

**Efficient Machine Learning Algorithms for Identifying Risk Factors of Prostate and
Breast Cancers among Males and Females**

by

Samaneh Rikhtehgaran

A Thesis Submitted to the Faculty of

The Charles E. Schmidt College of Science

In Partial Fulfillment of the Requirements for the Degree of

Professional Science Master

Florida Atlantic University

Boca Raton, FL

August 2021

Copyright 2021 by Samaneh Rikhtehgaran

Efficient Machine Learning Algorithms for Identifying Risk Factors of Prostate and Breast Cancers among Males and Females

by

Samaneh Rikhtehgaran

This thesis was prepared under the direction of the candidate's thesis advisor, Dr. Wazir Muhammad, Department of Physics, and has been approved by all members of the supervisory committee. It was submitted to the faculty of the Charles E. Schmidt College of Science and was accepted in partial fulfillment of the requirements for the degree of Professional Science Master.

SUPERVISORY COMMITTEE:



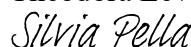
Wazir Muhammad, Ph.D.

Thesis Advisor



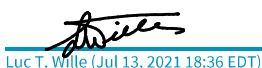
Theodora Leventouri (Jul 12, 2021 11:03 EDT)

Theodora Leventouri, Ph.D.



Silvia Pella (Jul 12, 2021 11:12 EDT)

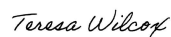
Silvia Pella, Ph.D., DABR



Luc T. Wille (Jul 13, 2021 18:36 EDT)

Luc Wille, Ph.D.

Chair, Department of Physics



Teresa Wilcox, Ph.D.

Interim Dean, Charles E. Schmidt College of Science



Robert W. Stackman Jr., Ph.D.

Dean, Graduate College

July 14, 2021

Date

Acknowledgements

I would like to express my thanks to my supervisor, Dr. Wazir Muhammad for his valuable advices and his support during my research.

I would like to thank Dr. Theodora Leventouri for all her support, guidance and help during my graduate studies.

I would like to thank Dr. Silvia Pella for her help with improving my clinical skills.

I would like to thank my parents for their unconditional love and support throughout my life. Thank you both for giving me strength to chase my dreams.

I would like to thank my sisters, Reyhaneh and Farinaz for their love, endless support and encouragement.

Abstract

Author Samaneh Rikhtehgaran
Title: Efficient Machine Learning Algorithms for Identifying Risk Factors of Prostate and Breast Cancers among Males and Females
Institution: Florida Atlantic University
Thesis Advisor: Dr. Wazir Muhammad
Degree: Professional Science Master
Year: 2021

One of the most common types of cancer among women is breast cancer. It represents one of the diseases leading to a high number of mortalities among women. On the other hand, prostate cancer is the second most frequent malignancy in men worldwide. The early detection of prostate cancer is fundamental to reduce mortality and increase the survival rate. A comparison between six types of machine learning models as Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, k Nearest Neighbors, and Naïve Bayes has been performed. This research aims to identify the most efficient machine learning algorithms for identifying the most significant risk factors of prostate and breast cancers. For this reason, National Health Interview Survey (NHIS) and Prostate, Lung, Colorectal, and Ovarian (PLCO) datasets are used. A comprehensive comparison of risk factors leading to these two crucial cancers can significantly impact early detection and progressive improvement in survival.

Dedication

To my family for their love and support.

**Efficient Machine Learning Algorithms for Identifying Risk Factors of
Prostate and Breast Cancers among Males and Females**

List of Tables	x
List of Figures	xi
1 Introduction	1
1.1 Importance of comparison between prostate and breast cancers	2
1.2 Machine learning algorithms.....	2
1.2.1 Supervised Learning	3
1.2.2 Unsupervised Learning	3
1.2.3 Reinforcement Learning	3
1.3 Logistic Regression	4
1.4 Decision Tree	6
1.5 Naïve Bayes.....	6
1.6 Random Forest	7
1.6.1 Random Forest performance.....	7
1.7 Gradient Boosting	8
1.7.1 Gradient Boosting performance	9
1.8 K Nearest Neighbor.....	10

1.9	Model Evaluation	11
1.9.1	AUC-ROC curve.....	12
1.9.2	Accuracy	12
1.9.3	Sensitivity	13
1.9.4	Specificity	13
1.9.5	PPV or Precision.....	13
1.9.6	NPV.....	14
1.9.7	P-value	14
2	Materials and Methods	15
2.1	Data	15
2.2	Data preprocessing	15
2.3	Data Acquisition.....	16
2.4	NHIS Dataset.....	16
2.5	PLCO Dataset.....	16
3	Results and Discussion	18
3.1	Six types of machine learning models.....	18
3.2	NHIS Results.....	19
3.2.1	NHIS-Breast Dataset.....	19
3.2.2	NHIS - Prostate Dataset	22
3.3	PLCO Results.....	25

3.3.1	PLCO – Breast Dataset	25
3.3.2	PLCO – Prostate Dataset	28
3.4	Gini Importance.....	31
3.5	Gini Importance-Breast Dataset	32
3.6	Gini Importance- Prostate Dataset	33
3.7	Comparison between risk factors	34
4	Conclusion.....	35
	References.....	36

List of Tables

Table 1 Confusion Matrix for binary classification problem.....	12
Table 2 Input variables for NHIS breast dataset.	19
Table 3 Accuracy, sensitivity, specificity, PPV and NPV for the NHIS breast dataset..	21
Table 4 Input variables for NHIS prostate dataset.	22
Table 5 Accuracy, sensitivity, specificity, PPV and NPV for the NHIS prostate dataset.	24
Table 6 Input variables for PLCO breast dataset.	25
Table 7 Accuracy, sensitivity, specificity, PPV and NPV for the PLCO breast dataset..	27
Table 8 Input variables for PLCO prostate dataset.	28
Table 9 Accuracy, sensitivity, specificity, PPV and NPV for the PLCO prostate dataset.	30

List of Figures

Figure 1 Sigmoid function.	5
Figure 2 Random Forest training and testing performance.....	8
Figure 3 Outline of the breast/prostate cancer using machine learning.....	15
Figure 4 ROC curve for machine learning models for breast cancer study with NHIS Dataset.....	20
Figure 5 ROC curve for machine learning models for prostate cancer study with NHIS Dataset.....	23
Figure 6 ROC curve for machine learning models for breast cancer study with PLCO Dataset.....	27
Figure 7 ROC curve for machine learning models for prostate cancer study with PLCO Dataset.....	29
Figure 8 Gini importance coefficient for each risk factors of the PLCO breast dataset..	33
Figure 9 Gini importance coefficient for each risk factors of the PLCO prostate dataset.	34

1 Introduction

One of the most common types of cancer among women is breast cancer[1]. It represents one of the diseases leading to a high number of mortalities among women. Breast cancer is the second leading cause of cancer deaths among US women[2]. On the other hand, prostate cancer is the second most frequent malignancy (after lung cancer) in men worldwide [2, 3]. It is the fifth leading cause of death worldwide [4], and its incidence increases with age [5]. The early detection of prostate cancer is fundamental to reduce mortality and increase the survival rate. With the rapid development of machine learning algorithms and with the surge of medical data, the application of big data analysis technology has changed our understanding and comprehension of the risk factors leading to cancers. One of the challenges of cancer prevention is the existing community skepticism that cancer can be prevented [6]; however only 5% of cancers are hereditary[6]. There is a lack of studies to understand what causes cancer and how to reduce the risk [7, 8]. Social marketing performs commercial marketing strategies to modify social behaviors related to public health [9]. Health-related social marketing campaigns use television, radio, digital media, and billboards with the intent of effecting voluntary change in health behavior [10, 11]. With this purpose, a comparison between six types of machine learning models as Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), k Nearest Neighbors (k-NN), and Naïve Bayes (NB) has been performed.

This research aims to identify the most efficient machine learning algorithms for identifying the most significant risk factors leading to prostate and breast cancers. For this

reason, National Health Interview Survey (NHIS) [12] and Prostate, Lung, Colorectal and Ovarian (PLCO) [13] datasets are used. A comprehensive comparison of risk factors leading to these two crucial cancers can significantly impact early detection and progressive improvement in survival.

