Breast Cancer Analysis using Machine Learning Methodologies

Mayank Semwal
M.Sc Computer Science
University of Windsor
Student Id: 104953967
Email: semwal@uwindsor.ca

Priyanka Motwani
M.Sc Computer Science
University of Windsor
Student Id: 105216601
Email: motwanip@uwindsor.ca

Vipul Malhotra
M.Sc Computer Science
University of Windsor
Student Id: 105111504
Email: malho117@uwindsor.ca

***Abstract*—Breast cancer (BC) is one of the most common cancers among women in the world, representing the majority of new cancer cases and deaths related to cancer according to global statistics, causing it a severe public health concern in today's society. For 2019, it was estimated earlier by the Canadian Cancer Statistics that 26,900 Canadian women will be diagnosed with breast cancer, and 5,000 will die of it [11]. Breast cancer accounts for approximately 25% of new cases of cancer and 13% of all cancer deaths in Canadian women. 1 in 8 women are expected to develop breast cancer during their lifetime, and 1 in 33 will die of it. While it can also be found in men, male breast cancer is an infrequent occurrence. Breast cancer starts in the cells of the mammary gland. Breast tissue covers a larger area than just the breast, extending up to the collarbone and from the armpit to the breastbone. A prediction of breast cancer in an initial stage provides a higher possibility of its cure. It needs a breast cancer prediction tool that can classify a breast tumor whether it is a malignant tumor or a benign tumor. Machine learning (ML) is widely recognised as a technique of choice in BC pattern classification and forecast modelling due to its unique advantages in critical feature detection from complex BC datasets. Classification and data mining methods are an effective way to classify data. This work aims to show the working of different machine learning algorithms and compare the results of their performance accuracy to present an effective method for the prediction of breast cancer.**

***Index Terms*—Supervised Learning, Support Vector Machine, Naïve Bayes, Feature selection, Breast Cancer Detection, XG-Boost, Random Forest Classifier, Logistic Regression, Gradient Boost, SelectKBest**

I. INTRODUCTION

Breast cancer starts in the breast cells. A cancerous tumour (malignant) is a group of cancer cells that can expand into adjacent tissue and damage it. It can also spread (metastasize) to other parts of the body. Breast cancer starts in cells that line the ducts, which are the tubes that carry milk from the glands to the nipple. This type of breast cancer is called ductal carcinoma. Cancer can also start in the cells of the lobules, which are the groups of glands that make milk. This type of cancer is called lobular carcinoma. Both ductal carcinoma and lobular carcinoma can be in situ, which means that the cancer is still where it started and has not grown into surrounding tissues. They can also be invasive, which means they have grown into surrounding tissues.

Breast Cancer is one of the most widespread and second most occurring cancer; at present, there are no effective ways to prevent breast cancer. However, efficient diagnosis in an early stage can increase the chance of full recovery which makes early detection and diagnosis a vital issue where currently mammography screenings are the primary imaging modality for early detection of breast cancer. [8]. In the follow-up process, early diagnosis of breast cancer is one of the most important works. Methods of data mining can help to reduce the number of false negative and positive decisions [7]. BC's early diagnosis will significantly improve the prognosis and chance of survival as it can allow patients to seek prompt medical care. More precise tumor classification can avoid unnecessary treatment of patients. The proper diagnosis of BC and the classification of patients' data into malignant or benign groups are therefore the subject of much research. Because of limitations, humans can make errors when diagnosing a disease and it depends on doctors expertise. Diagnosis can be done more concisely using machine learning algorithms (91.1%) compared to an experienced physician's diagnosis (79.97%)[4].

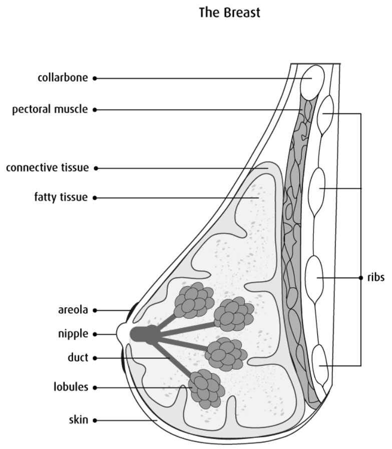


Fig 1. Breast Cancer

Why is Breast Cancer Analysis important?
In 2019, according to cancer studies' data provided by cBioPortal, 9203 breast cancer cases have been reported which is significantly higher in comparison to other types of cancers.

