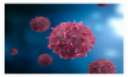
Leveraging genetic data for breast cancer recognition

By Salsigan Razvan-Dan

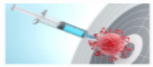
INTELLIGENT CLASSIFICATION OF BREAST CANCER USING GENETIC DATA

BACKGROUND



Breast cancer is a complex disease characterized by abnormal cell growth in breast tissue.
It is the most common cancer type and there is a need for accurate diagnosis and effective targeted treatment

IMPORTANCE



The study can provide a basis for developing targeted medicine which is crucial in personalised therapy and excludes the risk of impacting healthy ceells, which current methods do

OBJECTIVE

The objective of the study is to leverage genetic data and develop an intelligent method for distinguishing breast cancer patients from healthy individuals

DATA SOURCE

GWAS Catalog

The NHGRI-EBI Catalog of published genome-wide association studies

- LOGISTIC REGRESSION
- SGD
- DECISION TREE
- RANDOM FOREST

RESULTS

Random forrest classifier achieved an accuracy of 90%

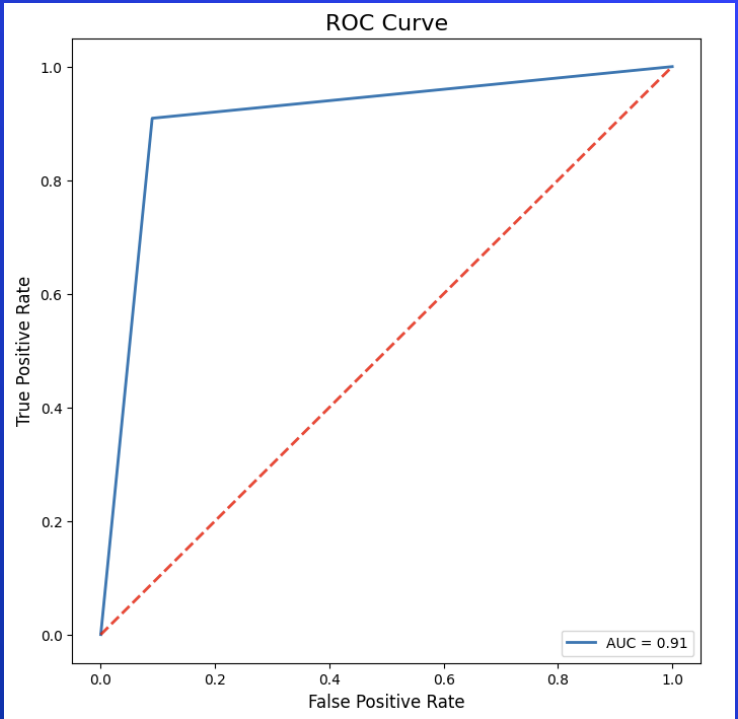
Dataset

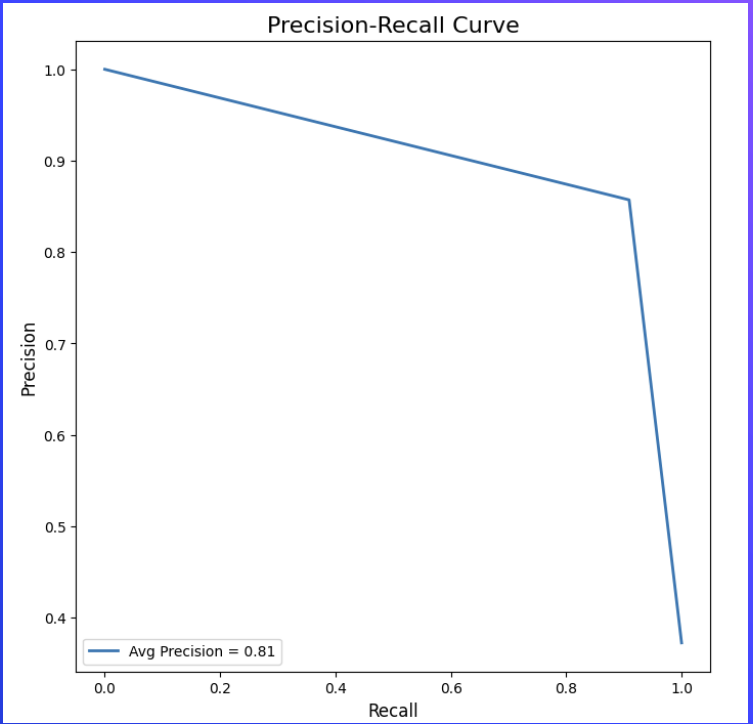
- The data source is from the GWAS study from NHGRI
- Features: 'Chromosome', 'Position', 'Arm', 'CHR_ID', 'CHR_POS', 'SNPS', 'RISK ALLELE FREQUENCY', 'PVALUE_MLOG', 'OR or BETA', 'CI'
- The target variable is extracted from the disease/trait column, which is a textual description of the disease or trait that specific individual has

