Machine Learning for Breast Cancer Risk Prediction in Indian Patients

Yukti Makhija
2019BB10067

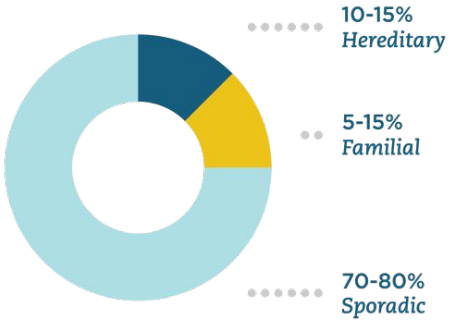
Samarth Bhatia
2019CH10124

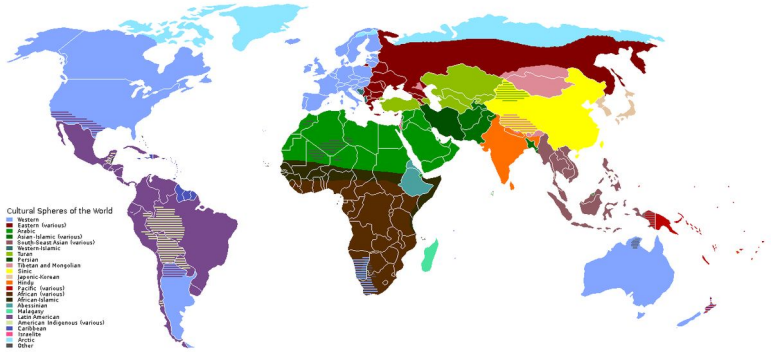
Prof. Ishaan Gupta, IIT Delhi
Medical Oncology Department, AIIMS

Summer Undergraduate Research Award 2021

Brief Overview of Breast Cancer

- About 5 to 10 percent are linked to gene mutations passed through generations. This makes the detection of **mutations** in genes like **BRCA1 and 2, PALB2, CHEK2 and ATM1** essential for breast cancer patients as well as their previvors to prevent cancer.
- The **prevalence** of these mutations varies with **geography** and **ethnicity**.
- If breast cancer is diagnosed or **predicted early**, those cases tend to have significantly **better prognosis** and a better chance of **recovery**.





Risk Prediction in Breast Cancer

- Recent advances in **personalized medicine** and **genetic testing** assist oncologists in **deciding the course of treatment** for diagnosed patients and **predicting the risk** of cancer in family members.
- Currently in labs, **National Comprehensive Cancer Network (NCCN)** guidelines are being used to advise **genetic testing**, but they have been shown by multiple studies to be **ineffective** and **miss out on many potential high risk patients** with mutations.
- **In India, we neither have reliable estimates about the prevalence of these mutations, nor any reliable fixed criteria (which needs to be adapted to India) regarding who is to be tested and who is not.**
- **Developing countries** use different ways to diagnose cancer which take **fewer resources**, preventing us from making direct comparisons between the indicators of breast cancer in developed countries and those in developing countries.

Existing Models: BOADICEA and BCRAT

- Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm **(BOADICEA)** calculate the risks of breast and ovarian cancer in women by computing the **probability of a patient being a carrier of mutations** in the susceptible ovarian and breast cancer genes and the age-specific risk of cancer.[5] It takes into account both **pathology data** and **population-specific cancer incidence history**.
- Breast Cancer Risk Assessment Tool **(BCRAT)** uses a woman's **personal medical** and **reproductive history** and the history of breast cancer among her **first-degree relatives** to estimate absolute breast cancer risk, the probability of developing invasive breast cancer in a defined age interval.

Immediate Gap in BOADICEA and BCRAT

- These models cater to the **western population**.
- Both the models assume a **linear relationship** between the risk factors and the risk of cancer development. This might not be true in general and contributes largely to the **poor accuracy** of these mathematical models compared to machine learning models.
- A study done on the **Thai population** shows that the BCRAT model **underestimates** the breast cancer risk. So, there is an urgent need for a model that can accurately predict the risk of breast cancer for Indian patients.

Model	BOADICEA	BCRAT	Machine Learning (vs BOADICEA)	Machine Learning (vs BCRAT)
Number of people	112,587(Swiss)	850 + 293 (first relative)	2500 (simulated) + 112,587 (Swiss)	1200 (simulated) + 1143 (US)
Number of families	2481	-	2481	-
Number of features	13	8	13	8
Pedigree Analysis	Yes	No	Yes	No
Simplicity of the model	Ⅱ	Ⅰ	Ⅳ	Ⅲ

Challenge : **Very less data** about breast cancer patients is available **for India**.

Objectives

- Collection of EHRs and data engineering.
- Develop machine learning based risk prediction model.
- Feature engineering and enhance model interpretability towards clinical deployment.

Data Curation

- We collected data for **236 breast cancer patients** in collaboration with the Oncology Department at AIIMS.
- The initial dataset contained 115 features for each patients.
- We categorized our data in five broad groups
  - ➔ **Clinical Data**
  - ➔ **Reproductive History**
  - ➔ **Psychological Factors**
  - ➔ **Family History**
  - ➔ **Genetic Mutations**

Clinical Data

- Age
- Weight
- Height
- T-stage, N-stage, M-stage
- TNBC
- Estrogen/Progesterone Levels
- Clinical Stage of cancer
- Two Breast Primaries
- RRSO Surgery Advised/Done
- RRM Surgery Advised/Done
- Type of cancer - Metachronous/Synchronous
- History of Hysterectomy

Reproductive History

- Age at Menarche
- Number of Children
- Gender of Children
- Duration of Breastfeeding
- Age at first Childbirth
- Usage of Oral Contraceptive Pill (OCP)
- Tubal Ligation

Psychological Factors

- Occupation
- Marital Status
- Any PTSD
- Distress levels
- Anxiety levels
- Alcohol use

The following scores were calculated on the basis of the stress/anxiety and depression levels and PTSD.

➔ **DASS-21(Depression, Anxiety and Stress Scale) score**

➔ **IES-R(Impact of Event Scale) score**

Family History

- Ethnicity
- Number of first degree relatives
- Number of affected first degree relatives
- Number of Family Members Tested
- Number of second degree relatives
- Number of affected second degree relatives

Genetic Mutations

Genetic Mutations can be categorized into 2 categories:

1. **Pathogenic Gene Mutations (~20%):** There is plenty of literature available which validates the role played by these genes in Breast Cancer.
   BRCA1, BRCA2, PALB2, RAD51d, RAD50, FANCI, ATM, TP53, MSH2, MUTYH
2. **VUS (Variants of Unknown Significance) (~50%):** Not much is known about these mutations and it is not known if they affect the risk of breast cancer either.

Data Preprocessing

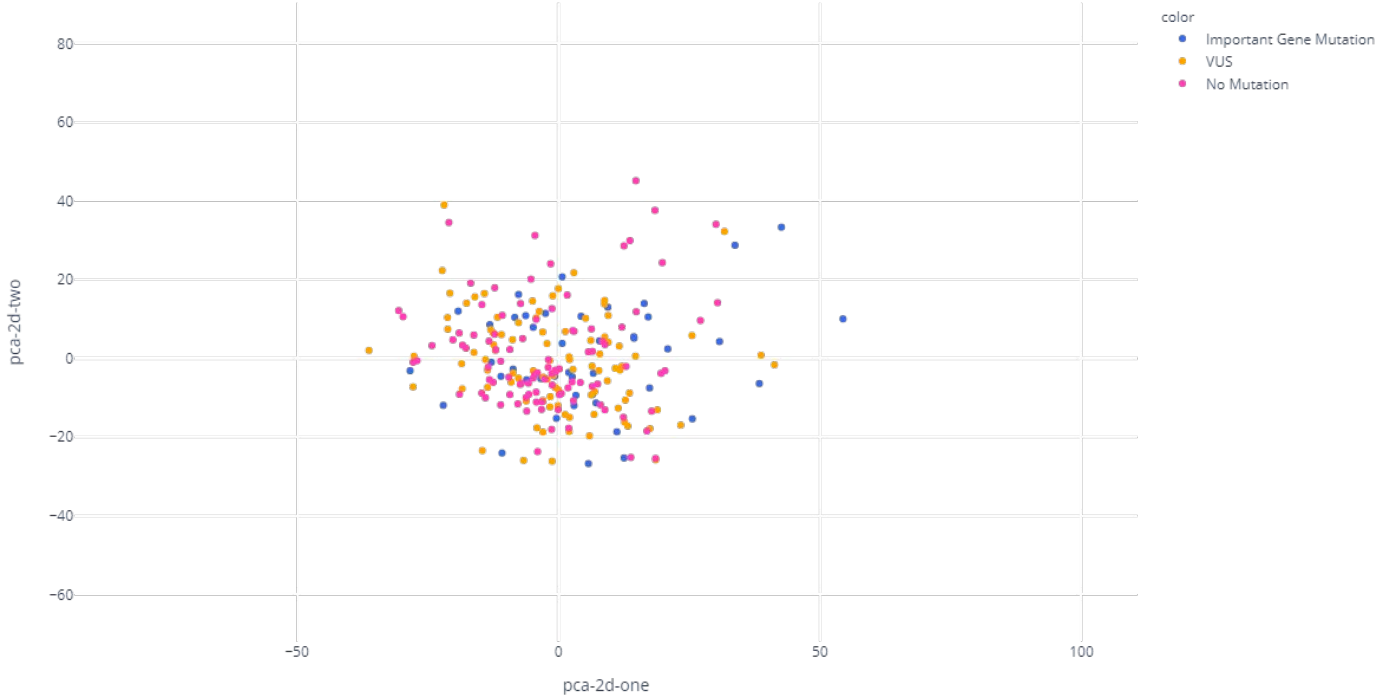
- All columns with more than 10% missing values were deleted.
- We imputed the missing values with mode for categorical variables.
- The data obtained after preprocessing contains **236 patients and 44 features**. This dataset is in a systematic form and has been used for further analysis and machine learning.

