Taylor & Francis
Taylor & Francis Group

Applied Artificial Intelligence
An International Journal

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/uaai20

Feature Selection-based Machine Learning Comparative Analysis for Predicting Breast Cancer

Chour Singh Rajpoot, Gajanand Sharma, Praveen Gupta, Pankaj Dadheech, Umar Yahya & Nagender Aneja

To cite this article: Chour Singh Rajpoot, Gajanand Sharma, Praveen Gupta, Pankaj Dadheech, Umar Yahya & Nagender Aneja (2024) Feature Selection-based Machine Learning Comparative Analysis for Predicting Breast Cancer, Applied Artificial Intelligence, 38:1, 2340386, DOI: 10.1080/08839514.2024.2340386

To link to this article: https://doi.org/10.1080/08839514.2024.2340386

Published online: 10 Apr 2024.

APPLIED ARTIFICIAL INTELLIGENCE
2024, VOL. 38, NO. 1, e2340386 (23 pages)
https://doi.org/10.1080/08839514.2024.2340386

Feature Selection-based Machine Learning Comparative Analysis for Predicting Breast Cancer

Chour Singh Rajpoot[a], Gajanand Sharma[b], Praveen Gupta[c], Pankaj Dadheech[d], Umar Yahya[e], and Nagender Aneja[f,g]

[a]School of Computing Science and Engineering, VIT Bhopal University, Bhopal, India; [b]Department of Computer Science and Engineering, JECRC University, Jaipur, Rajasthan, India; [c]Department of Computer Science and Engineering, GITAM School of Technology, GITAM (Deemed to be University), Visakhapatnam, India; [d]Department of Computer Science & Engineering, Swami Keshvanand Institute of Technology, Management & Gramothan (SKIT), Jaipur, India; [e]Department of Computer Science, Faculty of Science, Islamic University in Uganda, Kampala, Uganda; [f]Department of Computer Science, Purdue University, West Lafayette, Indiana, USA; [g]School of Digital Science, Universiti Brunei Darussalam, Darussalam, Brunei

ABSTRACT

Breast cancer is a serious disease, and therefore early detection is crucial for successful treatment and patient management. Unfortunately, globally, the number of breast cancer cases is increasing due to various multifaceted factors. It is currently one of the leading causes of cancer deaths in women, worldwide. Cancerous cells in the breast can form lumps that impact the patient's health, and even seemingly harmless tumors could be fatal if undiagnosed early enough. Fortunately, artificial intelligence techniques have proven effective in detecting diseases, and doctors can therefore use them to effectively and accurately diagnose breast cancer early. This paper explores the use of genetic algorithms, ant colony optimization, and Hybrid Hopfield Neural Network-E2SAT (HHNN-E2SAT) models, for breast cancer prediction. The HHNN-E2SAT models outperform standard algorithms like the Random Forest and Support Vector Machines, achieving over 98% on all performance metrics (i.e. Accuracy, F1-score, Sensitivity, Specificity, and Precision).

ARTICLE HISTORY
Received 31 July 2023
Revised 26 March 2024
Accepted 29 March 2024

Introduction

A fundamental challenge in bioinformatics or clinical research is the inability to accurately identify important information (Embi and Payne 2009). As such, diagnostics is an active and evolving field of medicine (Bolboacă 2019). Breast cancer is scored highly among various cancers because of its associated effects on a patient (Ataollahi et al. 2015). Many lives could potentially be saved from breast cancer deaths if intelligent data-driven methods are developed to a level of real-world application (Ahn et al. 2023). Previous studies highlight the use of feature-based data mining (DM) techniques in the prediction of various

CONTACT Umar Yahya umar.yahya@iuiu.ac.ug Department of Computer Science, Faculty of Science, Islamic University in Uganda, P.O Box 7689, Kampala, Uganda

diseases, breast cancer inclusive (Ahn et al. 2023; Amin, Kia Chiam, and Dewi Varathan 2019; Amrane et al. 2018; Ganggayah et al. 2019b). To analyze raw data such as the primary causes of mortality in cancer patients, and to offer fresh perspectives on illness prevention with precise forecasts, data mining methods generally offer a promising alternative (Ahn et al. 2023; Amin, Kia Chiam, and Dewi Varathan 2019; Ganggayah et al. 2019b). Being able to identify any type of cancer, particularly breast cancer, as early as possible is essential for proper patient management, thus minimizing cancer fatalities. Like most types of cancer, breast cancer, is a complex disease often influenced by a combination of various factors (Zhang et al. 2020). These include genetic, environmental, and lifestyle, among other factors. Table 1 shows key factors associated with cancer as extracted from (Ataollahi et al. 2015; CDC 2023; Petrucelli, Mary, and Tuya 1993).

As a result of advancements in digital data storage and processing, large amounts of clinical diagnostic data are nowadays available from various diagnostic centers, hospitals, research centers, as well as electronic repositories accessible via the world wide web. As data availability continues to be seamless digitally, it is imperative that intelligent data-driven methods for classification and rapid detection of diseases, breast cancer inclusive, be explored (Ahn et al. 2023). It should be noted, however, that medical diagnosis ought to be based

Table 1. A description of cancer-associated factors as established from literature.

S.No	Key factors	Category	Description
1.	Genetic Factors (Petrucelli, Mary, and Tuya 1993)	BRCA1 and BRCA2 Mutations	Inherited mutations in these genes significantly increase the risk of breast and ovarian cancers.
		Family History	A family history of breast cancer, especially in first-degree relatives (mother, sister, daughter), may elevate the risk.
2.	Environmental Factors (Parsa 2012)	Radiation Exposure	High levels of exposure to ionizing radiation, especially at a young age, can increase the risk of breast cancer.
		Hormone Replacement Therapy (HRT)	Long-term use of certain hormone replacement therapies, particularly with combined estrogen and progesterone, may elevate risk.
		Reproductive Factors	
3.	Lifestyle Factors (Anand et al. 2008; Khan, Afaq, and Mukhtar 2010)	Physical Inactivity	Lack of regular physical activity has been linked to a higher risk of breast cancer.
		Diet	A diet high in saturated fats and low in fruits and vegetables may contribute to increased risk.
		Alcohol Consumption	Regular and excessive alcohol consumption is associated with an elevated risk of breast cancer.
		Obesity	Being overweight or obese, especially after menopause, has been linked to an increased risk.
4.	Hormonal Factors (National Council Institute 2015)	Estrogen Exposure	Prolonged exposure to estrogen without the counterbalancing effects of progesterone (as seen in some hormone replacement therapies) can increase risk.
		Oral Contraceptives	Long-term use of oral contraceptives may slightly elevate the risk.
5.	Personal Health Factors (Ataollahi et al. 2015)	Breast Density	Women with dense breast tissue may have a higher risk.
		Previous Breast Cancer	Having had breast cancer in one breast increases the risk of developing it in the other breast or in a different part of the same breast.

not only on the medical practitioner's training but also their experience (Bolboacă 2019). This is a premise on which data-driven intelligent methods for diagnosis become a feasible alternative, as it is possible to aggregate both the expert knowledge and experience in the design of intelligent diagnostic systems (Arbaiy et al. 2017; Huang et al. 2023; Kattan 2001; Pietro 1985). Advancements in diagnostics notwithstanding, numerous difficulties, such as erroneous diagnosis, are still prevalent in intelligent medical diagnostic systems (Balogh, Miller, and Ball 2015). As the growing medical databases contain multidimensional heterogeneous data such as examination records, measurements, tests, prescriptions, etc., there is a need for more adaptive and advanced methods for extraction of meaningful information needed to achieve intelligent feature-based medical diagnosis (Ellis, Sander, and Limon 2022; Reyes et al. 2021). This work thus conducts a comprehensive comparative analysis of feature-based machine learning methods for breast cancer prediction.

Clinically, a multitude of procedures can be used to accurately identify breast cancer, including mammography, magnetic resonance imaging (MRI), breast examination, thermography, and tissue sampling, among new emerging others (Bethesda 2023). These clinical procedures generate data used for training Artificial Intelligence (AI) based breast cancer diagnosis systems. The need for AI-based methods can also be justified by the inherent changes faced by these traditional detection methods (Budh and Sapra 2024). For instance, whereas X-rays-based methods have commonly been used to detect cancer, because the number of extracellular carcinomas is so small, it is extremely difficult to diagnose breast cancer at an early phase (Jaglan, Dass, and Duhan 2019). On the other hand, mammography can detect cancer in its earliest stages, with the procedure taking only a few minutes (Takkar et al. 2017). Figure 1 shows the difference between normal and malignant images.

