Name of paper	Authors of paper	Published year	Datasets used	Name of the m	Accuracy	Error	Pros	Cons	Summary	Who did this
Forecast Model of Breast Cancer Diagnosis	Duan Yifan,Lu Jialin,Feng Boxi	2021	Wisconsin diagnostic breast cancer database	Random forest + AdaBoost	98.60%	not mentioned	GridSearchCV was used to automatically search for hyperparameters, avoiding the occurrence of overfitting and underfitting. Many experiments have shown	There's still room for accuracy improvement. The improved RF model based on Grid Search still has broader research prospects in terms of training	This paper did an integration of random forest and AdaBoost algorithms to predict and correctly diagnose if the tumor was benign or malignant. The final result was compared with single SVM, logistic regression, k-nearest neighbour, and decision tree algorithms Ensemble model's prediction accuracy - increased by 4.3% on average compared to single algorithm models, with highest accuracy increase up to 9.8% Dataset used: Wisconsin diagnostic breast cancer database To normalize the dataset, StandardScaler zero-mean	Sanjana S Murthy
Breast Cancer Prediction Based On Backpropagation Algorithm	Muhammad Sufyian Bin Mohd Azmi, Zaihisma Che Cob	2010	University of Wisconsin (UCI) Machine Learning Repository	A three-layered neural network neural network with Feed-forward Backpropagation Algorithm	96.63%	not mentioned	Best accuracy (96.63%) as compared to what would've been if the SVM classifier was used(96.19%). Best accuracy (96.63%) as compared to what would've been if the Decision Tree classifier was used(92.38%).	There still exists a higher possible value of accuracy that can be attained.	This paper outlines a system that can classify "Breast Cancer Disease" tumors using neural network with Feed-forward Backpropagation Algorithm to classify the tumor from a symptom that signifies the breast cancer disease. Breast cancer tumor database used for this purpose is from the University of Wisconsin (UCI) Machine Learning Repository. There are 699 records in this database with nine attributes graded on an interval scale from a normal state of 1–10. 241 (65.5%) records are malignant and 458 (34.5%) records are benign, represented by numbers between -1 and 1. The input layer consists of nine nodes that represent the nine attributes. The hidden layer consisted of 7 nodes. The output layer consisted of a single node representing diagnostic outcome; 0.0 for malignant and 1.0 for benign. Each of the iteration in backpropagation constitutes two sweeps: forward activation to produce a solution, and a backward propagation of the computer error to modify the weights repeatedly until the ANN solution agrees with the desired value within a prespecified tolerance. The training algorithm is divided into three main parts which are feed forward, error calculation and updating the weight. During the execution of the feed forward process which sends in the input signal, application algorithm for the neural network is used. This algorithm is applied to the real system to give the output and to classify the breast cancer tumor. A total of 7 hidden layers achieves the highest accuracy. The mean square error for this neural network model is small and towards 0.001. The feed-forward backpropagation algorithm is the best classifier to predict breast cancer disease with an accuracy of 96.63%.	Susan Mathew K
Breast Cancer Disease Prediction With Recurrent Neural Networks(RNN)	Sangapu Venkata Appaji, R Shiva Shankar, K.V.S. Murthy, Chinta Someswara Rao	2020	Breast cancer data belonging to UC Irvine repository	RNN	97%	not mentioned			This study explores deep learning techniques in conjunction with Recurrent Neural Networks (RNN) to predict the occurrence of breast cancer. To assess the efficiency of the proposed method, breast cancer data belonging to the UC Irvine repository were used. The dataset used consisted of 561 instances and 31 attributes, out of which 30 attributes are considered as input attributes and the 1st attribute is considered the target class. As a part of pre-processing, standardization of the data was performed. RNNs were preferred because RNN nodes are more dominant than other models for predicting the outcomes since these models use backpropagation. This method has one input layer consisting of 30 input nodes, three hidden layers consisting of 64, 128, and 256 nodes, and one output layer consisting of one node output of either 0 or 1. ReLU activation function is used in the hidden layer and dropout of about 0.25 is used. The dataset includes data from 569 instances with 31 characteristics. Based on experimental results, the RNN model exhibited the 97% of f1 score and an accuracy of 97.37%.	Susan Mathew K

