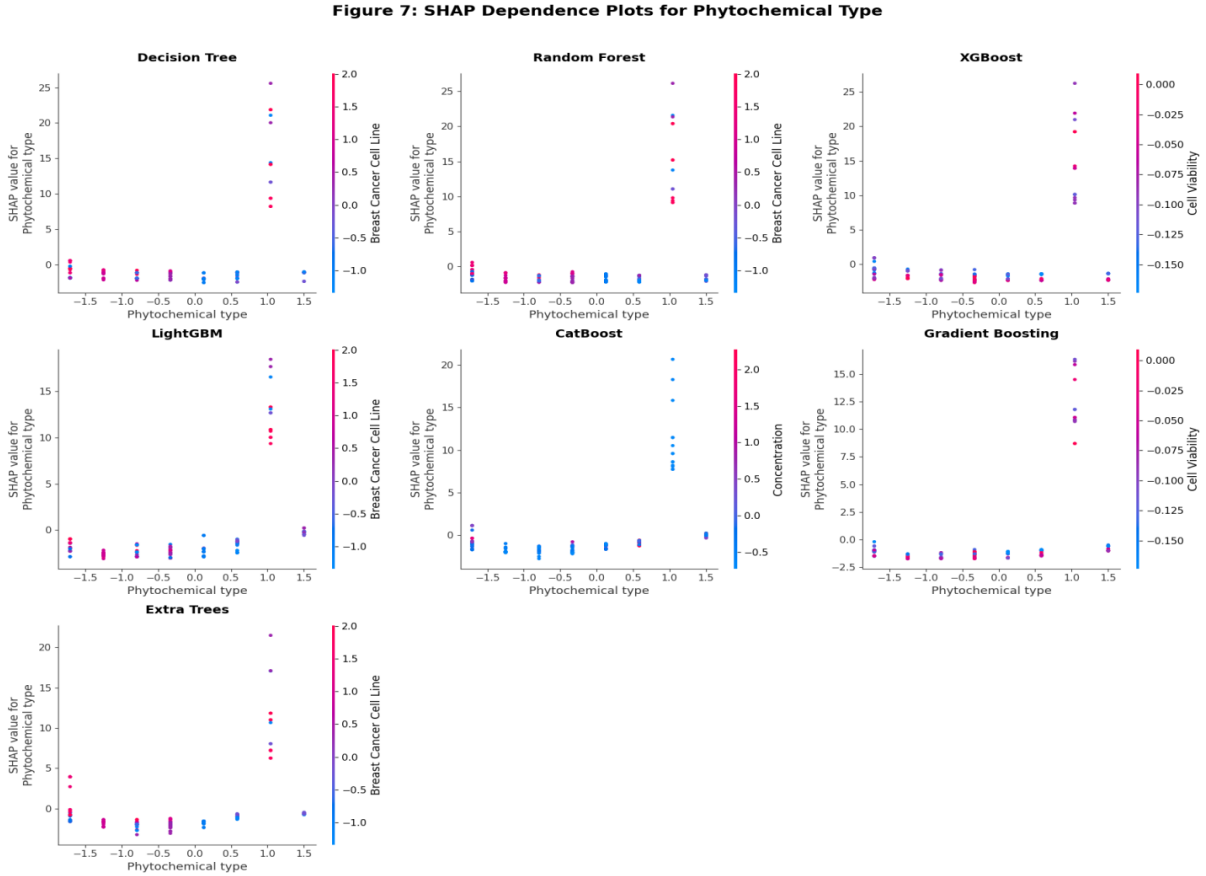


Figure 5. SHAP dependence plots illustrating the effect of phytochemical type on model predictions across the best-performing tree-based machine learning algorithms. Each point represents an individual observation, where the SHAP value quantifies the contribution of phytochemical type to the predicted apoptosis response. Point colors indicate interacting features, including breast cancer cell line, concentration, or cell viability, depending on the model. Across all models, specific phytochemical categories consistently produced larger positive SHAP values, confirming phytochemical type as the most influential predictor of apoptosis in the analyzed dataset.

DISCUSSION:

The present study shows that machine learning can be a useful tool for predicting phytochemical-induced apoptosis in breast cancer cell lines and for identifying compounds that may have stronger anticancer potential. By bringing together multiple regression models, phytochemical ranking, association rule mining, SHAP-based explainability, and statistical analysis, the proposed framework provides a broader understanding of how different experimental factors influence apoptosis. Overall, the findings suggest that tree-based ensemble models are better suited than simpler statistical models for handling the complex and nonlinear relationships present in phytochemical datasets.

