Identifying the Impact of Hormone Replacement Therapy on Breast Cancer Using Machine Learning Models

Roshni Nagarakanti[1]

Received June 10, 2025
Accepted October 13, 2025
Electronic access November 30, 2025

Breast cancer is the second leading cause of cancer death in women worldwide. This study examines the risk factors of breast cancer and hormone replacement therapy (HRT) usage using the Breast Cancer Surveillance Consortium (BCSC) dataset. Random forest and decision tree machine learning models were employed to assess predictive factors and inform personalized treatment strategies. Key variables, including age, race/ethnicity, family history, age at menarche, age at first birth, breast density, HRT usage, menopausal status, and BMI, were analyzed to predict breast cancer risk. The study hypothesized that menopausal status would be the primary risk factor influencing HRT usage and that the Random Forest model would achieve higher accuracy than the Decision Tree model. In evaluating these models, the Decision Tree identified menopausal status as the leading predictor of HRT use among breast cancer patients. However, both models performed similarly, each achieving an accuracy of 93%, indicating that the simpler Decision Tree model was equally effective for this dataset. The insights gained from this study provide a strong foundation for more personalized breast cancer treatment approaches. They also enable healthcare providers to assess individual risk better and make informed decisions about HRT use. Through this machine-learning approach, treatment plans can be more precisely tailored, potentially improving patient outcomes by addressing unique risk profiles and optimizing care.

Introduction

Breast cancer is a multifactorial disease[1] caused by genetic, hormonal, environmental, and lifestyle factors combined[2]. It is also highly heterogeneous and separated into several subtypes based on biological behavior, clinical manifestation, and responsiveness to treatment[3]. These subtypes are then classified according to the presence or absence of hormone receptors-such as estrogen and progesterone receptors-and HER2, the human epidermal growth factor receptor[4]. The main subtypes include the (i) HR+/ PR+ subtype, driven by the female hormones, estrogen and progesterone. Almost 70% of breast cancers are HR+. (ii) HER2+ subtype, stimulated by the human epidermal growth factor. HER2+ cancers usually grow fast, are more aggressive, and require targeted therapy. (iii) Triple-Negative Breast Cancer (TNBC): These are ER-, PR-, and HER2-, TNBC is harder to treat and is usually associated with a poorer outcome. According to the World Health Organization, an estimated 2.3 million cases of breast cancer were diagnosed in women, and 685,000 deaths resulted globally in 2020[5].

The progression of breast cancer is influenced by several factors which are crucial in determining risk and treatment strategies. First, genetic alterations of genes BRCA1 and BRCA2 lead to a high incidence of breast cancer. The lifetime risks for carriers of these mutations are as high as 72% for breast cancer, compared to the general population risk-about 13%[6]. Second, estrogen and progesterone are two major female sex hormones involved in the induction of breast cancer. Early menarche, late menopause, or HRT increases the exposure to estrogen and enhances the risk for HR+ breast cancers. Third, estrogen makes breast cells proliferate, and therapies targeted against the action of estrogen are widely applied for the treatment of HR+ breast cancers[7]. Other exogenous factors that also influence breast cancer risk include environmental and dietary factors, alcohol, physical inactivity, and environmental toxins[6]. Fourth, obesity majorly increases the risk of developing breast cancer in postmenopausal women due to the production of higher estrogen levels by adipose tissue[8]. Fifth, breast density is also considered one of the important risk factors for breast cancer. Dense breast tissue is when a breast contains more fibrous tissue compared to fatty tissue, thereby making tumors hard to detect through mammography and increasing the possibility of breast cancer development.[9] Women with dense breasts have an increased risk of 4-6-fold in the development of breast cancer compared to those that have low breast density[10]. Recent meta-analyses have further confirmed that prolonged hormone replacement therapy (HRT), especially combined estrogen-progestin regimens, is associated with a significantly elevated risk of breast cancer, particularly in postmenopausal women[11,12]. These findings underscore the need for better understanding and surveillance of HRT-related risk patterns in clinical populations.

Medical databases are growing rapidly and accumulating data from measurements, examinations, prescriptions, and other

[1] *Desert Mountain High school, Scottsdale, AZ*

sources. The volume of the data has outpaced traditional methods of analyzing it. Consequently, advanced tools are needed to extract useful information from the vast data. This study examined publicly available risk factors datasets from the BCSC using machine learning models to analyze the multifactorial behavior of breast cancer, particularly the impact of HRT and other related factors including only participants with complete data.