Breast Cancer Classification Using Deep Learning	Jasmir, Siti Nurmaini, Reza Firsandaya Malik, Dodo Zaenal Abidin, Ahmad Zarkasi, Yesi Novaria Kunang, Firdaus	2018	The data centers of Medical Center University, Institute Of Oncology, Ljublijana, Yugoslavia	Multilayer Perceptron (MLP) with Backpropagation learning rule	96.50%	not mentioned	Much better accuracy than what was observed when Bayesian Linear Discriminant Analysis was used. The results show an average classification accuracy of about 83.45%.	There is still further scope for improvement of the accuracy. Using DBN-NN gives an accuracy of 99.68%.	The classification based on recurrence and no-recurrence events uses datasets from the University of Medicine Center, Institute Of Oncology, Ljublijana,Yugoslavia. Out of the total 286 datasets, there were 201 No-Recurrences-Events classes, 85 Recurrences-events classes and 10 attributes. The algorithm used for breast cancer classification is the Multilayer Perceptron algorithm with the accuracy level of 96.5%. The dataset still had incomplete or missing values denoted by the "?". Data refinement is performed by filling in missing values on the dataset using the average attribute values of all the samples residing in the same class. General multilayer perceptron (MLP) with back propagation learning rule is used here. Each neuron in the hidden layer and output layer receives output vectors from the previous layer to evaluate the weighted sum and to achieve the output vectors by the activation functions. The dataset is divided into 10 partitions randomly. Then 10 experiments were conducted, each experiment using the 10th partition data for data testing the rest is used as training data. Hence, the number of input neurons is defined by the number of markers and the number of hidden neurons is optimized for each marker combination. The accuracy of classifying recurrent and no-recurrent is 96.5%.	Susan Mathew K
Research on the Detection Method of Breast Cancer Deep Convolutional Neural Network Based on Computer Aid	Mengfan Li	2021	collected from physical examination center of related affiliated hospitals, 145 ultrasound image samples of breast tumors, including 71 benign lesions and 74 malignant lesions	computer-based feature fusion convolution neural network	89%	not mentioned	Among all the traditional methods, the learning ability of fusion feature of CNN performed best for the considered ultrasound image dataset	If the training dataset is unbalanced, the prediction effect is very poor	Since traditional image classification methods require manual extra	Sanjana S Murthy
ESTROGEN RECEPTOR STATUS PREDICTION FOR BREAST CANCER USING ARTIFICIAL NEURAL NETWORK	GOPAL K. DHONDALAY, DONG L. TONG, GRAHAM R. BALL	2011	Gene signatures from a cohort of 278 breast cancer samples, labelled in ER positive and ER negative classes	Artificial neural network (ANN)	77.62%	not mentioned	This model showed the ability to identify the most significant gene subset from the breast gene microarray data. The model was able to select 3 most important predictor genes that were previously associated with breast cancer	Low sensitivity value of 58.78% in the ER negative class was obtained. This was due to the lack of standard threshold used by the immunohistochemistry to label ER status on the samples.	Estrogen receptor (ER) is a type of hormone receptor protein that a	Sanjana S Murthy

A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction	Adam Yala , Constance Lehman, Tal Schuster, Tally Portnoi, Regina Barzilay	2019	Command-Line version of the IBIS Breast Cancer Risk Evaluation Tool (version 8; IBIS, London, England, http://www.ems-trials.org/riskevaluator/)	Deep Learning	Hybrid DL showed a significantly higher AUC (0.70) than TC (0.62; P < .001) and RF-LR (0.67; P = .01)	not mentioned	Maximum utilization of unstructured data Elimination of the need for feature engineering Ability to deliver high-quality results Elimination of unnecessary costs Elimination of the need for data labeling	To deliver the best results, a deep learning algorithm requires massive amounts of data to get trained. If you are unable to feed them with enough data, the deep learning system is likely to fail. The term deep in deep learning doesn't refer to the level of learning, but it refers to architecture. So, deep learning algorithms don't really understand the context very well. With an increasing demand for real-time data analysis, it is required to quickly retrain deep learning models. There is no standardized theory to enable you to select the right deep learning tool as it requires you to have an understanding of the topology and other parameters. Deep learning is incredibly expert in providing cybersecurity. But the network of deep learning is itself prone to hacking. Deep learning requires numerous machines and expensive GPUs to work. So, it becomes really expensive to build deep learning architecture.	