The study also highlights the potential role of AI in reducing the amount of experimental screening required during early-stage phytochemical research. The final dataset contained 1,347 experimental observations collected from 54 peer-reviewed studies and included more than 16 phytochemical–drug combinations tested across different concentrations and exposure durations. By learning from previously reported experimental data, machine learning models may help researchers identify the compounds and conditions that are most promising before carrying out additional laboratory experiments.

Another strength of the study was the use of experiment-level data splitting. Rather than randomly dividing individual observations, all observations from the same experiment were kept together in either the training or testing dataset. This is important because measurements taken from the same experiment, even at different concentrations, may be highly related. If such observations were split between training and testing datasets, it could lead to data leakage and artificially high model performance. The approach used in this study therefore improves the reliability of model evaluation.

One of the main findings of this study was the strong performance of ensemble learning models. Among the ten models evaluated, Extra Trees achieved the highest training R^2 value of 0.9853, followed very closely by XGBoost ($R^2 = 0.9852$), Random Forest ($R^2 = 0.9839$), and Decision Tree regression ($R^2 = 0.9837$). In comparison, Linear Regression showed much lower performance, with an R^2 of 0.5005. This difference suggests that apoptosis induction cannot be explained well by simple linear relationships alone. Instead, it is likely influenced by several factors acting together, such as phytochemical type, concentration, treatment duration, breast cancer cell line, and combination treatment.

The better performance of tree-based models is understandable from a biological point of view. Phytochemicals do not always produce a uniform response across different cell lines or concentrations. Their effects may increase sharply beyond certain concentration levels or may vary depending on the biological characteristics of the cell line. Tree-based ensemble methods are able to capture these kinds of nonlinear relationships and interactions more effectively than traditional linear approaches. This may explain why Extra Trees, XGBoost, and Random Forest performed particularly well in this study.

The predicted-versus-actual apoptosis plots also supported the model performance results. Extra Trees, XGBoost, Random Forest, Decision Tree, and CatBoost showed predictions that were generally closer to the ideal 1:1 line, indicating better agreement between predicted and experimentally observed apoptosis values. In contrast, Linear Regression and SVR showed greater deviation from the reference line. KNN also showed larger prediction errors, especially at higher apoptosis values. These findings further indicate that tree-based models were more capable of capturing the variability present in the dataset.

The learning curves provided additional insight into how the models behaved as more training data were added. Extra Trees, CatBoost, XGBoost, Random Forest, Gradient Boosting, and Decision Tree showed relatively good convergence between training and cross-validation scores. Extra Trees, in particular, showed stable performance with only a small gap between the training and validation curves at higher sample sizes. This suggests that the model was able to learn consistent patterns from the available data. However, because the performance values reported in Table 1 are training results, the final selection of the best-performing model should also be supported by independent test-set performance and cross-validation results.

The statistical analysis provided additional support for the machine learning findings. The ANOVA results showed a significant difference in apoptosis across groups, with $F = 5.016$ and $p < 0.001$. Dosage also differed significantly between groups, whereas cell viability did not show a statistically significant difference, with $p = 0.316$. This suggests that apoptosis may be a more sensitive variable for distinguishing between the study groups than average cell viability in this dataset.

Interpretation and Decision Support

The phytochemical ranking also provided useful biological insights. Curcumin showed the highest average apoptosis value at 52.87%, followed by resveratrol at 46.13%. Quercetin and genistein showed moderate apoptotic activity, while fisetin, kaempferol, and sulforaphane showed comparatively lower average apoptosis values. Within the present dataset, curcumin and resveratrol therefore appeared to have the strongest apoptosis-inducing potential.

These findings should, however, be interpreted carefully. The average apoptosis value of a phytochemical can be influenced by several experimental factors, including concentration, treatment duration, cell line, and whether the compound was used alone or in combination with another therapeutic agent. Therefore, a lower-ranked phytochemical should not automatically be considered biologically less effective. It may still show strong anticancer activity under specific conditions.

The observed effects are also consistent with the known ability of phytochemicals to act on multiple cancer-related signaling pathways. The study highlights pathways such as PI3K/Akt/mTOR, MAPK/ERK, NF-κB, Wnt/β-catenin, JAK/STAT, and p53-related apoptotic mechanisms. Because breast cancer is biologically heterogeneous, compounds that act on multiple pathways may be especially useful for further investigation. Nevertheless, the present ranking should be considered a method for prioritizing compounds for additional laboratory studies rather than as direct evidence of clinical effectiveness.

