Machine Learning on METABRIC Breast Cancer Dataset

Nimish Verma, Rida Zaidi, Eisa Adil
School of Computer Science
University of Windsor
Windsor, Canada
(verma11t, zaidir, adile) @uwindsor.ca
105210324, 105160249,11001266

***Abstract*—Breast Cancer is the most common type of cancer amongst women.It starts in the cells of breast tissues and spreads into the body destroying the nearby tissue.Due to a considerable support for breast cancer research funding and for spreading awareness related to it, we are able to diagnose and treat it much faster than before. Thus, increasing the rate of survival and decreasing the number of deaths caused by breast cancer. We are only able to achieve it due to the through research advancements that lead to earlier detection and designing a better personalized procedure to treat cancer. Our project basically focuses on a dataset related to Breast Cancer provided METABRIC.This study aims to predict progesterone-receptor-positive (PR+),tumour stage and OncoTree Code,which indicates the type of cancer a patient is suffering from, through various classification and regression techniques.**

***Index Terms*—Breast Cancer, Classification, Feature Selection, Oversampling, Dimensionality Reduction**

I. Introduction

A. Breast Cancer

The most common invasive cancer after skin cancer among women is breast cancer. It is a tumour that forms in the tissues of breasts mostly in women but could also affect men. There are various kinds of breast cancer. The most common are Ductal Carcinoma and Lobular Carcinoma. Ductal Carcinoma is a cancer that starts in the cells that line the walls of a duct that transfers milk from glands to the nipple. Lobular Carcinoma is a cancer that starts in the lobules which are the glands producing milk [1].Figure 1 shows the labeled sructure of a breast.

It's estimated more than 600,000 breast cancer deaths among women and men worldwide occurred in 2018. [19].Figure 2 shows the occurrences of breast cancer worldwide.

B. Symptoms

Following are the most symptoms of breast cancer that should not be ignored in any case possible:

- A lump or a hard knot in the breast or the underarm area,
- Swelling,redness or darkness of the breast skin,
- Changes in size or shape of the breast,
- Itchy sensation or rash on the nipple,
- Sudden nipple discharge could contain blood,
- Feeling pain in a particular spot in breast,
- Dimpling of breast skin,

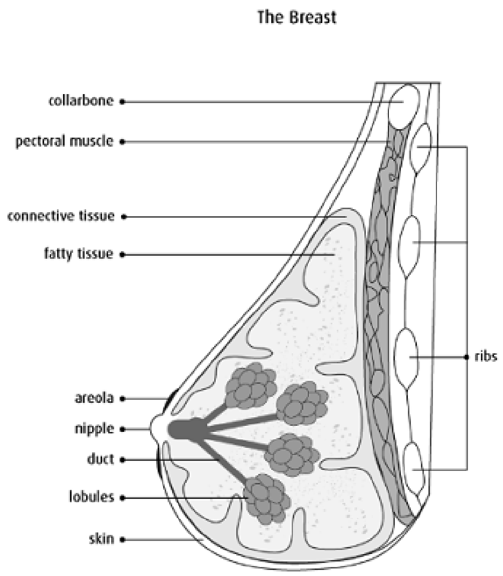


Fig. 1. Structure of a Breast(from [1]).

- Pulling in of a nipple or breast skin

These symptoms may not always mean the presence of a breast cancer but most definitely should be checked. [17]

C. Causes

The doctors are still baffled about the causes pertaining to breast cancer.According to the studies following are the main factors that could lead to breast cancer:

- Family history of breast cancer,
- Older age in women,
- Higher consumption of alcohol,
- Prolonged usage of birth control pills,
- Higher breast density,
- Prolactin is a hormone that is responsible for breast growth and the production of milk during breastfeeding.Women with higher blood levels of prolactin may have a higher risk of breast cancer.
- Exposure to radiation,
- Older age at first child birth,

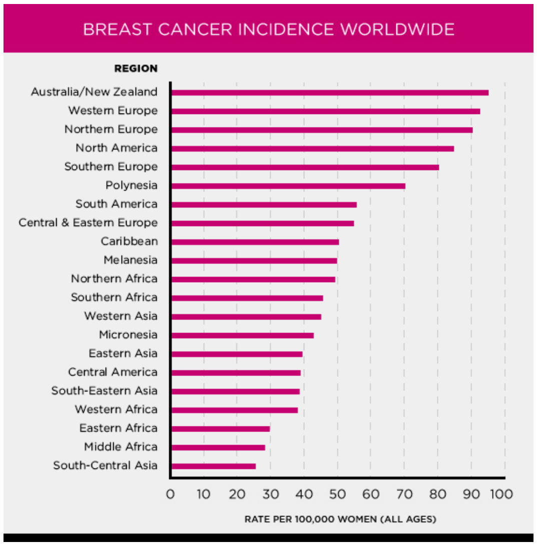


Fig. 2. Cases of occurrence of Breast Cancer world wide(from [20])

- Insulin-like growth factor 1 (IGF-1) a hormone that is responsible for growth.Higher levels of IGF-1 in the blood stream can lead to increase in the risk of breast cancer,
- Inherited gene Mutation [18]

A detailed study of gene expression in the patients suffering from breast cancer showed that the most important reason behind causing it includes the inheritance of mutated genes passing through the family named Breast Cancer Gene 1(BRAC1) and Breast Cancer Gene 2 (BRAC2).In older women of about age 70, with BRAC1 gene mutation, has about 55 to 65 percent chance of developing cancer.Older women with BRAC2 gene mutation has about 45 percent chance of being diagnosed with breast cancer [19].

D. *Diagnosis and Treatment*

Breast Cancer can be diagnosed through Mammograms,Breast MRI, and Biopsies.