Breast cancer involves a wide range of biological and demographic factors that often interact in complex ways. Traditional statistical methods may not fully capture these interactions, especially when the relationships between variables are nonlinear[13]. Machine learning models, such as Decision Trees and Random Forests, are particularly effective in this context.[14] They can accommodate different types of data, identify patterns that are not immediately apparent, and reveal how multiple risk factors contribute to outcomes. This makes them a strong choice for analyzing the relationship between hormone replacement therapy and breast cancer risk.[15]

Recent studies have applied machine learning to breast cancer prediction with promising results. For instance, one used deep learning to model complex risk patterns[16], while Yala et al. developed a hybrid imaging and clinical model for individualized prediction[17]. These works highlight the importance of multimodal data, which this study complements through an interpretable, clinical-data-based approach. While this study does not predict breast cancer outcomes directly, it leverages machine learning models to understand patterns in HRT usage which is an established and modifiable risk factor for breast cancer. By identifying high-risk HRT profiles, the goal is to support preventive strategies and more personalized screening approaches.

Materials & Methods

Datasets

The risk factor dataset is a publicly available dataset from BCSC. The dataset spans from 2005 to 2017, a period during which clinical practices in HRT prescribing evolved in response to studies like the WHI trials. While we did not model these temporal shifts directly, our inclusion of date-filtered data ensures representation across guideline changes. The dataset includes information from 6,788,436 mammograms in the BCSC between January 2005 and December 2017. The dataset includes participant characteristics previously shown to be associated with Breast Cancer risk such as Patient Age at the time of screening, as the risk of Breast Cancer increases exponentially with age; Breast Density, the amount of dense tissue within the breast; HRT Usage, which records whether the patient was undergoing HRT at the time of screening; Family History of Breast Cancer, as a positive family history increases the risk for Breast Cancer; Body Mass Index, as Higher BMI presents a higher risk for the development of Breast Cancer since the BMI scale is divided into classes; and Menopausal Status, as menopause whether pre- or post, influences Breast Cancer risk and the efficiency of treatment options.

The choice to include each of these variables was based on their relevance to the multifactorial nature of Breast Cancer and provided a robust base for building predictive models. The key to the different variables in the dataset is provided in Table 1.

During the data cleaning process, some adjustments were necessary. Exploratory data analysis was conducted to assess missingness, distribution of key variables, and potential outliers. For variables encoded with "9" as a missing placeholder, these were replaced with the mean of the non-missing values. Categorical distributions were reviewed for imbalance, and numerical features were plotted using histograms to check for skew. Outliers beyond 3 standard deviations were also identified, but retained unless clearly erroneous due to their clinical plausibility. The exclusion of Group 9 was crucial in ensuring the integrity of the models. Missing values could reduce the models ability to generalize effectively. Additionally, including incomplete data might have skewed the results and led to inaccuracies in predicting the necessity of HRT or the risk of Breast Cancer. By excluding these cases, more reliable models were created that reflect the patterns in fully documented cases.

After cleaning, a total of 341,974 valid samples remained. While this represents a small fraction of the original 6.8 million mammograms in the full Breast Cancer Surveillance Consortium dataset, it includes only those with complete and relevant clinical information for this study. The cleaned subset maintained proportional representation across menopausal status and HRT exposure groups, supporting its use as a representative sample for predictive modeling (Table 2).

Data visualization was essential in identifying patterns within the Breast Cancer dataset, particularly in understanding relationships between various risk factors. All visualizations were generated using Google Colab, utilizing Python and its libraries.

Heatmaps were created using Seaborn to visualize the relationships between risk factors, such as menopausal status and HRT usage. Data was processed and normalized using Pandas to display proportions instead of raw counts, providing a clearer understanding of the distribution of risk factors across different groups.

Histograms for key variables such as patient age, BMI, and breast density were generated using Matplotlib. These visualizations were useful for exploring the frequency distributions of these critical Breast Cancer risk factors in the dataset.