Our approach for such type of classification problem is using supervised learning classifiers. The primary goal of

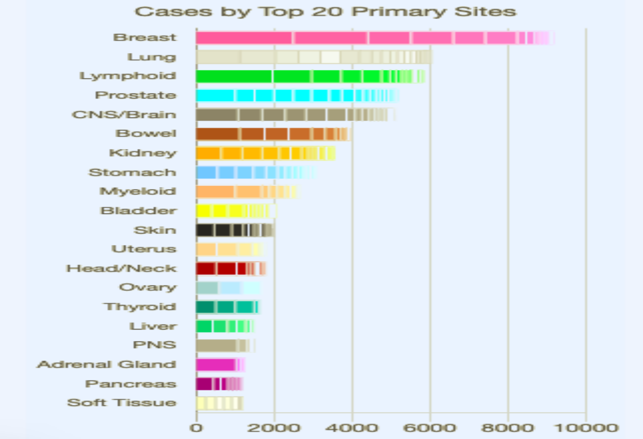


Fig 2. Types of Cancers Reported in 2019

supervised learning is to train a model from labeled training data (Figure 1) that allows us to make predictions about future data. Here, the term supervised refers to a classification of the desired result.

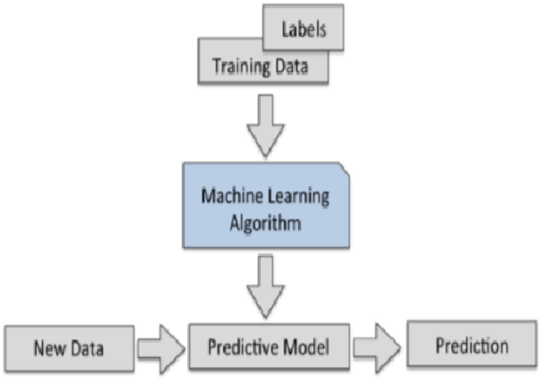


Fig 3. Machine Learning Model

We have used six famous classifiers like Random Forest Classifier (RF), Naïve Bayes Classifier (Gaussian NB), Support Vector Machine (SVM), Gradient Boost Classifier(GBoost), Extreme Gradient Boosting(XGBoost) and Logistic Regression(LR). Each classifier is compared with other on the basis of performance metrics. Important feature selection and extraction is done using SelectKBest technique with chi squared score. Distribution of all sections of this paper is as follows: Section II describes about the dataset used. Section III gives an insight to the problem statement. Section IV describes the related work done in this area. Section V narrates the methodology with details of classifiers implementation. Section VI describes the algorithms used. Section VII shows the experimental setup. Section VII discusses the results obtained. Section IX briefs about the ethical, legal and societal impacts. Section X refers the conclusion and future works. Section XI consists of contribution.

II. DATASET DESCRIPTION

The classification has been done on “Breast Cancer (METABRIC, Nature 2012 & Nat Commun 2016)”[3] dataset which is imported from cBioPortal. The dataset has 1125 samples of patients and 16,384 samples of genes. From a total of 16,384 features, important features are extracted using ”SelectKBest’ technique and are classified into types of classes based on the results. Label Encoding is applied to convert categorical labels into numerical values.

III. PROBLEM STATEMENT

The problem in the dataset is the multi-class classification problem which involves a large number of features. The input data is the gene id’s of patients, while the output is multi-class in nature depending on the type of the labels, i.e., cellularity, cancer type, chemotherapy and overall survival status of the patient. So to understand the significance of the features, we have to perform some strategies for feature selection and extraction, apply a multi-class classification algorithm and iterate this process, until performance saturates. The objective of our project is to build a predictive model that will improve the accuracy, objectivity and predictability of breast cancer diagnosis.

IV. RELATED WORK

Classification is one of the most important and essential tasks in machine learning and data mining. A lot of research has been conducted to apply data mining and machine learning on different medical datasets to classify Breast Cancer data. Many of such experiments result in good classification accuracy.

In the analysis on the Wisconsin Breast Cancer dataset [9], the authors(Musjtaq, Yaqub & Hassan, 2019) have focused on recognizing tumorous(malignant) and non-tumorous (benign) from the dataset. They have included various supervised machine learning classifiers and their performance measures. Separate kernel PCA-based techniques were applied to these algorithms after data scaling and fitting. Different performance metrics were exercised later. From the tests, quality differences by classifiers were shown in these techniques. Naive Bayes with sigmoid PCA performed well with 99.2% accuracy but SVM with sigmoid showed worse performance with 83.5% accuracy and overall, KNN with Linear and RBF kernels having 97.80% accuracy performed the best among the other supervised machine learning algorithms with this breast cancer dataset.

The researchers(Al-Shargabi & Al-Shami, 2019) performed an experimental analysis to identify breast cancer algorithms for three major ML algorithms such as K-Nearest Neighbor (KNN), Random Forest (RF) and Multilayer Perceptron (MLP)[2]. The tests were carried out on the WDBC dataset in order to obtain the best accuracy algorithms. Ultimately, through more than one feature selection algorithms, classify the most basic and important characteristics of malignant tumour classification. The best result for breast cancer classification was 100% accuracy for KNN and RF and

97.19% accuracy for the original MLP.

In the paper[7], researchers(Lg & At, 2013) analyzed breast cancer data of Iranian Center for Breast Cancer (ICBC) dataset using three classification techniques namely Support Vector Machine(SVM), Artificial Neural Network(ANN) and Random Forest Classifier(RF) to predict the recurrence of the cancer and then compared the results. The results indicated that SVM is the best classifier predictor with 95% accuracy, followed by ANN and RF having 94% and 93% accuracy.