The course of treatment usually depends on the severity of the cancer cells spread into the body, the size of the tumour, the stage of the cancer and of course the medical history of a patient [2]. Following are the different ways doctors could treat the breast cancer:

- Surgery
- Radiation therapy
- Chemotherapy
- Hormone therapy
- HER2-targeted therapy
- Oral cancer drugs [21]

Every treatment has its own side effects mostly include fatigue, loss of appetite, nausea,constipation or diarrhoea, hair loss, mouth sores, skin and nail problems. For this disease to be caught as early as possible posses an immense importance because the earlier the diagnosis the more chances of survival there are for the patient and the course of treatment is much less painful.

II. Project Overview

A. *Dataset*

Our dataset is provided by METABRIC(Molecular Taxonomy of Breast Cancer International Consortium). It is a Canada-UK based study group which aims to classify breast tumour into further categories, on the basis of which doctors can determine the course of treatment [3]. The dataset consists of two files. One file containing the gene information of about 24,368 genes and the other file containing the clinical data of the 2,173 sample patients.

B. *Target Values*

We aim to predict tumor size, tumor stage and OncoTree Code. We chose these three labels after studying breast cancer meticulously to determine which label would be the most beneficial for the doctors. We picked tumor stage as it determines the chances of survival of the patient.We selected OncoTree code because it explains the type of breast cancer present in the patient's body. We selected progesterone-receptor-positive (PR+) because the course of treatment that the doctor takes for a certain patient depends on it.

The possible values of OncoTree Code are:

- IDC - Invasive Ductal Carcinoma
- ILC - Invasive Lobular Carcinoma
- BREAST - BREAST Carcinoma
- MDLC - Breast Mixed Ductal and Lobular Carcinoma
- IMMC - Breast Invasive Mixed Mucinous Carcinoma

The possible values of Tumor Stage are:

- Stage 0
- Stage 1
- Stage 2
- Stage 3
- Stage 4

The possible values of PR Status are:

- Positive
- Negative

We will be using the aforementioned abbreviation in the paper later on.

C. *Work Flow*

We designed an efficient procedure considering our large dataset to predict these labels accurately as shown in Figure 3. Firstly, data processing was done in order to merge the two dataset files and perform data cleaning. Secondly, we did feature selection using Chi-Square for Classification problem. Thirdly, resampling was done which comprised of oversampling and undersampling techniques using SMOTE. Fourthly, we did classification using SVM, Gaussian Naïve Bayes, and Logistic Regression. After classification, we performed voting scheme on all the models using a Voting Classifier combining their predictions. Lastly, we evaluated the result of classification through confusion matrices.

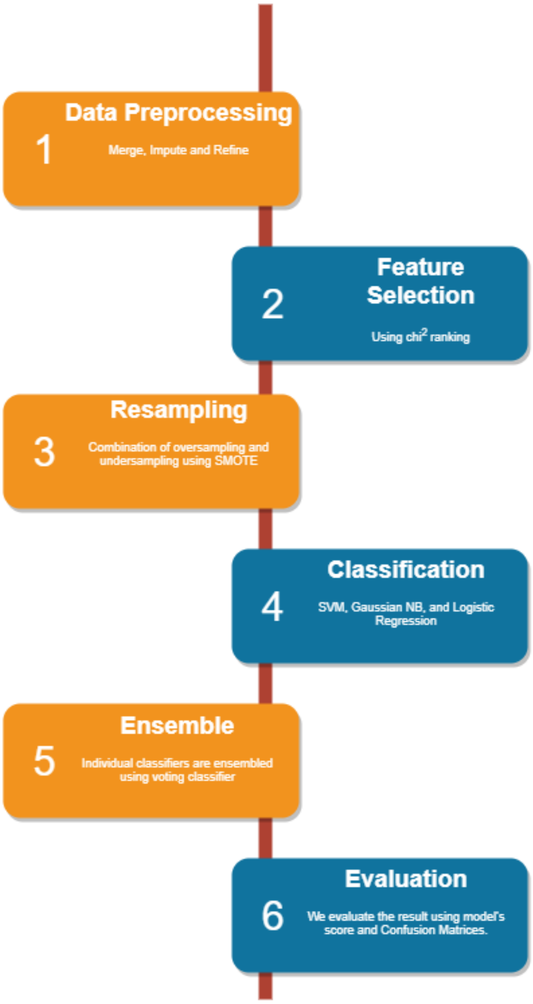

Data Preprocessing
1
Merge, Impute and Refine
Feature Selection
2
Using chi^2 ranking
Resampling
3
Combination of oversampling and undersampling using SMOTE
Classification
4
SVM, Gaussian NB, and Logistic Regression
Ensemble
5
Individual classifiers are ensembled using voting classifier
Evaluation
6
We evaluate the result using model's score and Confusion Matrices.

Fig. 3. Project flowchart.

III. Proposed Solution

We used Scikit-learn libraries which are implemented in Python for our procedure.

A. Data Preprocessing

1) Data Merging: We had two dataset files one containing the gene information and the other containing clinical data of patients. Both files contained PATIENT_ID on which we merged the two files. We faced a problem in the process of merging when it comes to the gene expression dataset because it had all the patient ids in columns rather than rows. As a solution we transposed the gene data so that all the genes(features) are in columns. After then we finally merged the dataset using pd.merge() taking a joint on PATIENT_ID.

2) Data Cleaning: After we get our final dataset we performed cleaning. In this step we dropped all the redundant columns containing same values which do not contribute into prediction.For example we dropped Metaplastic breast cancer (MBC) because it only contained one value "Breast Cancer".

3) Data Imputing: After cleaning our dataset we handled the NaN and missing values by replacing them with the mean.

B. Feature Selection

We performed feature selection because we had about 24,368 genes features. So, we removed irrelevant and redundant features from our dataset which resulted in better performance of the classifier. We use Chi-Square for feature selection for Classifying Tumor Stage, and OncoTree Code. Scikit learn library provides us sklearn.feature_selection.chi2(X, y) function to perform feature selection. It basically works by selecting n best features based on the values of the chi-squared statistic test. The chi-squared test basically indicates the dependence of the variables, so we select the features with the highest scores and removes all the features that seem to be independent from class thus does not contribute to classification. We first performed classification by selecting 2000 features and then later by 5000 features which resulted in giving much more better accuracies.