Dimensionality Reduction and Visualisation

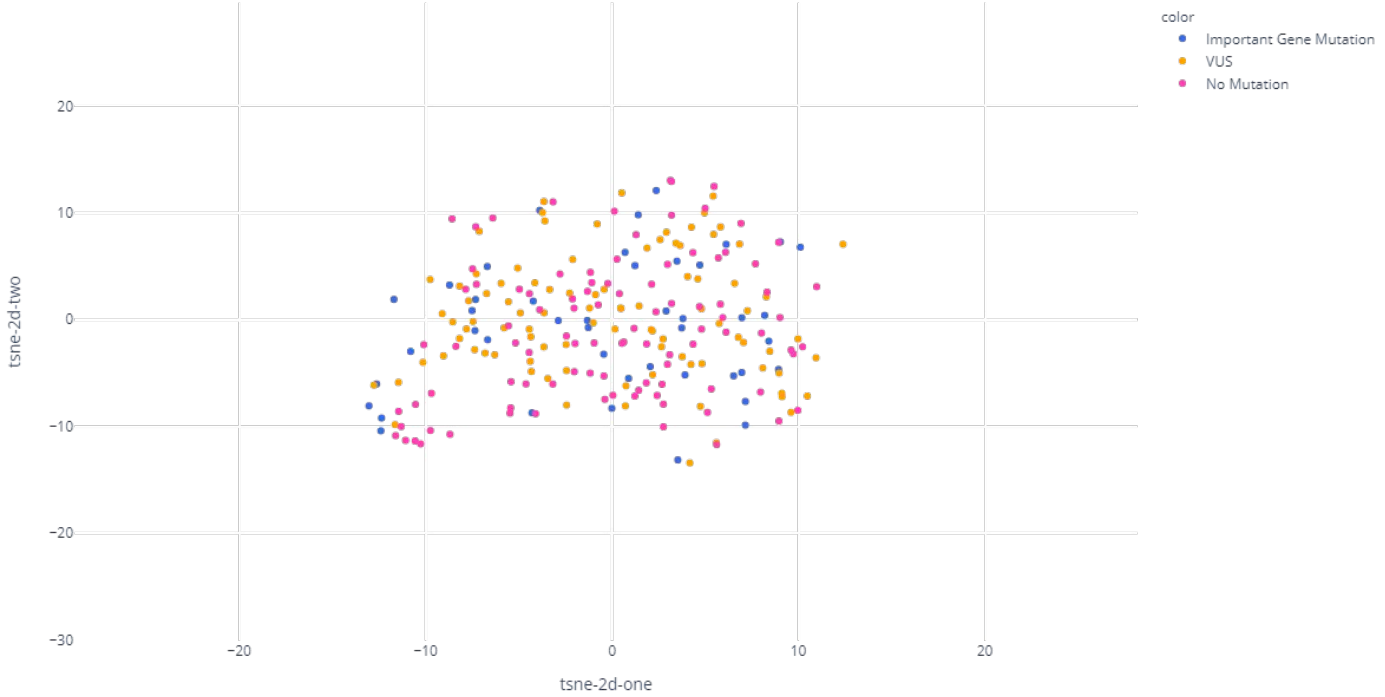
We used various state-of-the-art algorithms for this purpose and also compared their results.

- PCA (Principal Component Analysis)
- tSNE (t-Distributed Stochastic Neighbor Embedding)
- LDA (Linear Discriminant Analysis)
- NCA (Neighborhood Component Analysis)
- UMAP (Uniform Manifold Approximation and Projection)

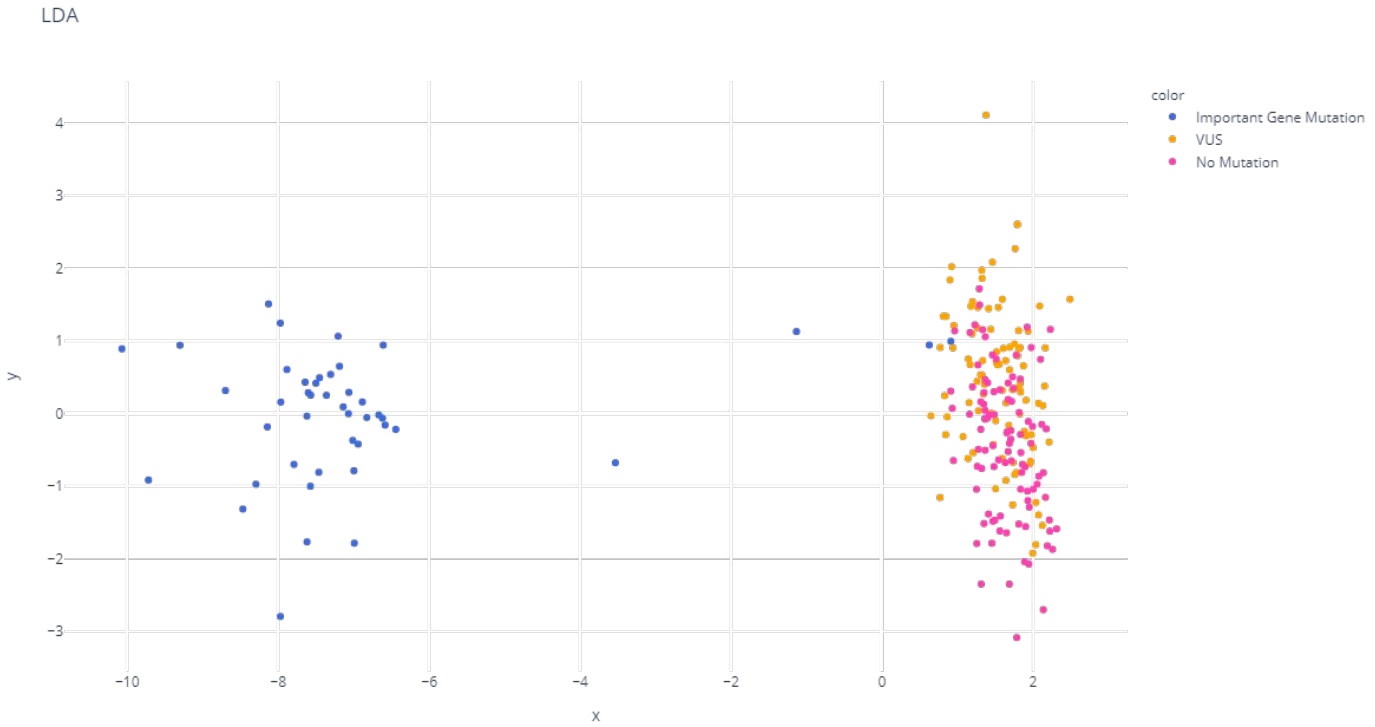
Principal Component Analysis (PCA)