1.1 Importance of comparison between prostate and breast cancers

The importance of comparison between prostate and breast cancers is that they are among the most common cancer diagnoses among males and females, worldwide [14]. Moreover, breast and prostate cancer co-occur in families, and therefore, women with a family history of prostate cancer are at increased breast cancer risk [15]. It is shown that those with familial breast cancer had a 21% greater risk of prostate cancer overall and a 34% greater risk of lethal disease. Furthermore, family history of prostate cancer alone was associated with a 68% increased risk of total disease, and 72% increased risk of lethal disease. It is demonstrated that men with a family history of both cancers were also at higher risk. Thus, a comprehensive comparison of risk factors can be beneficial to understand the correlation between these two crucial cancers.

1.2 Machine learning algorithms

Machine learning (ML) has grown rapidly in recent years in the form of data analysis which allows the applications to function in an intelligent manner [16]. Machine learning models enable the system to learn and enhance from experience automatically, and they are generally referred to as the most popular technologies.

These learning algorithms can be categorized into different types based on their purposes. The main categories can be listed as supervised learning, unsupervised learning, and reinforcement learning [17].

1.2.1 Supervised Learning

This algorithm consists of an outcome variable (dependent variable) which is to be predicted from a given set of predictors (independent variables). Therefore, using these variables a function can be generated to map the inputs to desired outputs [18]. The training process continues until the model reaches the desired level of accuracy on the training data. Some examples of supervised learning are regression, decision tree, random forest, k nearest neighbor and, logistic regression.

1.2.2 Unsupervised Learning

This algorithm does not include any outcome variable to predict. It is used for clustering populations in different groups, which is widely used for segmenting customers in different groups for specific intervention. Some popular unsupervised models consist of k-d trees, random projection (RP trees), and clustering trees [19].

1.2.3 Reinforcement Learning

This algorithm is used to train the machine to make specific decisions. Thus, in this algorithm, the machine uses observations gathered from the interaction with the environment to take actions that would maximize the reward or minimize the risk. Therefore, machine learns from the past experience and strives to capture the best possible knowledge to make accurate decisions. What distinguishes reinforcement learning from supervised learning is that only partial feedback is given to the learner's predictions [20]. Reinforcement Learning is a type of Machine Learning and, therefore a branch of Artificial Intelligence.

In this study, we will focus only on supervised machine learning algorithms.

1.3 Logistic Regression

Logistic regression is a machine learning algorithm which is used for classification problems. Like all regression analysis, logistic regression is a predictive analysis. Logistic regression is a linear regression with a more complex cost function which can be defined as ‘Sigmoid function’ or ‘logistic function’. Figure 1 shows sigmoid function. Logistic regressions are commonly used, interpretable, and make no assumptions about the explanatory data [21]. However, logistic regressions lack statistical sophistication since these models assume that inputs are linearly related to the log odds of the outcome [22].

The cost function is limited between 0 and 1 (see Eq (1)) by the hypothesis of logistic regression. Thus, it is not possible to explain it with linear functions.

$$0 \leq h_{\theta}(x) \leq 1 \tag{1}$$

The sigmoid function is used to depict the predicted values to probabilities, Figure 1. This function maps any real value into a value between 0 and 1 and therefore, it is used to map predictions to probabilities, see Eq (2).

$$\sigma(x) = \frac{1}{1 + e^{-(x)}} \quad (2)$$

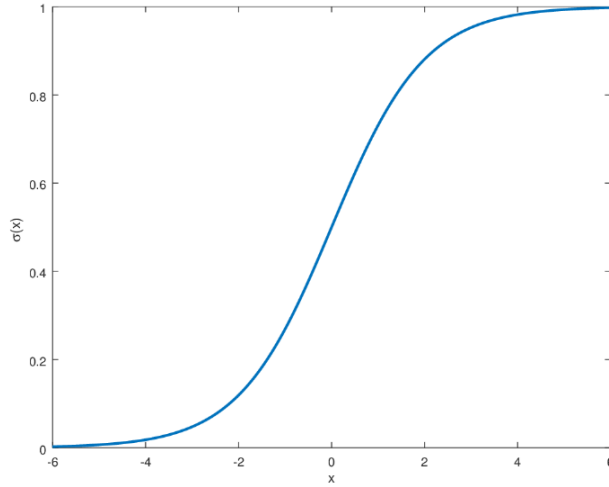


Figure 1 Sigmoid function.

Logistic regression is very similar to linear regression but with a binomial response variable[23]. Logistic regression will model the chance of an outcome based on individual characteristics. Since this chance is a ratio, the logistic regression will model the logarithm of the chance, and it is given by Eq (3).

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_m x_m \quad (3)$$

Where p indicates the probability of an event and β_i are the regression coefficients associated with the reference group and x_i explanatory variables. The reference group, represented by β_0 , is constituted by those individuals presenting the reference level of each variable $x_{1...m}$.

1.4 Decision Tree

Decision tree methodology is a data mining method used commonly for classification systems for developing prediction algorithms for a target variable [24]. Tree models where the target variable can take a discrete set of values are called classification trees. Decision trees where the target variable can take continuous values, typically real numbers, are called regression trees. Decision trees classify the data by asking recursive questions about predictor variables [25]. Decision trees have nodes that represent a test for a particular input, branches that represent responses to the nodes, and leaves which are nodes at the bottom of the tree that provide ultimate classification [26]. Decision trees are highly interpretable, and such an important interpretable model could be used in a clinical setting.

1.5 Naïve Bayes

Naïve bayes models are probabilistic classifiers [25] which are constructed based on applying Bayes' theorem with strong independence assumptions between the features. Generally, they need less training data than other classifiers and have fewer parameters than models such as neural networks and support vector machines [27]. Naïve bayes models assume that input variables are independent [28], which is rarely true in classification tasks. For some types of probability models, naïve bayes classifiers can be trained very efficiently in a supervised learning setting. In practical applications, parameter estimation for naïve bayes models uses maximum likelihood method; Which means one can work with naïve Bayes model without accepting Bayesian probability or using any Bayesian methods. Despite their naïve design and oversimplified assumptions, Naïve Bayes classifiers are well-worked in many complex real-world situations.

1.6 Random Forest

Random forest is a supervised learning algorithm that is used for both classification as well as regression. Although, it is mostly used for classification problems. Generally, a forest is made up of trees, and more trees mean more robust forest. Similarly, a random forest algorithm generates decision trees on a data sample, and it receives predictions from each of them. Consequently, it chooses the best solution using voting. The superiority of random forest to decision trees is that it reduces overfitting by averaging the results. The random forest used in this study is made of randomly selected features or a combination of features at each node to grow a tree [29]. Bagging is a method to generate a training dataset by randomly drawing with the replacement of N examples, where N is the size of the original training set [30] was used for each feature combination selected. Then, any examples are classified by taking the most popular voted class from all the tree predictors in the forest.

1.6.1 Random Forest performance

In the first place, start with the selection of random samples from a given dataset. Then, this algorithm will create a decision tree for every sample. Afterward, it will get the prediction result from every decision tree. In the next step, voting will be applied for every predicted result. Finally, select the most voted prediction result as the final prediction result. The following diagram depicts how the random forest works, see Figure 2.

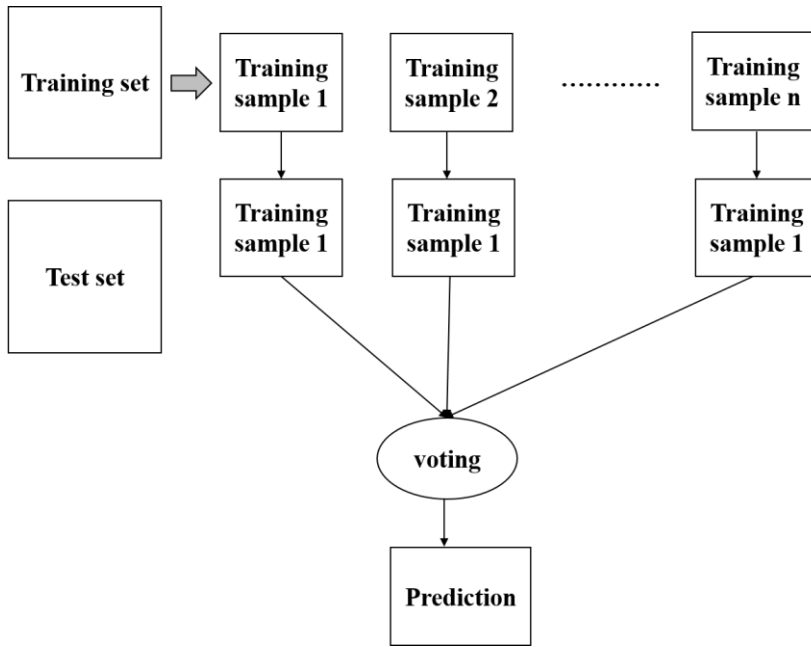


Figure 2 Random Forest training and testing performance.