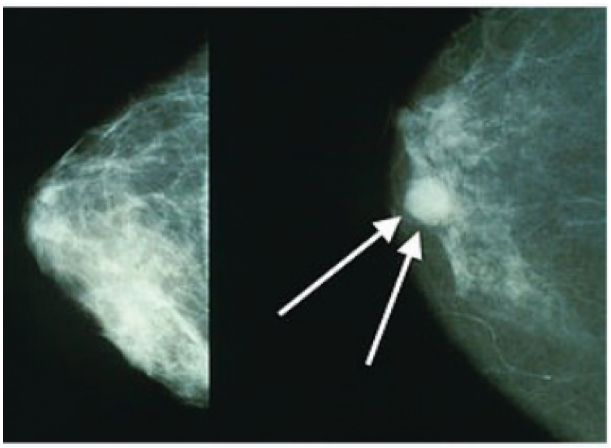

Figure 1. (Left side) Normal breast and (Right side) cancerous breast.

Early discovery of malignant breast cancer can lead to better treatment and reduce fatalities in breast cancer patients. This study explores various machine learning techniques for breast cancer prediction, through the extraction of significant traits (features) from complex data sets.

Using PCA (Principal Component Analysis) and hybrid machine learning classifiers, this study aims to reduce the dataset's size by focusing on the most important aspects for accurate diagnosis. This study aims to detect breast cancer in the early stage in order to improve survival chances of the patient. The remainder of the paper is arranged as follows: Section 2 presents a review of previous related work, while Section 3 explains the methodology proposed in the current study. Section 4 describes the validation of the proposed hybrid machine learning classifier using existing techniques on the breast cancer dataset. Section 5 concludes with the study's findings and recommendations for further research.

Literature Review

As the quantity and scope of databases that store medical data continues to expand quickly, conventional learning approaches can no longer suffice to analyze and look for unique patterns and knowledge concealed inside this large volume of medical data. Hence, there is an increasing need for innovative tools and methods to find meaningful information from large data repositories. Previous studies have utilized various machine learning methods in the detection or forecast or classification of breast cancer (Hajiabadi et al. 2020; Mümine 2019; Serhat et al. 2020; Uddin et al. 2023; Vikas, Pal, and Tiwari 2018; Vivek et al. 2020). By eliminating the less useful factors in detection, classification, and forecasting methods (Arunadevi and Ganeshamoorthi 2019), PCA reduces the number of unrelated parameters into a collection of principal variables (Amin, Kia Chiam, and Dewi Varathan 2019). Variable selection is therefore important in prediction and decision analysis, particularly when working with medical information. Rather than predicting cancer patient survival using all dataset's features, which can result in complex and unintelligible outcomes and representations, parameter selection is crucial to creating a better prediction model that only uses an integral parameter set (Amrane et al. 2018; Dana and Shubair 2016; Ganggayah et al. 2019b).

To identify breast cancer, popular classifiers like random forest (RF), support vector machines (SVM) and Decision Tree (DT) have been commonly employed (Dana and Shubair 2016; Mümine 2019; Uddin et al. 2023). Additionally, dimensionality reduction using PCA and other methods has also been previously used for selecting salient features (Khandezamin, Naderan, and Javad Rashti 2020). In a previous study (Khandezamin, Naderan, and Javad Rashti 2020), researchers developed a system that used SVM, distinguishing between benign and malignant

tumors with excellent classification accuracy and little computing effort. The authors utilized hybridized PCA (principal component analysis) and a variety of classifiers on several datasets of breast cancer that obtained great performance. Similarly, to detect breast cancer, the author in presented a method built on memetic Pareto artificial neural networks (Muhammad et al. 2019).

Artificial Neural Networks (ANN) are a computing method created to simulate how the nervous system executes a function using widely dispersed multiprocessing of the different basic nodes. Such components, also known as nodal or nerve cells for computing, are characterized neurologically because they retain facts or develop capabilities and make them accessible to the user by modifying values. Several experts suggested and invented various ways to use ANN to detect and treat cancer in the breast due to the efficient measurement that it provided and then for their good precision toward an accurate diagnosis of disease (Jalal Abdullah, Mohammed Hasan, and Waleed 2019). By their fundamental design, neural networks are particularly adept at processing intricate datasets, such as those from breast cancer research, in which the input data and goal predictions interact unpredictably (Uddin et al. 2023).

Support Vector Machine (SVM) is a supervised training technique for developing regression and classification principles using information. It is a supervised machine learning system that performs well in pattern recognition applications. If there are numerous characteristics and samples, SVM performs well in classification tasks. The SVM algorithm creates a binary classifier. The SVM method has frequently been utilized for prediction and classification of breast cancer (Akay 2009; Huang et al. 2017). The common application use-case is for the SVM classifier to identify malignant growths as part of the cancer diagnosis process, particularly of the breast in females (Akay 2009; Huang et al. 2017; Jalal Abdullah, Mohammed Hasan, and Waleed 2019).

Random Forest (RF) algorithm (RFA) has been described as a generic rule for randomized predictors (Tomislav et al. 2018). The binary tree is typically divided into identical vertices by iteratively updating to create an RF. The information propagation from the root node increases the success or node's resemblance to the parent node. The source information is gathered using bootstrapping sample size to produce many trees for growing RF. Regarding collecting predictive variables as feed to some trees, every branch inside the RF leads to activation. The dataset's independent variables can be effectively handled by RF (Dana and Shubair 2016; Uddin et al. 2023). A group of various separate randomized decision trees collaborate to form RF. These trees are produced by bootstrapping given information. Any individual tree inside the random forest throws one value based on the set of predictor values entered. The category that receives the most results determines the model's forecast in categorizing

a training set. RF is produced by repeatedly splitting the tree structure into similar nodes. Through transmission, the tree's root influences the child node's resemblance.

Decision Tree (DT), a non-linear supervised algorithm, could be applied to classification and regression problems and is often employed in health care (Elhazmi et al. 2022). The root node of a Decision Tree (DT) architecture is one of the multiple vertices. All vertices are connected by edges, except for the root, and each node contains a single initial. Certain vertices, known as internal nodes or sample vertices, provide one or even more external connections, but others do not. Vertices, often known as exit or endpoint, are all in this category. Every intermediate node in a decision tree is responsible for dividing the example space into two or more subsets, and it does so by a specific discontinuity function determined by the input similarity measure. Many diseases, like cancer, diabetes, and heart disease, have previously been diagnosed using DT (Fatin Kadhim 2022). DT is a robust and accurate training algorithm that may address classification and regression issues. That employs the highest, tree-based advancement technique. A layered division approach is used to split the data among several groups at every stage to ensure that the information within every category is similar. Each internal network of the DT ties to a trial parameter, ties to something like the results obtained at each route, and ties to a distinct category for every sub-tree. A tree can grow from either parent node by choosing an "optimal feature" or "perfect feature" from the available characteristics utilizing unpredictability or knowledge to obtain measurements before performing "dividing".

David (Omondiagbe, Veeramani, and Sidhu 2019) proposed breast cancer diagnosis based on feature selection and an SVM Classifier. Haji Abadi et al (Hajiabadi et al. 2020) suggested that linear discrimination analysis was applied using a typical attitude to minimize the quantity of features, before ANN applying three loss functions (i.e. the current cup, a hinge, and cross-entropy) was trained to evaluate the dataset at different volume levels. SVM, Logistic regression, and KNN classifiers have equally been recommended for the analysis of breast cancer (Omondiagbe, Veeramani, and Sidhu 2019; Rabiei et al. 2022). SVM was found to be one of the best classifiers. Breast cancer's prognostic qualities have not changed at all. Mogana Darshini Ganggayah et al. (2019a) investigated using supervised machine-learning algorithms to identify breast cancer, making use of breast cancer datasets from the UCI repository (UCI 2019). On their dataset, they used Logistic Regression, SVM, and KNN. In addition, their efficiency was estimated and compared. They determined that SVM was the best predictive analytic technique, with a 92.7% accuracy.