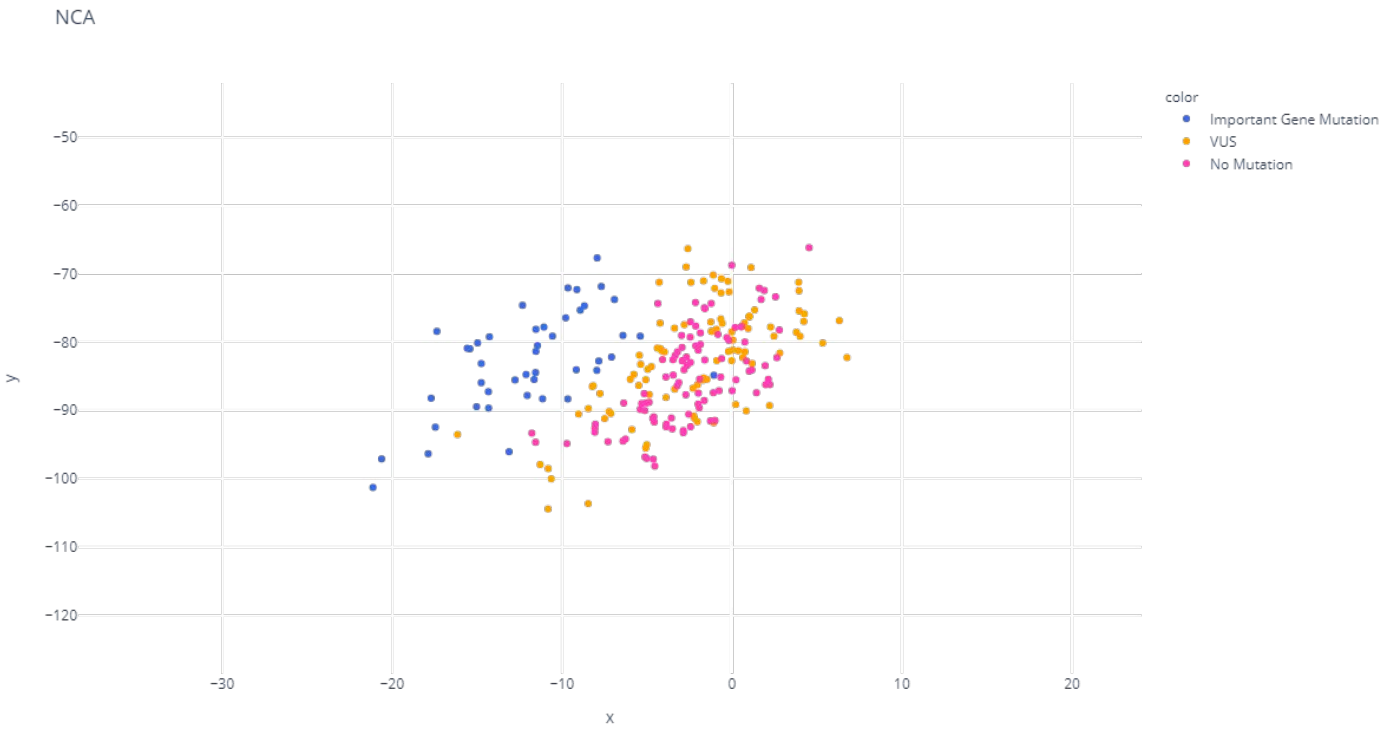
t-Distributed Stochastic Neighbor Embedding (t-SNE)



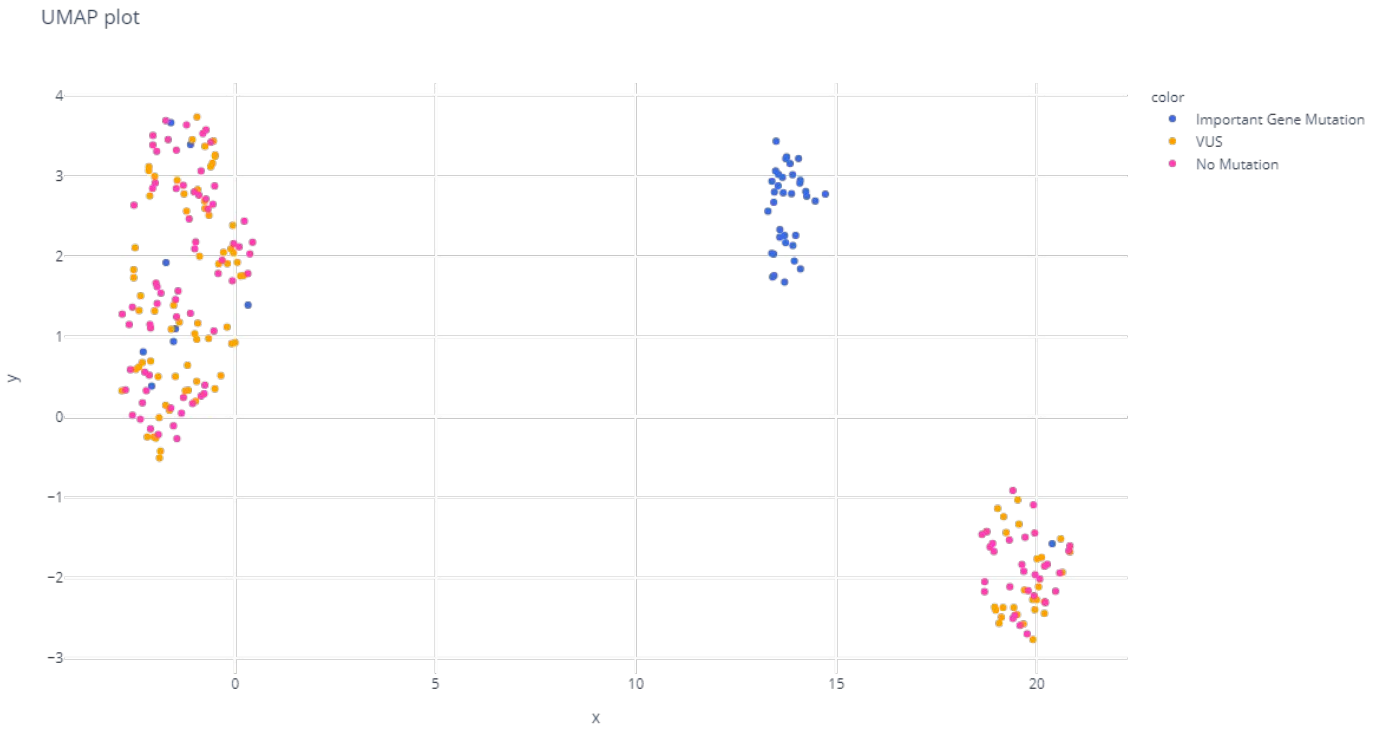
Linear Discriminant Analysis (LDA)



Neighborhood Component Analysis (NCA)

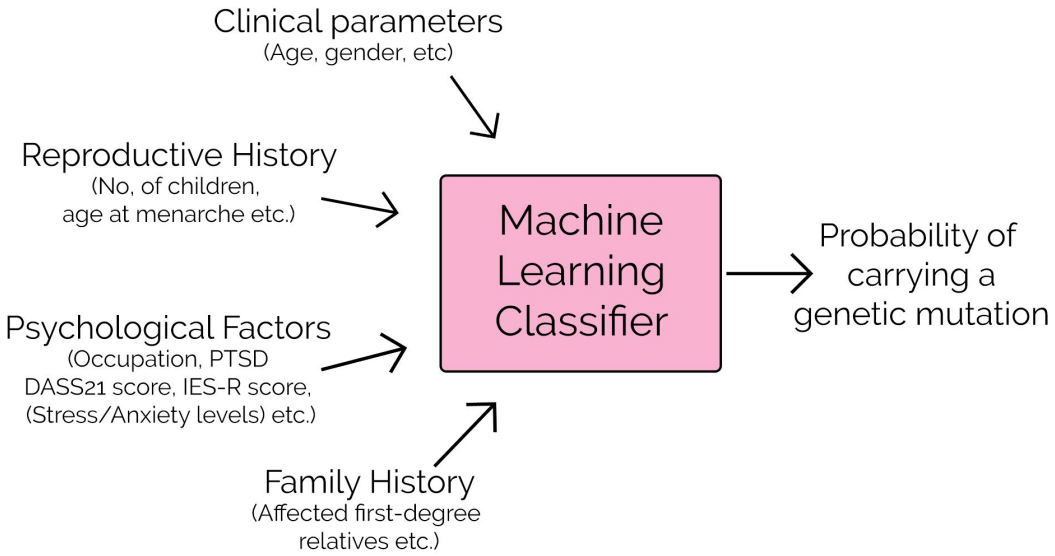


Uniform Manifold Approximation and Projection (UMAP)



Development of the models

We developed prediction models on the curated data using frameworks such as XGBoost, AdaBoost, LightGBM, SVMs (Support Vector Machines) etc.



