BABEŞ BOLYAI UNIVERSITY, CLUJ NAPOCA, ROMÂNIA
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

Breast cancer recognition based on genetic factors

– DS - Intelligent Modeling report –

Author
Razvan-Dan Salsigan

2022-2023

Abstract

Breast cancer is a complex disease that is challenging to diagnose and effectively treat. In this study, I aimed to leverage genetic data to develop an intelligent method for finding whether various features can be differentiated between breast cancer patients and healthy patients.

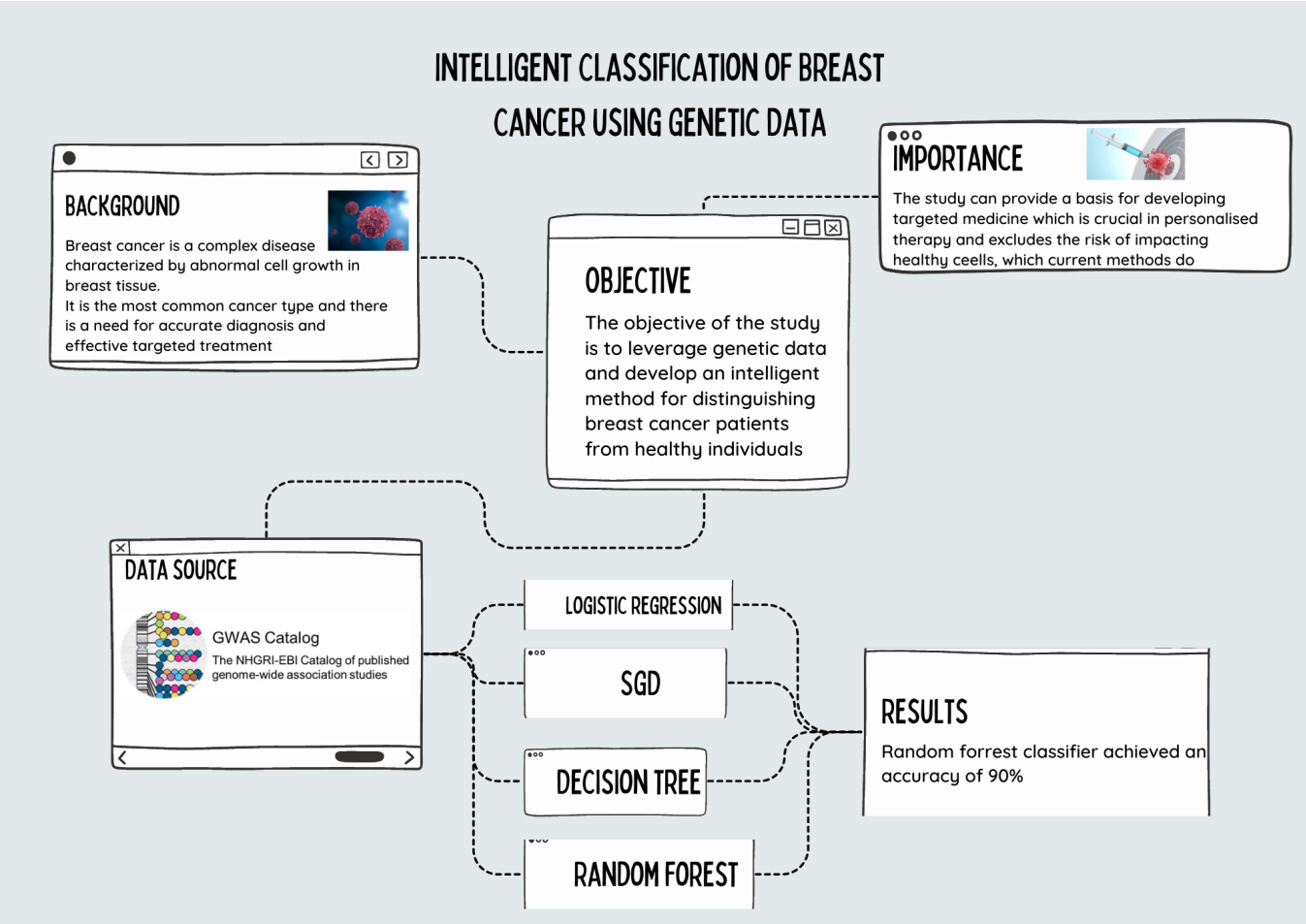
The study is important as there is a need for improved diagnostic tools for detecting breast cancer cases accurately.

Multiple types of classification methods were used in order to obtain the best results.

The data used in this study comes from the Genome-Wide Association Study (GWAS) project from the National Human Genome Research Institute (NHGRI).

The results of the study demonstrate that there is potential in accurately classifying breast cancer cases against non-cancer cases from genetic data. The best among the classifiers trained obtained an accuracy of over 90%.

Figure 1: Graphic abstract



Contents

List of Tables

List of Figures

Chapter 1

Introduction

Breast cancer (BC) is the most common cancer type and the second leading cause of cancer death among women.