More recently, away from classical machine learning, application of deep learning methods in the diagnosis of cancer and other human-related health illnesses has gained increasing popularity (Anari et al. 2022).

Recent literature indicates that Convolutional Neural Networks (CNN) are particularly better fitted model for cancer diagnosis tasks compared to other deep learning models (Kasgari et al. 2023; Nazanin et al. 2022; Ramin et al. 2023). CNN has been successful for breast tumor segmentation and detection using mammograms (Ramin et al. 2023). Brain tumor localization and segmentation from magnetic resonance imaging (MRI) are difficult and crucial challenges for a variety of medical analysis applications (Nazanin et al. 2022)

Table 2 summarizes the literature reviewed, taking into account; datasets utilized, number of classification classes, methods and techniques used, as well as the accuracy or outcome obtained from the classification/prediction/forecast task.

Materials And Methods

Description of Dataset

The breast cancer database is taken from the UCI repository (UCI 2019). A total of 699 tumor cases in the database were reduced to 659 after data preprocessing. Benign tumor cases are 458 (65.5%), and malignant cases are 241 (34.5%). The nine dataset attributes are presented in Table 3, excluding code number and category level. Each feature was measured in the domain range from 1 to 10, where value 1 represents benign and 10 represents malignant cases.

In the current research work, instances are classified into two types: benign as positive and malignant patients as negative. Linear correlation describes straight-line correlations in the range 1 to +1 for two variables, where 1 represents the ideal negative association and +1 represents the perfect positive relationship. The Pearson correlation between positive and negative classes is established by determining the link between nine features of benign and malignant classes.

Data Pre-Processing

Data pre-processing replaces missing values, identifies and eliminates external factors, and resolves subjective discrepancies. For example, the code number form has been deducted from the database because it does not affect diseases. As a result, the database has 16 missing value values. Medium replaces attributes that do not exist for that class. In addition, the database uses random selection to ensure proper data rotation.

Table 2. Summary of related literature reviewed.

Author/ Year	Dataset	Number of Classes	Methodology/ Techniques	Accuracy/Outcome
(Wang, Cao, and Yu 2022)	Data taken in this research paper ST (Spatio-Temporal) data types, ST data instances, and ST data formats are used	There are three classes Local class, High class, and Low-class prediction-done based on DL(Deep Learning)Model.	According to the authors, deep learning models such as RNN, CNN, LSTM and spatio-temporal data mining are offered for setting the objectives.	Learning techniques for applications like flexibility, and on-demand services, including logistics and crime analysis.
(Ali et al. 2021)	The data in this study include a. event data, b. reference data, c. trajectory data points d. and raster data, among other things.	This work finds three classes Class A(square), Class B(rectangle), and Class C(triangle), etc.	Techniques used in this paper are artificial intelligence, machine learning, data mining, etc.	Contains research from more than a few centuries.
(Dana and Shubair 2016)	The data set used in this study was obtained from http://archive.ics.uci.edu/ml. (William and Mangasarian 1993)	In this work, classes 0 and 1, class' 0'mean No heart disease, and class'1' mean 'presence of heart disease.'	Authors implemented algorithms and compared their results of nave bayes and logistic regression. Author also worked with neural network with fuzzy decision tree, extreme learning machine and decision tree. The author also worked on support vector machine, nave bayes cart neural network with a genetic algorithm. (Dana and Shubair 2016)	Accuracy of Naïve Bayes and Logistic Regression 87.41%, SVM 86.76%, Extreme Learning Machine 86.50%.
(Hartama, Perdana Windarto, and Wanto 2019)	In this research work, data was taken from an education organization.	The author describes two classes in this working class: state-owned and privately owned.	In this work author implemented 1. knowledge discovery in database process, 2. data selection, 3. pre-processing, cleaning, 4. transformation, data mining, 5. interpretation evaluation etc.	Accuracy: 81.71%
(Mughal 2018)	This research paper uses data sets like web data, etc.	Structured information and unstructured information	Decision tree, naive bayes, support vector machine, neural network	Identified significant information for cancer diagnoses.
(Sohail et al. 2019)	In this research work, 2800 research articles were reviewed.	Five disease classes: - a. Heart disease, b. Breast cancer, c. lung cancer, Diabetes, d. Skin cancer, etc.	Data mining and machine learning technique are implemented as stated in this review article.	Retrieved relevant information on cancer diagnoses.

(*Continued*)

Table 2. (Continued).

Author/ Year	Dataset	Number of Classes	Methodology/ Techniques	Accuracy/Outcome
(Yusupova et al. 2020)	Dataset taken from hospital.	Poisoning substance, non-poisoning substance.	Medical data processing, data mining, complex analysis, etc.	Death ratio diagnosis based on toxicological data
(Yang et al. 2020)	In this research study data set was taken from UK Biobank (http://www.ukbiobank.ac.uk).	Catalogue of Somatic Mutations (COSMIC) cancer, Human Gene Mutation Database (HGMD) cancer.	a. Partition-based algorithm, b. Hierarchical clustering algorithm, c. Density-based algorithm, d. grid-based algorithm, Apriori algorithm, etc.	Efficient outcome based on health diagnosis
(Shadi et al. 2019)	This research work implemented on the real-time dataset.	Two classes of healthy patients, and epileptic patients.	Decision support systems	KDD (Knowledge Discovery in Database) of classification accuracy 99%
(Vikas, Pal, and Tiwari 2018)	Dataset taken from the repository of UCI Machine Learning.	Two classes here 1. Benign 2. Cancerous	(|NB) Naïve Bayes Algorithm, J48 Decision Tree and RBF network	Naïve Bayes accuracy is: 97.36% RBF Network accuracy is 96.77% J48 accuracy: 93.41%
(Vivek et al. 2020)	A dataset with 627,000 samples taken from the UCI repository.	1. Benign cell 2. Malignant cell	Author adopted techniques of Ada Boost with Decision Table[11]. J-Rip, J48, Lazy IBK, Lazy K-star, Multiclass Classifier, Multilayer Perceptron, R Forest with R Tree [11].	Scores outcome above 94%. Naïve Bayes accuracy is : 73.21% for Tree secondly Lazy classifier algorithms obtained accuracy of 99%.
(Serhat et al. 2020)	In this research work dataset taken from medical organization.	1. Death 2. Survivor	1. ANN (Artificial Neural Networks), 2. Logistic Regression, 3. Information fusion	Outcome survival time of breast cancer.
(Muhammad et al. 2019)	In this research work data set was taken from x-ray samples from a medical organization.	1. Breast Cancer 2. Non- Breast Cancer	1. Data Mining, 2. Bagging Algorithm, 3. IBk Algorithm, 4. Random Forest (RF) Algorithm, 5. Random Committee Algorithm, 6. Classification Algorithm	Accuracy:90%
(Omondiagbe, Veeramani, and Sidhu 2019)	Dataset was obtained from the WDBC.	1. benign tumor 2. malignant tumor	SVM, Naïve Bayes classifier	Classification accuracy: 98.82%, specificity: 99.07% sensitivity: 98.41%
(Tan et al. 2021)	This study uses data from the WDBC.	1. Cancer 2. Non-Cancer	Artificial Neural Network Loss function Robust loss function	K-Nearest Neighbour: 87% Kernel SVM:93% Random forest: 91% Support Vector Machine:94%

(*Continued*)

Table 2. (Continued).