C. Sampling Data

Before diving into the Classification, we studied our data to check if it is balanced.For each label, we found out the number of rows. For OncoTree Code, following are the number of rows in each class:

- BREAST - 17 rows
- IDC - 1499 rows
- ILC - 142 rows
- IMMC - 22 rows
- MDLC - 207 rows

For Tumor Stage, we have the following:

- Stage 0 - 4 rows
- Stage 1 - 474 rows
- Stage 2 - 800 rows
- Stage 3 - 115 rows
- Stage 4 - 9 rows

For PR Status, we have the following:

- Positive - 1009 rows
- Negative - 894 rows

For our dataset, the dimensionality is high and since the classes are highly imbalanced in Oncotree Code and Tumor Stage, the accuracy of our classifier will be reduced. There are four ways to deal with this problem:

1) Synthesise of new minority class instances
2) Over-sampling of minority class
3) Under-sampling of majority class
4) Tweak the cost function such as misclassifying minority class instances has more penalty than majority class instances.

To avoid this problem, we resampled our data using the package - *SMOTE.*

SMOTE stands for Synthetic Minority Oversampling Technique. It is one of the many available techniques to tackle the problem of class imbalance.

SMOTE synthesises new minority instances between real minority instances [4] . This implementation of SMOTE makes

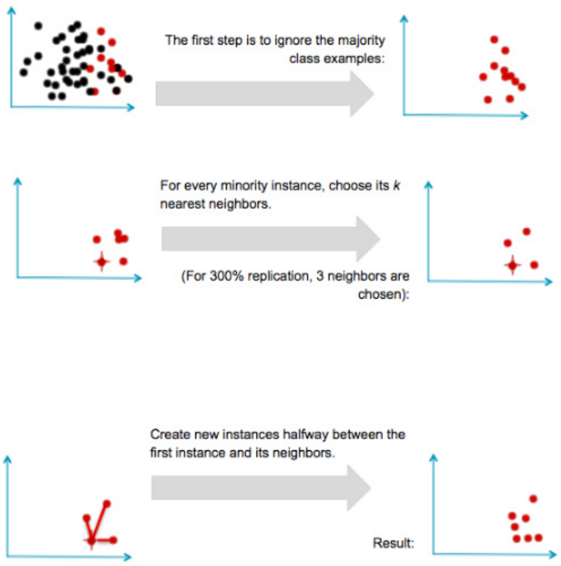

The first step is to ignore the majority class examples:
For every minority instance, choose its *k* nearest neighbors.
(For 300% replication, 3 neighbors are chosen):
Create new instances halfway between the first instance and its neighbors.
Result:

Fig. 4. Generating new minority samples using SMOTE (from [5])

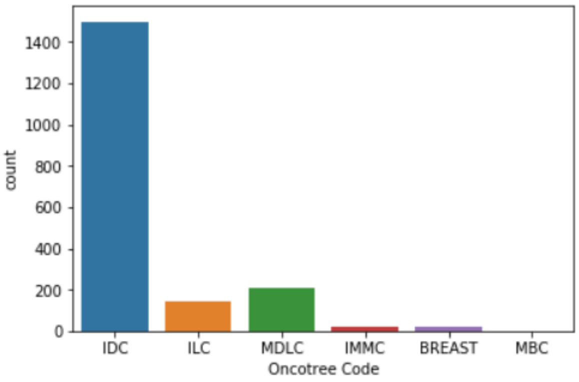


Fig. 5. Data Imbalance in OncoTree Code as target

the number of minority instances same as number of majority instances. It applies K-NN algorithm to join the existing instances, and creates synthetic sample in that space. The algorithm takes samples of the feature space for each target class and generates new examples that satisfy the feature space of its neighbors.

Figures 5,7, and 9 show the data imbalance for each of our target variables. Figures 6, 8, and 10 show the number of instances for each class after applying SMOTE for resampling.

D. Classification

After feature selection and resampling we finalized on a version of dataset ready for classification. We have used four different classification techniques i.e. Logistic Regression, Support Vector Machines (SVM) with Linear Kernel, Support Vector Machines (SVM) with RBF Kernel, and Random Forest. The detailed working of each algorithm is explained blow.