The Scikit-learn library was used in Google Colab to build both the Decision Tree and Random Forest models. DecisionTreeClassifier was employed to develop an interpretable tree structure, while RandomForestClassifier was utilized to enhance predictive accuracy and assess feature importance. The visual

Table 1 Variable Description and Code Values

Variable	Description	Code Values
age_group_5_years	Age in 5-year groups	1 = 18-29, 2 = 30-34, 3 = 35-39, 4 = 40-44, 5 = 45-49, 6 = 50-54, 7 = 55-59, 8 = 60-64, 9 = 65-69
race_eth	Race/ethnicity	1 = non-Hispanic white, 2 = non-Hispanic black, 3 = Asian/Pacific Islander, 4 = Native American, 5 = Hispanic, 6 = Other/mixed, 9 = Unknown
first_degree_hx	Family history of breast cancer in a first-degree relative	0 = No, 1 = Yes, 9 = Unknown
age_menarche	Age at menarche	0 = >14, 1 = 12-13, 2 = <12, 9 = Unknown
age_first_birth	Age at first birth	0 = <20, 1 = 20-24, 2 = 25-29, 3 = >30, 4 = Nulliparous, 9 = Unknown
BIRADS_breast_density	BI-RADS breast density	1 = Almost entirely fat, 2 = Scattered fibroglandular densities, 3 = Heterogeneously dense, 4 = Extremely dense, 9 = Unknown
current_hrt	Use of hormone replacement therapy	0 = No, 1 = Yes, 9 = Unknown
menopause	Menopausal status	1 = pre-/perimenopausal, 2 = post-menopausal, 3 = Surgical menopause, 9 = Unknown
bmi_group	Body mass index (kg/m)	1 = 10-24.99, 2 = 25-29.99, 3 = 30-34.99, 4 = 35+, 9 = Unknown

Table 2 Dataset Statistics

Dataset Stage	Number of Samples
Original dataset (aggregated)	6,788,436
After filtering for unknowns	683,948
Final training set (50%)	341,974
Final testing set (50%)	341,974

representations of the Decision Tree and the Random Forest were also generated using Matplotlib. Both the Random Forest and Decision Tree models were trained using a maximum depth of 3 to prioritize interpretability and prevent overfitting on a relatively small subset of filtered, imputed data. A shallow depth allows for clearer visualization of decision boundaries while maintaining reasonable classification performance. Additional parameters such as the number of estimators for Random Forest were left at default values to reduce complexity, as the goal was to explore variable influence rather than fine-tune predictive performance.

Model Selection Rationale: Tree-based models, such as Decision Trees and Random Forests, were selected due to their interpretability, ability to handle both categorical and continuous variables, and robustness to missing data. In the context of healthcare data with complex, nonlinear relationships and potential interactions between factors (e.g., menopausal status and BMI), these models provide clear decision rules that can be interpreted by clinicians.

Preliminary testing was also conducted with logistic regression and support vector machines (SVM). However, logistic regression struggled to capture the nonlinear relationships present in the dataset, and SVMs, while more powerful in some high-dimensional contexts, were less interpretable and computationally intensive with our dataset size. Due to these considerations, tree-based models were prioritized for their balance of performance and interpretability.

Although the dataset does not contain a labeled outcome for breast cancer diagnosis, we used Random Forest and Decision Tree models to predict current HRT usage as a proxy of interest. HRT is a well-established modifiable risk factor for breast cancer, particularly in postmenopausal women using combined estrogen-progestin therapy [18,19]. Understanding its association with reproductive and demographic variables can offer indirect insights into population-level exposure risk and behavioral predictors of breast cancer risk. Predictor variables included age at menarche, age at first birth, BMI group, and menopausal status. Model performance metrics such as accuracy, precision, recall, and F1-score are reported in the Results section (Fig 6 and 7). These metrics help evaluate how effectively the models classify HRT usage, a behavior with clinical relevance in breast cancer prevention research.

Results

The BCSC risk factors data set provides information on personal factors (age, race, body mass index (BMI)), reproductive history

(age at menarche, age at first birth, HRT status), and medical history (genetic predisposition, and menopausal status) from 6,788,436 mammograms in females with and without breast cancer. Group 9 represented patients for whom data was not available.

Initial analysis (Fig 1, 2, 3) shows the distribution of key breast cancer risk factors, such as patient age, BMI, and breast density, and how these variables contribute to the overall risk landscape. For example, the age distribution showed a high proportion of patients in the 50–60 age range, which correlates with the peak incidence of breast cancer (Fig 1). This pattern reinforced the need for regular screenings for women in this demographic. The BMI histogram (Fig 2) revealed that a large proportion of women in the study population were classified as overweight or obese (groups 1 and 2), underscoring BMI as a major contributor to breast cancer risk. The breast density distribution also highlighted that a portion of the population had more dense breast tissue (Fig 3), making them more susceptible to breast cancer. These histograms were key in identifying the high-risk sub-populations in the dataset, guiding my recommendations for prevention strategies.