1.7 Gradient Boosting

Gradient boosting is a type of machine learning algorithm used for regression and classification. Gradient boosting is widely used due to its efficiency, accuracy, and interpretability [31]. The family of boosting methods is based on a different, constructive strategy ensemble formation. The main idea of boosting is to add new models to the ensemble sequentially [32]. This model aims to set the target outcomes for this next model to minimize the error. Thus, weak learners can convert into strong learners. Gradient boosting produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. Gradient boosting trains many models in a gradual, additive, and sequential manner. Like other boosting techniques, gradient boosting combines weak learners into one single strong learner in an iterative fashion.

1.7.1 Gradient Boosting performance

1.7.1.1 Loss Function

The loss function must be differentiable, and it depends on the type of problem being solved. It can be a squared error for regression and logarithmic loss for classification.

1.7.1.2 Weak Learner

Generally, in gradient boosting, decision trees are used as the weak learner. Regression trees are specifically used to output real values for splits whose outputs can be added together, allowing subsequent model outputs to be added and correct the residual in predictions. Trees are created in a greedy manner, and therefore, the best split points can be chosen to minimize the loss. There are different ways to constrain the weak learners, such as a maximum layer, nodes, splits, or leaf nodes.

1.7.1.3 Additive Model

In gradient boosting, trees are added one at a time, and the trees existing in the model are not changed. A gradient descent procedure is used to minimize the loss when adding the trees. A set of parameters such as coefficients in a regression equation or weights in a neural network can minimize the gradient descent. After calculating the error or loss, the weights are updated to minimize that error. This procedure is the traditional path.

Alternatively, we have weak learner sub-models or, more specifically, decision trees. When the loss is calculated, we must add a tree to the model that reduces the loss. This step should be done by parametrizing the tree, then modifying the parameters of the tree and, therefore, reducing the residual loss. This approach is called functional gradient descent or gradient descent with functions.

1.8 K Nearest Neighbor

A k Nearest Neighbor or k-NN is an approach to data classification that estimates how likely a data point is to be a member of one group or the other depending on what group the data points nearest to it are in. It is one of the top 10 techniques for data mining [33]. It is a well-known decision rule that is widely used in pattern classification [34]. Since this algorithm relies on distance, it is essential to keep the physical units consistent or normalize the training data to improve its accuracy dramatically. For both classification and regression, it is beneficial to assign weights to the contributions of the neighbors; therefore, the nearer neighbors contribute more to the average than the more distant ones.

K nearest neighbor algorithm stores all available cases and then classifies new cases based on a similarity measure (e.g., distance functions). This technique has been employed in statistical estimation and pattern recognition as a non-parametric technique.

A case is classified based on the majority vote of its neighbors, with the case being assigned to the class most common amongst its k nearest neighbors, measured by a distance function. Assume $k = 1$; then the case is assigned to the class of its nearest neighbor.

Euclidean, Manhattan, and Minkowski are given distance functions respectively in Eq (4), (5) and (6).

$$f = \sqrt{\sum_{i=1}^k (x_i - y_i)^2} \quad (4)$$

$$f = \sum_{i=1}^k |x_i - y_i| \quad (5)$$

$$f = (\sum_{i=1}^k (|x_i - y_i|^q))^{1/q} \quad (6)$$

These measures are only valid for continuous variables. If categorical variables are subject to study, the Hamming distance must be used. Therefore, it causes the issue of standardization of the numerical variables between 0 and 1 when a mixture of categorical and numerical data exists in the dataset. Thus, the Hamming distance equation is given in Eq (7).

$$D_H = \sum_{i=1}^k |x_i - y_i| \quad (7)$$

If $x = y$, then $D = 0$. If $x \neq y$ then $D = 1$. In the first step, the data should be inspected to select the optimal value for k . Generally, a large k value is more precise since it reduces the overall noise, but it cannot be promised. Alternatively, cross-validation is another way to determine a good k value by using an independent dataset to validate the k value.

1.9 Model Evaluation

A confusion matrix, also known as error matrix, is a specific table layout used to visualize the algorithm's performance and is mostly used in supervised learning algorithms [35]. This confusion matrix is shown in Table 1. Each row of the matrix represents the

instances in an actual class, while each column shows the instances in a predict class (or vice versa).

Table 1 Confusion Matrix for binary classification problem.

<i>True Class</i>	<i>Predicted Class</i>	
	<i>Positive</i>	<i>Negative</i>
<i>Positive</i>	<i>TP</i>	<i>FN</i>
<i>Negative</i>	<i>FP</i>	<i>TN</i>

1.9.1 AUC-ROC curve

The performance of a machine learning algorithm is a significant and essential task. Thus, when it comes to a classification problem, an AUC-ROC curve can be employed to investigate the performance of a machine learning algorithm. A receiver operating characteristic curve or ROC curve is a two-dimensional measure of classification performance [36]. The ROC curve is generated by plotting the true positive rate (TPR) against false positive rate (FPR) at various threshold settings. The true positive rate is also known as sensitivity or recall. The false positive rate is known as specificity. AUC ranges in value from 0 to 1. If the model's predictions are 100% wrong the AUC equals 0. If the model's predictions are 100% correct, it has an AUC value of 1.

1.9.2 Accuracy

In medical studies, diagnostic tests are used to identify the presence or absence of diseases in the study subjects [37]. An example includes testing for the presence or absence of cancer. The labels positive and negative refer to the presence or absence, respectively,

of the condition of interest (having cancer in this study). In this study, the accuracy of a test is defined as test's ability to differentiate the patient and healthy cases correctly. The test's accuracy is calculated as the proportion of true positive and true negative in all evaluated cases. It is defined in Eq (8).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

1.9.3 Sensitivity

The sensitivity of a test is defined as the test's ability to determine the patient cases correctly. Therefore, the proportion of true positive patient cases need to be calculated as in Eq (9).

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (9)$$

1.9.4 Specificity

The specificity of a test is defined as the ability of a test to determine the healthy cases correctly. Thus, the proportion of true negative in healthy cases needs to be calculated as in Eq (10).

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (10)$$

1.9.5 PPV or Precision

Positive predictive value or precision is defined as the fraction of correctly classified instances among the ones classified as positive. It is defined in Eq (11).

$$\text{PPV} = \frac{TP}{TP+FP} \quad (11)$$

1.9.6 NPV

Negative predictive value is defined as the proportions of negative classified instances that are true negative. It is defined in Eq (12).

$$\text{NPV} = \frac{TN}{TN+FN} \quad (12)$$

1.9.7 P-value

The p-value of calculated probability is the probability of finding the observed or more extreme results when the null hypothesis (H_0) of a study question is true. The null hypothesis is defined as a hypothesis of “no difference” between two groups of study. It is recommended to set a level of significance (a theoretical p-value) that acts as a reference point to identify significant results, that is, to identify results that differ from the null hypothesis of no effect. Fisher recommended using $p = 0.05$ judge whether an effect is significant or not and we will follow this in this study [38]. The alternative hypothesis (H_1) is the opposite of the null hypothesis. The null hypothesis states that there is no relationship between the two variables being studied (one variable does not affect the other). However, the alternative hypothesis states that the independent variable did affect the dependent variable, and it happens if the null hypothesis is concluded to be untrue.

2 Materials and Methods

2.1 Data

In this section, we describe our data acquisition methods and the approval of the data collection details. In addition, we describe the data preprocessing and feature selection methods. Furthermore, we discuss the risk factors related to breast and prostate cancers. Figure 3 shows the process of data acquisition, data preprocessing and data evaluation.

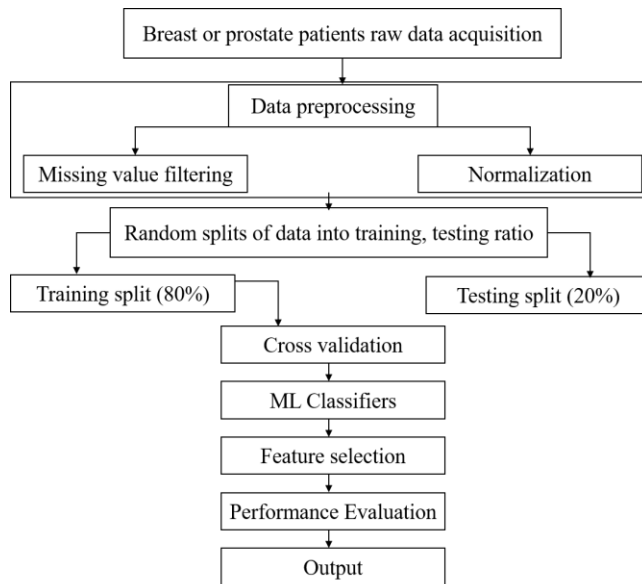


Figure 3 Outline of the breast/prostate cancer using machine learning.