Effective Disease Management

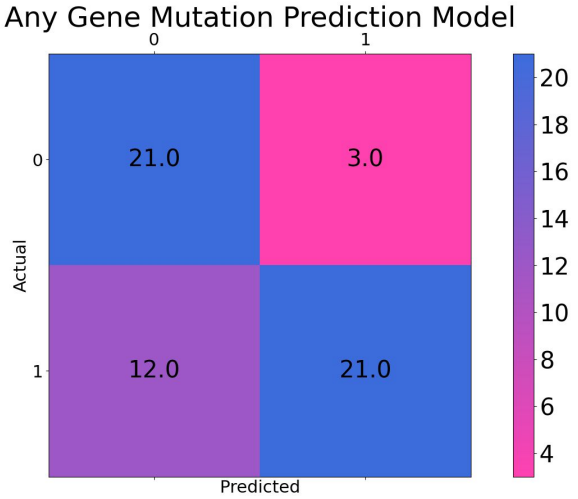
Gene Mutation Prediction Models

We have built three different gene mutation prediction models:

- Important gene mutation prediction
- VUS (Variants of Unknown Significance) Prediction
- Any gene mutation prediction (Important gene + VUS)

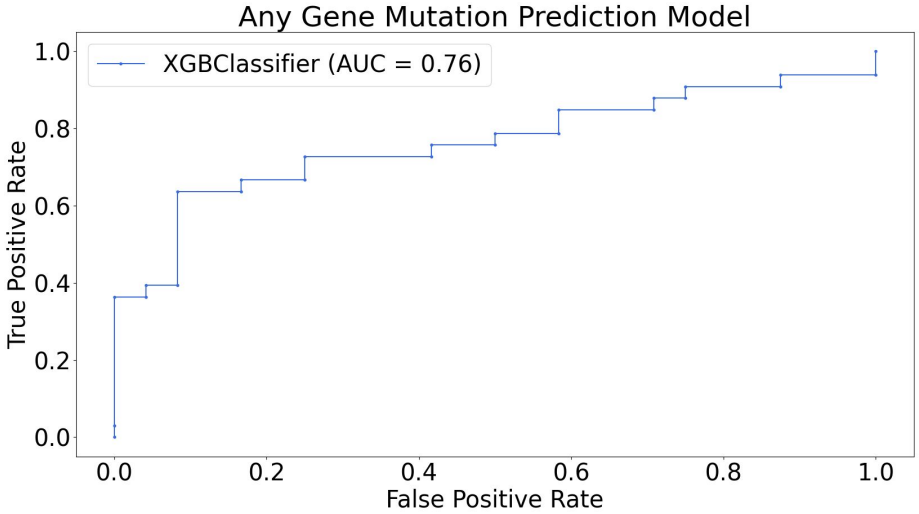
Any Gene Mutation Prediction Model (XGBoost)

Training Dataset	179 patients (105 have mutations + 74 don't have any mutations)
Validation Dataset	57 patients (33 have mutations + 24 don't have any mutations)
Training Accuracy	75.42%
Test Accuracy	**73.68%**
F-score	**0.7368**
Sensitivity (Recall)	0.6364
Specificity	0.875
PPV (Precision)	0.875
NPV	0.875
AUC-ROC	**0.76**
AUC-PR	0.85

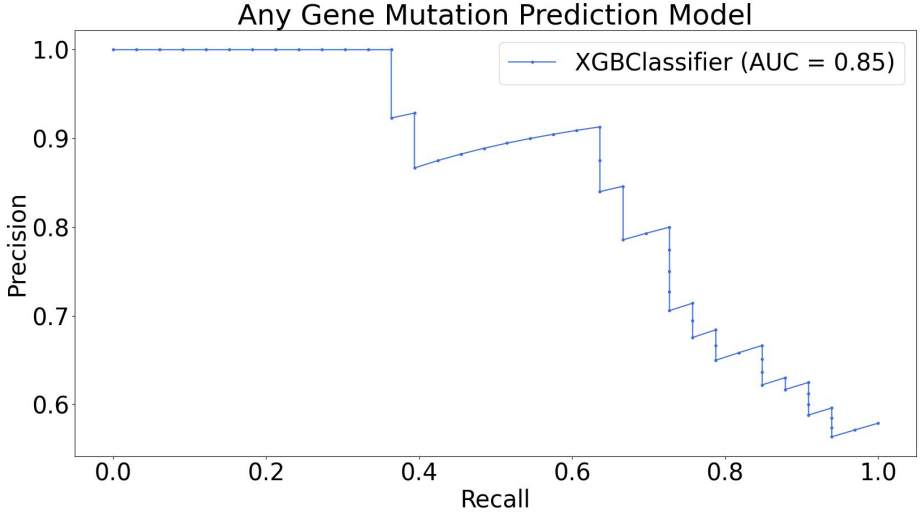


Confusion Matrix

ROC Curve

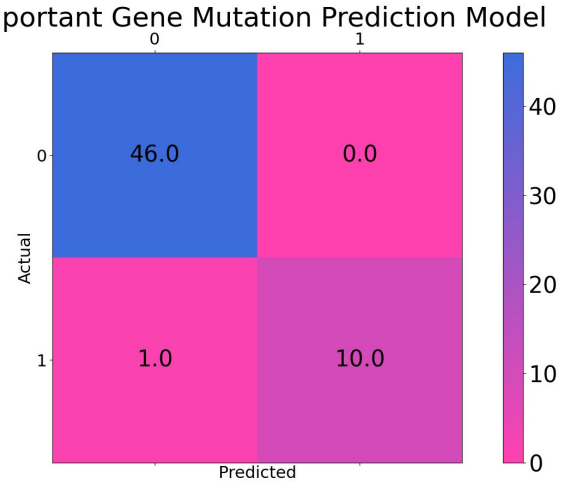


Precision-Recall Curve



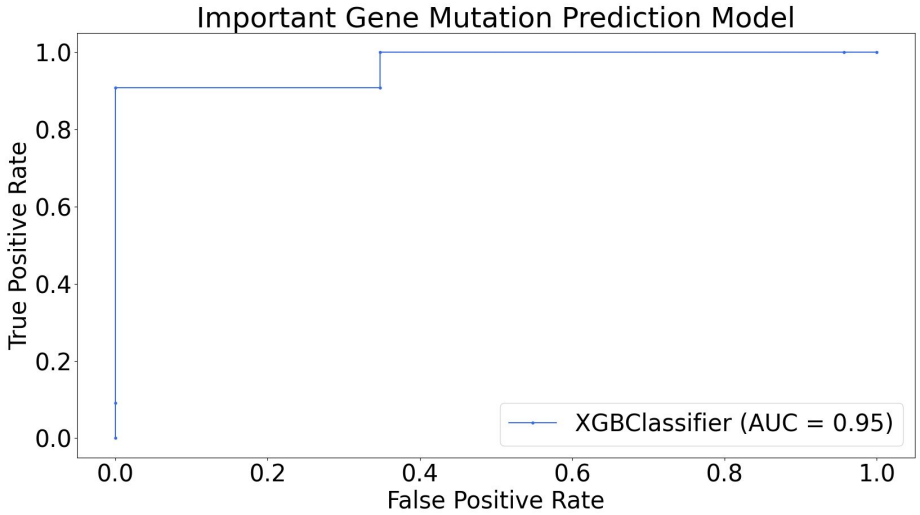
Important Gene Mutation Prediction Model (XGBoost)

Training Dataset	179 patients (33 have mutations + 146 don't have any mutations)
Validation Dataset	57 patients (11 have mutations + 46 don't have any mutations)
Training Accuracy	100%
Test Accuracy	**98.25%**
F-score	**0.95**
Sensitivity (Recall)	0.91
Specificity	1.0
PPV (Precision)	1.0
NPV	1.0
AUC-ROC	**0.95**
AUC-PR	0.96