Association Rule Mining provided another layer of interpretation by identifying common relationships between concentration, dosage, viability, and apoptosis. The most frequent pattern was the combination of low apoptosis with low concentration, with a support of 38.8%. High concentration was also frequently associated with high apoptosis and low cell viability, with support values of 34.0% and 33.8%, respectively. These patterns suggest a general concentration-related trend in which higher phytochemical exposure is often associated with increased apoptosis and reduced viability.

However, these association rules should not be interpreted as evidence of cause and effect. A high concentration occurring together with high apoptosis does not necessarily mean that concentration alone is responsible for the response. Different phytochemicals may behave differently, and each breast cancer cell line may have a different level of sensitivity. The main value of Association Rule Mining in this study is that it helps identify useful patterns that can guide future laboratory experiments.

The SHAP analysis made the machine learning results easier to interpret. Phytochemical type was found to be the most influential predictor of apoptosis, followed by breast cancer cell line, cell viability, and concentration. Dosage showed the smallest relative contribution in the SHAP summary plot. This suggests that the biological identity of the phytochemical itself may have a stronger influence on apoptosis than dosage considered on its own. Breast cancer cell line was also an important predictor, which is biologically meaningful. Different breast cancer cell lines represent different molecular characteristics and may respond differently to the same compound. The SHAP dependence plots further showed that certain phytochemical categories consistently contributed positively to predicted apoptosis across several tree-based models. This supports the idea that phytochemical effects should be evaluated in the context of specific breast cancer cell types rather than assuming that all cell lines will respond in the same way.

Cell viability also contributed to apoptosis prediction. This relationship is expected because reduced cell viability and increased apoptosis often occur together following effective anticancer treatment. However, the two outcomes are not identical. Cell viability reflects the overall number of surviving or metabolically active cells, whereas apoptosis represents programmed cell death. Including cell viability in the model may therefore help capture the overall response to treatment, although future studies should examine how strongly this variable is related to the target outcome.

The combination of machine learning, Association Rule Mining, and explainable AI is an important strength of the study. A model that provides only high predictive accuracy may have limited value in biomedical research if researchers cannot understand why a particular prediction was made. By adding SHAP analysis and association rules, the framework provides information not only about which phytochemicals are predicted to produce higher apoptosis, but also about the factors that contribute to those predictions.

Limitations and Future Directions

At the same time, several limitations need to be considered. First, the dataset was collected from different published studies. As a result, there may be differences in laboratory protocols, cell culture conditions, assay methods, treatment duration, compound purity, and reporting practices. Although data preprocessing and standardization were carried out, some level of variation between studies is likely to remain.

Second, the study is based on in-vitro breast cancer cell-line experiments. While these models are useful for initial screening, they cannot fully represent the complexity of living organisms. Factors such as metabolism, bioavailability, tumor microenvironment, immune response, and systemic toxicity cannot be assessed from cell-line data alone. Therefore, strong predicted apoptosis should not be interpreted as direct evidence of preventive or therapeutic benefit in patients.

Third, several tree-based models showed very high training R^2 values, approaching 0.99. Although this reflects strong fitting to the training dataset, it may also raise concerns about overfitting if similar performance is not observed on the independent testing dataset. The use of experiment-level splitting and cross-validation helps reduce this risk, but reporting complete test-set results would provide stronger evidence of model generalizability.

Another point that should be addressed before publication is the consistency of the reported performance metrics. Some of the RMSE and MAE values in Table 1 appear unusual relative to one another and should be rechecked to ensure that all metrics were calculated using the same scale and the same set of observations. This is especially important when comparing the performance of different models.

The ANOVA analysis also appears to use a smaller number of observations than the full machine learning dataset. The manuscript should clearly explain which subset of the data was used for the ANOVA analysis and why. This would make it easier for readers to understand how the statistical analysis relates to the overall machine learning results.