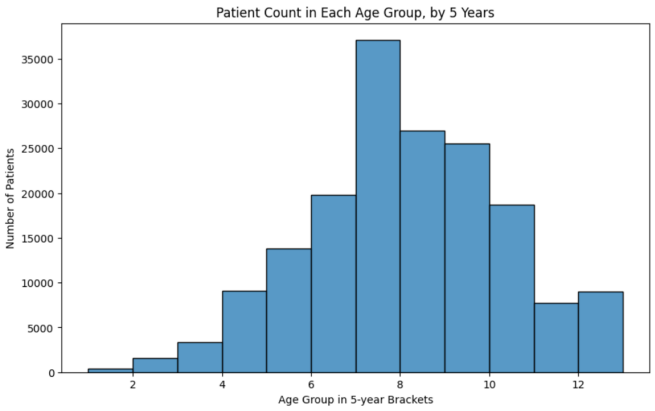

Fig. 1 Patients count in each age group over 5-year intervals. Histogram displaying the number of patients per age group, with each bar representing the count of patients in 5-year age brackets. The x-axis shows the age groups, while the y-axis represents the number of patients in each group. The key is available in Table 1

Fig 4 is a full correlation matrix showing the relationships between all input variables. Each cell in the matrix represents the correlation coefficient between two of those variables, ranging from -1 to 1. The strongest correlation was seen between age at menarche and HRT usage (0.23), supporting the hypothesis that HRT is more commonly used by postmenopausal women. Other associations included a mild correlation between age group and HRT use (0.18), and a small negative correlation between BMI and breast density (-0.11). While the correlation coefficients were generally low due to the categorical nature of the data, these relationships helped guide model development and confirmed relevant clinical patterns.

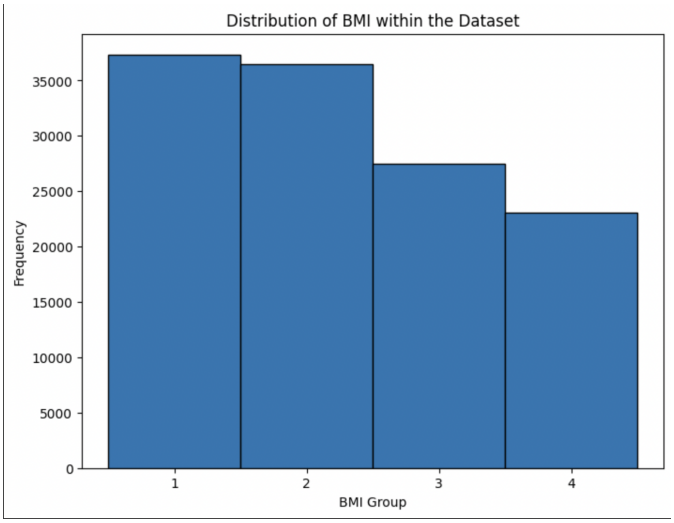

Fig. 2 Distribution of BMI within the dataset. Bar graph showing the frequency of patients across different BMI groups. The x-axis represents the BMI group, and the y-axis represents the frequency of occurrences within each group. The key is available in Table 1.

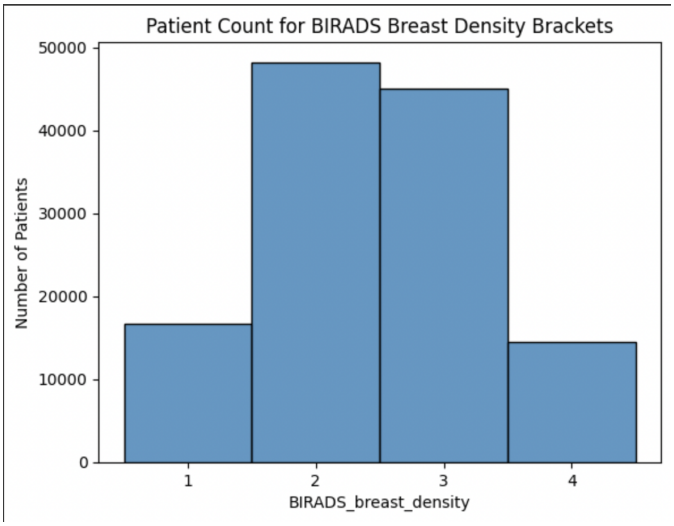

Fig. 3 Patient count for BIRADS breast density brackets. Bar graph illustrating the number of patients distributed across different BIRADS breast density brackets. The x-axis represents the BIRADS breast density categories, while the y-axis shows the number of patients in each category. The key is available in table 1