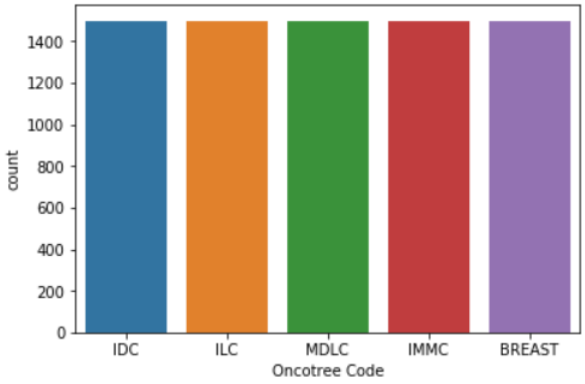


Fig. 6. Data Resampling for OncoTree Code

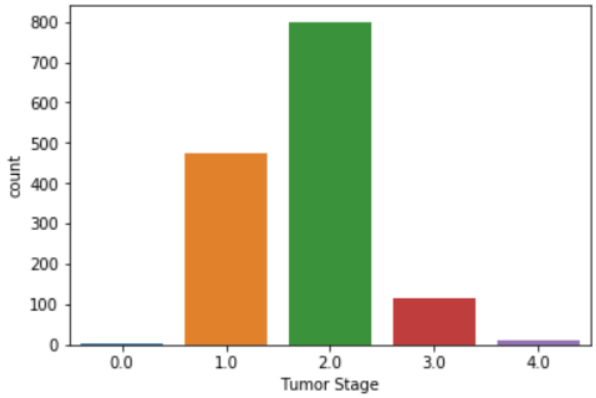


Fig. 7. Data Imbalance in Tumor Stage as target

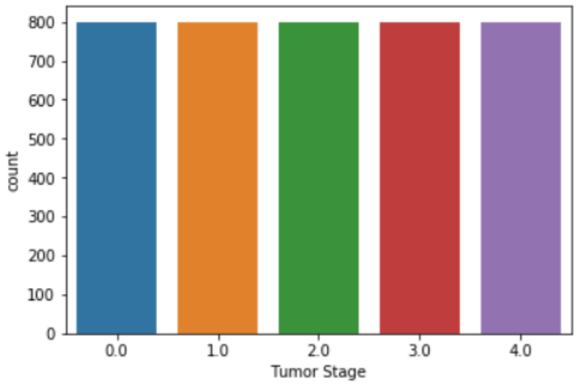


Fig. 8. Data Resampling for Tumor Stage

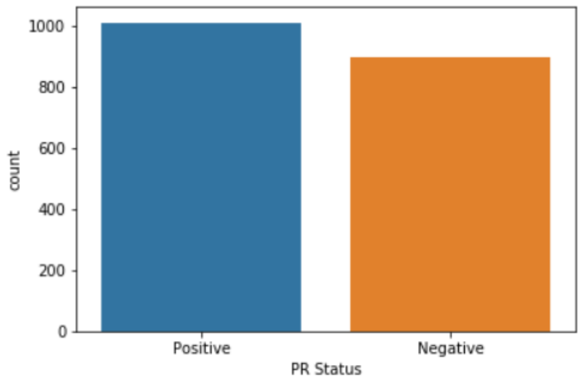


Fig. 9. Data Imbalance in PR status as target

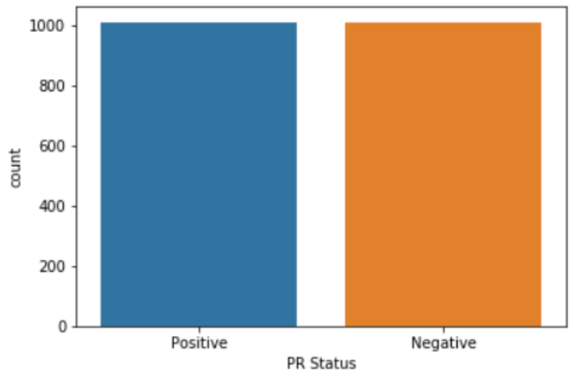


Fig. 10. Data Resampling for PR Status

1) Logistic Regression: Logistic Regression is the most preferable technique to use when our target variables are categorical. There are three types of Logistic regression:

1) Binary Logistic Regression: The class labels have only two possible outcomes.Example: PR status has only two values Positive and Negative.
2) Multinomial Logistic Regression: The class labels have three or more possible outcomes.Example: OncoTree Code has about six possible values which are IDC, ILC, MDLC, IMMC, BREAST, and MBC.
3) Ordinal Logistic Regression: The class labels have three or more categories but they are in order.Example: Tumor Stage: 0,1,2,3, and 4.

We used scikit library's function *sklearn.linear_model.LogisticRegression* for predictive analysis and chose the value 'auto' for the argument 'multi-class' which decides which type of logistic regression to apply.Logistic Regression works by predicting the probability of an occurrence of target variable by fitting the data to a logit (sigmoid) function [10].

$$h_\theta(x) = \frac{1}{1 + e^{-}\theta^T x}$$

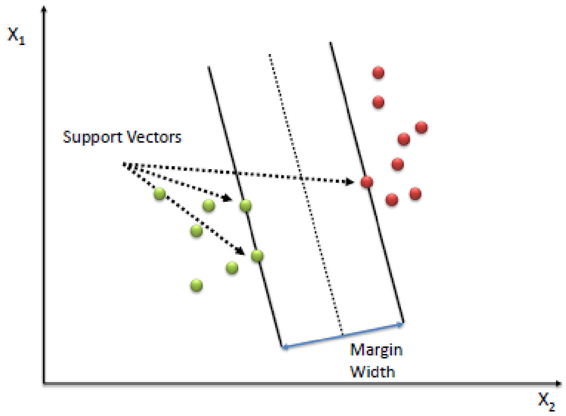


Fig. 11. Support Vector Machine(SVM) (from [9])

2) Support Vector Machine(SVM): Support Vector Machine is a supervised machine learning algorithm with an aim to create a hyperplane that successfully classifies all its data points in a n-dimensional feature space where n is the number of features. As shown in Figure 11 support vectors are the datapoints closest to the hyperplane. SVM draws a hyperplane in such a way that maximizes the margin, meaning the distance between points that are closest to the other class points [11]. Support Vector Machines are very powerful when it comes to classification. As we have studied that SVM draws a linear hyperplane between the classes but what if the classes aren't linearly separable. SVM uses a trick called Kernel trick. These functions map the datapoints to higher dimension where the problem is linearly separable and SVM can draw a hyperplane correctly classifying our target variables [12]. We used RBF kernel for our problem as it maps data to infinite higher dimensions giving us a very good accuracy that would be explained in the next section. RBF kernel function is shown below:

$$e^{\frac{(-|u-v|^2)}{\sigma^2}} \tag{1}$$

3) Random Forest: Random Forest is a powerful yet very flexible machine learning technique. It is basically a large group of relatively uncorrelated decision trees that are merged together to get a more accurate result. The key reason behind the good accuracy is the uncorrelation between the trees, when one tree is wrong the others might be right so their combined effect is always right. Random Forest usually produces the correct result even without a hyper tuning parameter.The higher the number of uncorrelated trees in the model the more better would be our accuracy. [13]

E. Ensemble

In Machine Learning, ensemble learning uses multiple learning algorithms to obtain better predictive performances than those that could be obtained from any of its constituting learning algorithms [6].