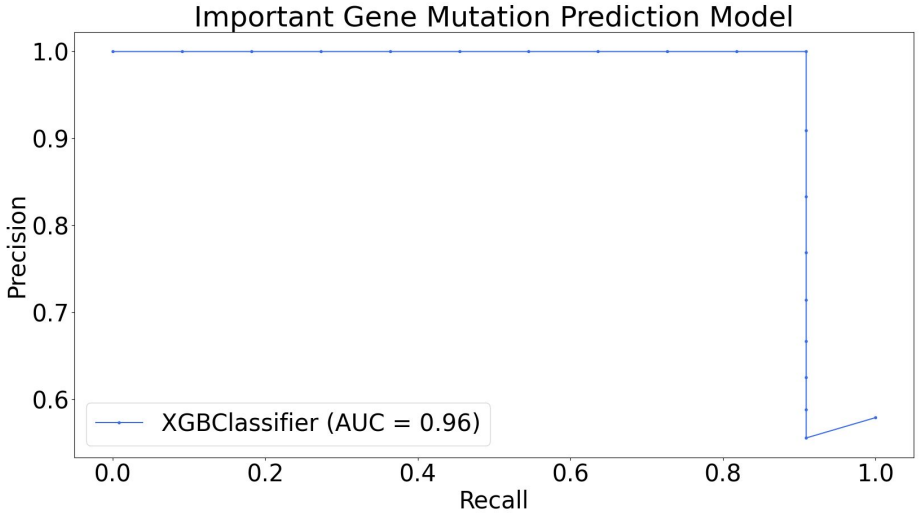


Confusion Matrix

ROC Curve

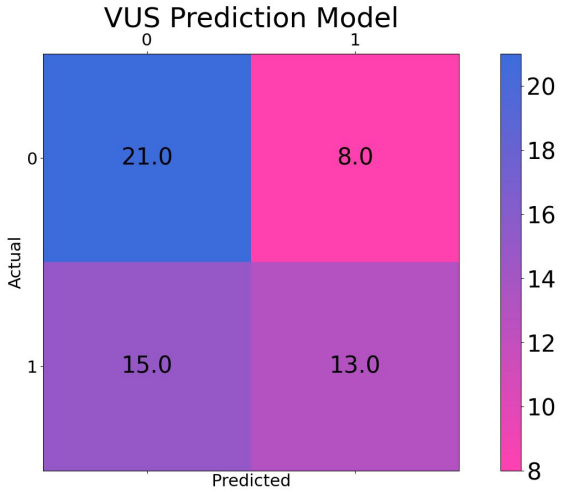


Precision-Recall Curve



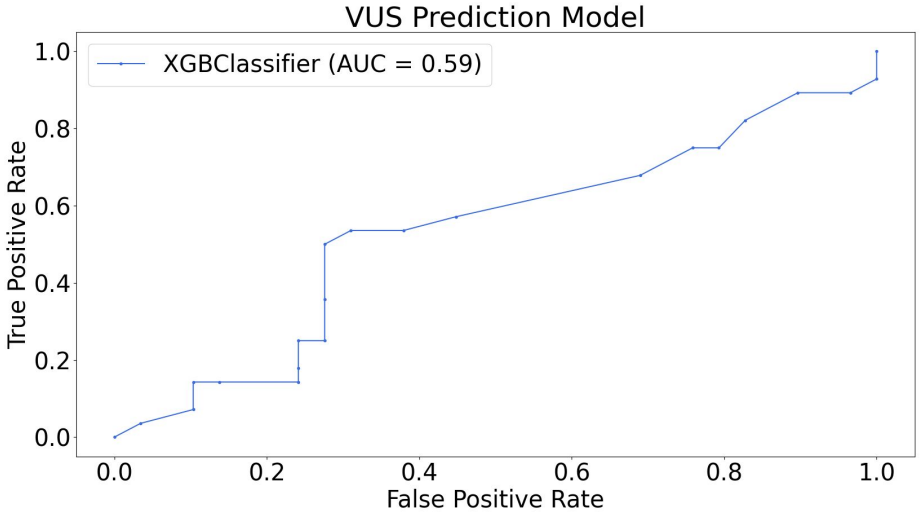
VUS Prediction Model (XGBoost)

Training Dataset	179 patients (87 have mutations + 92 don't have any mutations)
Validation Dataset	57 patients (28 have mutations + 29 don't have any mutations)
Training Accuracy	66.48%
Test Accuracy	**59.65%**
F-score	**0.5306**
Sensitivity (Recall)	0.4643
Specificity	0.7241
PPV (Precision)	0.619
NPV	0.5833
AUC-ROC	**0.59**
AUC-PR	0.52