2.2 Data preprocessing

Before proceeding further for the analysis of the data, preprocessing of the data was required. Since the missing values were less than one percent, then missing values were removed from the data. In addition, the data were normalized using Scikit-learn in order to

bring all the collected data in the same range. Rescaling machine learning model inputs was very important since it helped the models to train more quickly, and results would be improved enormously [39].

Data was labeled with 7 for responses of “Refused”, 8 for “Not ascertained” and 9 “Don’t know” in NHIS were considered as missing values. PLCO used the same way of labeling of the data. These are different from data not missing at random.

2.3 Data Acquisition

We have investigated two different datasets, National Health Interview Survey (NHIS) and Prostate, Lung, Colorectal, and Ovarian (PLCO). In this section, we described each of these datasets.

2.4 NHIS Dataset

The National Health Interview Survey (NHIS) [12] dataset is a cross-sectional study of the overall health status of the United States. Each year, roughly 30000 adults are interviewed on a range of current and past personal conditions. NHIS data are freely downloadable by the public and generally are available in June or July for the preceding year’s dataset. The first survey of the NHIS after a significant revision was administered in 1997 and so data from years 1997-2019 was used. Due to Coronavirus Pandemic, the data for 2020 has not been released yet and it is not included in this thesis.

2.5 PLCO Dataset

The Prostate, Lung, Colorectal, and Ovarian (PLCO) [13] dataset is a randomized, controlled longitudinal study on the efficacy of screening for prostate, lung, colorectal and ovarian cancer. Approximately 155000 participants were enrolled between November

1993 and July 2001. Participants were randomized, entered into the trial, and answered a baseline questionnaire (BQ). Participants were followed for up to 14 years, exiting the trial early if they were diagnosed with any cancer or died.

3 Results and Discussion

3.1 Six types of machine learning models

After data preprocessing, we could proceed with the further steps. Models were trained and evaluated on the NHIS and PLCO datasets for prostate and breast cancers. We initially downloaded the data for the case study. We trained a set of machine learning models, including logistic regression, decision tree, random forest, k nearest neighbor, naïve Bayes, and gradient boosting for prostate and breast datasets. The input variables for breast study included age, prior personal history of cancer, history of breast cancer, pack-years of cigarettes smoked, BMI, and number of live births, number of relatives with breast cancer, diabetes, heart attack, high blood pressure and miscarriage, pregnancy, race, etc.

Furthermore, for prostate cancer, the inputs include the risk factors of age, prior personal history of cancer, history of prostate cancer, pack-years of cigarettes smoked, BMI, enlarged prostate, number of relatives with prostate cancer, diabetes, heart attack, high blood pressure, stroke, race, etc. Some of these risk factors might be missing in the PLCO dataset. Therefore, there was a slight difference in risk factors between these two datasets. Before input selection variables, we implemented a study on the p-value of the selected variables to make sure they significantly impact the performance of the machine learning model.

For each NHIS and PLCO, we randomly split the dataset into 70% training data and 30% testing the data. All machine learning models were trained on the same training dataset of

subjects and were generated using Python (version 3.7.7) [40]. We used Python Scikit-learn package (version 0.24.1) to perform the models [41]. For logistic regression, the “*linear_model.Logistic Regression*” was used. For naïve Bayes, we used the function “*naive_bayes.GaussianNB*”. The “*tree.DecisionTreeClassifier*” was employed to create decision tree. For k nearest neighbor, we used “*neighbors KNeighborsClassifier*”. In this study, positive is considered that the participant has cancer and negative means that the participant is healthy.

3.2 NHIS Results

3.2.1 NHIS-Breast Dataset

For the NHIS dataset, we initially downloaded the data for all of the women. NHIS dataset consists of 31133 women with 21641 healthy and 9492 breast cancer cases. The missing values were discarded from the data. The input variables for the NHIS breast dataset are shown in Table 2.

Table 2 Input variables for NHIS breast dataset.

<i>Input variables</i>
<i>Age</i>
<i>Personal prior history of cancer</i>
<i>BMI</i>
<i>Race</i>
<i>Hispanic origin or ancestry group</i>
<i>Ever smoked at least 100 cigarettes in entire life</i>
<i>At least 12 drinks alcoholic beverages in any one year</i>

The results of performing six types of machine learning algorithms for the NHIS breast dataset are shown in Figure 4. The AUCs are shown in Figure 4 and Table 3. The results indicate that logistic regression and gradient boosting with AUCs 0.624 and 0.620 show the best performances. We cannot only trust AUCs for comparing the performance of ML algorithms. Therefore, accuracy, sensitivity, specificity, PPV, and NPV are calculated in Table 3.

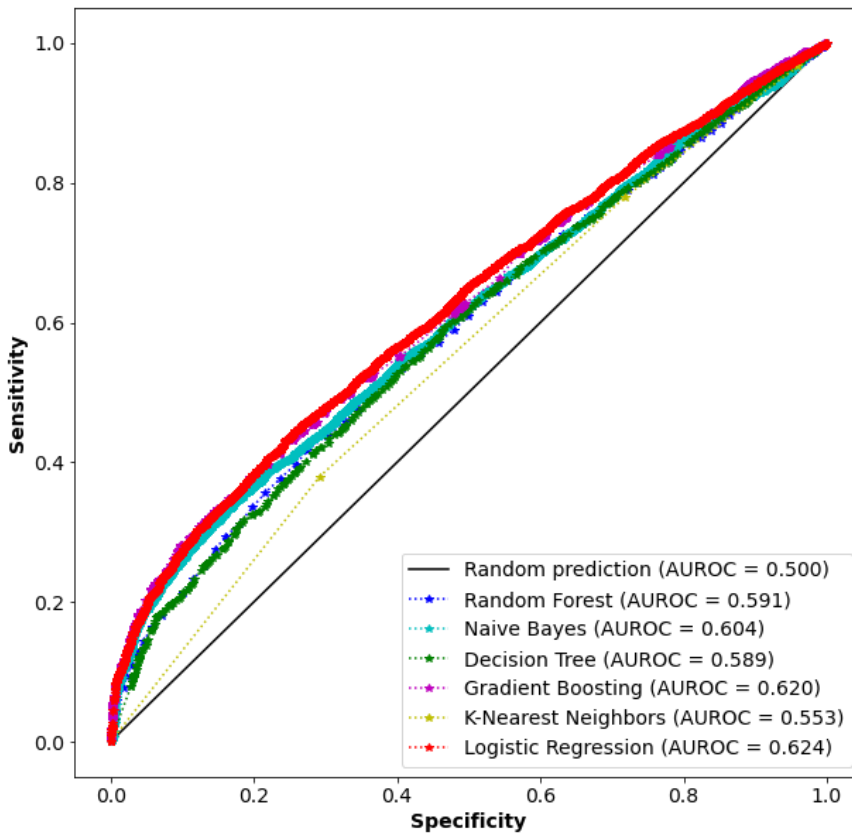


Figure 4 ROC curve for machine learning models for breast cancer study with NHIS Dataset.

From Table 3, it can be concluded that among all ML models, the random forest has the best performance. It has the accuracy of 65%, which means that random forest can differentiate the cancer patients from the healthy cases by 65%. In addition, RF has a sensitivity of 37%, which means that RF can predict the patient cases correctly by 37%. Furthermore, the specificity of the RF model is 71% which means the ability of the RF model to determine the healthy cases correctly is 71%. The PPV or precision for RF is 21% which indicates that among all cases identified as cancer cases, RF can classify 21% correctly as cancer cases.

Finally, the negative predictive value (NPV) for RF is 85%, indicating that RF can classify 85% correctly as healthy cases among all cases identified as healthy cases. Therefore, the random forest has the best performance among these ML algorithms.

Table 3 Accuracy, sensitivity, specificity, PPV and NPV for the NHIS breast dataset.