To avoid overfitting in a single classifier, we have used ensemble learning. In our project, we use *VotingClassifier*, an ensemble learning technique in sklearn. The idea behind the VotingClassifier is to combine conceptually different machine

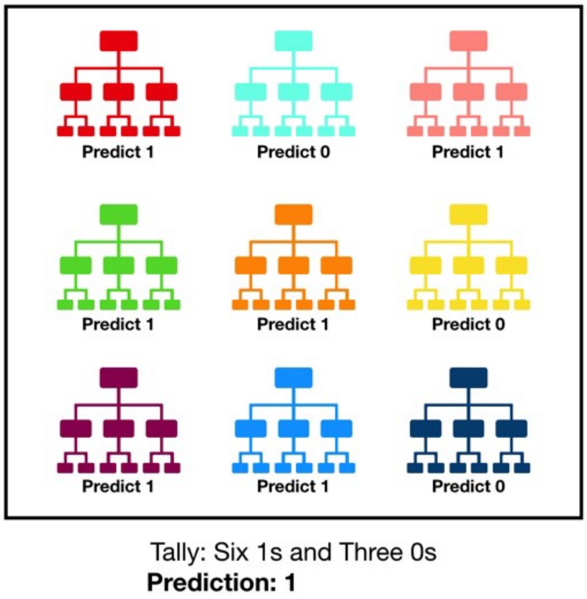


Fig. 12. Random Forest (from [13])

learning classifiers and use a majority vote or the average predicted probabilities (soft vote) to predict the class labels [7].

This model is used for equally well performing models so that they cover their weaknesses when used together in a voting ensemble. We have used a **hard** voting strategy in the VotingClassifier and find the average accuracy. Ensemble allows us to have a much more flexible model to exist amongst all the other alternatives.

F. Evaluation

After performing classification and obtaining the predicted labels, we evaluate the models using the *score* attribute of sklearn models, that returns the mean accuracy on the given test labels [8].

For multi-class prediction, which is the case in Tumor Stage and OncoTree Code, just knowing the mean accuracy is a harsh metric, since we require for each sample, each label is predicted correctly. Therefore, we also print the confusion matrix.

Confusion Matrix is generally used in a binary classification, but for multi-class classification, we treat each label as a one-vs-all case,and get a confusion matrix for each label. We used *multilabel_confusion_matrix* package from sklearn. We defined our output label in the parameter, and got the confusion matrices in that particular order.The confusion matrices for each target variable is discussed in detail in the Section V of this paper.

G. Visualization using Dimensionality Reduction

Data Visualization is required in order to view the number of possible clusters and identify different patterns in data. Our dataset has about 24,368 features, so how do we visualize our data?

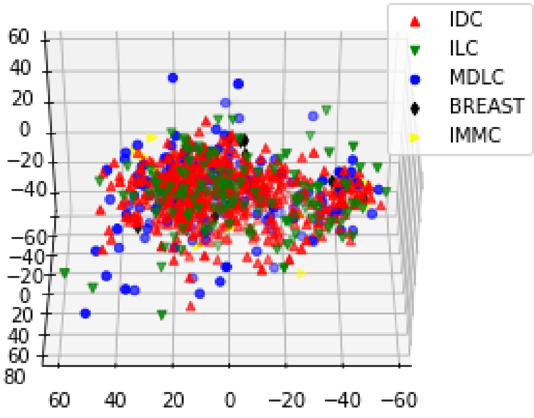


Fig. 13. PCA on OncoTree Code Before Resampling

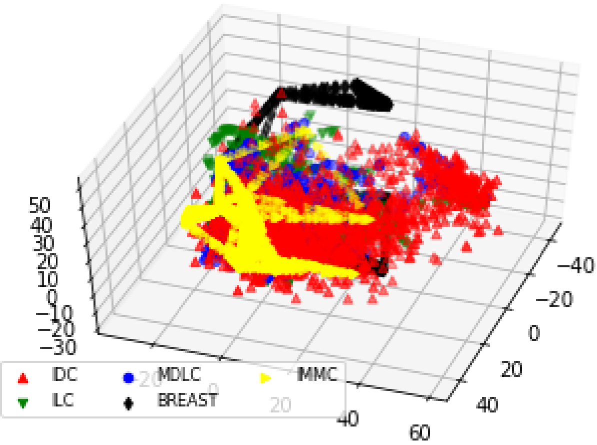


Fig. 14. PCA on OncoTree Code After Resampling

To solve this problem, we use *Principal Component Analysis* (PCA). PCA detects the correlation between variables. In simple words, PCA finds the direction (eigenvector) for which the variance is maximum. This is named as the *First Component*. For our purpose, we choose 3 components, as we can visualize our data in at most 3 dimensions. Therefore, as a result, we get 3 new dimensions with the highest variances. That is $\sigma_{c1} > \sigma_{c2} > \sigma_{c3}$. We then plot our new reduced dataset and observe the patterns by labelling the points according to the different label values predicted in this research - *Tumor Stage, OncoTree Code and PR Status* Figures 13,15,and 17 show the plot of the visualization done for each target variable before SMOTE resampling. Figures 14,16,and 18 shows the visualization after SMOTE resampling.