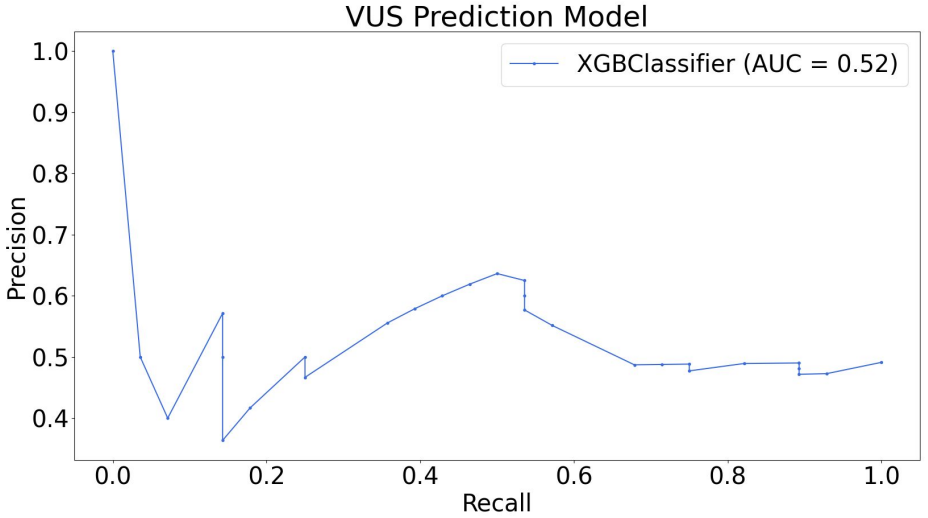


Confusion Matrix

ROC Curve



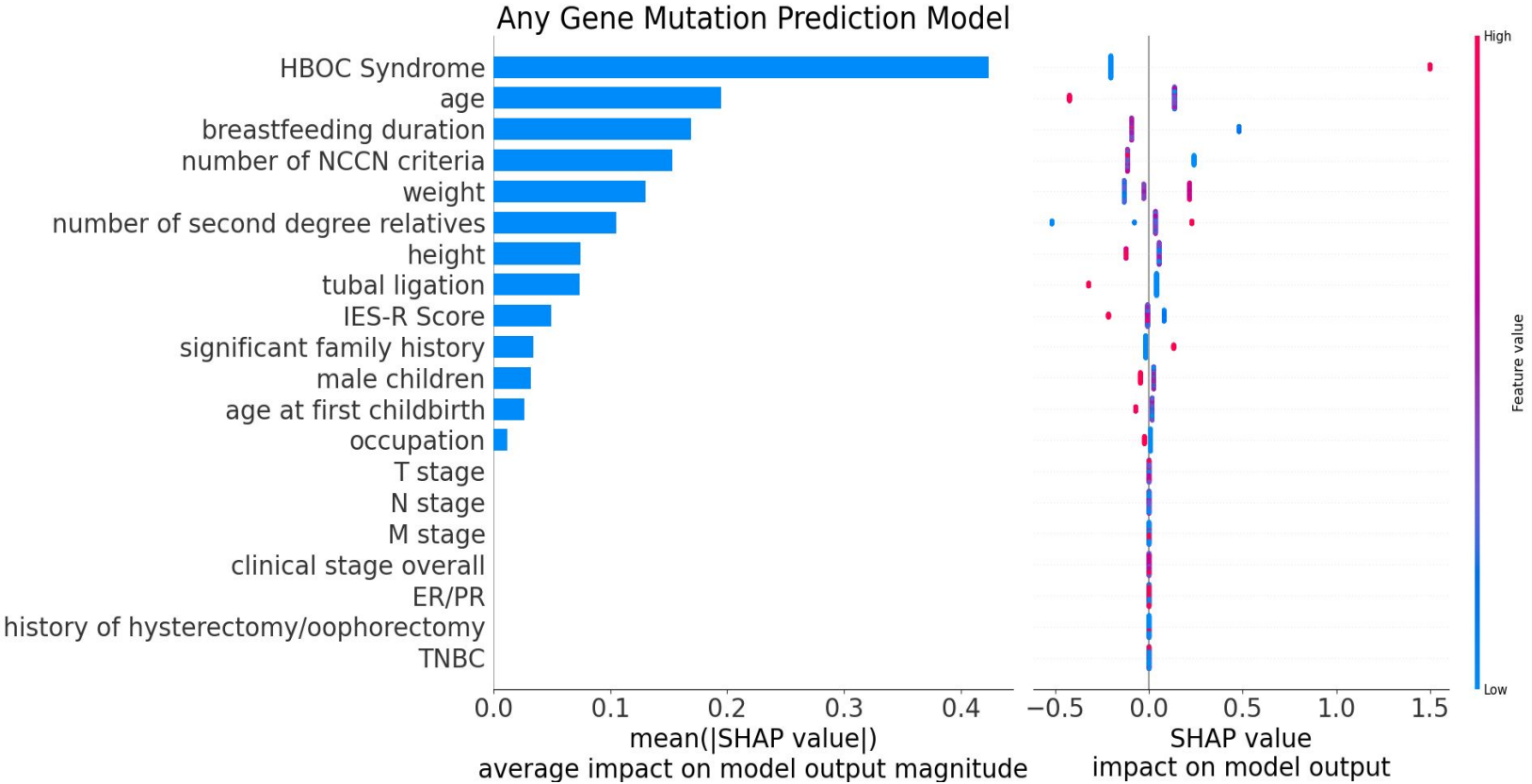
Precision-Recall Curve



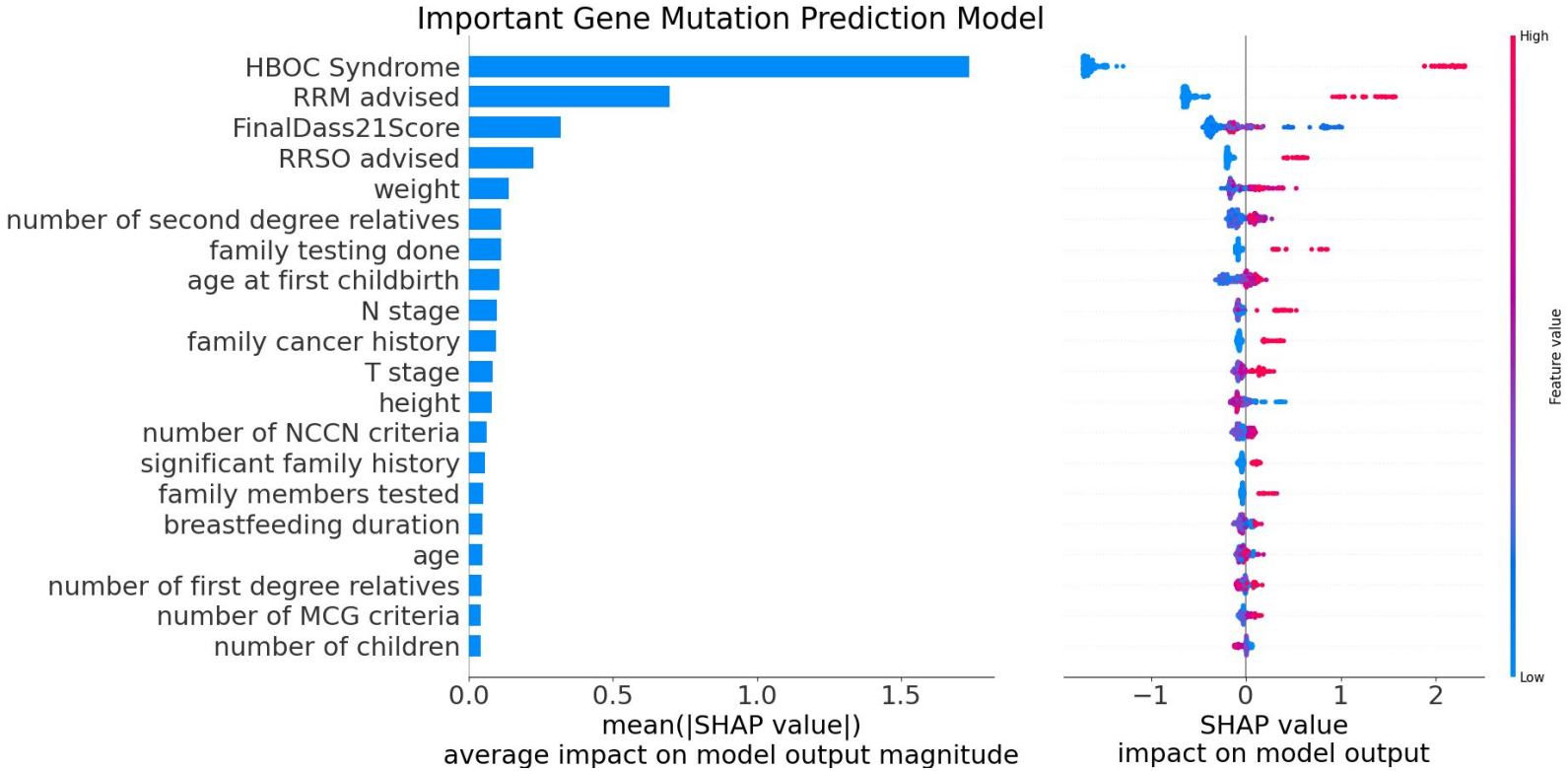
Feature Engineering

- We performed feature engineering and use techniques like **SHAP** to explain the working of ML models.
- The **top ten features** obtained after feature engineering were used to build **reduced models** and compare their performance to the original model.
- We also integrated the models with an interactive calculator in a web-app.