Materials and Methods: This retrospective study included 88 994 consecutive screening mammograms in 39 571 women. Cancer outcomes were obtained through linkage to a regional tumor registry. Comparisons were made to an established breast cancer risk model that included breast density. We excluded 21328 women because they lacked sufficient follow-up or had another form of cancer in their breast.Of the 80 243 mammographic examinations used for training and validation, 3045 were followed by a cancer diagnosis within 5 years. Of the 8751 mammographic examinations used for testing, 269 were followed by a cancer diagnosis within 5 years. In each cell, we reported the fraction of examinations that developed cancer within 5 years, assessing the risk of breast cancer 3–5 years after mammography. This can be especially beneficial to patients who do not know their family history of breast or ovarian cancer. In the States, almost half of all women screened are told that they are at increased risk of breast cancer on the basis of their dense breast tissue. At the same time, this practice can mislead women who do not have dense breast tissue to believe they are not at increased risk for breast cancer.	Toshani Rungta

Prospective Breast Cancer Risk Prediction Model for Women Undergoing Screening Mammography	William E. Barlow, Emily White, Rachel Ballard-Barbash, Pamela M. Vacek, Linda Titus-Ernstoff, Patricia A. Carney, Jeffrey A. Tice, Diana S. M. Buist, Berta M. Geller, Robert Rosenberg, Bonnie C. Yankaskas, Karla Kerlikowske	2006	The particular dataset used in this study was a large cross-classification of risk factors by cancer outcome.	Logistic Regression	risk of 4.6 (95% CI = 1.7 to 12.6)	not mentioned	Logistic regression is easier to implement, interpret, and very efficient to train. It makes no assumptions about distributions of classes in feature space. It can easily extend to multiple classes (multinomial regression) and a natural probabilistic view of class predictions. It is very fast at classifying unknown records.	If the number of observations is lesser than the number of features, Logistic Regression should not be used, otherwise, it may lead to overfitting. The major limitation of Logistic Regression is the assumption of linearity between the dependent variable and the independent variables. It can only be used to predict discrete functions. Hence, the dependent variable of Logistic Regression is bound to the discrete number set. Logistic Regression requires average or no multicollinearity between independent variables.	The predictive accuracy of the model is highest in annually screened non-Hispanic white women and is lowest in women with different demographic characteristics than the population from which the model was developed. Women with high Gail scores have been encouraged to be screened, to undergo genetic or biomarker evaluation, and to participate in intervention trials. Women with previous breast cancer were excluded. Women with breast augmentation were also excluded because augmentation decreases breast cancer detection by mammography. Screening examinations had to be designated as bilateral screening by the radiology facility and needed to be done at least 9 months after any preceding breast imaging to ensure an accurate designation as a screening mammogram. Patient information was primarily obtained from self-report at the time of the screening mammogram. Age at menarche was often not collected or not reported, but it was tested as a risk factor in women who reported it. We evaluated the probability of a cancer diagnosis within 1 year of each screening mammogram by use of logistic regression analysis in SAS, Version 9.0, and Stata, Version 8 . The primary goal was to find a model that best predicted a diagnosis of breast cancer separately in premenopausal and postmenopausal women with a minimal number of predictors. The c statistic ranges from 0.50 to 1.00, with a higher score indicating better prediction for an individual woman. The particular dataset used in this study was a large cross-classification of risk factors by cancer outcome. This procedure required excluding 7.6% of the mammograms from women aged 45 – 54 years with unknown menopausal status.	Toshani Rungta