IV. Results and Observation

1) Individual Classifiers: After finishing classification with SVM using Linear and RBF kernel, Logistic Regression and Random Forest. We performed classification using 2000 and 5000 *K-Best* Features. Upon observing the results we decided to consider 5000 features for our final evaluation as they gave

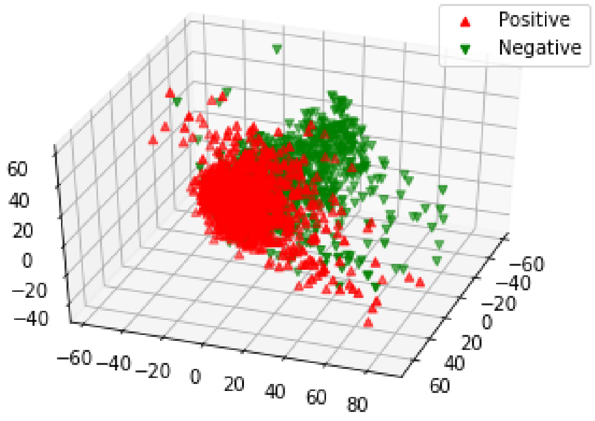


Fig. 15. PCA on PR Status Before Resampling

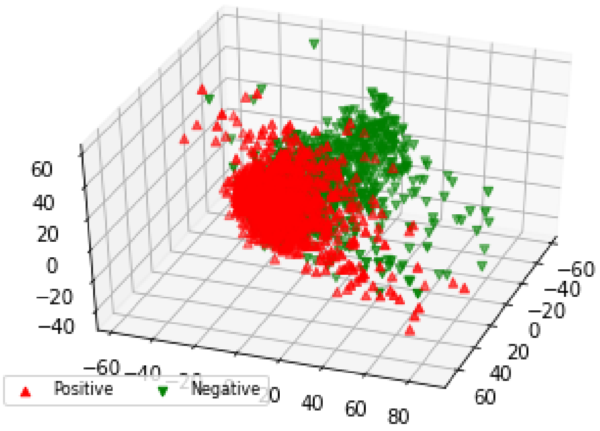


Fig. 16. PCA on PR Status After Resampling

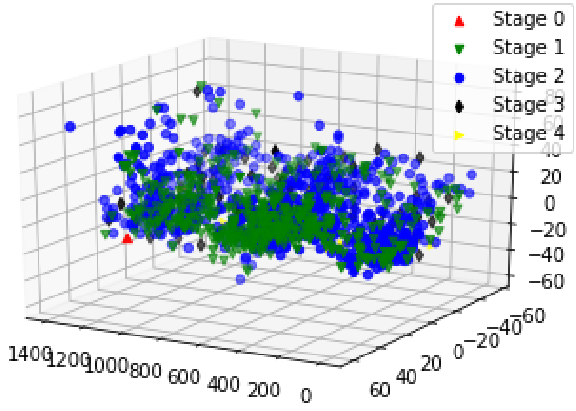


Fig. 17. PCA on Tumor Stage Before Resampling

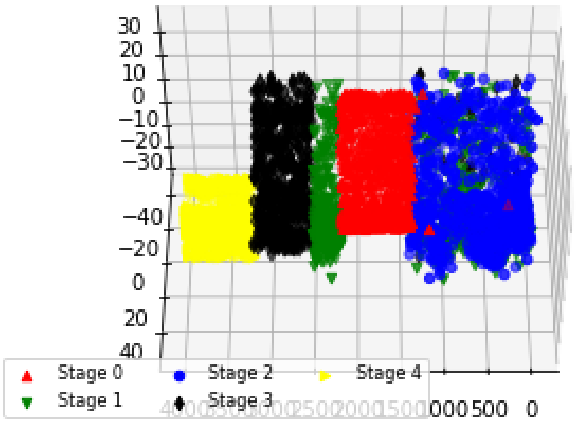


Fig. 18. PCA on Tumor Stage After Resampling

much better accuracies. The results are shown in Table I, II, and III.

TABLE I
OncoTree Code Accuracy

Classifier	Accuracy %
SVM (linear)	97.93
SVM (rbf γ 0.001)	99.79
Logistic Regression	89.45
Random Forest	97.59

TABLE II
Tumor Stage Accuracy

Classifier	Accuracy %
SVM (linear)	89.5
SVM (rbf γ 0.001)	93.25
Logistic Regression	81.75
Random Forest	88.62

TABLE III
PR Status Accuracy

Classifier	Accuracy %
SVM (linear c= 0.0025)	87.37
SVM (rbf γ 0.0002)	88.81
Logistic Regression	88.81
Random Forest	89.10

As we see from Table 1, SVM with RBF kernel and gamma = 0.001 gives the highest accuracy in predicting the OncoTree Code. Other classifiers have good performance except Logistic Regression, which has a comparatively lower accuracy - 89.45 %. In Table 2, we see the accuracy of each model while predicting Tumor Stage. For this target value, Random Forest with Linear Kernel, Logistic Regression, and Random Forest perform poorly, while SVM with RBF kernel and gamma = 0.001 has the highest accuracy - 93.25%.

Similarly, in Table 3 we have compared the accuracy while predicting PR Status. For this, we see that the accuracy of all models is almost similar. We have used RBF kernel with gamma 0.00045 which achieved 88.86% accuracy.

We see that RBF kernel on SVM gives us better accuracy than the linear kerenl RBF and Logistic Regression (except in PR status, where it is equal). This is because, the RBF kernel performs *kernel trick*, by finding the dot product of two input vectors and then calculating their projection in a higher (infinite) dimension, in such a way that it becomes linearly separable in that dimension.

We also see, the accuracy of Random Forest is high, and in the case of PR Status even higher, this is because Random Forest is a kind of *ensemble*, and as we have discussed, ensembles avoid overfitting by increasing varianace and including different models. In a random forest, there are many base estimators which are nothing but decision trees generated randomly. The final prediction of the Random Forest model is a majority vote of the individual trees. Therefore, sometimes it can perform better in cases where even RBF fails, as we see in Table III.

2) Ensemble Learning: Results after combining the individual classifiers using a *Voting Classifier* are shown in Table IV

TABLE IV
ENSEMBLE ACCURACY

Target Label	Accuracy %
OncoTree Code	98.20
Tumor Stage	90.25
PR Status	89.60

As we can see, the ensemble accuracy is better than the individual classifiers. Ensemble always performs better than the individual classifiers, when the individual classifiers have similar accuracy. But in the case of Tumor Stage and Oncotree Code, we have individual classifiers with somewhat varying accuracies.

In Oncotree Code and Tumor Stage, we have RBF kernel of SVM classiifer outperforming others by a considerable margin, while in PR Status the accuracy of each model is somewhat similar, therefore the **accuracy of ensemble in PR status is higher than the individual classifiers**.