<i>Breast</i>	<i>LR</i>	<i>DT</i>	<i>RF</i>	<i>GB</i>	<i>KNN</i>	<i>NB</i>
<i>AUC</i>	0.624	0.589	0.591	0.620	0.553	0.604
<i>Accuracy</i>	70%	65%	65%	70%	63%	31%
<i>Sensitivity</i>	44%	37%	37%	51%	36%	31%
<i>Specificity</i>	70%	71%	71%	70%	71%	65%
<i>PPV</i>	4%	21%	21%	1%	28%	100%
<i>NPV</i>	98%	84%	85%	100%	78%	1%

3.2.2 NHIS - Prostate Dataset

For the NHIS dataset, we initially downloaded the data for all of the men. NHIS dataset consists of 21005 men with 15011 healthy and 5994 prostate cancer cases. The missing values were discarded from the data. The input variables for the NHIS prostate dataset are shown in Table 4.

Table 4 Input variables for NHIS prostate dataset.

<i>Input variables</i>
<i>Age</i>
<i>Personal prior history of cancer</i>
<i>BMI</i>
<i>Race</i>
<i>Hispanic origin or ancestry group</i>
<i>Ever smoked at least 100 cigarettes in entire life</i>
<i>At least 12 drinks alcoholic beverages in any one year</i>

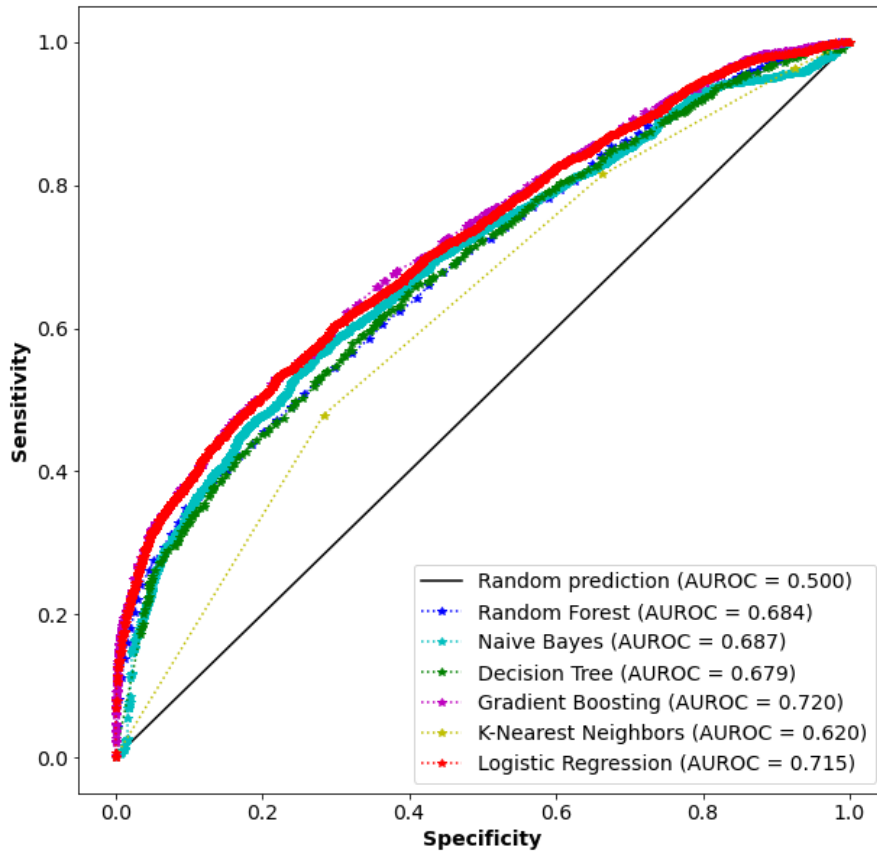


Figure 5 ROC curve for machine learning models for prostate cancer study with NHIS Dataset.

The results of performing six types of machine learning algorithms for the NHIS prostate dataset are shown in Figure 5. The AUCs are shown inside the Figure 5 and Table 5. The results indicate that gradient boosting and logistic regression with AUCs 0.72 and 0.715 show the best performances. We cannot only trust AUCs for comparing the performance of ML algorithms. Therefore, accuracy, sensitivity, specificity, PPV, and NPV are calculated in Table 5.

Table 5 Accuracy, sensitivity, specificity, PPV and NPV for the NHIS prostate dataset.

<i>Prostate</i>	<i>LR</i>	<i>DT</i>	<i>RF</i>	<i>GB</i>	<i>KNN</i>	<i>NB</i>
<i>AUC</i>	0.715	0.679	0.684	0.720	0.620	0.687
<i>Accuracy</i>	73%	70%	71%	73%	68%	29%
<i>Sensitivity</i>	60%	46%	49%	66%	42%	29%
<i>Specificity</i>	75%	75%	75%	74%	75%	70%
<i>PPV</i>	20%	27%	26%	15%	34%	99%
<i>NPV</i>	94%	87%	89%	97%	81%	1%

From Table 5, gradient boosting shows the best performance. It has the accuracy of 73%, which means that gradient boosting can differentiate the cancer patients from healthy cases by 73%. In addition, GB has a sensitivity of 66%, which means that GB predicts the patient cases correctly by 66%.

Furthermore, the specificity of GB is 74% which shows the ability of GB to determine the healthy cases correctly by 74%. The PPV or precision for GB is 15%, which indicates that GB can classify 15% correctly as cancer cases among all cases identified as cancer cases. Finally, the negative predictive value (NPV) for GB is 97%, which indicates that GB can classify 97% correctly as healthy cases among all cases identified as healthy cases.

Logistic regression has a very similar performance, but gradient boosting shows better performance with higher AUC, sensitivity, and negative predictive value. Therefore, gradient boosting shows the best performance.

3.3 PLCO Results

3.3.1 PLCO – Breast Dataset

For the PLCO dataset, we initially downloaded the data for all of the women. PLCO dataset consists of 72080 women with 67854 healthy and 4226 breast cancer cases. The missing values were discarded from the data since it was much less than the whole dataset, and therefore it did not impact our interpretation. The input variables for the PLCO breast dataset are shown in Table 6.

Table 6 Input variables for PLCO breast dataset.

<i>Input variables</i>
<i>Age</i>
<i>Personal prior history of cancer</i>
<i>History of breast cancer</i>
<i>Age at birth of first child</i>
<i>Pack years of cigarettes smoked</i>
<i>BMI</i>
<i>Number of live births</i>
<i>Number of relatives with breast cancer</i>
<i>Diabetes</i>
<i>Heart attack</i>
<i>High blood pressure</i>
<i>Stroke</i>
<i>Miscarriage</i>
<i>Pregnancy</i>
<i>Age at menopause</i>
<i>Race</i>
<i>Removed ovarian</i>
<i>Number of pregnancies</i>
<i>Ever used females' hormones</i>

After analyzing the dataset, we understood that we have the imbalanced data since the number of cancer patients or positive instances was much lower than the healthy

participants or negative instances. If we build a classifier on such imbalanced data, it may be biased towards negative prediction, and therefore, it will generate high value for false negative predictions, which is not correct. Therefore, we used a simple random sampling model to choose a random sample from the data and performed the machine learning models on the selected portion of the data. The results of performing six types of machine learning algorithms for PLCO breast cancer are shown in Figure 6. The AUCs are shown inside the Figure 6 and Table 7.

The results indicate that the random forest and gradient boosting with AUCs 0.868 and 0.861 show the best performances. However, it is not recommended to only consider AUC values for comparing the performance of different ML models. Therefore, accuracy, sensitivity, specificity, PPV, and NPV are calculated in Table 7.

From Table 7, it can be concluded that among all ML models, the random forest has the best performance. It has the accuracy of 95%, which means that random forest can differentiate the cancer patients from the healthy cases by 95%. In addition, RF has the sensitivity of 89%, which means that RF can predict the patient cases correctly by 89%. Furthermore, the specificity of the RF model is 95% which means the ability of the RF model to determine the healthy cases correctly by 95%. The PPV or precision for RF is 17% which indicates that among all cases identified as cancer cases, RF can classify 17% correctly as cancer cases.

Finally, the negative predictive value (NPV) for RF is 100%, which means that RF can classify 100% correctly as healthy cases among all cases identified as healthy cases. Overall, the RF shows the best performance for the PLCO dataset for breast study.