							Advantages of Support Vector Machine (SVM) 1. Regularization capabilities: SVM has L2 Regularization feature. So, it has good generalization capabilities which prevent it from over-fitting. 2. Handles non-linear data efficiently: SVM can efficiently handle non-linear data using Kernel trick. 3. Solves both Classification and Regression problems: SVM can be used to solve both classification and regression problems. SVM is used for classification problems while SVR (Support Vector Regression) is used for regression problems. 4. Stability: A small change to the data does not greatly affect the hyperplane and hence the SVM. So the SVM model is stable. Advantages of Random Forest: It can come out with very high dimensional (features) data, and no need to reduce dimension, no need to make feature selection It can judge the importance of the feature Can judge the interaction between	Disadvantages of Support Vector Machine (SVM) 1. Choosing an appropriate Kernel function is difficult: Choosing an appropriate Kernel function (to handle the non-linear data) is not an easy task. It could be tricky and complex. In case of using a high dimension Kernel, you might generate too many support vectors which reduce the training speed drastically. 2. Extensive memory requirement: Algorithmic complexity and memory requirements of SVM are very high. You need a lot of memory since you have to store all the support vectors in the memory and this number grows abruptly with the training dataset size. 3. Requires Feature Scaling: One must do feature scaling of variables before applying SVM. 4. Long training time: SVM takes a long training time on large datasets. 5. Difficult to interpret: SVM model is difficult to understand and interpret by human beings unlike Decision Trees. Disadvantages of Random Forest: Random forests have been shown to fit	Cancer is one of the deadliest diseases in the world. There are over 100 different types of cancer that affect humans. However, this study, aims to analyse breast cancer, a disease in which cells grow out of control to form a tumour which tends to affect another part of the body. There are three common parts of the breast whose cells have the ability to turn into cancer namely lobules, ducts, and the connective tissue. The exact causes of breast cancer are still not known, but experts are of the opinion that an interaction between genes with lifestyle, environment, and hormone, tends to provoke abnormal cell growth. There are several factors that increase the risk of getting breast cancer such as age. According to research, in most cases people are diagnosed after the age of 50. Men still have a risk of getting breast cancer even though it is a lot lower than women. The random forest algorithm follows the steps below: − Choose a feature to be named as root node and make a branch for all possible features. Boruta feature selection is built around	

Breast cancer risk prediction model based on C5.0 algorithm for postmenopausal women	a) Xia Zhang School of Information Science and engineering University of Jinan Jinan, China b)Yingming Sun School of Information Science and engineering University of Jinan Jinan, China (corresponding author)	2018	1031 postmenopausal women (≥43 years old) with breast cancer data, including 26 patients in the disease group and 1005 in the control group	a) Default_C5.0 Model	The accuracy of predicting is 97.33%. However the senistivity is 0% implying 100% of the diseases are incorrectly predicted as NOT positive	Since the sensitivity of model to predict is 0%, the error of prediction is 100%	The accuruacy of predicting is as high as 97.33%. Lower prediction error rate of 12.34% compared to Costmatrix_C5.0 model	Since the sensitivity of model to predict is 0, chances of NOT predicting a disease is 100% which is quite risky	Performace of C5.0 model with cost matrix enhances the breast cancer risk prediction model based on C5.0 algorithm. When comared to other models, since the sensitivity of costmatrix based model is higher the chance of not diagnosing the breast cancer had greatly reduced.	Sunaina
				b) Adaptiveboosting_C5.0 Model	The accuracy of predicting is 97.33%. However the senistivity is 0% implying 100% of the diseases are incorrectly predicted as NOT positive	Since the sensitivity of model to predict is 0%, the error of prediction is 100%	The accuruacy of predicting is as high as 97.33%. Lower prediction error rate of 12.34% compared to Costmatrix_C5.0 model	Since the sensitivity of model to predict is 0, chances of NOT predicting a disease is 100% which is quite risky		
				c) Costmatrix_C5.0 Model	The accuracy of predicting is 85.39%. But the senistivity is as high as 85.71% implying the diseases are more correctly predicted	Since the sensitivity of model to predict is 85.71%, the error of prediction is 14.29%	The sensitivity of model is pretty high (85.71%) meaning the probability of NOT diagnosing a disease is greatly reduced. Though the accuracy is reduced, the proability of not missing a disease is greatly reduced	The accuracy of prediction slighty reduces to 85.39%		