Author/ Year	Dataset	Number of Classes	Methodology/ Techniques	Accuracy/Outcome
(Shravya, Pravalika, and Subhani 2019)	This study uses data from UCI ML Repository.	1. Cancer 2. Non-Cancer	1. SVM(Support Vector Machine) 2. KNN(K-Nearest Neighbor) 3. Logistic Regression 4. PCA(Principal Component Analysis) [17]	SVM(Support Vector Machine) : 92.7%
(Aavula and Bhramaramba 2019)	In this research paper, the dataset was taken from the National Cancer Institute (NCI) in the USA.	1. Cancer 2. Non-Cancer	1.Decision Tree Induction 2. Logistic Regression 3. SVM-RFSS 4. Neural Network	1. Naïve Bayes 95.85% 2. SVM-RFSS 98.90%
(Gopal et al. 2021)	Dataset from repository of the UCI used for Machine Learning repository.	1. Cancer 2. Non-Cancer	1. LASSO Logistic Regression (LLR) 2. Multilayer Perception (MLP) 3. Linear Regression	Accuracy: 94.05%
(Abdar et al. 2020)	This study is based on WDBC Cancer dataset.	1. Cancer 2. Non-Cancer	1. SV-Naive Bayes-Meta Classifier	SV-Naive Bayes-3-Meta Classifier accuracy: 98.07%
(Amrane et al. 2018)	Work uses dataset on WDCD(Wisconsin breast Cancer dataset).	1. Cancer 2. Non-Cancer	Naïve Bayesian Classifier (NBC), Cross-validation, Machine Learning (ML) technique, k-nearest neighbor (KNN)	KNN(K-nearest neighbor) accuracy :97.51% NB Classifier accuracy:96.19%
(Ganggayah et al. 2019a)	The dataset used is of University of Malaya Medical Center in Malaysia.	1. Cancer 2. Non-Cancer	1. Random forest 2. Decision tree 3. Support vector machine 4. Extreme boost 5. Neural networks.	Decision tree results : 79.8% Random forest algorithm accuracy: 82.7%

Table 3. Dataset attributes.

Attributes	Stage of Breast Cancer	Domain Range	Symbol
Mitosis	Malignant	1 to 10	X_9
Marginal adhesion	Benign	1 to 10	X_4
Cell extent uniformity	Benign	1 to 10	X_2
Cell form uniformity	Malignant	1 to 10	X_3
Nuclei (Bare)	Malignant	1 to 10	X_6
Bland chromatin	Malignant	1 to 10	X_7
Nuclei (normal)	Malignant	1 to 10	X_8
Cell extent for Single epithelial	Malignant	1 to 10	X_5
Thickness for clump	Malignant	1 to 10	X_1

Correlation Function in Feature Selection

The proposed model deals with 569 breast cancer samples with 32 attributes, of which only 11 are isolated. The features are designated using correlation-based feature selection (CFS) based on PCA-based trait assessment. This attribute removes duplicate or inappropriate features in the

validation dataset. PCA also solves the redundant matching problem by removing additional variables or alternatives that combine two or more variables. The scale from 0 to 1 identifies the variables that need not be removed. It improves the model's effectiveness and precision. The research examined 32 attributes, of which 11 were considered key attributes are mentioned in Table 4.

The extraction of 11 attributes was achieved using the WEKA tool's multifactor method. Training and test data are in the data set, with the covariate control to be performed as shown in the equation. (1) Principal component analysis (PCA) is used for statistical analysis and effective pretreatment to assess the scattering plot of breast cancer cells and analyze their characteristics. The process is performed using the call function training and testing data sets.

$$\sigma\ \mathrm{trt} = \mathrm{Cov}(\mathrm{tr} - \mathrm{T}) = \sum_{j=1}^{n} P[(rj - \mathrm{E}[\mathrm{tr}](\mathrm{tj} - \mathrm{E}(\mathrm{T})))] \tag{1}$$

Here, t is the test data, E is the anticipated sample, and T is the breast cancer data set. To identify which qualities are appropriate, it compares training and testing occurrences. Two data sets are correlated linearly, and Pearson's correlation coefficient is calculated using the covariance between the two models multiplied by the constant variance of each sample of genetic data. This process of normalization is denoted by Eq. (2).

$$\mathrm{CorCoeff}_{\mathrm{tr,t}} = \mathrm{Cov}(\mathrm{T}, \mathrm{tr})/(\mathrm{stdv}(\mathrm{tr})) \tag{2}$$

HYBRID Hopfield Neural Network-E2SAT MODELS for Classification

For achieving the correct synaptic weight in HNN-E2SAT models, HNN-E2SATGA and HNN-E2SATACO (Kasihmuddin, Sathasivam, and Asyraf Mansor 2017) are used in the learning phase and each neuron will receive a consistent interpretation, with minimal clausal inconsistencies used to

Table 4. Selected features from the breast cancer dataset.

Attribute	Description for attributes
A	Smoothness_worst
B	Symmetry_se
C	Concavity_se
D	Perimeter_se
E	Smoothness_se
F	Texture_se
G	Texture_worst
H	Fractal_dimension_mean
I	Concave points_se
J	Symmetry_mean
K	Concave points _mean

describe the cost function for the HNNE2SAT model (Hikmatul et al. 2020). HNN – the following algorithms replace HNN-E2SATGA, E2SATES, and HNN-E2SATACO (Hikmatul et al. 2020). The satisfiability problem is explained in detail with relevant examples in (Ali et al. 2017; Barrett, Dill, and Stump 2002; Sathasivam, Tajuddin, and Abdullah 2011; Velev 2004).

Algorithm of the Hybrid Hopfield Neural Network-E2SAT Model

(1) Translate E2 Satisfiability clauses into the Boolean calculation
(2) Assign neurons in E2 Satisfiability clauses to each variable
(3) Initialize to zero any synaptic weight (Hikmatul Fadhilah Sianipar et al. 2020).
(4) Check the E2 Satisfiability logic's inconsistency.
(5) Derive E2 Satisfiability Cost Function by assigning.

$$\mathrm{X} = \frac{1}{2}(1 + \mathrm{SX}) \text{and } \mathrm{X} = \frac{1}{2}(1 + \mathrm{SX})$$

The neuron states that it is true when $S_x = 1$ and false when $S_x = -1$

Derive all cost functions of E2 Satisfiability clauses on behalf of multiplication (Hikmatul et al. 2020).

(6) Check clauses on gratification through the use of EA, GA, and ACO (Hikmatul Fadhilah Sianipar et al. 2020).
(7) Obtained synaptic weight relates to the E2 Satisfiability logic.
(8) Calculate equation H to calculate the prediction of breast cancer.

$$P^{\theta}_{2\mathrm{SAT}} = -\frac{1}{2}\left(\sum \mathrm{i} \sum \mathrm{j} B^{(2)}_{ij} \mathrm{Si}\ \mathrm{Sj} - \sum \mathrm{i} B^{(1)}_{i} \mathrm{Si}\right)$$

(8) Apply Equation $\mathrm{hi} = \sum_j B^{(2)}_{ij} \mathrm{Sj} + B^{(1)}_{i}$ To find the neuron state in manipulation of the corresponding local field.
(9) Final Energy inspection means a prediction of breast cancer.

Experiment Results And Discussion

We individually used five machine learning methods: RF, DT, SVM, and ANN, to predict whether a cell is harmful or normal (i.e. benign or malignant). Processing was carried out using an Intel Core i7 with 32GB of RAM, using Python and its associated open-source libraries, in Jupiter Notebook.

In this particular way, taxonomic analysis is based on the exact traits that define the benign and fatal categories of tumor cases. In the early stages, symptoms are divided into seven stages and nine bad aspects of breast cancer.

Principal Component Analysis (PCA) is used for dimensionality reduction, and Classification techniques are used to anticipate breast cancer cases. The results of the experiments demonstrate that the suggested model properly classified 521 of the 569 occurrences while wrongly classifying 48 of them.

Performance Measure Parameters

A few performance parameters can be used to gauge how well machine learning approaches perform. Four types of confusion matrixes are used to evaluate the parameters: one for actual data and one for predicted data. The following is a breakdown of the meanings of the terms:

True Positive = TP
False Positive = FP
True Negative = TN
False Negative = FN

Our study employs the following parameters extensively to evaluate some phrases using their related formulae to measure the study's performance. The following formulae are used to gauge the effectiveness of the different classification methods in this current comparative study:

Accuracy (AC) is the ratio of properly classified samples from total samples:

$$\text{Accuracy(Acc)} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

Sensitivity (SE): Recall is another term for sensitivity. Percent of all positive cases that are considered to be "perceived" as positive:

$$\text{Sensitivity(Sen)} = \frac{TP}{TP + FN} \tag{4}$$

Specificity (SP): The rate at which breast cancer is projected to be present in all examples is known as specificity, defined as the correlation between any one set of observed negative criteria and any other group of observed negative examples.

$$\text{Specificity(Spec)} = \frac{TN}{TN + FP} \tag{5}$$

Precision (P):

$$\text{Precision(Prec)} = \frac{TP}{TP + FP} \tag{6}$$

Negative predictive value (NPV): The percentage of negative situations that remain true negatives is called NPV:

$$\text{Negative Predictive value(NPV)} = \frac{TN}{TN + FN} \tag{7}$$

F1 score: Harmonic mean of precision and sensitivity is distinct as F1 score:

$$\text{F1 Score} = \frac{2 * Precision * Recall}{Precision + Recall} \tag{8}$$

Table 4 provides the prediction outcome of techniques, respectively.