Family linkage studies have identified several high-penetrance genes, such as BRCA1, BRCA2, PTEN and TP53, that are responsible for inherited BC syndromes. Other studies suggest that genes involved in DNA repair are associated with moderate risk. Genome wide association studies (GWAS) revealed single nucleotide polymorphisms (SNPs) in several novel genes associated to breast cancer susceptibility.

The GWAS studies use a large number of common SNPs to identify associations with disease that rely upon patterns of linkage disequilibrium in the human genome. Those studies provide a powerful tool for identifying novel markers for susceptibility or prognosis of disease. [1]

1.1 What? Why? How?

The main purpose of this study is to establish if there is the possibility of differentiating between breast cancer subjects and healthy ones by leveraging the data from the GWAS study from NHGRI.

The importance of the study consists in the need for improved diagnostic tools in breast cancer, providing ways of computing the risk of individuals to develop the disease and create a foundation for developing targeted medicine.

To achieve the objective, multiple machine learning techniques were employed, from data preprocessing and feature engineering to training state-of-the art classification algorithms, such as stochastic gradient descent, k nearest neighbors, decision tree and random forest classifiers in order to achieve a good accuracy.

Chapter 2

Scientific Problem

The problem addressed is the accurate classification of breast cancer patients. Current diagnostic methods rely on a combination of clinical assessment, imaging techniques and histopathological examination. Integrating genetic information in the diagnostic process has great potential in enhancing personalized treatment strategies.

2.1 Problem definition

The problem consists in utilising NHGRI GWAS data for accurately classifying breast cancer cases. The said dataset has the following features: chromosome, position, arm, chromosome id, chromosome position, risk allele frequency, pvalue mlog, OR or BETA, confidence interval and the SNP.

The use of an intelligent algorithm is motivated by the fact that it can effectively handle large scale and high dimensional data, and can identify relevant genetic features and patterns that are not easy to determine by traditional statistic methods. Moreover, it has the potential to learn and adapt from the data, improving its results the more data it receives.

The inputs of the method are the above mentioned features, and the results is binary, whether or not the data suggests breast cancer risk.

The problem is very important and interesting, as it serves as the basis of genetic screening, and it also provides a foundation for developing targeted medicine to fit the individual's needs.

Chapter 3

State of the art/Related work

3.0.1 "Genome-wide association study identifies novel breast cancer susceptibility loci" [2]

This study aimed to identify susceptibility genes for breast cancer through a genome-wide association study. The researchers conducted a three-stage study involving a total of 26,258 breast cancer cases and 27,894 controls from multiple studies. In the first stage, they genotyped 227,876 single nucleotide polymorphisms (SNPs) and identified five novel independent loci associated with breast cancer risk, four of which contained plausible causative genes (FGFR2, TNRC9, MAP3K1, and LSP1). In the second stage, they selected 1,792 significant SNPs and confirmed their associations. In the third stage, they tested 30 of the most significant SNPs in additional case-control studies and identified six SNPs that showed strong associations with breast cancer risk.

The study findings suggest that breast cancer susceptibility is likely to be polygenic, with multiple loci each contributing a small effect on risk. The known susceptibility genes, BRCA1 and BRCA2, account for less than 25% of the familial risk of breast cancer, indicating that there are many additional common susceptibility alleles yet to be identified. The researchers utilized the advances in genotyping technology to analyze hundreds of thousands of SNPs and identified several loci associated with breast cancer risk.

The most significant SNP identified in this study, rs2981582, is located within intron 2 of the FGFR2 gene. Other significant SNPs were found within or near genes TNRC9, MAP3K1, LSP1, and H19. These findings provide new insights into the genetic basis of breast cancer and suggest potential target genes for further research.

Overall, this study highlights the importance of large-scale association studies in identifying common genetic variants associated with breast cancer risk. The findings contribute to our understanding

of the genetic architecture of breast cancer and may have implications for risk assessment, prevention, and targeted therapies in the future.

3.0.2 "Design and Multiseries Validation of a Web-Based Gene Expression Assay for Predicting Breast Cancer Recurrence and Patient Survival" [3]

The study aims to develop and validate a prognostic algorithm for breast cancer patients using gene expression analysis. The researchers analyzed genomic and clinical data from 477 patients with breast cancer to identify genes associated with disease outcome. They used Cox regression models to identify genes that were independently associated with the risk of disease recurrence and overall survival.

The researchers developed a "metagene" algorithm based on the identified genes, which could stratify patients into high or low-risk groups for recurrence. They then applied this classifier to an additional 1016 patients from five independent series to validate its performance. The results showed that the algorithm successfully stratified patients into risk groups with significant differences in recurrence-free and overall survival.

The study also found that the prognostic classifier was the strongest predictor of outcome in each validation series when compared to standard prognostic factors. In node-negative patients who did not receive treatment, the algorithm showed 88% sensitivity and 44% specificity for 10-year recurrence-free survival.