Methodology

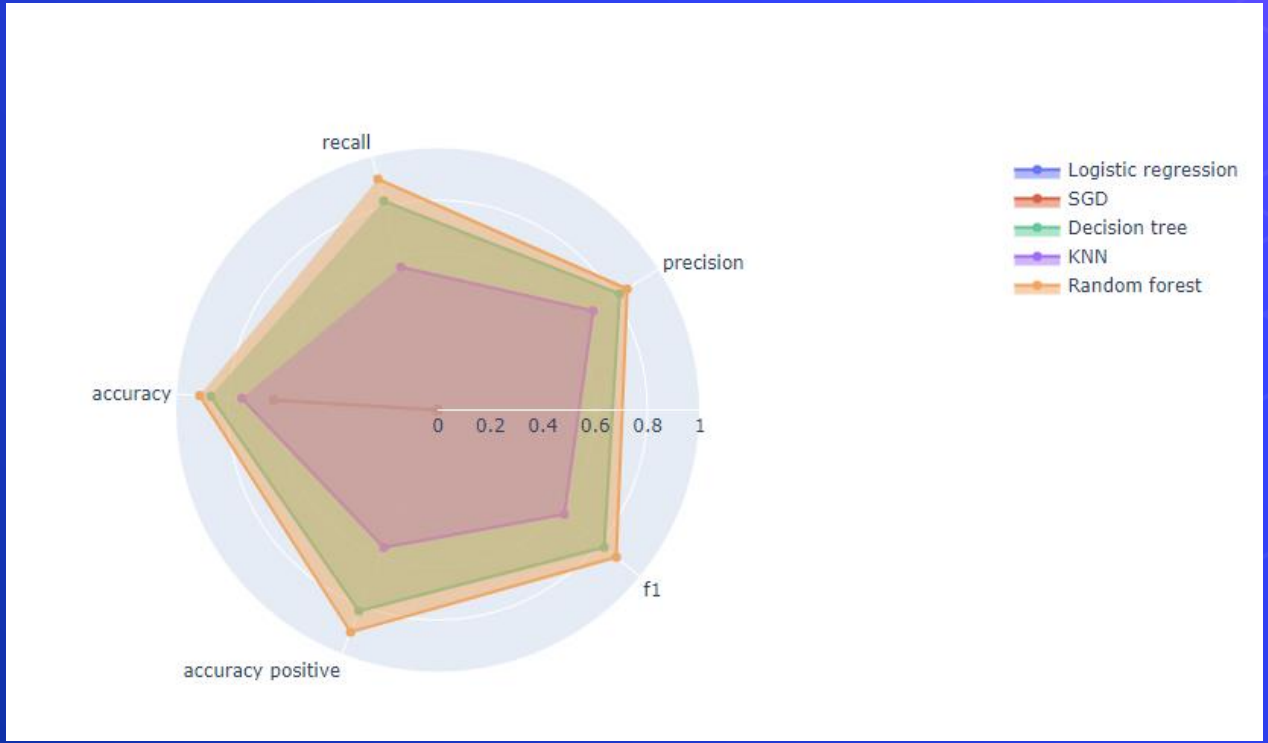
- The classification algorithms trained and tested on the dataset are logistic regression, SGD, decision tree, KNN and random forest
- The original dataset was downsampled, as there were only around 1200 breast cancer cases
- Dropped unimportant columns, split intervals into two columns, split region into chromosome, arm and position, label encoded string values, removed outliers
- Training (20%) and testing (80%)

Performance metrics





Results


recall
precision
accuracy
f1
accuracy positive
0
0.2
0.4
0.6
0.8
1
Logistic regression
SGD
Decision tree
KNN
Random forest

Discussion

- The study successfully classified breast cancer cases with high accuracy, supporting the potential of genetic data in accurate classification
- The random forest algorithm played a significant role in achieving the high accuracy by considering multiple features simultaneously
- The presence of relevant and informative genetic markers likely contributed to the algorithm's ability to discriminate between breast cancer and non-cancer cases

Future work

- Future work should focus on expanding the dataset and integrating clinical data to enhance classification accuracy
- Exploring additional features and advanced techniques, such as deep learning, could further improve the model's performance
- Identifying specific genetic markers associated with breast cancer would contribute to precision medicine and personalized treatment strategies

Conclusion

- 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

References

- https://iris.unipa.it/retrieve/handle/10447/101841/374977/Breast\%20cancer\%20genome-wide\%20association\%20studies\%20There\%20is\%20strength\%20in\%20numbers.pdf
- https://scikit-learn.org/stable/
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4231885/

THANK YOU!