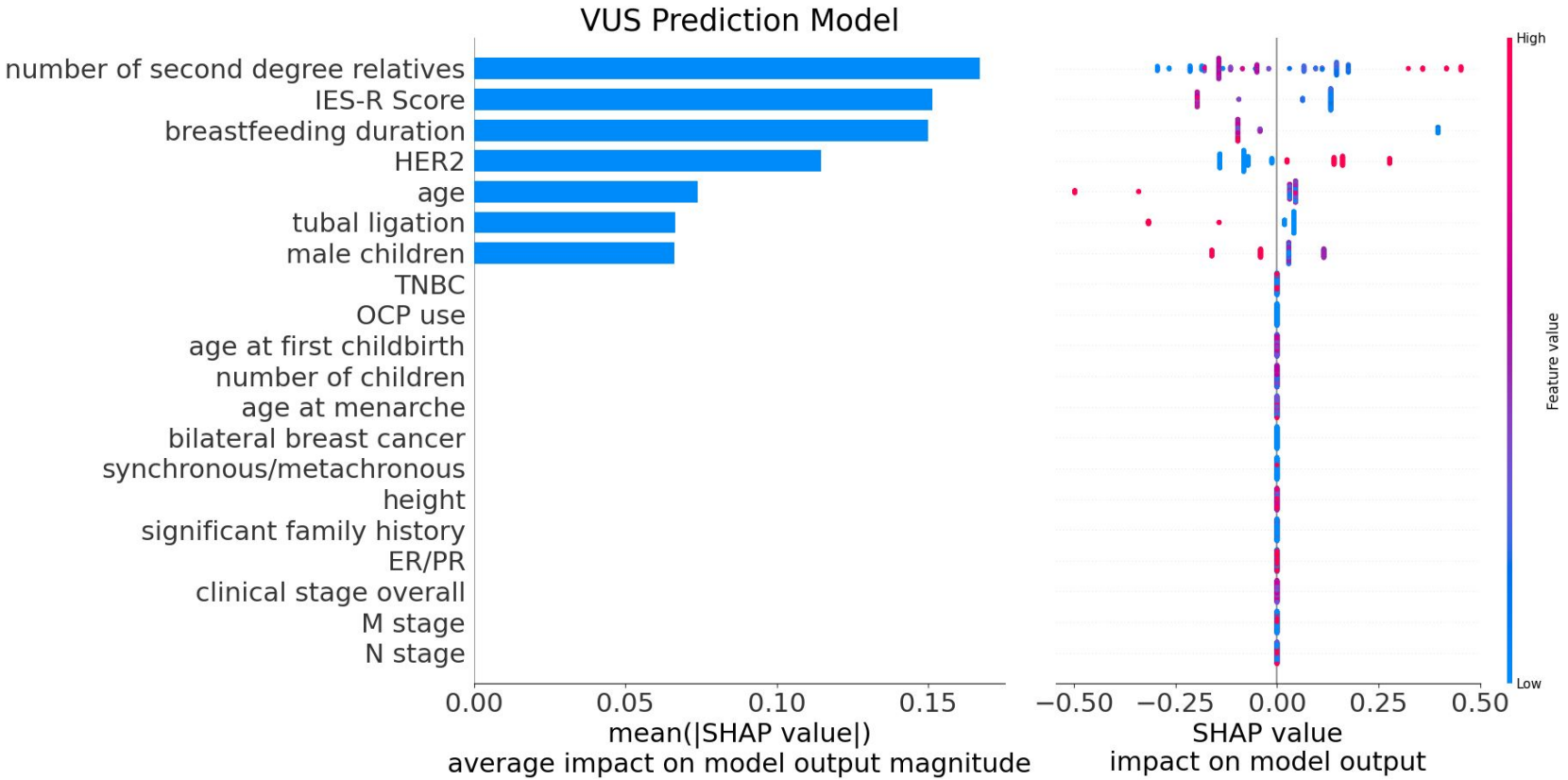
SHAP Feature Importance



SHAP Feature Importance

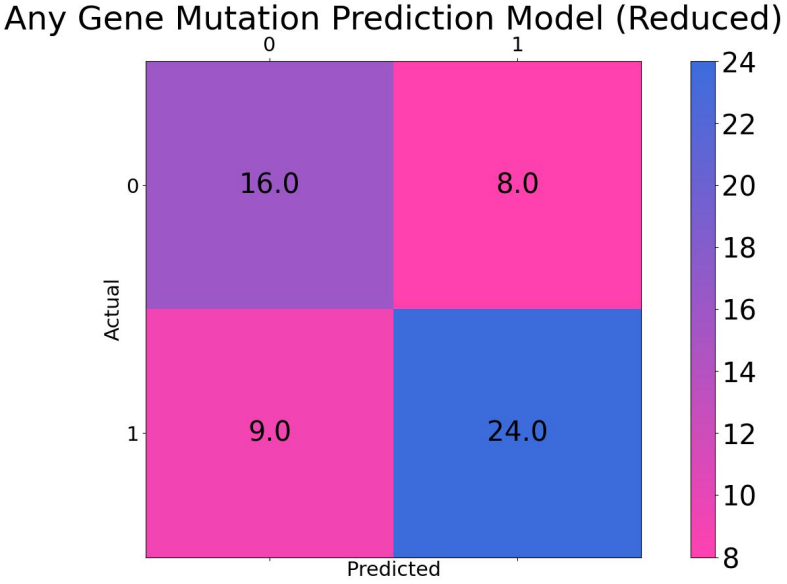


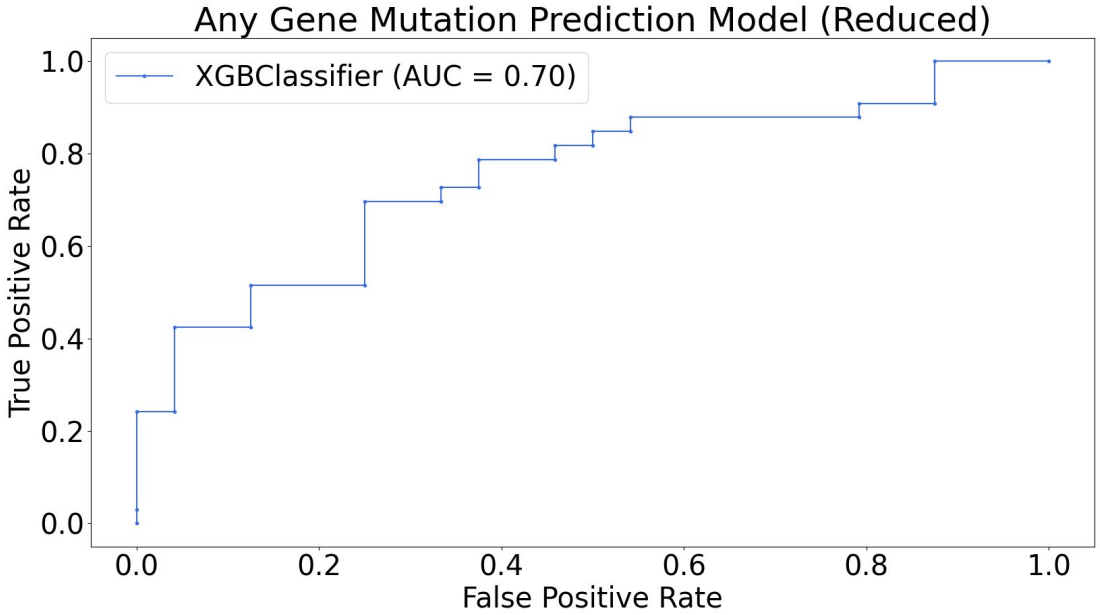
SHAP Feature Importance



Any Gene Prediction Reduced Model

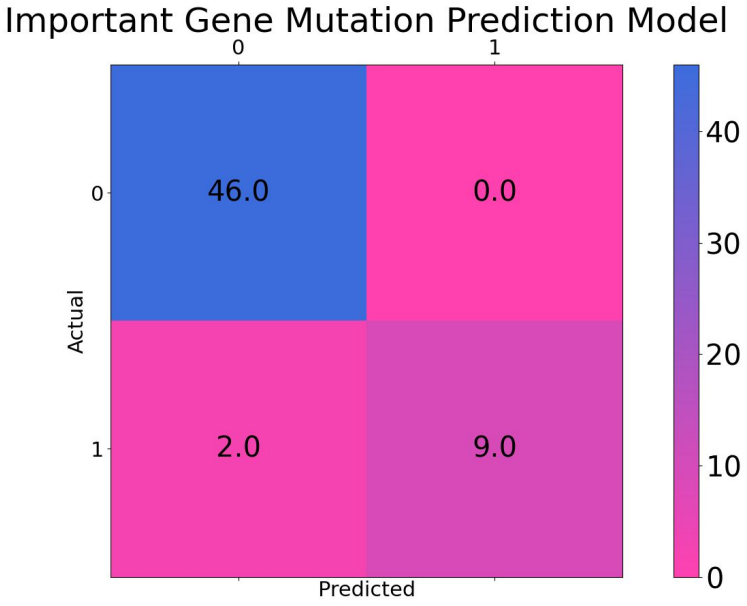
Training Dataset	179 patients (105 have mutations + 74 don't have any mutations)
Validation Dataset	57 patients (33 have mutations + 24 don't have any mutations)
Training Accuracy	78.21%
Test Accuracy	**70.18%**
F-score	**0.7384**
Sensitivity (Recall)	0.7273
Specificity	0.6667
PPV (Precision)	0.75
NPV	0.64
AUC-ROC	**0.70**

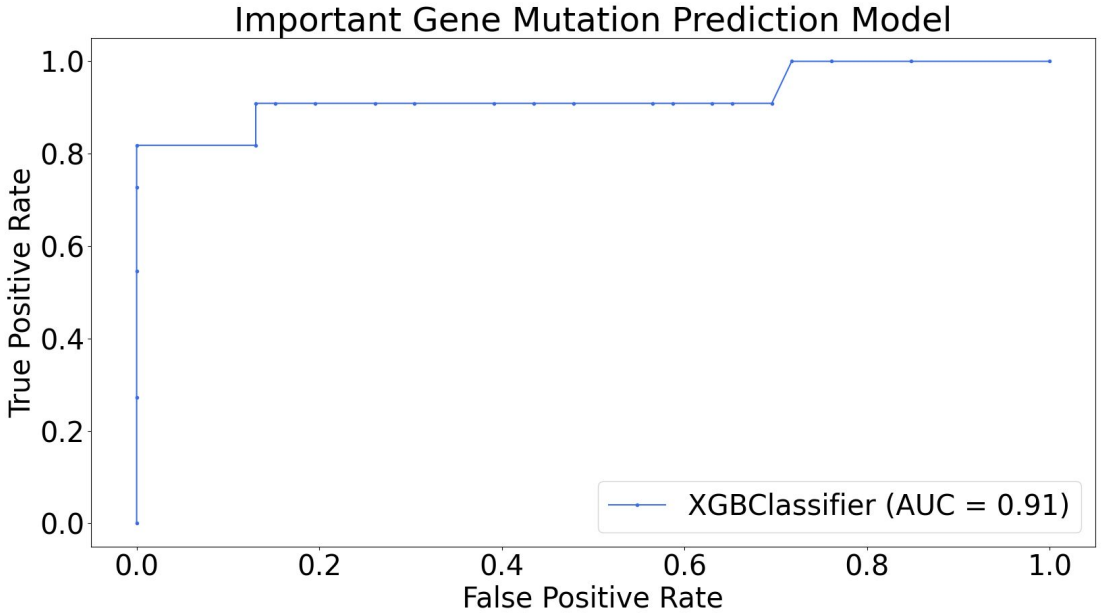

Any Gene Mutation Prediction Model (Reduced)
0
1
0
1
16.0
8.0
9.0
24.0
Actual
Predicted
24
22
20
18
16
14
12
10
8


Any Gene Mutation Prediction Model (Reduced)
XGBClassifier (AUC = 0.70)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