The researchers concluded that the 200-gene prognosis signature they developed and validated could serve as a strong independent predictor of patient outcome in a range of breast cancer subtypes. They implemented the algorithm within an online analysis environment called ChipDX, which was made available for research use.

The study highlights the potential of genomic profiling and multigene algorithms in predicting the risk of disease recurrence and overall survival in breast cancer patients. By incorporating gene expression analysis into clinical management, it may be possible to tailor treatment strategies based on individual patient's risk profiles.

3.0.3 "The role of gene expression profiling by microarray analysis for prognostic classification of breast cancer" [5]

The study used gene expression profiling to analyze breast cancer samples and identify gene expression patterns associated with different outcomes. Two different microarray platforms were used, one with 25,000 oligonucleotide probes and another with 18,000 cDNA probes.

The researchers collected tumor samples from 295 patients younger than 53 years with stage I and

II breast cancer who were treated at their institute between 1984 and 1993. They isolated RNA from these tumors and assessed the expression of 25,000 genes using the microarray platforms. They then used various statistical approaches to correlate gene expression with distant metastasis-free probability and overall survival of the patients.

In addition, a randomized phase II trial was conducted on patients with locally advanced breast cancer. These patients received neoadjuvant chemotherapy with either AC (n = 24) or AD (n = 24) regimens. Core needle biopsies were obtained from these patients before treatment, and gene expression profiles for 18,000 genes were generated from these samples. The researchers correlated these profiles with the response of the primary tumor to the chemotherapy administered. They also compared the gene expression profiles before and after chemotherapy.

The results of the study showed that a 70-gene expression profile was associated with an increased risk of developing distant metastases within 5 years. The researchers also identified a Wound Signature in the same tumors, which could further divide the tumors into 'activated' and 'quiescent' subgroups. By combining the 70-gene expression profile and the Wound Signature, patients with different prognoses could be identified. The researchers are continuing to test additional gene expression signatures in these tumors to develop an optimal prognostic classifier and gain a better understanding of breast cancer biology.

In terms of predicting the response to neoadjuvant chemotherapy, the study found that 20% of the patients achieved (near) pathological complete remission of the primary tumor after treatment. However, no gene expression pattern correlating with response could be identified for all patients or for the AC or AD treated groups separately. The researchers concluded that gene expression profiles predicting the response to specific chemotherapy regimens are likely to be subtle and may require larger patient cohorts for detection.

Overall, the study suggests that gene expression profiling can be useful in predicting the prognosis of breast cancer and guiding clinical decision-making in the treatment of primary breast cancer. However, further research and validation are needed to develop reliable prognostic and predictive tests based on gene expression profiling.

3.1 Proposed approach - methodology

3.1.1 Exploratory data Analysis

The data preprocessing step involves data cleaning, such as removing outliers and overflowing values, filling the missing values and encoding categorical features. Another important preprocessing step

consists of downsampling, as the whole dataset has a major difference in the number of breast cancer entries and other types of cases.

The feature selection step consists in dropping the columns that are ids or irrelevant data for such a study.

The data visualization step consists in creating general statistic analysis of the data, correlation matrix and a pairwise relationship scatterplot.

3.1.2 Training and testing the model

For training, the dataset was split into subsets of 20% for testing and 80% for training.

Several state-of-the classification algorithms from scikit learn [4] were trained, namely LogisticRegression, SGDClassifier, KNeighborsClassifier, DecisionTreeClassifier and RandomForestClassifier.

Each model was tested by calculating the accuracy for all the data and for the positive class (the breast cancer cases) in particular, and the confusion matrix.

The model with the best results was further evaluated using precision, recall and ROC curve and the differences between the correctly and incorrectly classified entries were analysed.

3.1.3 Methodology

3.1.3.1 Evaluation criteria

- Accuracy: The proportion of correctly classified instances
- Precision: The ability to correctly identify breast cancer cases
- Recall: The ability to capture all actual breast cancer cases
- Area under the ROC curve (AUC-ROC): A metric that evaluates the model's ability to distinguish between breast cancer and non-breast cancer instances

3.1.3.2 Hypothesis and experimental methodology

The hypothesis is that by utilizing NHGRI GWAS data and an intelligent algorithm, we can accurately classify breast cancer versus non-breast cancer patients.

The experiment involves training and testing the selected machine learning algorithm on the available NHGRI GWAS data. The aim is to evaluate the performance of the algorithm in accurately classifying breast cancer cases.

3.1.3.3 Training/test data

- The training data consist of NHGRI GWAS data, including genetic markers and corresponding labels indicating breast cancer or non-breast cancer. This data is used to train the machine learning algorithm to learn the patterns and relationships between genetic markers and breast cancer.
- The test data are separate instances with NHGRI GWAS data, similar to the training data but unseen by the model during training. These instances are used to evaluate the algorithm's performance in accurately classifying breast cancer cases.

3.1.4 Data

The NHGRI GWAS (Genome-Wide Association Study) project data is a comprehensive collection of genetic information gathered from numerous individuals across different populations. GWAS is a study design that aims to identify genetic variations associated with diseases, traits, or other phenotypes.