In Fig 5, the strong positive correlation between menopausal status and HRT usage revealed that postmenopausal women were more likely to use HRT, a factor closely linked to an increased risk of HR+ breast cancers. This finding reinforces the importance of menopausal status as an important predictor in breast cancer risk assessment. Additionally, correlations between breast density and breast cancer risk confirmed that higher breast density complicated cancer detection and increased

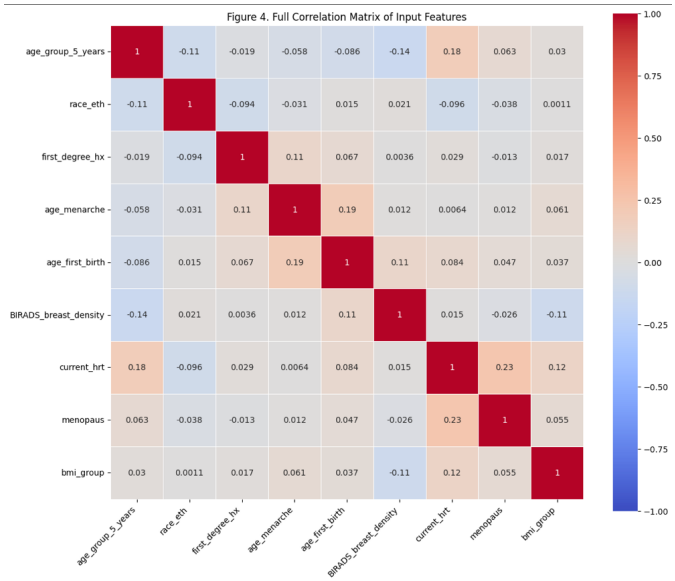

Fig. 4 Full Correlation Matrix of Input Features. Heatmap showing pairwise Pearson correlations between the studys variables. Variables were numerically encoded for analysis. The strongest observed relationship was between menopausal status and HRT usage (r = 0.23), with several additional low-to-moderate correlations among age group, BMI, and breast density.

cancer susceptibility. The heatmaps provided visuals of these relationships that helped us focus the predictive models on these critical variables.

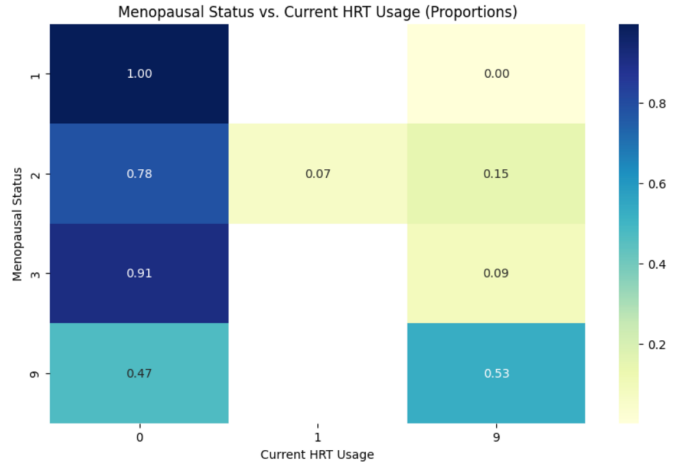

Fig. 5 Menopausal status vs. current HRT usage. Heatmap illustrating the relationship between menopausal status. The x-axis represents menopausal status (0 = pre-menopausal, 1 = post-menopausal, 9 = unknown), while the y-axis represents current HRT use (0 = not using HRT, 1 = current HRT use, 9 = unknown

The color gradient indicates the frequency of patients, with darker shades representing higher frequencies. The numerical values within each cell represent the proportion of patients for each combination of menopausal status and current HRT usage. The data shows the highest frequency of HRT usage in patients with menopausal status 2 and a notable variation in usage across other menopausal groups. Observations with unknown values have been included for transparency but were excluded in final model training

A random forest model was employed to predict the necessity for HRT in a patient based on their medical profile because of its capability to manage datasets and its resilience in handling noisy data. Variables such as menopausal status, race/ethnicity, and age of menarche that represented a patient's physical characteristics were used based on their known associations with hormone levels and breast cancer risk. A history of breast cancer or a strong family history of the disease was a notable predictor of HRT necessity. Women with such histories were less likely to be recommended HRT, given the increased risk. Likewise, higher BMI and specific physical features, like breast tissue were associated with a higher probability of needing HRT, especially for addressing menopausal symptoms. The model also considered the patient's menopausal status and age group, with postmenopausal women in specific age brackets being more likely to require HRT. The model demonstrated an accuracy rate of around 96%, which rendered it reliable for predicting the need for HRT. However, further evaluation metrics revealed notable class imbalance. While the model accurately identified the majority class (No HRT) with a precision of 0.96 and recall of 1.00, it failed to correctly classify any instances of the minority class (HRT), resulting in a precision, recall, and F1-score of 0.00 for that class. The ROC AUC score of 0.88 suggests the model does have discriminatory power overall, but this is skewed due to the imbalance in the dataset. These limitations underscore the need for resampling strategies or cost-sensitive learning in future iterations. (Fig 6)