Important Gene Mutation Prediction Reduced Model

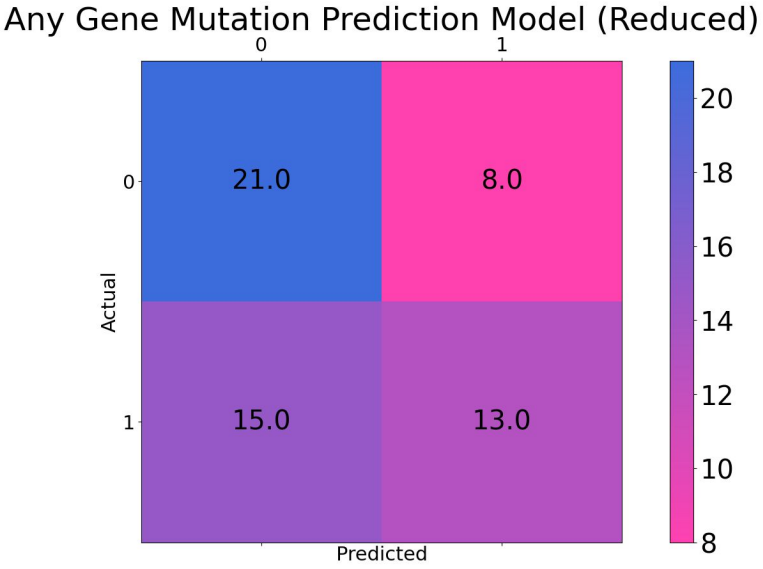
Training Dataset	179 patients (33 have mutations + 146 don't have any mutations)
Validation Dataset	57 patients (11 have mutations + 46 don't have any mutations)
Training Accuracy	100%
Test Accuracy	**96.49%**
F-score	**0.9**
Sensitivity (Recall)	0.8182
Specificity	1.0
PPV (Precision)	1.0
NPV	0.9583
AUC-ROC	**0.91**

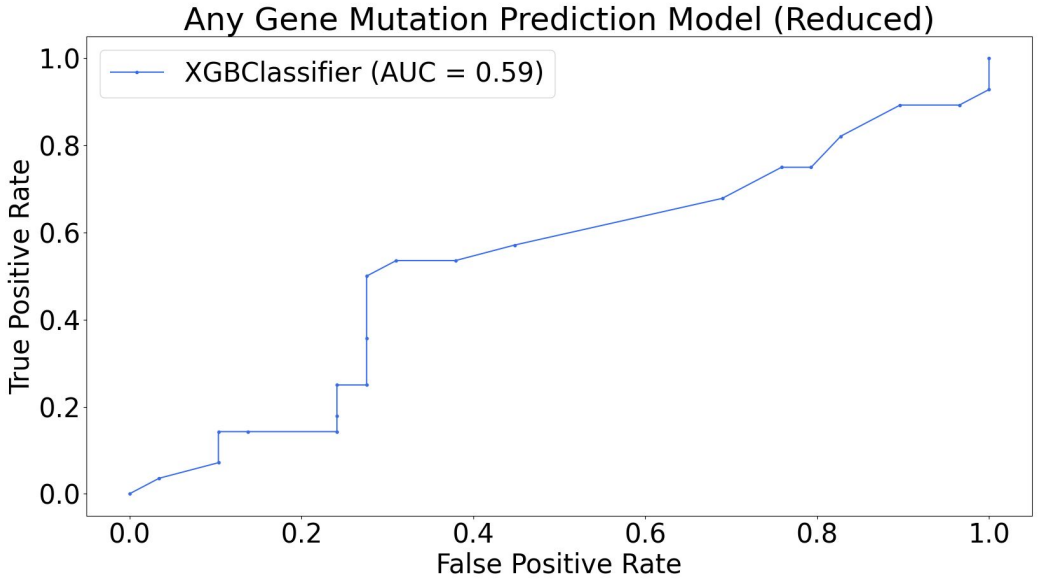

Important Gene Mutation Prediction Model
0
1
0
1
46.0
0.0
2.0
9.0
Actual
Predicted
40
30
20
10
0


Important Gene Mutation Prediction Model
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
XGBClassifier (AUC = 0.91)

VUS Prediction Reduced Model

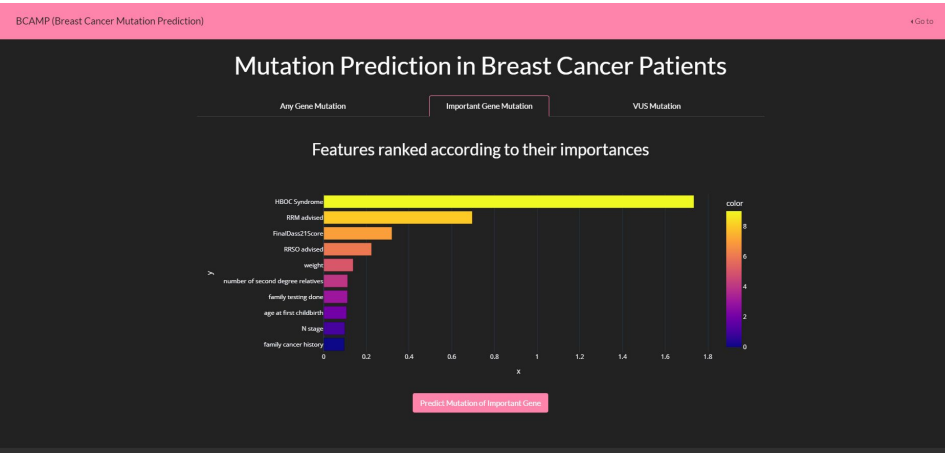
Training Dataset	179 patients (87 have mutations + 92 don't have any mutations)
Validation Dataset	57 patients (28 have mutations + 29 don't have any mutations)
Training Accuracy	66.48%
Test Accuracy	**59.65%**
F-score	**0.5306**
Sensitivity (Recall)	0.4643
Specificity	0.7241
PPV (Precision)	0.619
NPV	0.5833
AUC-ROC	**0.59**

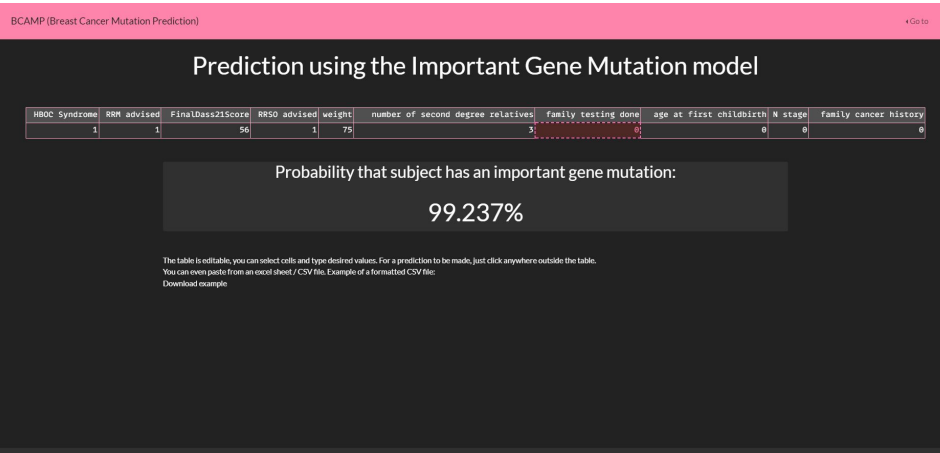

Any Gene Mutation Prediction Model (Reduced)
0
1
0
1
21.0
8.0
15.0
13.0
Actual
Predicted
20
18
16
14
12
10
8


Any Gene Mutation Prediction Model (Reduced)
XGBClassifier (AUC = 0.59)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

Web-app/Online Calculator

We have made an online web-app/calculator to present our work and enable everyone to use our model to calculate the risk of mutation.
https://bcampred.team1719.repl.co/





Future Work

1. We had a small training dataset, and **adding more data** will improve the performance of the models and keep a check on overfitting.
2. Most of the data was for **patients living near Delhi**, inducing a bias on the model, as it was collected in AIIMS. We would be able to make a more general model with the inclusion of more varied data.
3. Data collection only happened for the breast cancer patients and **not for their relatives**, which could be an important feature. A **pedigree analysis** could also be done if this data was collected.

References

1. da Costa Vieira, R. A., Biller, G., Uemura, G., Ruiz, C. A. & Curado, M. P. Breast cancer screening in developing countries. Clinics 72, 244 (2017).
2. Cancer Statistics, 2020: Report From National Cancer Registry Programme, India. https://ascopubs.org/doi/full/10.1200/GO.20.00122
3. BOADICEA - Centre for Cancer Genetic Epidemiology. https://ccge.medschl.cam.ac.uk/boadicea/.
4. Lee, A. et al. BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. Genet. Med. 21, 1708–1718 (2019)
5. Ming, C. et al. Machine learning techniques for personalized breast cancer risk prediction: comparison with the BCRAT and BOADICEA models. Breast Cancer Res. 21, 1–11 (2019)