Each entry in the NHGRI GWAS dataset typically includes information such as the disease or trait being studied, the chromosome and position of the genetic marker, the allele frequencies (frequency of different alleles observed), statistical measures like p-values or odds ratios indicating the strength of association, and confidence intervals.

The following features were used in the analysis:

- 'DISEASE/TRAIT': represents the specific disease or trait associated with each genetic marker. For the breast cancer class, columns where "breast cancer" appeared in the description were considered, and the others were labeled as non-breast cancer
- 'Chromosome': denotes the chromosome number where the genetic marker is located.
- 'Position': indicates the specific position of the genetic marker on the chromosome.
- 'Arm': represents the arm or region of the chromosome where the genetic marker is located.
- 'CHR_ID': serves as an identifier for the chromosome.
- 'CHR_POS': provides the position of the genetic marker in terms of base pairs (bp).
- 'SNPS': contains the specific genetic marker, represented by its rsID (Reference SNP ID).
- 'RISK ALLELE FREQUENCY': represents the frequency of the risk allele (the allele associated with an increased risk of the disease) in the population.

- 'PVALUE_MLOG': indicates the negative logarithm of the p-value associated with the association between the genetic marker and breast cancer.
- 'OR or BETA': represents the odds ratio (OR) or beta coefficient associated with the genetic marker. These values provide information about the strength and direction of the association between the marker and the disease.
- 'CIleft' and 'CIright': denote the lower and upper bounds of the confidence interval for the OR or beta coefficient.

3.1.5 Results

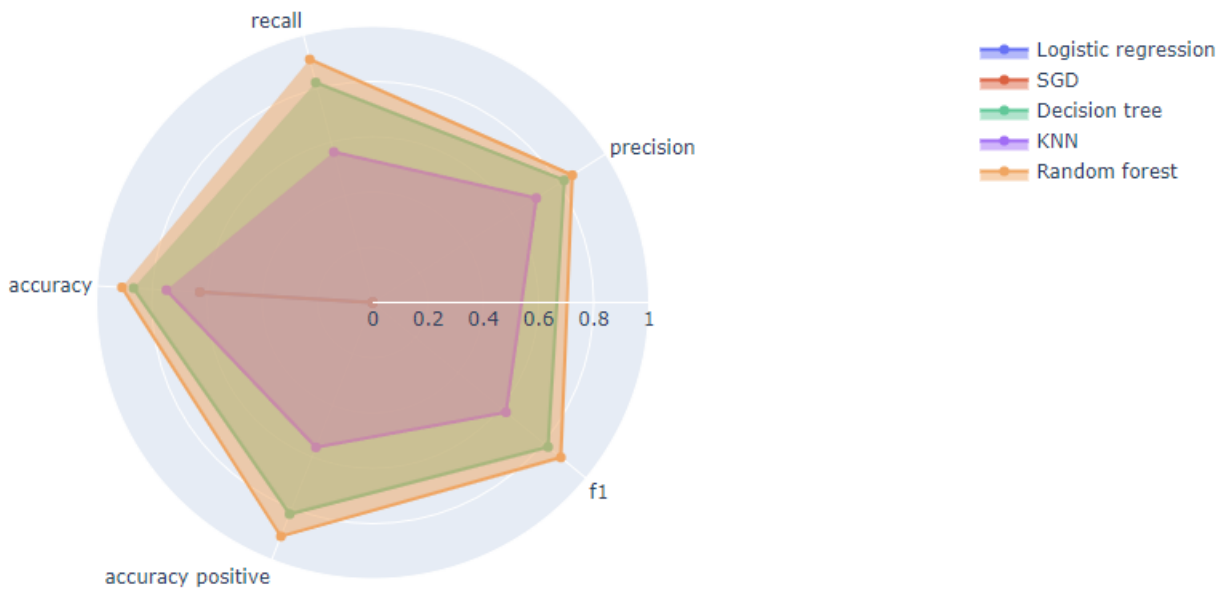


Figure 3.1: Result comparison

As can be seen from the radar chart, Logistic regression and SGD were not able to correctly classify any positive classes, and therefore they do not appear on the graph.

The best result was obtained by the random forest classifier, achieving the following results.