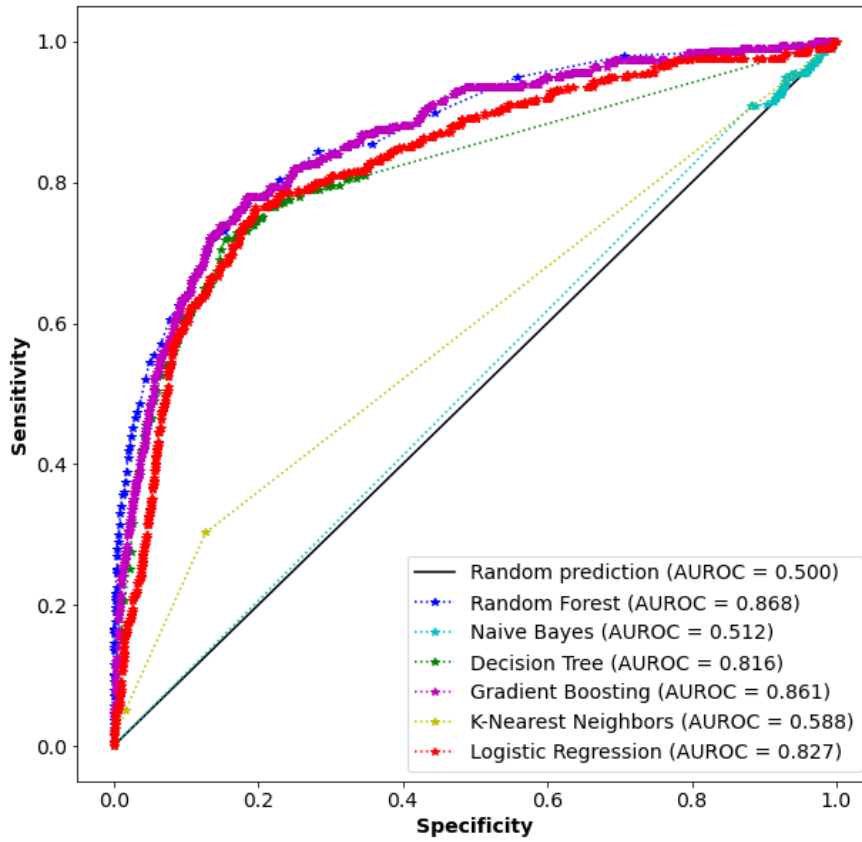


Figure 6 ROC curve for machine learning models for breast cancer study with PLCO Dataset.

Table 7 Accuracy, sensitivity, specificity, PPV and NPV for the PLCO breast dataset.

<i>Breast</i>	<i>LR</i>	<i>DT</i>	<i>RF</i>	<i>GB</i>	<i>KNN</i>	<i>NB</i>
<i>AUC</i>	0.827	0.816	0.868	0.861	0.588	0.512
<i>Accuracy</i>	94%	94%	95%	95%	93%	7%
<i>Sensitivity</i>	38%	45%	89%	62%	15%	5%
<i>Specificity</i>	95%	95%	95%	95%	95%	96%
<i>PPV</i>	5%	12%	17%	12%	5%	99%
<i>NPV</i>	99%	99%	100%	100%	98%	1%

3.3.2 PLCO – Prostate Dataset

We initially downloaded the PLCO data for all men to investigate prostate cancers. PLCO dataset consists of 70395 men with 62235 healthy and 8160 prostate cancer cases. By these statistics, again, we have to deal with imbalanced data. Thus, we used a simple random sampling model to select random data among the whole data, and all ML models were investigated on that portion of the data. The missing values were discarded from the data since it was much less than the whole dataset, and therefore, we could safely discard them. The input variables for the PLCO prostate dataset are shown in Table 8.

Table 8 Input variables for PLCO prostate dataset.

<i>Input variables</i>
<i>Age</i>
<i>Personal prior history of cancer</i>
<i>History of prostate cancer</i>
<i>Pack years of cigarettes smoked</i>
<i>BMI</i>
<i>Enlarged prostate</i>
<i>Number of relatives with prostate cancer</i>
<i>Diabetes</i>
<i>Heart attack</i>
<i>High blood pressure</i>
<i>Stroke</i>
<i>PSA test</i>
<i>Rectal exam</i>
<i>Prostate surgery</i>
<i>Race</i>

The results of performing six types of machine learning algorithms for PLCO prostate cancer are shown in Figure 7. The AUCs are shown in Figure 7 and Table 9. The results indicate that random forest and logistic regression with AUCs 0.859 and 0.850 show the best performances. However, it is not recommended to only consider AUC values for comparing the performance of different ML models. Therefore, accuracy, sensitivity, specificity, PPV, and NPV are calculated in Table 9.

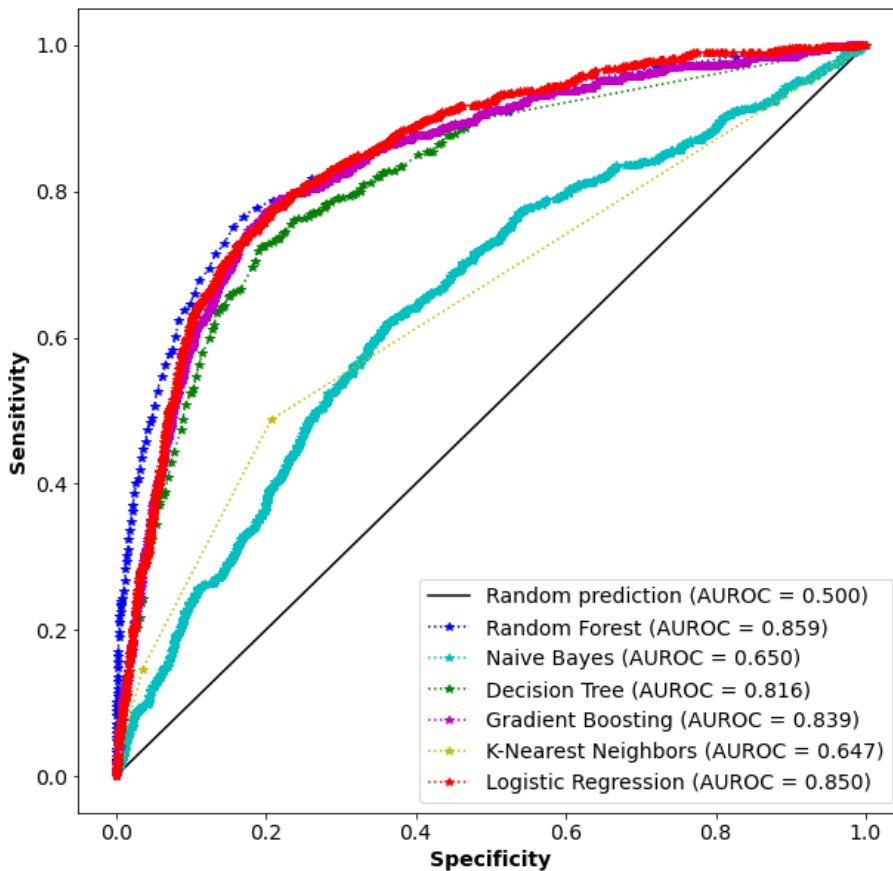


Figure 7 ROC curve for machine learning models for prostate cancer study with PLCO Dataset.

Based on Table 9, random forest (RF) shows the best performance among all ML algorithms with an AUC of 0.859. RF has the accuracy of 90%, which means that it can

differentiate the cancer patients from the healthy participants by 90%. In addition, RF has a sensitivity of 91%, which indicates that it determines the healthy cases correctly by 91%. Furthermore, the specificity of the RF model is 76% which means the ability of RF to predict the patient cases correctly is 76%. The positive predictive value (PPV) or precision for RF is 25%, which shows that RF can classify 25% correctly as cancer cases among all cases identified as cancer cases.

Finally, the negative predictive value (NPV) for RF is 99%, which indicates that RF can classify 99% correctly as healthy cases among all cases identified as healthy cases. Overall, the RF shows the best performance for the PLCO dataset for prostate cancer cases.

Table 9 Accuracy, sensitivity, specificity, PPV and NPV for the PLCO prostate dataset.

<i>Prostate</i>	<i>LR</i>	<i>DT</i>	<i>RF</i>	<i>GB</i>	<i>KNN</i>	<i>NB</i>
<i>AUC</i>	0.850	0.816	0.859	0.839	0.647	0.650
<i>Accuracy</i>	88%	88%	90%	88%	87%	16%
<i>Specificity</i>	54%	51%	76%	56%	36%	12%
<i>Sensitivity</i>	90%	91%	91%	89%	89%	93%
<i>PPV</i>	25%	30%	25%	15%	15%	97%
<i>NPV</i>	97%	96%	99%	98%	96%	5%

This study indicates that the proposed ML models have similar or higher AUC values than previous publications [42]. In addition, other parameters such as sensitivity, specificity, PPV, NPV values reported in this thesis showed the efficiency of the proposed ML models

compared to reported publications [43]. Therefore, they validate the accuracy of the results reported in this study.