V. EVALUATION

We have used Confusion Matrices to evaluate our results. For multi-class classification, we have used the class *multilabel_confusion_matrix* from sklearn [14]. The confusion matrices can be seen in Figures 19, 20, and 21.

The confusion matrices are for the **ensemble VotingClassifier** of each target variable, and not for the individual classifiers. The individual classifier confusion matrices can be seen in the Jupyter Notebooks attached with this code.

Our model has achieved **0 mis-classified samples** in Stage 4 Tumor Stage and in IMMC OncoTree Code. Also, the model only misclassified 4 ILC OncoTree Code Samples and 3 BREAST OncoTree Code Cancer.

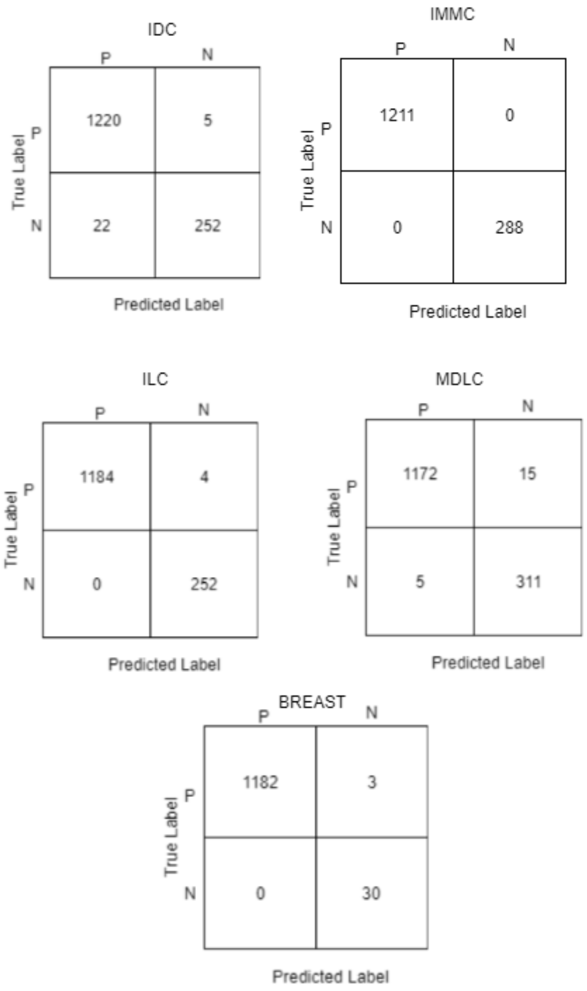


Fig. 19. Confusion Matrices for OncoTree Code

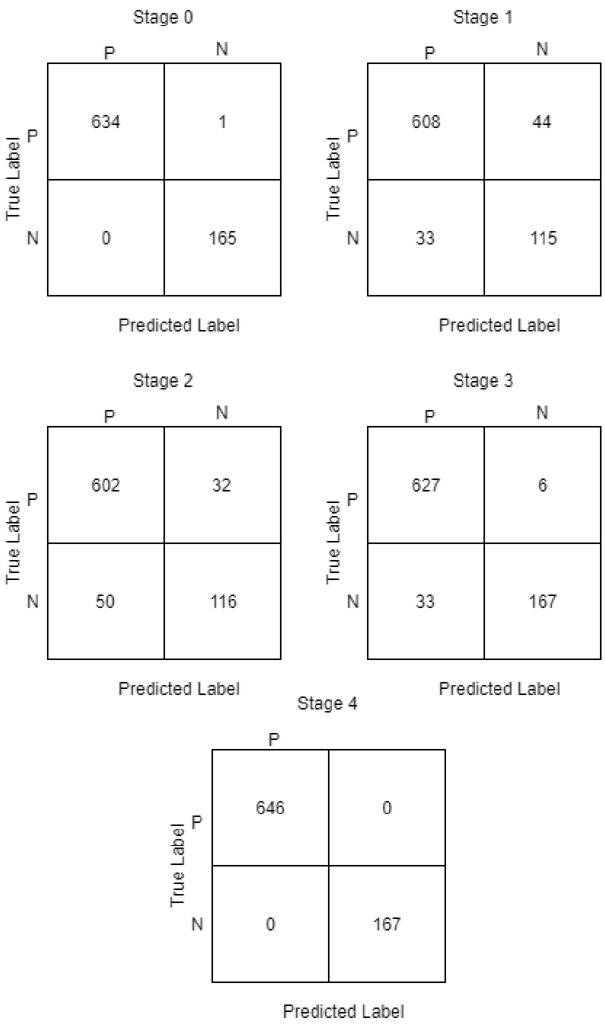
VI. ETHICAL, LEGAL AND SOCIETAL ASPECTS

A. Ethical

According to IEEE [15], there are three main pillars of the Ethically Aligned Design (EAD) Conceptual Framework - Universal Human values, Political Self-Determination, and Technical Dependability. We will now discuss these aspects related to our research.

1) Universal Human Values: Our AI system should be designed to protect human rights, values and well-being. The system that we have created should be equally accessible to all groups of people and not only for the small group of rich people. Our system should also safeguard the natural resources and environment. By this we imply that the amount of electricity and other resources spend on this AI system should not be deteriorating to the natural resources and should be optimized in terms of the use of such resources.