Performance Analysis of Proposed Model without Feature Selection

Here, two types of analyses (i.e., with feature and without feature selection) are carried out to test the efficiency of the suggested model with functioning techniques. Table 5 provides the validation analysis of all methods without feature selection in terms of Precision, Sensitivity, and Specificity. Figures 2 and 3 show the graphical comparison of techniques based on Specificity (SP) and Precision (P).

Considering precision, RF, and DT achieved nearly 74%, ANN reached 69.33%, SVM achieved 85.52%, and the proposed model achieved 87%. DT and SVM techniques achieved nearly 67% on sensitivity, while RF & ANN attained 70% and 76%, respectively, on sensitivity, and the proposed model

Table 5. LR-confusion matrix for ten-fold cross-validation.

	Benign	Malignant
Benign	TP = 45 (95.74%)	FP = 22 (32.84%)
Malignant	FN = 2 (4.26%)	TN = 45 (67.16%)

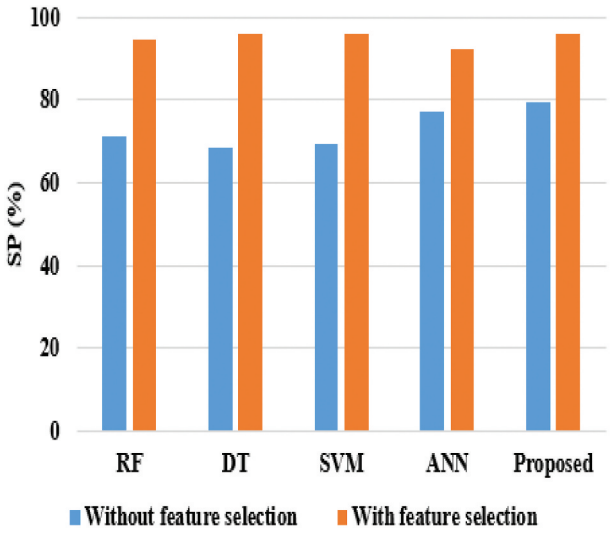

Figure 2. Comparative analysis in terms of Specificity (SP).

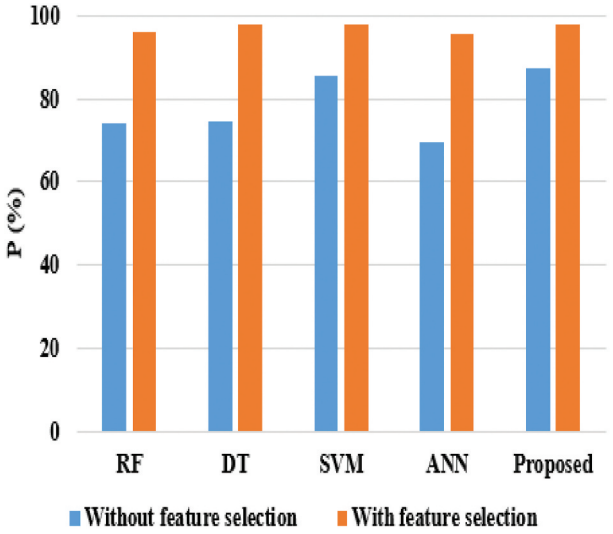

Figure 3. Comparative Analysis of Techniques in terms of Precision (P).

achieved the highest sensitivity score of 80.21%. The proposed model's better performance is that the HNN-E2SAT uses ACO and GA for gaining the synaptic weight, where DT is unstable, i.e., the structure of the optimal tree is highly affected by small changes in the training data. Hence, it shows low performance. In the analysis of specificity, DT and SVM achieved nearly 69%, RF gained 71.33%, ANN reached 77.27%, and the proposed model achieved 79.27%. Table 6 shows the experimental evaluation of the suggested model with existing techniques in terms of AC and F1-score.

In the analysis of the F1-score, the SVM, ANN, and RF achieved nearly 71% to 74%, DT reached 70.85%, and the proposed model achieved only 79.88%. The reason for the poor results of the proposed model is that it is tested without feature selection, and all attributes are considered. However, the proposed model achieved better performance than existing techniques. While analyzing the experiments on AC, the RF achieved low performance, i.e., 79.20%, DT, SVM, and ANN achieved nearly 84% to 86% of AC, and the proposed model achieved 87.45% of AC. The RF requires more sophisticated techniques for high classification accuracy, which is usually inferior to gradient-boosted trees. The comparative graphical representation of all classifiers in terms of SE is shown in Figure 4.

Table 6. Analysis of the model without feature selection.

Classifier	Precision (%)	Sensitivity (%)	Specificity (%)
RF	74.18	70.01	71.33
DT	74.64	66.67	68.27
SVM	85.52	68.84	69.62
ANN	69.33	76.40	77.27
Proposed	**87.13**	**80.21**	**79.27**

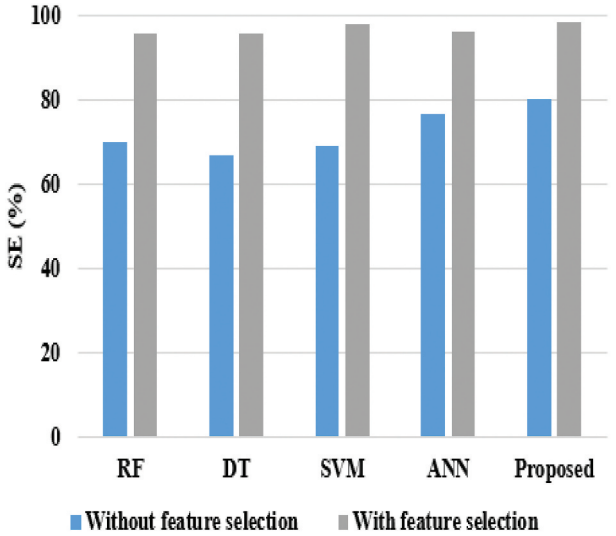

Figure 4. Comparative analysis terms of sensitivity (SE).

Performance Analysis of Proposed Model with Feature Selection

All the ML techniques, including the proposed model, are tested with feature selection, and the results are considered in Table 7.

When comparing Tables 5 and 6, the performance of all classifiers achieved better performance, as shown in Table 7. This is achieved because the feature selection technique plays a major role and uses only 11 attributes for breast cancer classification. In the analysis of sensitivity, RF and DT achieved nearly 95%, ANN and SVM achieved almost 97%, and the proposed model achieved 98.27%. When considering specificity, DT and SVM gained nearly 95%, ANN earned 92%, RF achieved 94%, and the proposed model reached 96%. When compared with all techniques, ANN and RF achieved nearly 95% at precision, SVM and DT achieved nearly 97%, while the proposed model achieved 98.02%. Even though SVM achieved better performance than other existing techniques, it equally performs lower than the proposed model. The reason is that SVM does not perform well when the data is large and contains more noise, i.e., overlapping target classes. Table 8 provides the comparative results of the proposed model in terms of Accuracy and F1-score. Figures 5 and 6 provide the comparison results of the proposed classifier in terms of AC and F1-score.

Table 7. Comparison of the proposed model without feature selection.

Classifier	F1-Score (%)	Accuracy (%)
RF	71.73	79.20
DT	70.85	82.98
SVM	75.82	85.49
ANN	72.64	86.34
Proposed	**79.88**	**87.45**

Table 8. Performance evaluation of various classification techniques with feature selection.

Classifier	Sensitivity (%)	Specificity (%)	Precision (%)
RF	95.74	94.65	95.83
DT	95.65	95.83	97.77
SVM	97.82	95.83	97.82
ANN	96.12	92.3	95.65
Proposed	98.27	96	98.02

On accuracy comparison, the proposed model achieved 98.57%, RF and DT achieved nearly 95%, while SVM and ANN achieved almost 97% as per Table 9. It shows that feature selection plays a major role in refining the performance of the all-classifier technique in diagnosing breast cancer. In the analysis of the F1-score, the SVM and ANN achieved nearly 97%, RF and DT reached almost 96%, and the proposed model achieved 98.90% as per Table 9.