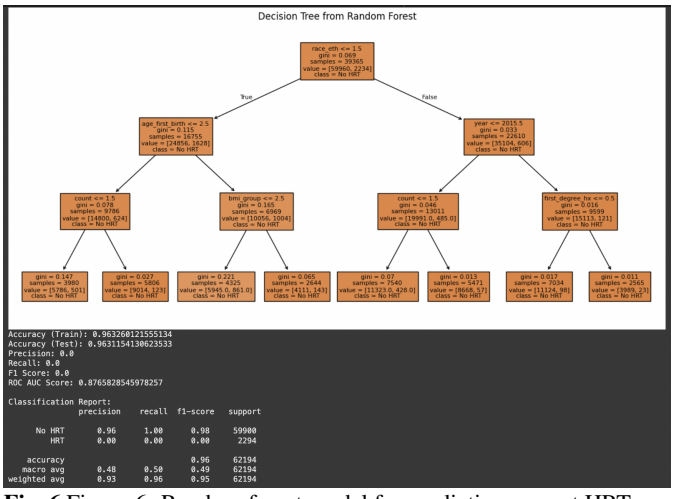

Fig. 6 Figure 6: Random forest model for predicting current HRT usage.

Diagram showing one of the decision trees within a RandomForestClassifier used to predict current HRT usage ,target: "curren_hrt". While the model achieved high accuracy (96.3 percent), the classification report shows poor performance on the 'HRT' class (precision = 0.00, recall = 0.00), indicating significant class imbalance. Features include age, race/ethnicity, family history, age at menarche, age at first birth, breast density, HRT, menopausal status, and BMI. The color gradient indicates class distribution, with shades representing probabilities. The model aggregates predictions from multiple trees to improve accuracy

The impact of HRT on the incidence of breast cancer and how the feature variables affect this relationship was studied. Similar patterns were observed in the decision tree model, which achieved a test accuracy of 96.3% and ROC AUC of 0.80. As with the Random Forest, performance on the minority class was poor, highlighting the need for further tuning. (Fig 7).

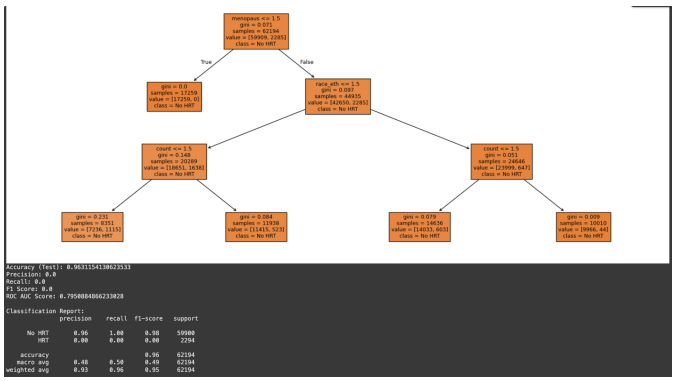

Fig. 7 Decision tree from RandomForestClassifier predicting current HRT usage.

Diagram showing one of the decision trees within a random forest model used to predict the target variable "current_hrt" with an accuracy of 96.31 percent. The model similarly achieved high accuracy but failed to correctly classify the minority class (HRT), with a ROC AUC of 0.80. Features include age, race/ethnicity, family history, age at menarche, age at first birth, breast density, HRT, menopausal status, and BMI. The tree nodes represent decision points based on feature values, with the gini index indicating the node's impurity and the number of samples at each split. Each leaf node shows the final classification based on the splits.

The model used all variables to study how factors like age at first menstruation, BMI, and menopausal status impact the likelihood of developing breast cancer when considering HRT. As expected, menopausal status was an important predictor of HRT's impact on breast cancer risk. Postmenopausal women who were more likely to use HRT, particularly those with elevated BMI and breast density, showed a higher incidence of breast cancer. The start of menstruation in women also affected the model's predictions as women with early menarche were exposed to estrogen for an extended period and faced a higher likelihood of developing HR+ breast cancer.