3.4 Gini Importance

Gini importance or Mean Decrease in Impurity (MDI) calculates each feature importance as the sum over the number of splits across all trees that include the feature, proportionally to the number of samples it splits [44].

Random forest applies an implicit feature selection, using a small subset of “strong variables” for classification [45], which leads to its outstanding performance. Gini Importance can be used to visualize the outcome of the implicit feature selection of the random forest [46], and it can be used as a feature selection.

This feature importance score can provide a ranking of spectral features, and it comes from the training of random forest classifier: at each node τ within the binary trees T of the random forest, the optimal split obtained using Gini impurity $i(\tau)$ which measures how good a split is separating the samples of the two classes in this specific node. Where $p_k = \frac{n_k}{n}$ is the fraction of the n_k samples from class $k = \{0,1\}$ out of the total of n samples at node τ , the Gini impurity $i(\tau)$ is calculated as Eq (13).

$$i(\tau) = 1 - p_1^2 - p_0^2 \quad (13)$$

It causes Δi to decrease that results from splitting and sending the samples to two sub-nodes τ_l and τ_r (sample fractions $p_l = \frac{n_l}{n}$ and $p_r = \frac{n_r}{n}$) by a threshold t_θ on variable θ is defined in Eq (14).

$$\Delta i(\tau) = i(\tau) - p_l i(\tau_l) - p_r i(\tau_r) \quad (14)$$

By searching over all variables θ available at the node and all possible threshold t_θ , the pair $\{\theta, t_\theta\}$ leading to a maximal Δi is determined. Then, the decrease in Gini impurity is recorded and summed over all nodes τ in all trees T in the forest, individually for all variables θ .

$$I_G(\theta) = \sum_T \sum_\tau \Delta i_\theta(\tau, T) \quad (15)$$

The Gini importance I_G in Eq (15) indicates how often a particular feature θ was selected for a split, and how large its overall discriminative value was for the classification problem under study.

3.5 Gini Importance-Breast Dataset

Figure 8 shows the Gini importance of a random forest for the PLCO dataset for breast study. This figure indicates that some features such as ‘age at which breast cancer happened’, ‘pack years cigarettes’, ‘personal history of cancer’, ‘BMI’, ‘age at birth of first child’, ‘hypertension’, ‘miscarriage’, ‘age’, ‘race’, ‘removed ovarian’, ‘number of pregnancies’ are the most influential risk factors of the breast cancer. Therefore, Gini importance can provide a significant insight for feature selection.

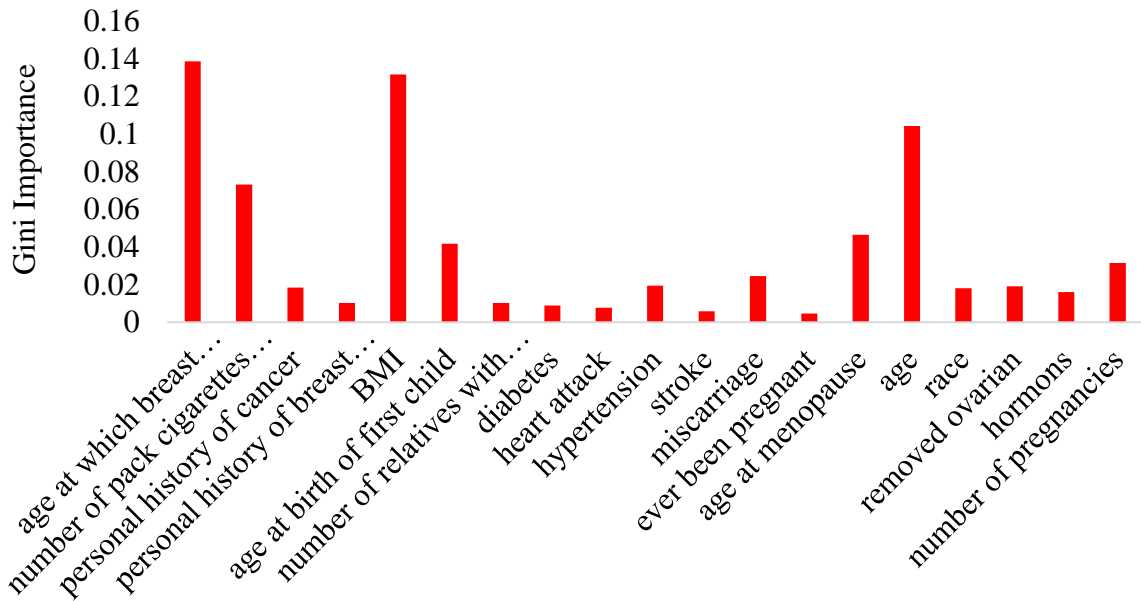


Figure 8 Gini importance coefficient for each risk factors of the PLCO breast dataset.

3.6 Gini Importance- Prostate Dataset

Figure 9 shows the Gini importance for the PLCO prostate dataset. This figure indicates that ‘age at which prostate cancer happened’, ‘number of packed-cigarettes smoked per day’, ‘personal history of cancer’, ‘BMI’, ‘hypertension’ and ‘PSA test’, ‘rectal exam’, ‘race’ and ‘age’ are the most significant features based on the Gini importance of the random forest classifier.

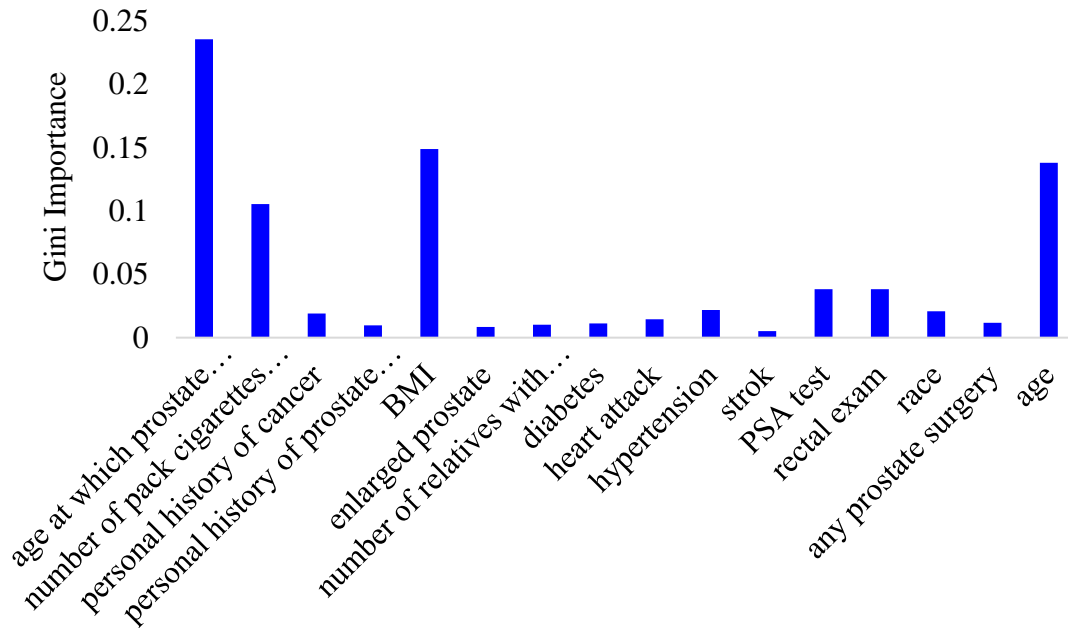


Figure 9 Gini importance coefficient for each risk factors of the PLCO prostate dataset.

3.7 Comparison between risk factors

The results of Figure 8 and Figure 9 show that some risk factors of prostate and breast cancers are similar such as ‘age at which cancer happened’, ‘number of packed-cigarettes per day’, ‘BMI’, ‘age’, ‘race’, ‘hypertension’ and ‘personal history of cancer’. Furthermore, the influence of these risk factors in men is more intense than in women based on the Gini importance coefficient studied in this thesis. Therefore, social marketing campaigns can consider these risk factors and their influence to decide whether to target the general population or tailor messages for different audiences with differing demographic, cultural, or behavioral characteristics.