While comparing with previous related studies utilizing feature-based machine learning methods (Dana and Shubair 2016; Omondiagbe, Veeramani, and Sidhu 2019), the results of the current study emphasize the

Table 9. Evaluation of the proposed model with feature selection.

Classifier	Accuracy (%)	F1-score (%)
RF	95.71	96.77
DT	95.71	96.72
SVM	97.14	97.83
ANN	97.14	97.77
Proposed	98.57	98.90

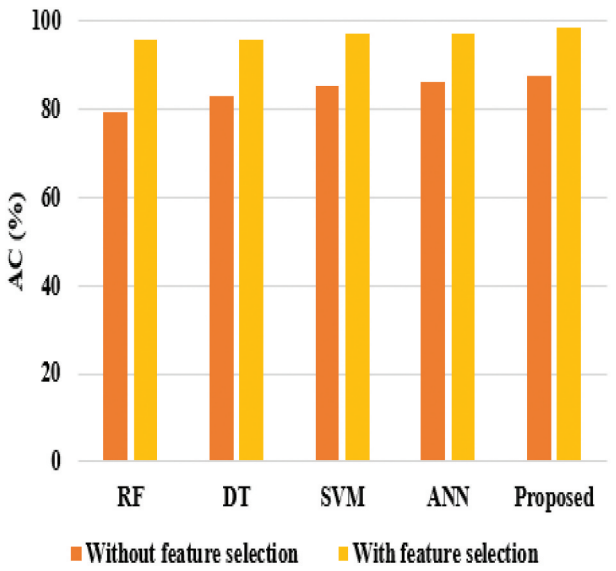

Figure 5. Comparative analysis in terms of accuracy (AC).

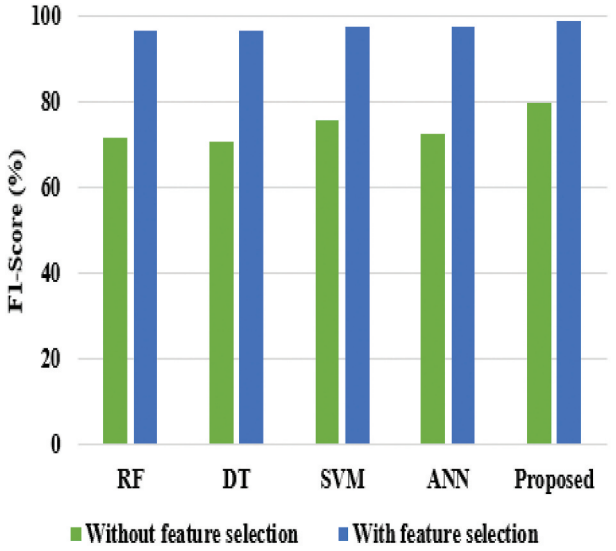

Figure 6. Comparative analysis in terms of F1-score.

fact that optimizing input data through feature-selection enhances the performance of classification and prediction models of breast cancer. Additionally, the proposed HNN-E2SAT algorithm outperforms several classical machine learning methods with and without feature selection as demonstrated in the current study as well as in previous related studies (Akay 2009; Ganggayah et al. 2019b; Huang et al. 2017; Muhammad et al. 2019). Moreover, future work could further explore hyperparameter optimization in breast cancer diagnostic systems as experimented in (Ogundokun et al. 2022) in order to further achieve real-time breast cancer diagnosis, cloud computing platforms could be leveraged as explored in (Lahoura et al. 2021).

Conclusion And Future Scope Of Work

Diagnosis methods in the medical industry are both expensive and time-consuming. Machine-learning approaches can be used as a clinical aid to detect breast cancer, especially early on in order to increase survival chances. This can be particularly helpful for new physicians and medical practitioners as misdiagnosis is quite common in the absence of highly experienced personnel. The main goal of this research was to explore feature-based machine learning techniques for early detection of breast cancer, especially in its early stages. Salient feature selection was performed on the basis of correlation coefficients. The attribute evaluator then used a PCA-based ranking algorithm to determine relevant characteristics, with the top-ranking attributes being selected for use in breast cancer categorization. The suggested classifier was used to improve the accuracy of the

prediction approach. Out of 569 cases, 559 were accurately categorized. Overall, the suggested model proved effective in identifying benign and malignant breast cancer class labels, as confirmed by statistical analysis of all comparative methodologies. The proposed model (HNN-E2SAT) achieved an accuracy of 98.57% and a precision of 98.02%, whereas the SVM approach achieved a precision of 97.82% and an accuracy of 97.14%. However, this study is limited to a single dataset. In the future, it is recommended that extension of this study involves conducting experiments with larger datasets and combining deep learning methods to further optimize breast cancer prediction and classification methods. Moreover, extended experiments to determine the robustness and interpretability of the proposed HNN-E2SAT model ought to be conducted, for the model to be deemed generalizable beyond the present evaluation scope.

Acknowledgement

Our sincere thanks to VIT Bhopal University, Bhopal (MP) for providing infrastructure, lab facilities and resources to pursue this research work.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

ORCID

Umar Yahya http://orcid.org/0000-0002-4255-0364

References

Aavula, R., and R. Bhramaramba. 2019. XBPF: An extensible breast cancer prognosis framework for predicting susceptibility, recurrence and survivability. *International Journal of Engineering and Advanced Technology* 8(5):159–166. https://www.ijeat.org/portfolio-item/C5808028319/.

Abdar, M., M. Zomorodi-Moghadam, X. Zhou, R. Gururajan, X. Tao, P. D. Barua, and R. Gururajan. 2020. A new nested ensemble technique for automated diagnosis of breast cancer. *Pattern recognition letters* 132 (April):123–31. doi:10.1016/j.patrec.2018.11.004.

Ahn, J. S., S. Shin, S.-A. Yang, E. Kyung Park, K. Hwan Kim, S. Ick Cho, C.-Y. Ock, and S. Kim. 2023. Artificial intelligence in breast cancer diagnosis and personalized medicine. *Journal of Breast Cancer* 26 (5):405–35. doi:10.4048/jbc.2023.26.e45.

Akay, M. F. 2009. Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Systems with Applications* 36 (2, Part 2):3240–47. doi:10.1016/j.eswa.2008.01.009.

Ali, M., R. Ali, M. Ali, and R. Ali. 2017. Satisfiability in big boolean algebras via boolean-equation solving. *Journal of King Abdulaziz University Engineering Sciences* 28 (1):3–18. doi:10.4197/Eng.28-1.1.

Ali, H., K. Shaban, A. Erradi, A. Mohamed, S. Khan Rumi, and F. D. Salim. 2021. 'Spatiotemporal data mining: A survey on challenges and open problems'. *Artificial Intelligence Review* 55 (2):1441–88. https://link.springer.com/article/10.1007/s10462-021-09994-y .

Amin, M. S., Y. Kia Chiam, and K. Dewi Varathan. 2019. Identification of significant features and data mining techniques in predicting heart disease. *Telematics and Informatics* 36 (March):82–93. doi:10.1016/j.tele.2018.11.007.

Amrane, M., S. Oukid, I. Gagaoua, and T. Ensari. 2018. Breast cancer classification using machine learning. In *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*, 1–4. Istanbul, IEEE. doi:10.1109/EBBT.2018.8391453.

Anand, P., A. B. Kunnumakara, C. Sundaram, K. B. Harikumar, S. T. Tharakan, O. S. Lai, B. Sung, and B. B. Aggarwal. 2008. Cancer is a preventable disease that requires major lifestyle changes. *Pharmaceutical Research* 25 (9):2097–116. doi:10.1007/s11095-008-9661-9.

Anari, S., T. S. Nazanin, M. Negin, D. Shadi, and R. Amirali. 2022. Review of deep learning approaches for thyroid cancer diagnosis. In *Mathematical Problems in Engineering*, ed. A. Darba, (August): 1–8. doi:10.1155/2022/5052435.

Arbaiy, N., S. Eliza Sulaiman, N. Hassan, and Z. Afizah Afip. 2017. Integrated knowledge based expert system for disease diagnosis system. In *IOP Conference Series: Materials Science and Engineering* 226 (August):012097. doi:10.1088/1757-899X/226/1/012097.