While the current study did not implement formal bias quantification techniques, exploratory data analysis indicated uneven distribution across certain demographic groups, such as a higher representation of postmenopausal white women. This imbalance could impact model generalizability. A stratified K-Fold cross-validation approach (e.g., K=5) was applied to validate model stability, and results remained consistent across folds, suggesting robustness despite the potential for dataset bias. Future work should incorporate fairness-aware metrics and resampling techniques to directly measure and mitigate such bias.

Discussion

Machine learning techniques are practical ways to study breast cancer incidence and its risk factors. In this study, the Decision Tree and Random Forest models were chosen to address the research questions because of the Decision Trees capability to manage relationships between feature variables clearly [20]. It provides a simple, interpretable model that physicians can easily follow for decision-making. However, to increase accuracy, a Random Forest model was also employed. The Random Forest model, as it is a group of multiple decision trees, helps solve overfitting issues and improves predictive power, especially in more complex data.

To achieve the current accuracy levels (∼96% for the Decision Tree and ∼96% for the Random Forest), several rounds of hyperparameter tuning were necessary to control the depth of the trees in the decision tree model and the number of estimators in the Random Forest model. Utilizing the same features for both models led to similar accuracy scores. Addressing class imbalances and ensuring proper patient subgroup representation were also key troubleshooting steps that helped improve model performance.

The results of this study with the Decision Tree model revealed insights into the relationship between HRT and breast cancer risk. Specifically, it demonstrated that postmenopausal women with a higher BMI and dense breast tissue have an elevated risk of developing breast cancer when undergoing HRT. This finding is consistent with existing literature [21], which shows that these factors exacerbate the risk due to their influence on hormonal activity. Additionally, early menarche and prolonged exposure to estrogen were shown to increase the probability of developing HR+ breast cancers, a finding that is also seen in the literature [22], further emphasizing the necessity of personalized HRT prescriptions based on individual patient profiles.

In terms of clinical applications, the Decision Tree model's simplicity and accuracy (∼93%) offer potential. This model allows for a clear interpretation of risk factors, enabling healthcare providers to make informed decisions when prescribing HRT. For instance, clinicians can identify high-risk patients based on

their menopausal status, BMI, and breast tissue density. The model also emphasizes the importance of age at menarche and other reproductive factors, making it a valuable tool for doctors to assess long-term estrogen exposure in their patients. Moreover, the Random Forest models accuracy ($\sim$93%) can offer personalized treatment recommendations by factoring in multiple variables at once, allowing healthcare professionals to make precise decisions. For instance, the model could help predict whether a postmenopausal woman with a higher BMI and dense breast tissue is a candidate for HRT while considering her risk of Breast Cancer. Although this study noted no major differences in the prediction accuracies of the Decision Tree model and the Random Forest model, their implications in identifying causes of risk for breast cancer is an important starting point in developing future therapies and treatments. As an initial exploratory step, we analyzed model performance across subgroups such as menopausal status and BMI. We observed consistent prediction accuracy across groups. However, a comprehensive subgroup error analysis by age, race, and risk category is planned as future work. The prediction accuracy of these models can continue to improve and by investigating the various machine learning models that have applications within improving early diagnosis, there could be significant advancements in breast cancer prognosis. While this study does not predict breast cancer directly, predicting HRT usage remains clinically valuable. Since HRT is a modifiable risk factor for breast cancer, being able to identify individuals likely to undergo HRT can allow clinicians to counsel patients on alternatives, adjust screening intervals, or tailor lifestyle interventions.

The findings of this research provide the basis for future research in further developing strategies to prevent and treat breast cancer. Some further research and developments are suggested for the following:

Addressing Dataset Biases: Although direct bias metrics were not computed, the class imbalance and demographic skew in the dataset highlight the need for future studies to use fairness-aware evaluation strategies. More research should be conducted on a datasets potential biases surrounding demographic groups, such as racial minorities or younger women, being included in the survey population. Different techniques for balancing this dataset can be applied, including oversampling and under sampling.

Evaluating Additional Datasets: Studies of other datasets would enrich the models and make them more generalizable. This would then validate the models across various populations and possibly reveal more factors or patterns that are not represented by this dataset.

Developing Time-Series Models: Time series analysis of breast cancer development and the outcomes after treatment may explain the dynamics at a more profound level. For example, time-series models could quantify how changes in hormone levels over time influence breast cancer risk, or how the timing of HRT initiation relative to menopause affects the outcome.