Table 3.1: The evaluation metrics of the random forest classifier

Metric	**Value**
Accuracy	90.96%
Accuracy for the positive class	90.9%
Precision	0.8571
Recall	0.909
F1	0.8823

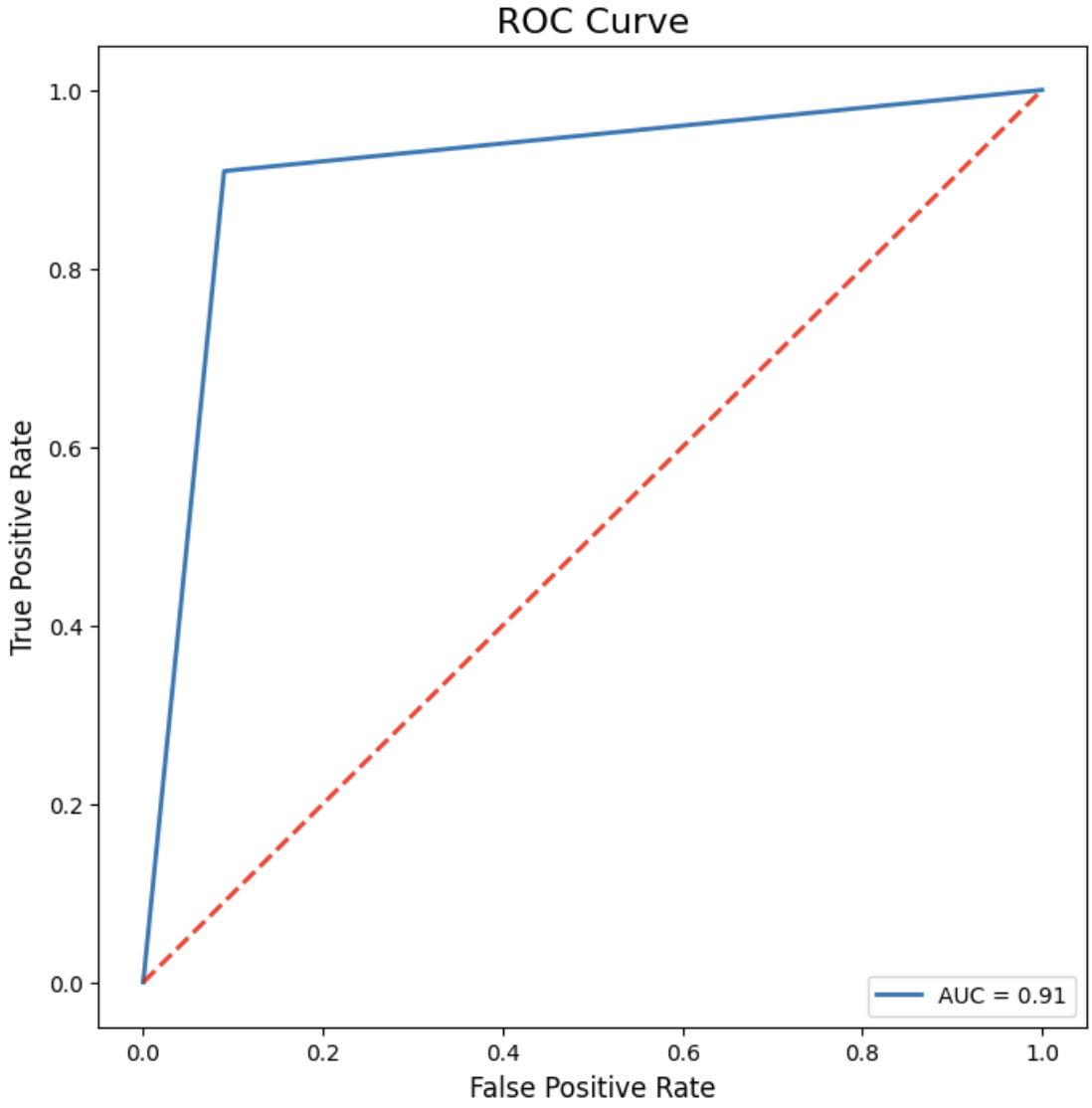


Figure 3.2: ROC curve

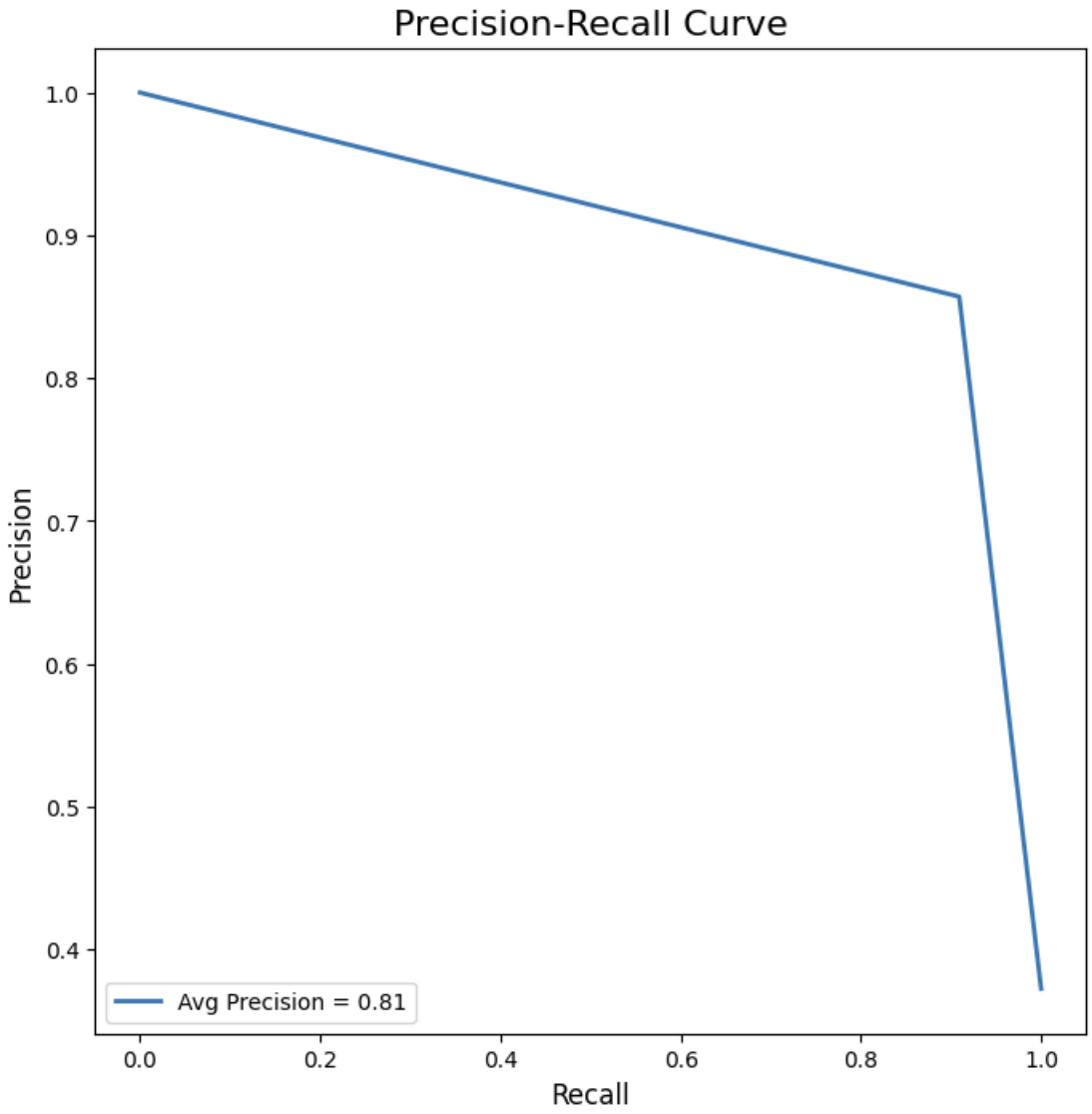


Figure 3.3: Precision recall curve

3.2 Discussion

The hypothesis stated that leveraging genetic data and using an intelligent method can accurately classify breast cancer cases. The obtained results, with an accuracy of 90.96% and high precision and recall values, support the hypothesis. The method successfully differentiated between breast cancer and non-cancer cases, indicating the potential of genetic data in accurately classifying breast cancer patients.

The obtained results suggest several strengths of the method:

- High Accuracy: The method achieved an accuracy of 90.96%, indicating its ability to make correct predictions.
- Good Precision and Recall: The precision and recall values of 0.8571 and 0.909, respectively, indicate a relatively low rate of false positives and false negatives in identifying breast cancer cases.

The random forest algorithm, known for its ability to handle high-dimensional data and capture complex relationships, likely played a significant role in achieving the high accuracy and discriminatory power observed. The algorithm's ensemble of decision trees and its ability to consider multiple features simultaneously may have facilitated effective breast cancer classification based on the provided genetic data. The data itself, sourced from the NHGRI GWAS project, contains genetic markers associated with breast cancer. The presence of informative markers and their relationship with breast cancer likely contributed to the algorithm's ability to discriminate between breast cancer and non-cancer cases.

Moreover, an analysis on the differences between the correctly and incorrectly classified entries was conducted, and the below details were noticed.

- Chromosome: The mean value for both groups is similar, with the correctly predicted instances having a slightly lower mean (9.496) compared to the incorrectly predicted instances (9.755). The range of values (min and max) is similar for both groups.
- Position: The mean position value is slightly higher for the incorrectly predicted instances (20.891) compared to the correctly predicted instances (20.622). The standard deviation is higher for the correctly predicted instances (7.161) compared to the incorrectly predicted instances (6.429). The range of positions is similar for both groups.
- Arm: The arm values are categorical (1 or 2), so the descriptive statistics provide information on the count and distribution of each arm in the groups. The majority of instances in both groups have arm value 1, with a slightly higher count in the correctly predicted instances.