Arunadevi, J., and K. Ganeshamoorthi. 2019. Feature selection facilitated classification for breast cancer prediction. In *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, 560–63. doi:10.1109/ICCMC.2019.8819752.

Ataollahi, M. R., J. Sharifi, M. R. Paknahad, and A. Paknahad. 2015. Breast cancer and associated factors: A review. *Journal of Medicine and Life* 8 (Spec Iss 4):6–11.

Balogh, E. P., B. T. Miller, and J. R. Ball. 2015. Committee on diagnostic error in health care, board on health care services, institute of medicine, and engineering the national academies of sciences. In *Improving Diagnosis in Health Care*, National Academies Press (US). https://www.ncbi.nlm.nih.gov/books/NBK338594/. 'Overview of Diagnostic Error in Health Care'

Barrett, C. W., D. L. Dill, and A. Stump. 2002. Checking satisfiability of first-order formulas by incremental translation to SAT. In *Computer Aided Verification*, ed. E. Brinksma and K. G. Larsen, 236–49. US: Springer. doi:10.1007/3-540-45657-0_18.

Bethesda. 2023. 'PDQ Breast Cancer Screening'. pdqCancerinfosummary. *Nciglobal,ncienterprise*. 24 November 2023. https://www.cancer.gov/types/breast/patient/breast-screening-pdq.

Bolboacă, S. D. 2019. Medical diagnostic tests: A review of test anatomy, phases, and statistical treatment of data. *Computational and Mathematical Methods in Medicine* 2019 (May):1891569. doi:10.1155/2019/1891569.

Budh, D. P., and A. Sapra. 2024. Breast cancer screening. In *StatPearls*, US: StatPearls Publishing. http://www.ncbi.nlm.nih.gov/books/NBK556050/.

CDC. 2023. Breast and Ovarian Cancers and Family Health History | CDC. Accessed October 11, 2023. https://www.cdc.gov/genomics/disease/breast_ovarian_cancer/breast_ovarian_cancer.htm.

Dana, B., and R. Shubair. 2016. Comparative study of machine learning algorithms for breast cancer detection and diagnosis. 2016. https://ieeexplore.ieee.org/document/7818560.

Elhazmi, A., A. Al-Omari, H. Sallam, H. N. Mufti, A. A. Rabie, M. Alshahrani, A. Mady, A. Alghamdi, A. Altalaq, M. H. Azzam, et al. 2022. Machine learning decision tree algorithm role for predicting mortality in critically Ill adult COVID-19 patients admitted to the ICU. *Journal of Infection and Public Health* 15(7):826–34. doi:10.1016/j.jiph.2022.06.008.

Ellis, R. J., R. M. Sander, and A. Limon. 2022. Twelve key challenges in medical machine learning and solutions. *Intelligence-Based Medicine* 6 (January):100068. doi:10.1016/j.ibmed.2022.100068.

Embi, P. J., and P. R. O. Payne. 2009. Clinical research informatics: challenges, opportunities and definition for an emerging domain. *Journal of the American Medical Informatics Association* 16 (3):316–27. doi:10.1197/jamia.M3005.

Fatin Kadhim, N. 2022. 'Breast cancer detection using decision tree and k-nearest neighbour classifiers'. 2022. https://ijs.uobaghdad.edu.iq/index.php/eijs/article/view/5386.

Ganggayah, M. D., N. Aishah Taib, Y. Cheng Har, P. Lio, and S. Kaur Dhillon. 2019a. Predicting factors for survival of breast cancer patients using machine learning techniques. *BMC Medical Informatics & Decision Making* 19 (1):48. doi:10.1186/s12911-019-0801-4.

Ganggayah, M. D., N. Aishah Taib, Y. Cheng Har, P. Lio, and S. Kaur Dhillon. 2019b. Predicting factors for survival of breast cancer patients using machine learning techniques. *BMC Medical Informatics & Decision Making* 19 (1):48. doi:10.1186/s12911-019-0801-4.

Gopal, V. N., F. Al-Turjman, R. Kumar, L. Anand, and M. Rajesh. 2021. Feature selection and classification in breast cancer prediction using IoT and machine learning. *Measurement* 178 (June):109442. doi:10.1016/j.measurement.2021.109442.

Hajiabadi, H., V. Babaiyan, D. Zabihzadeh, and M. Hajiabadi. 2020. Combination of loss functions for robust breast cancer prediction. *Computers & Electrical Engineering* 84 (June):106624. doi:10.1016/j.compeleceng.2020.106624.

Hartama, D., A. Perdana Windarto, and A. Wanto. 2019. The application of data mining in determining patterns of interest of high school graduates. *Journal of Physics: Conference Series* 1339 (1):012042. doi:10.1088/1742-6596/1339/1/012042.

Hikmatul, F. S., N. Nazifah Yu Zaini, M. Shareduwan, M. Kasihmuddin, M. A. Mansor, and S. Sathasivam. 2020. Hybrid ant colony optimization for even-2 Satisfiability Programming in Hopfield Neural Network | AIP Conference Proceedings, https://pubs.aip.org/aip/acp/article/2266/1/040014/1007336/Hybrid-ant-colony-optimization-for-even-2.

Huang, S., C. NIanguang, P. P. Pacheco, S. Narandes, Y. Wang, and X. Wayne. 2017. Applications of Support Vector Machine (SVM) learning in cancer genomics. *Cancer Genomics & Proteomics* 15 (1):41–51. doi:10.21873/cgp.20063.

Huang, X., X. Tang, W. Zhang, S. Pei, J. Zhang, M. Zhang, Z. Liu, R. Chen, and Y. Huang. 2023. A generic knowledge based medical diagnosis expert system. *arXiv*. doi:10.48550/arXiv.2110.04439.

Jaglan, P., R. Dass, and M. Duhan. 2019. Breast Cancer Detection Techniques: Issues and Challenges. *Journal of the Institution of Engineers (India): Series B* 100 (March):379–86. doi:10.1007/s40031-019-00391-2.

Jalal Abdullah, A., T. Mohammed Hasan, and J. Waleed. 2019. An expanded vision of breast cancer diagnosis approaches based on machine learning techniques. https://ieeexplore.ieee.org/document/8950530.

Kasgari, A. B., S. Safavi, M. Nouri, J. Hou, N. Tataei Sarshar, and R. Ranjbarzadeh. 2023. Point-of-interest preference model using an attention mechanism in a convolutional neural network. *Bioengineering* 10 (4):495. doi:10.3390/bioengineering10040495.

Kasihmuddin, M. S. M., S. Sathasivam, and M. Asyraf Mansor. 2017. Hybrid genetic algorithm in the hopfield network for maximum 2-satisfiability problem' 1870. (August):050001. doi:10.1063/1.4995911.

Kattan, M. W. 2001. Expert systems in medicine. In *International Encyclopedia of the social & behavioral sciences*, ed. J. S. Neil and B. B. Paul, 5135–39. US: Pergamon. doi:10.1016/B0-08-043076-7/00556-8.

Khan, N., F. Afaq, and H. Mukhtar. 2010. Lifestyle as risk factor for cancer: evidence from human studies. *Cancer Letters* 293 (2):133–43. doi:10.1016/j.canlet.2009.12.013.

Khandezamin, Z., M. Naderan, and M. Javad Rashti. 2020. Detection and classification of breast cancer using logistic regression feature selection and GMDH Classifier. *Journal of Biomedical Informatics* 111 (November):103591. doi:10.1016/j.jbi.2020.103591.

Lahoura, V., H. Singh, A. Aggarwal, B. Sharma, M. Abed Mohammed, R. Damaševičius, S. Kadry, and K. Cengiz. 2021. Cloud computing-based framework for breast cancer diagnosis using extreme learning machine. *Diagnostics (Basel, Switzerland)* 11 (2):241. doi:10.3390/diagnostics11020241.

Mughal, M. J. H. 2018. Data mining: Web data mining techniques, tools and algorithms: An overview. *International Journal of Advanced Computer Science and Applications (IJACSA)* 9 (6). doi:10.14569/IJACSA.2018.090630.

Muhammad, H. M., J. Ping Li, A. Ul Haq, M. Hunain Memon, and W. Zhou. 2019. Breast cancer detection in the iot health environment using modified recursive feature selection'. https://www.hindawi.com/journals/wcmc/2019/5176705/.