4 Conclusion

In this study, six types of machine learning algorithms were investigated to identify the most significant risk factors of prostate and breast cancers. For this reason, National Health Interview Survey (NHIS) and Prostate, Lung, Colorectal, and Ovarian (PLCO) datasets were used. The results show that random forest and gradient boosting were the most efficient ML algorithms for classifying the cases into healthy or cancer cases. Moreover, Gini importance coefficient was used to identify the most significant prostate and breast cancer risk factors. A comparison between the most significant risk factors of prostate and breast cancers shows that ‘age at which cancer happened’, ‘number of packed cigarettes per day’, ‘BMI’, ‘age’, ‘race’, ‘hypertension’ and ‘personal history of cancer’ are common factors in these two types of cancers. In addition, the influence of these risk factors in men is more intense than in women based on the Gini importance coefficient obtained in this study. Social marketing campaigns can consider these risk factors and their influence to decide whether to target the general population or tailor messages for different audiences with differing demographic, cultural, or behavioral characteristics.

References

1. Pfeiffer, R.M., et al., *Risk prediction for breast, endometrial, and ovarian cancer in white women aged 50 y or older: derivation and validation from population-based cohort studies*. PLoS Med, 2013. **10**(7): p. e1001492.
2. Bray, F., et al., *Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries*. CA Cancer J Clin, 2018. **68**(6): p. 394-424.
3. Siegel, R.L., K.D. Miller, and A. Jemal, *Cancer statistics, 2018*. CA Cancer J Clin, 2018. **68**(1): p. 7-30.
4. Rawla, P., *Epidemiology of prostate cancer*. World journal of oncology, 2019. **10**(2): p. 63.
5. Ferlay, J., et al., *Global cancer observatory: cancer today*. Lyon, France: International Agency for Research on Cancer, 2018.
6. Colditz, G.A., K.Y. Wolin, and S. Gehlert, *Applying What We Know to Accelerate Cancer Prevention*. Science Translational Medicine, 2012. **4**(127): p. 127rv4.
7. Willcox, S.J., B.W. Stewart, and F. Sitas, *What factors do cancer patients believe contribute to the development of their cancer? (New South Wales, Australia)*. Cancer Causes & Control, 2011. **22**(11): p. 1503.
8. Thomson, A.K., et al., *Beliefs and perceptions about the causes of breast cancer: a case-control study*. BMC Research Notes, 2014. **7**(1): p. 558.
9. Evans, W.D., *How social marketing works in health care*. BMJ (Clinical research ed.), 2006. **332**(7551): p. 1207-1210.
10. Gordon, R., et al., *The effectiveness of social marketing interventions for health improvement: What's the evidence?* Public Health, 2006. **120**(12): p. 1133-1139.
11. Wakefield, M.A., B. Loken, and R.C. Hornik, *Use of mass media campaigns to change health behaviour*. The Lancet, 2010. **376**(9748): p. 1261-1271.
12. Lynn A. Blewett, J.A.R.D., Miriam L. King and Kari C.W. Williams, *IPUMS Health Surveys: National Health Interview Survey, Version 6.4 [dataset]*. , M.I. Minneapolis, Editor. 2019.
13. NCI, *Cancer Data Access System (CDAS): Prostate, Lung, Colorectal and Ovarian (PLCO)*. 2020.
14. Howlader, N., et al., *SEER cancer statistics review, 1975–2017*. National Cancer Institute, 2020.
15. Barber, L., et al., *Family History of Breast or Prostate Cancer and Prostate Cancer Risk*. Clinical Cancer Research, 2018. **24**(23): p. 5910-5917.
16. Sarker, I.H., M.H. Furhad, and R. Nowrozy, *AI-Driven Cybersecurity: An Overview, Security Intelligence Modeling and Research Directions*. SN Computer Science, 2021. **2**(3): p. 173.
17. Mohammed, M., M.B. Khan, and E.B.M. Bashier, *Machine learning: algorithms and applications*. 2016: Crc Press.

18. Nasteski, V., *An overview of the supervised machine learning methods*. Review Scientific Paper, 2017.
19. Tavallali, P., P. Tavallali, and M. Singhal, *K-means tree: an optimal clustering tree for unsupervised learning*. The Journal of Supercomputing, 2021. **77**(5): p. 5239-5266.
20. Szepesvári, C., *Algorithms for reinforcement learning*. Synthesis lectures on artificial intelligence and machine learning, 2010. **4**(1): p. 1-103.
21. Maja Pohar, M.B., and Sandra Turk, *Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study*. Metodološki zvezki, 2004. **1**(1): p. 143–161.
22. Tu, J.V., *Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes*. Journal of Clinical Epidemiology, 1996. **49**(11): p. 1225-1231.
23. Sperandei, S., *Understanding logistic regression analysis*. Biochem Med (Zagreb), 2014. **24**(1): p. 12-8.
24. Song, Y.-Y. and L. Ying, *Decision tree methods: applications for classification and prediction*. Shanghai archives of psychiatry, 2015. **27**(2): p. 130.
25. Lorena, A.C., et al., *Comparing machine learning classifiers in potential distribution modelling*. Expert Systems with Applications, 2011. **38**(5): p. 5268-5275.
26. Miguel-Hurtado, O., et al., *Comparing Machine Learning Classifiers and Linear/Logistic Regression to Explore the Relationship between Hand Dimensions and Demographic Characteristics*. PLOS ONE, 2016. **11**(11): p. e0165521.
27. Al-Aidaros, K., A. Bakar, and Z. Othman, *Naïve bayes variants in classification learning*. 2010 International Conference on Information Retrieval & Knowledge Management (CAMP), 2010: p. 276-281.
28. Rish, I. *An empirical study of the naive Bayes classifier*. in *IJCAI 2001 workshop on empirical methods in artificial intelligence*. 2001.
29. Pal, M., *Random forest classifier for remote sensing classification*. International journal of remote sensing, 2005. **26**(1): p. 217-222.
30. Breiman, L., *Bagging predictors*. Machine Learning, 1996. **24**(2): p. 123-140.
31. Ke, G., et al., *LightGBM: a highly efficient gradient boosting decision tree*, in *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017, Curran Associates Inc.: Long Beach, California, USA. p. 3149–3157.
32. Natekin, A. and A. Knoll, *Gradient boosting machines, a tutorial*. Frontiers in Neurorobotics, 2013. **7**(21).
33. Mucherino, A., P.J. Papajorgji, and P.M. Pardalos, *K-nearest neighbor classification*, in *Data mining in agriculture*. 2009, Springer. p. 83-106.
34. Wu, Y., K. Ianakiev, and V. Govindaraju, *Improved k-nearest neighbor classification*. Pattern Recognition, 2002. **35**(10): p. 2311-2318.
35. Haghghi, S., et al., *PyCM: Multiclass confusion matrix library in Python*. Journal of Open Source Software, 2018. **3**(25): p. 729.
36. Marzban, C., *The ROC curve and the area under it as performance measures*. Weather and Forecasting, 2004. **19**(6): p. 1106-1114.

37. Wong, H.B. and G.H. Lim, *Measures of diagnostic accuracy: sensitivity, specificity, PPV and NPV*. Proceedings of Singapore healthcare, 2011. **20**(4): p. 316-318.
38. Fisher, R.A., *Statistical methods for research workers*, in *Breakthroughs in statistics*. 1992, Springer. p. 66-70.
39. Saleh, H., *Machine Learning Fundamentals: Use Python and scikit-learn to get up and running with the hottest developments in machine learning*. 2018: Packt Publishing Ltd.
40. Van Rossum G, D.F., *Python 3 Reference Manual*. CreateSpace, 2009.
41. Pedregosa, F., et al., *Scikit-learn: Machine learning in Python*. the Journal of machine Learning research, 2011. **12**: p. 2825-2830.
42. Stark, G.F., et al., *Predicting breast cancer risk using personal health data and machine learning models*. PloS one, 2019. **14**(12): p. e0226765-e0226765.
43. Barlow, H., S. Mao, and M. Khushi, *Predicting High-Risk Prostate Cancer Using Machine Learning Methods*. Data, 2019. **4**(3): p. 129.
44. Menze, B.H., et al., *A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data*. BMC bioinformatics, 2009. **10**(1): p. 1-16.
45. Breiman, L., *Consistency for a simple model of random forests*. 2004.
46. Breiman, L., *Random forests*. Machine learning, 2001. **45**(1): p. 5-32.