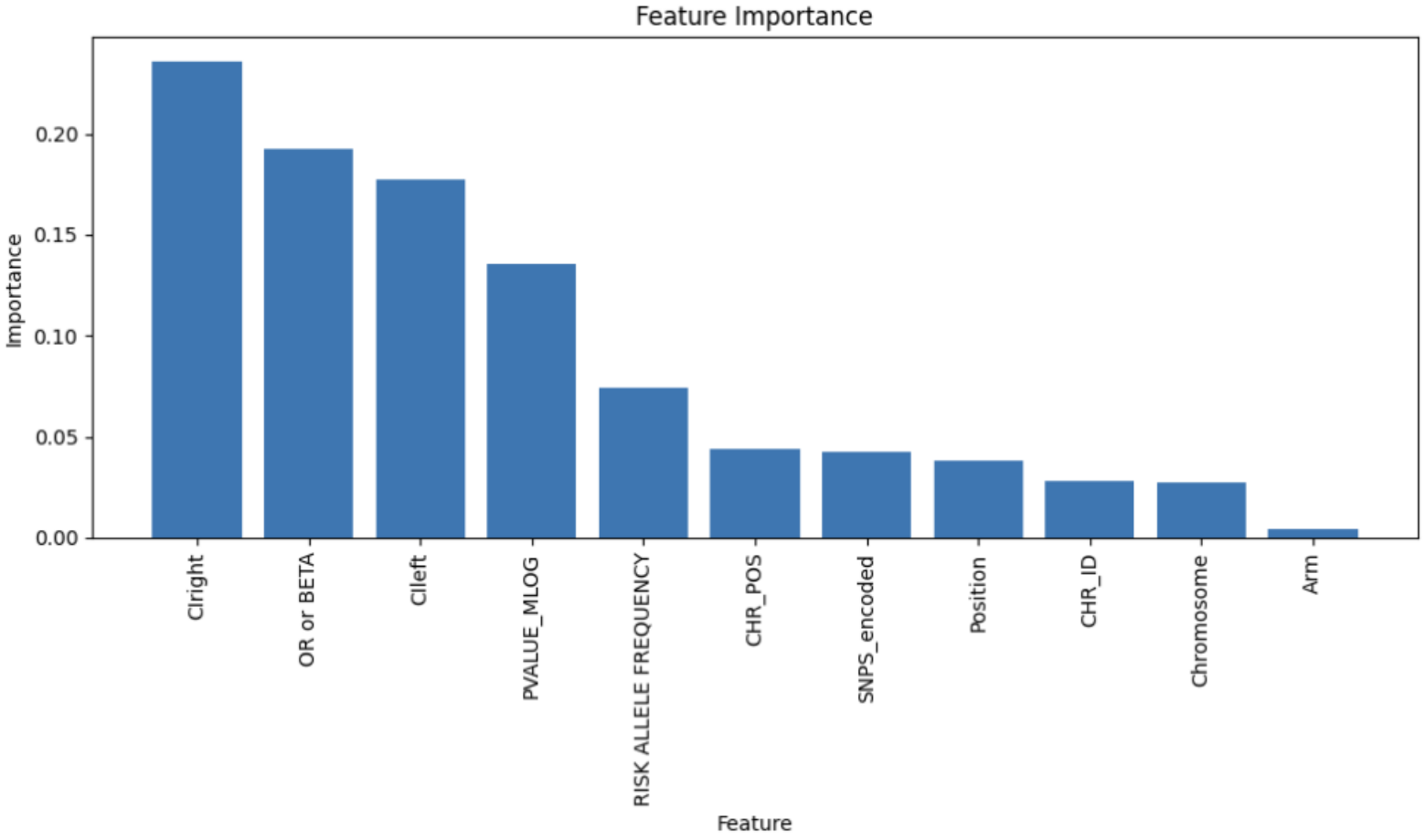


Figure 3.4: Feature importance analysis

- CHR_ID and CHR_POS: Both features have the same mean and range values for both groups, indicating similar distributions.
- RISK ALLELE FREQUENCY: The mean RISK ALLELE FREQUENCY value is slightly lower for the correctly predicted instances (0.395) compared to the incorrectly predicted instances (0.446). The standard deviation is slightly higher for the incorrectly predicted instances (0.253) compared to the correctly predicted instances (0.225).
- PVALUE_MLOG: The mean PVALUE_MLOG value is higher for the correctly predicted instances (20.788) compared to the incorrectly predicted instances (12.750). The standard deviation is also higher for the correctly predicted instances (73.338) compared to the incorrectly predicted instances (10.524). The range of PVALUE_MLOG values is larger for the correctly predicted instances.
- OR or BETA: The mean OR or BETA value is slightly lower for the correctly predicted instances (0.413) compared to the incorrectly predicted instances (0.640). The standard deviation is higher for the correctly predicted instances (0.971) compared to the incorrectly predicted instances (0.552).
- CIleft and CIright: The mean and range values for both CIleft and CIright are similar for both groups, indicating similar distributions.

Chapter 4

Conclusion and future work

The approach successfully managed to correctly classify breast cancer cases and controls using genetic data with high accuracy, which is the aim of the study.

The precision and recall of the model also support the hypotesis.

By utilizing genetic data from the NHGRI GWAS project, the method taps into valuable information encoded in the genome, which can provide insights into the genetic factors associated with breast cancer.

4.0.1 Future work

Some shortcomings of the approach is the fact that it only uses a relatively small amount of data (around 3000 entries), which is both a positive aspect, as it managed to become accurate even on the limited data size, but also a negative one, as more data might provide even more insight or findings.