Mümine, K. K. 2019. Breast cancer prediction and detection using data mining classification algorithms: A comparative study. *Tehnicki Vjesnik – Technical Gazette* 26:1. doi:10.17559/TV-20180417102943.

National Council Institute. 2015. Risk factors: Hormones – NCI. cgvArticle. Nciglobal,ncienterprise. Accessed April 29, 2015. https://www.cancer.gov/about-cancer/causes-prevention/risk/hormones.

Nazanin, T. S., R. Ranjbarzadeh, S. Jafarzadeh Ghoushchi, G. G. de Oliveira, S. Anari, M. Parhizkar, and M. Bendechache. 2022. 'Glioma brain tumor segmentation in four MRI modalities using a convolutional neural network and based on a transfer learning method'. 2022. https://link.springer.com/chapter/10.1007/978-3-031-04435-9_39.

Ogundokun, R. O., S. Misra, M. Douglas, R. Damaševičius, and R. Maskeliūnas. 2022. Medical Internet-of-things based breast cancer diagnosis using hyperparameter-optimized neural networks. *Future Internet* 14 (5):153. doi:10.3390/fi14050153.

Omondiagbe, D. A., S. Veeramani, and A. S. Sidhu. 2019. Machine learning classification techniques for breast cancer diagnosis. In *IOP Conference Series: Materials Science and Engineering* 495 (June):012033. doi:10.1088/1757-899X/495/1/012033.

Parsa, N. 2012. Environmental factors inducing human cancers. *Iranian Journal of Public Health* 41 (11):1–9. doi:10.1186/1479-5876-4-14.

Petrucelli, N., B. D. Mary, and P. Tuya 1993. BRCA1- and BRCA2-associated hereditary breast and ovarian cancer. In *GeneReviews®*, ed. P. Margaret. J. F. Adam, G. M. Mirzaa, R. A. Pagon, S. E. Wallace, L. J. Bean, K. W. Gripp, and A. Amemiya, US: University of Washington, Seattle. http://www.ncbi.nlm.nih.gov/books/NBK1247/.

Pietro, T. 1985. Knowledge based expert systems for medical diagnosis. *Statistics in Medicine* 4 (3):317–25. doi:10.1002/sim.4780040311.

Rabiei, R., S. Mohammad Ayyoubzadeh, S. Sohrabei, M. Esmaeili, and A. Atashi. 2022. Prediction of breast cancer using machine learning approaches. *Journal of Biomedical Physics & Engineering* 12 (3):297–308. doi:10.31661/jbpe.v0i0.2109-1403.

Ramin, R., S. Jafarzadeh Ghoushchi, N. Tataei Sarshar, E. Babaee Tirkolaee, S. Samar Ali, T. Kumar, and M. Bendechache. 2023. 'ME-CCNN: Multi-encoded images and a cascade convolutional neural network for breast tumor segmentation and recognition'. *Artificial Intelligence Review* 56 (9):10099–136. https://link.springer.com/article/10.1007/s10462-023-10426-2 .

Reyes, L. T., J. Klöckner Knorst, F. Ruffo Ortiz, and T. Machado Ardenghi. 2021. Scope and challenges of machine learning-based diagnosis and prognosis in clinical dentistry: A literature review. *Journal of Clinical and Translational Research* 7 (4):523–39.

Sathasivam, S., W. A. Tajuddin, and W. Abdullah. 2011. Logic mining in neural network: Reverse analysis method. *Computing* 91 (2):119–33. doi:10.1007/s00607-010-0117-9.

Serhat, S., U. Kursuncu, E. Kibis, M. Anis Abdellatif, and A. Dag. 2020. 'A hybrid data mining approach for identifying the temporal effects of variables associated with breast cancer survival'. *Expert Systems with Applications* 139:112863. https://www.sciencedirect.com/science/article/abs/pii/S0957417419305731.

Shadi, A., A. Anguera, J. William Atwood, J. A. Lara, and D. Lizcano. 2019. Particularities of data mining in medicine: lessons learned from patient medical time series data analysis. *EURASIP Journal on Wireless Communications and Networking* 2019 (1):260. doi:10.1186/s13638-019-1582-2.

Shravya, C., K. Pravalika, and S. Subhani. 2019. Prediction of breast cancer using supervised machine learning techniques. *International Journal of Innovative Technology and Exploring Engineering* 8(6):1106–1110. https://www.ijitee.org/portfolio-item/F3384048619/ .

Sohail, M. N., R. Jiadong, M. Musa Uba, and M. Irshad. 2019. A comprehensive looks at data mining techniques contributing to medical data growth: a survey of researcher reviews. In *Recent developments in intelligent computing, communication and devices*, ed. S. Patnaik and V. Jain, 21–26. US: Springer. doi:10.1007/978-981-10-8944-2_3.

Takkar, N., S. Kochhar, P. Garg, A. K. Pandey, U. R. Dalal, and U. Handa. 2017. Screening methods (clinical breast examination and mammography) to detect breast cancer in women aged 40–49 years. *Journal of Mid-Life Health* 8 (1):2–10. doi:10.4103/jmh.JMH_26_16.

Tan, X., A. T. Su, H. Hajiabadi, M. Tran, and Q. Nguyen. 2021. Applying machine learning for integration of multi-modal genomics data and imaging data to quantify heterogeneity in tumour tissues. In *Artificial neural networks*, ed. H. Cartwright, 209–28. US: Springer US. doi:10.1007/978-1-0716-0826-5_10.

Tomislav, H., M. Nussbaum, M. N. Wright, G. B. M. Heuvelink, and B. Gräler. 2018. 'Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables'. 2018. https://pubmed.ncbi.nlm.nih.gov/30186691/.

UCI. 2019. 'UCI machine learning repository'. https://archive.ics.uci.edu/.

Uddin, K., M. Mohi, N. Biswas, S. Tasmin Rikta, and S. Kumar Dey. 2023. Machine learning-based diagnosis of breast cancer utilizing feature optimization technique. *Computer Methods and Programs in Biomedicine Update* 3 (January):100098. doi:10.1016/j.cmpbup.2023.100098.

Velev, M. N. 2004. Efficient translation of boolean formulas to cnf in formal verification of microprocessors. In *Proceedings of the 2004 Asia and South Pacific Design Automation Conference*, 310–15. ASP–DAC '04. Yokohama, Japan, IEEE Press.

Vikas, C., S. Pal, and B. B. Tiwari. 2018. 'Prediction of benign and malignant breast cancer using data mining techniques'. *SSRN Electronic Journal* 2018. https://journals.sagepub.com/doi/10.1177/1748301818756225.

Vivek, K., B. Kishore Mishra, M. Mazzara, D. N. H. Thanh, and A. Verma. 2020. 'Prediction of malignant and benign breast cancer: A data mining approach in healthcare applications'. 2020. https://link.springer.com/chapter/10.1007/978-981-15-0978-0_43.

Wang, S., J. Cao, and P. S. Yu. 2022. 'Deep learning for spatio-temporal data mining: a survey |. *IEEE Journals & Magazine | IEEE Xplore*' 2022. https://ieeexplore.ieee.org/document/9204396.

William, W., and O. Mangasarian. 1993. 'Breast cancer Wisconsin (Diagnostic)'. [object Object]. doi:10.24432/C5DW2B.

Yang, J., Y. Li, Q. Liu, L. Li, A. Feng, T. Wang, S. Zheng, A. Xu, and J. Lyu. 2020. Brief introduction of medical database and data mining technology in big data era. *Journal of Evidence-Based Medicine* 13 (1):57–69. doi:10.1111/jebm.12373.

Yusupova, N., O. Smetanina, A. Agadullina, and E. Sazonova. 2020. *Knowledge Identification by structured data for decision making in project teams* 385–90. Atlantis Press. doi:10.2991/aisr.k.201029.072.

Zhang, Y.-B., X.-F. Pan, J. Chen, A. Cao, Y.-G. Zhang, L. Xia, J. Wang, H. Li, G. Liu, and A. Pan. 2020. Combined lifestyle factors, incident cancer, and cancer mortality: a systematic review and meta-analysis of prospective cohort studies |. *British Journal of Cancer*' 2020. https://www.nature.com/articles/s41416-020-0741-x.