Predicting Breast Cancer Risk Using Subset of Genes	Tahsien Al-Quraishi, Jemal H. Abawajy, Naseer Al-Quraishi, Ahmad Abdalrada, Lamyaa Al-Omairi Deaken University, Geelong, Australia	2019	Microarray breast cancer gene expression data containing 24,481 scanned gene expressions with 97 patients	Four feature selection algorithms namely a) Correlation-based filter method (FCBF) b) Regularized Random Forest algorithm (RRF) c) Decision Tree algorithm(DT-FWD) and d) Symmetrical Uncertainty Criteria(SUC) function are applied on four common ensemble models named 1) Deep Neural Network and Support Vector Machine (DNN+SVM) 2)Deep Neural Network and Recursive Feature Elimination (DNN+RF) 3)Deep Neural Network and AdaBoost (DNN+AdaBoost) 4) Support Vector Machine and Recursive Feature Elimination (SVM+RF).	a) Accuracy of FCBF on any of the models is ranging from 92% to 97% b) Accuracy of RRF on any of the models is ranging from 78% to 86% c) Accuracy of DT-FWD on any of the models is ranging from 80% to 86% d) Accuracy of SUC on any of the models is ranging from 90% to 93%	No data	The accuracy of FCBF feature selection is best when applied on DNN+SVM model. It also fairs good with SVM+RF model	FCBF feature selection does not work so efficiently with models like DNN+RF and DNN+AdaBoost.	Models based on ensemble classifier techniques Deep Neural Network and Support Vector Machine (DNN+SVM), Deep Neural Network and Recursive Feature Elimination (DNN+RF), Deep Neural Network and AdaBoost (DNN+AdaBoost) and Support Vector Machine and Recursive Feature Elimination (SVM+RF) are taken. On these models, predication is done using feature selection algorithms Correlation-based filter method (FCBF), Regularized Random Forest algorithm (RRF), Decision Tree algorithm(DT-FWD) and Symmetrical Uncertainty Criteria (SUC) function. Results are compared against combination of models and feature selections. Results show that prediction based on genes with feature selection FCBF gave best results for their DNN+SVM model.	Sunaina

Predicting breast cancer risk using personal health data and machine learning models	Gigi F. StarkID, Gregory R. HartID, Bradley J. NartowtID, Jun Deng	2019	All data used in this work has been downloaded from the publicly available data sets provided by the National Cancer Institute (https://www.cdc.gov/nchs/nhis/about_nhis.htm , https://biometry.nci.nih.gov/cdas/plco/).Approval to data access (PLCO-392) has been granted by NCI.	Logistic regression, Gaussian naive Bayes, Decision tree, Linear discriminant analysis, Support vector machine and feed-forward Artificial neural network.	The sensitivity of logistic regression, naive Bayes, linear discriminant analysis, and neural network models were approximately 0.6 (values ranged from 0.587 for the linear discriminant analysis to 0.621 for the neural network). The specificities of these models were all around 0.5 (values ranged from 0.474 for the neural network to 0.512 for the linear discriminant analysis). The precisions of these models were all relatively low, hovering around 0.025.	Since the specificities of these models were all around 0.5 (values ranged from 0.474 for the neural network to 0.512 for the linear discriminant analysis), the error of prediction is 50%.	With the PLCO inputs being broken down to training and testing datasets, the result suggests that the incorporation of additional inputs, not the implementation of more complex machine learning models, has led to improved breast cancer risk prediction.	Due to the lack of access t other datasets, they were forced to split the PLCO data set into training and testing portions. Further work could examine whether, if when trained on the entire PLCO data set, these machine learning models generalize well to new data sets. Another potential limitation was the lack of available biopsy or atypical hyperplasia data in the PLCO data set which would increase the percentage of correct prediction.	The paper uses six models : logistic regression, Gaussian naïve Bayes, decision tree, linear discriminant analysis, support vector machine, and feed-forward artificial neural network and compares it with a traditional BCRAT model (implementation of GAIL model). Training is done with two sets of inputs.The first set contains only the 5 BCRAT model inputs, whereas the second model contains 5+8 predictors. While using only the BCRAT inputs none of the machine learning models were significantly better than BCRAT but when a boarder set of predictors were used it was observed that the logistic regression, linear discriminant analysis, and neural network model effectively predicted five-year breast cancer risk better than the BCRAT model.	Sunaina