2) Political Self-Determination: Our system should also respect the right of a patient over his data. According to IEEE standards, " Agency in the digital arena enables an individual to make informed decisions where their own terms and condi-


True Label
Predicted Label

Stage 0

	P	N
P	634	1
N	0	165

Stage 1

	P	N
P	608	44
N	33	115

Stage 2

	P	N
P	602	32
N	50	116

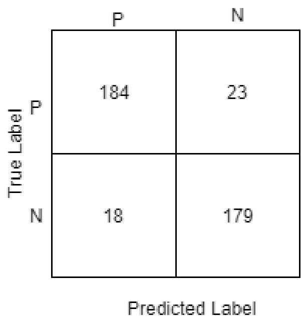
Stage 3

	P	N
P	627	6
N	33	167

Stage 4

	P	
P	646	0
N	0	167

Fig. 20. Confusion Matrices for Tumor Stage


True Label
Predicted Label

	P	N
P	184	23
N	18	179

Fig. 21. Confusion Matrices for PR Status

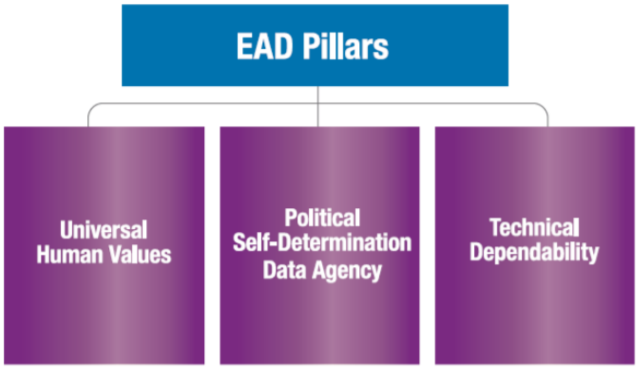


Fig. 22. Three Pillars of EAD [15]

tions can be recognized and honored at an algorithmic level. ” [16] . Therefore, our AI system is completely **transparent** and **prohibits misuse of data**.

The system provides the users with full access and control to their own data, that they provided to us, and is in accordance cultural precepts.

We will perform **data anonymization** to anonymize the patient name while collecting the samples. This will ensure that there is no misuse of the data. Our AI system will also make us aware of misuse by maintaining logs of the queries made.

3) Technical Dependability: Lastly, we ensure **competence** of our system, which is one of the most important ethical aspect of an AI system. As discussed in the Section IV, our system has achieved very high accuracy in OncoTree Code, and considerably high in PR Status, and Tumor Stage.

The AI we have designed delivers services that can be relied upon, and does not incline to use the data provided in any other sense that is not implied.

The system is *trustworthy*, and accomplishes objectives for which it is designed.

B. Legal

Irrespective of the application of an AI system, there is a need to discuss the legal aspects. This is because the law does not cover much of AI issues (as of yet). As the use of AI is increasing rapidly, general norms and laws will be adopted in the areas of data protection, intellectual property and negligence.

The question arises, ”Who is responsible for the misuse of data?”. As the developers of this AI system, it is our duty to ensure security and anonymization of data to prevent as much misuse as we could.

Moreover, this AI system does not provide a 100% confident prediction. It is still a *prediction*, and may be wrong or right. But we provide good accuracy which can be considered reliant to provide an initial **prognosis** to doctors and patients, so that they can perform *actual tests* accordingly.

C. Societal

Considering Societal aspects can be a bit complicated. Will our system take over the jobs of doctors? No. Our system is designed in an aim to assist the doctors in making decisions and going in the right direction for prognosis. Our system relies on these experts, as they are the ones that can understand, provide, and improve the system. Instead, we will be needing doctors to label future data and enrich our dataset, by providing more *real* instances of minority classes.

Our AI system should not be used towards undesirable ends, and even though the prediction accuracy is not 100% it should not deterioriate the reliablity of AI in general.

For liability, as we have already mentioned, our system is just an assistance for the doctors' prognosis, and should not be solely considered to make decisions while diagnosing as it may cause drastic damage if it is wrong (which it can be sometimes). Therefore, it should be the duty of doctors to perform necessary tests and double check, before confirming the diagnosis.

VII. Conclusion

Good results with all the four classifiers used, justify the fact that the 5000 features(genes) returned by the feature selection are the primary genes responsible for the OncoTree Code, Tumor Stage, and PR Status of the breast cancer, and if required, the doctors and researchers can further research into these features(genes).

We have also found out the top 10 important features using the *features_importances_* attribute of the RandomForest classifier. The features and their importances are stored in the excel file named 'Top 10 important features.xls' attached with the project.

We have discussed the Ethical, Legal, and Societal Aspects of our AI system. We concluded that it is our duty to safeguard these aspects for not just humans but also the environment. Also, that a machine can not be relied upon so much that it is the sole judge to make decisions of life and death.

References

[1] https://www.cancer.ca/en/cancer-information/cancer-type/

[2] https://www.mayoclinic.org/diseases-conditions/breast-cancer

[3] http://molonc.bccrc.ca/aparicio-lab/research/metabric/

[4] http://rikunert.com/SMOTE_explained SMOTE EXPLAINED

[5] https://datasciencecampus.ons.gov.uk/projects/synthetic-data-for-public-good/ Synthetic Data for public good

[6] https://en.wikipedia.org/wiki/Ensemble_learning Ensemble Learning Wikipedia

[7] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble. VotingClassifier.html Voting Classifier SKLEARn

[8] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html SVM score

[9] https://discuss.analyticsvidhya.com/t/smote-implementation-in-python/19740

[10] https://towardsdatascience.com/logistic-regression-for-dummies-a-detailed-explanation-9597f76edf46

[11] https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[12] https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

[13] https://towardsdatascience.com/understanding-random-forest-58381e0602d2

[14] https://scikit-learn.org/stable/modules/generated/sklearn.metrics. multilabel_confusion_matrix.html Multi label Confusion AMtrix

[15] https://ethicsinaction.ieee.org/ EAD IEEE

[16] https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e_personal_data.pdf Personal Data and Individual Agency

[17] https://ww5.komen.org/BreastCancer/WarningSigns.html

[18] https://ww5.komen.org/BreastCancer/BreastCancerRiskFactorsTable.html

[19] https://ww5.komen.org/BreastCancer/InheritedGeneticMutations.html

[20] https://ww5.komen.org/BreastCancer/Statistics.html

[21] https://ww5.komen.org/BreastCancer/TreatmentIntroduction.html