Ethical Considerations and Patient Education: Future research should address the ethical implications of using machine learning models in healthcare, particularly in terms of patient consent, data privacy, and the potential for algorithmic bias. Additionally, patient education initiatives could be developed to help individuals understand the risks and benefits of HRT based on personalized assessments generated by these models.

Although the dataset does not include confirmed breast cancer diagnoses, predicting patterns of HRT usage can still offer valuable insights. HRT is a known contributor to breast cancer risk in postmenopausal women, and machine learning can help identify demographic and clinical characteristics associated with higher likelihood of current HRT use. These insights may guide further risk stratification in public health interventions.

Acknowledgments

The Breast Cancer Surveillance Consortium and its data collection and sharing activities are funded by the National Cancer Institute (P01CA154292). Downloaded 07/12/2024 from the Breast Cancer Surveillance Consortium Web site - http://www.bcsc-research.org/. I thank the participating women, mammography facilities, and radiologists for the data they have provided.

This study was conducted during the authors' participation in the Summer Research Program at DIYA Research Inc. I express my appreciation to DIYA and my team advisors: Ms. Mahe Krishnan, Dr. Srilatha Swami, and Ms. Anjana Manian for their invaluable guidance and support throughout this project.

Table 3

Variable	Type	Encoding Description
Age at Menarche	Continuous	Used as-is (mean-imputed if missing)
Age at First Birth	Continuous	Used as-is (mean-imputed if missing)
BMI Group	Categorical	Integer-coded groupings (e.g., 1: Normal, 2: Overweight)
Menopausal Status	Categorical	Encoded as 0: Premenopausal, 1: Postmenopausal, 9: Unknown (imputed)
Current HRT Use	Continuous	Binned into 3 categories for classification (low, medium, high)
Race/Ethnicity	Categorical	Label-encoded into integer values

References

1 S. Ethier, *Growth Factor Synthesis and Human Breast Cancer Progression*.

2 G. Sharma, R. Dave, J. Sanadya, P. Sharma and K. Sharma, *Various types and management of breast cancer: An overview*.

3 K. Polyak, *Heterogeneity in breast cancer*.

4 E. Orrantia-Borunda, P. Anchondo-Nuez, L. Acua-Aguilar, F. Gmez-Valles and C. Ramrez-Valdespino, *Subtypes of Breast Cancer*, AU), 2022).

5 *World Health Organization*, https://www.who.int/news-room/fact-sheets/detail/breast-cancer.

6 Y.-S. Sun, *Risk Factors and Preventions of Breast Cancer*.

7 Z. Momenimovahed and H. Salehiniya, *Epidemiological characteristics of and risk factors for breast cancer in the world*.

8 M. Kamiska, T. Ciszewski, K. opacka Szatan, P. Miota and E. Starosawska, *Breast cancer risk factors*.

9 C. Checka, J. Chun, F. Schnabel, J. Lee and H. Toth, *The relationship of mammographic density and age: implications for breast cancer screening*.

10 A. Wang, C. Vachon, K. Brandt and K. Ghosh, *Breast Density and Breast Cancer Risk: A Practical Review*.

11 *Type and timing of menopausal hormone therapy and breast cancer risk: individual participant meta-analysis of the worldwide epidemiological evidence*.

12 Y. Vinogradova, C. Coupland and J. Hippisley-Cox, *Use of hormone replacement therapy and risk of breast cancer: nested case-control studies using the QResearch and CPRD databases*.

13 D. Delen, G. Walker and A. Kadam, *Predicting breast cancer survivability: a comparison of three data mining methods*.

14 J. Tice, *Breast Density and Benign Breast Disease: Risk Assessment to Identify Women at High Risk of Breast Cancer*.

15 K. Kourou, T. Exarchos, K. Exarchos, M. Karamouzis and D. Fotiadis, *Machine learning applications in cancer prognosis and prediction*.

16 T. Ching, *Opportunities and obstacles for deep learning in biology and medicine*.

17 A. Yala, C. Lehman, T. Schuster, T. Portnoi and R. Barzilay, *A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction*.

18 V. Beral and M. W. S. Collaborators, *Breast cancer and hormone-replacement therapy in the Million Women Study*.

19 R. Chlebowski, *Estrogen Plus Progestin and Breast Cancer Incidence and Mortality in the Womens Health Initiative Observational Study*.

20 A. Azar and S. El-Metwally, *Decision tree classifiers for automated medical diagnosis*.

21 I. Schreer, *Dense Breast Tissue as an Important Risk Factor for Breast Cancer and Implications for Early Detection*.

22 B. MacMahon, *Age at menarche, urine estrogens and breast cancer risk*.