The major shortcoming of the study is that is stops at classifying the two classes of cases and controls, while there is a massive need for understanding precisely which genetic markers are associated with breast cancer, this being the next step in achieving the higher goal of heading towards precision medicine.

4.0.2 SWOT

4.0.2.1 Strengths

- Relevant features The selected features are specifically chosen for their relevance regarding the study of breast cancer from a genetic perspective
- Balanced dataset After downsampling, the dataset consists of a considerable number of both non breast cancer classes and healthy classes, allowing for balanced training and evaluation

- Use of random forest classifier The adoption of a random forest classifier is a notable strength, as this machine learning method is known for its robustness and ability to handle high-dimensional datasets
- High accuracy Achieving an accuracy of 90.96% and a positive class accuracy of 90.9% demonstrates the effectiveness of the classification model in accurately distinguishing between breast cancer and non-breast cancer cases.

4.0.2.2 Weaknesses

While the chosen features are relevant, the dataset may not capture the full spectrum of breast cancer subtypes or genetic variations, limiting the generalizability of the results as there might be other important genetic or clinical variables that are not part of the dataset and could contribute to a more comprehensive analysis and understanding of the underlying factors of breast cancer.

4.0.2.3 Opportunities

Integrating clinical data, expanding the training data with other similar datasets, exploring additional features or deriving new ones from the existing dataset could provide more information and improve the classification accuracy and enhance the model's predictive capabilities. The dataset also offers an opportunity to explore genetic associations with breast cancer, leveraging the genetic variant information.

4.0.2.4 Threats

As the data the model uses is relatively large, more data should be employed in order to ensure the generalizability of the classifier and prove its reliability across different genetic data from multiple individuals and demographic areas.

Bibliography

[1] Breast cancer genome-wide association studies: There is strength in numbers. 2011.

[2] Douglas F. Easton, Karen A Pooley, Alison M. Dunning, Paul D. P. Pharoah, Deborah J. Thompson, Dennis Ballinger, Jeffery P. Struewing, Jonathan Morrison, Helen I. Field, Robert N. Luben, Nicholas J. Wareham, Shahana Ahmed, Catherine S. Healey, Richard Bowman, Kerstin B. Meyer, Christopher A. Haiman, Laurence Kolonel, Brian E. Henderson, Loic Le Marchand, Paul J. Brennan, Suleeporn Sangrajrang, Valerie Gaborieau, Fabrice Odefrey, Chen-Yang Shen, PeiâEi Wu, Hui-Chun Wang, Diana M. Eccles, D. Gareth R. Evans, Julian Peto, Olivia Fletcher, Nichola Johnson, Sheila Seal, Michael R. Stratton, Nazneen Rahman, Georgia Chenevix-Trench, Stig Egil Bojesen, Børge Grønne Nordestgaard, Christen Kirk Axelsson, Montserrat García-Closas, Louise A. Brinton, Stephen J. Chanock, Jolanta Lissowska, Beata PepÅoÅska, Heli Nevanlinna, Rainer Fagerholm, Hannaleena Eerola, Daehee Kang, Keun-Young Yoo, Dong-Young Noh, Sei Hyun Ahn, David J. Hunter, Susan E Hankinson, David G. Cox, Per Hall, S Wedrén, Jianjun Liu, Yen Ling Low, Natalia V Bogdanova, Peter SchuÌrmann, Thilo Dörk, Rob Aem Tollenaar, Catharina E. Jacobi, Peter Devilee, Jan G. M. Klijn, Alice J. Sigurdson, Michele Morin Doody, Bruce H. Alexander, Jinghui Zhang, Angela Cox, Ian Wallace Brock, Gordon R. Macpherson, Malcolm WR. Reed, Fergus J. Couch, Ellen L. Goode, Janet E. Olson, Hanne E J Meijers-Heijboer, Ans van den Ouweland, André G. Uitterlinden, Fernando Rivadeneira, Roger L. Milne, Gloria Ribas, Anna González-Neira, Javier Benítez, John L. Hopper, Margaret Mccredie, Melissa C. Southey, Graham G. Giles, Christopher J Schroen, Christina Justenhoven, Hiltrud B Brauch, Ute Hamann, Yon D. Ko, Amanda B. Spurdle, Jonathan Beesley, Xiaoqing Chen, Arto Mannermaa, VâM. Kosma, Vesa V. Kataja, Jaana M. Hartikainen, Nicholas E. Day, David Cox, and Bruce A. J. Ponder. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, 447:1087–1093, 2007.

[3] Ryan K. Van Laar. Design and multiseries validation of a web-based gene expression assay for predicting breast cancer recurrence and patient survival. *The journal of molecular diagnostics*,

2011.

[4] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[5] M. J. van de Vijver. The role of gene expression profiling by microarray analysis for prognostic classification of breast cancer. *Breast Cancer Research*, 7(1):S1, May 2005.