Future studies could improve the framework by adding more detailed molecular information about the phytochemicals. For example, chemical descriptors, molecular fingerprints, physicochemical properties, target proteins, pathway information, gene-expression profiles, and other omics data could be incorporated. This may

allow the models to move beyond predicting the responses of phytochemicals already present in the dataset and potentially estimate the activity of newly identified compounds.
Independent external validation will also be important. The best-performing model should ideally be tested on a completely separate dataset that was not used during model development. A further step would be prospective laboratory validation, in which the model is used to identify promising phytochemicals and the predicted apoptosis values are then compared with experimentally measured results.
In the future, the framework could also be expanded to include concentration-response modeling, breast cancer subtype-specific predictions, combination therapy analysis, uncertainty estimation, and multi-parameter ranking based on apoptosis, viability, toxicity, and selectivity. Integrating machine learning with molecular docking, genomics, proteomics, and pathway analysis may also provide deeper biological insights.

CONCLUSION :

This study suggest that machine learning can effectively capture complex patterns in phytochemical-based breast cancer experiments. Extra Trees, XGBoost, and Random Forest showed particularly strong predictive performance, while curcumin and resveratrol ranked highest for average apoptosis within the analyzed dataset. Association Rule Mining identified meaningful concentration-related patterns, and SHAP analysis showed that phytochemical type and breast cancer cell line were major contributors to apoptosis prediction. Together, these findings support the use of interpretable AI as a supportive tool for phytochemical screening and prioritization. Further experimental and external validation will be necessary before the framework can be applied to translational or clinical breast cancer prevention research.

REFERENCES:

1. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2021;71(3):209–249.
2. World Health Organization. *Breast cancer*. Geneva: World Health Organization; 2024.
3. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2024. *CA Cancer J Clin.* 2024;74(1):12–49.
4. Loibl S, Poortmans P, Morrow M, Denkert C, Curigliano G. Breast cancer. *Lancet.* 2021;397(10286):1750–1769.
5. Waks AG, Winer EP. Breast cancer treatment. *JAMA.* 2019;321(3):288–300.
6. Harbeck N, Penault-Llorca F, Cortes J, et al. Breast cancer. *Nat Rev Dis Primers.* 2019;5:66.
7. Sharma P, McClees SF, Afaq F. Pleiotropic effects of phytochemicals in cancer prevention and therapy. *Biomed Pharmacother.* 2017;96:505–526.
8. Farooqi AA, Butt G, Razzaq Z. Alleviation of breast cancer through phytochemicals. *Semin Cancer Biol.* 2021;69:233–245.
9. Ko EY, Moon A. Natural products for chemoprevention of breast cancer. *Arch Pharm Res.* 2015;38(1):11–28.
10. Aumeeruddy MZ, Mahomoodally MF. Phytochemicals in breast cancer prevention. *Molecules.* 2019;24(22):4295.
11. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. *Nat Med.* 2019;25:24–29.
12. Vamathevan J, Clark D, Czodrowski P, et al. Applications of machine learning in drug discovery and development. *Nat Rev Drug Discov.* 2019;18(6):463–477.
13. Stokes JM, Yang K, Swanson K, et al. A deep learning approach to antibiotic discovery. *Cell.* 2020;180(4):688–702.
14. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 2016. p. 785–794.
15. Breiman L. Random forests. *Mach Learn.* 2001;45(1):5–32.
16. Cortes C, Vapnik V. Support-vector networks. *Mach Learn.* 1995;20(3):273–297.
17. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* 2015;521(7553):436–444.
18. Bommasani R, Hudson DA, Adeli E, et al. On the opportunities and risks of foundation models. arXiv. 2021.
19. Wu Z, Pan S, Chen F, et al. A comprehensive survey on graph neural networks. *IEEE Trans Neural Netw Learn Syst.* 2021;32(1):4–24.
20. Arrieta AB, Díaz-Rodríguez N, Del Ser J, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion.* 2020;58:82–115.
21. Zhavoronkov A, Vanhaelen Q, Oprea TI. Will Artificial Intelligence for Drug Discovery deliver on its promise? *Nat Rev Drug Discov.* 2020;19(6):367–368.
22. Paul D, Sanap G, Shenoy S, et al. Artificial intelligence in drug discovery and development. *Drug Discov Today.* 2021;26(1):80–93.
23. Schneider P, Walters WP, Plowright AT, et al. Rethinking drug design in the artificial intelligence era. *Nat Rev Drug Discov.* 2020;19(5):353–364.
24. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nat Rev Genet.* 2015;16(6):321–332.

25. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature.* 2021;596(7873):583–589.
26. Pratheeksha S, Sindhu R, Haripriya R, Dhamodhar D, Prabu D, Rajmohan M, et al. Transforming fibromyalgia management: a systematic review of AI and ML driven approaches. IJDSIR. 2026 Mar;9(2):114-130.