With respect to all related work mentioned above, our work compares the behaviour of six machine learning algorithms including SVM, NB, RF, GBoost, XGBoost and LR using Breast Cancer dataset in both diagnosis and analysis to make decisions. The aim is to achieve the best accuracy with the lowest error rate in analysing data. To do so, we compare efficiency and effectiveness of those approaches in terms of many criteria, including: accuracy, precision, sensitivity and specificity, correctly and incorrectly classified instances and time to build model, among others.

V. METHODOLOGY

This study involves a comparison of various supervised machine learning classification algorithms. The dataset imported is used for model-training purpose. We are extracting the features and labels from the dataset, and then the processed data is used for training and testing phases. Various methodologies have been applied for feature extraction and selection. Data is divided into 70% training and 30% testing parts. Different methodologies have been applied for classification of the patients' data samples, and so the data is divided into multiple classes based on the type of labels. Performance measures like accuracy, confusion matrix and the number of misclassified samples have been determined.

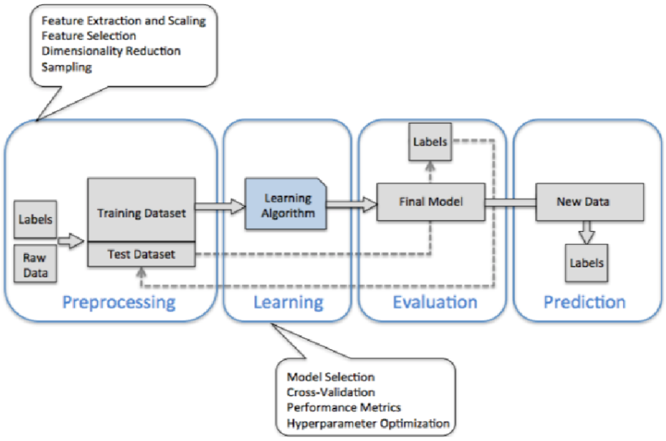


Fig 4. Proposed Breast Cancer Analysis Model

A. Data Inspection

The dataset "Breast Cancer(METABRIC, Nature 2012 & NAT Commum 2016)"[3] has 1125 samples of patients data and 16,384 samples of genes data which is taken from cBioPortal.

The HUGO Gene Nomenclature Committee (HGNC) is a committee of the Human Genome Organisation (HUGO) that sets the standards for human gene nomenclature. The HGNC approves a unique and meaningful name and symbol for every known human gene, based on a query of experts.
i) Features: The dataset has a file named "data_expression_median" that is considered for features of the data. Over here, there are two columns namely, "Hugo_Symbol" and "Entrez_Gene_Id". The latter one is dropped and the former one, "Hugo_Symbol" is used as a feature name because it consists of a unique and meaningful name for every known human gene and corresponding to that, "Patient_Id" is spread along the rows. Hence, the data of each patient having any particular type of gene is available in the dataset.
ii) Labels: The dataset has two important files namely "data_clinical_sample" and "data_clinical_patient" that are considered for the labels of data. The labels' dataframe is prepared by applying join on the above two files using "Patient_Id" as a key which is common amongst the two.

The four important labels considered for classification are:
a) Cellularity: Cancer Cellularity is defined as the proportion of cancer within the residual tumor bed. In clinical practice, the largest cross-sectional area of the preidentified tumor bed is divided into multiple slides, which are stained with hematoxylin and eosin (H&E) and then reviewed with a microscope.Few studies have considered tumour cellularity, which is defined as the percentage of invasive tumour, which was comprised of tumour cells, in the assessment of response to therapy in breast carcinoma. Cellularity is divided into three categories that include high, moderate and low, we have disregarded NA values for our models.

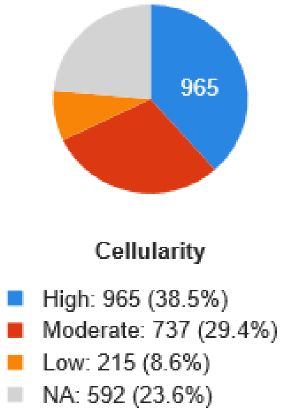


Fig 5. Cellularity

b) Chemotherapy: This label provides the information that if the patient is required to have chemotherapy treatment or not. The label will have a binary classification output that will be either a yes or a no.

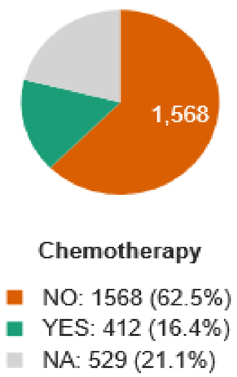


Fig 6. Chemotherapy

c) Overall Survival Status: This field determines the overall survival status of the patients suffering from Breast Cancer. The two types of this label include living and deceased.
d) Cancer Type Detailed: This field describes the type of breast cancer detected in the human body. The data set has six labels and few none values, we have removed none values and considered four of these labels reason being less than 10 records exists for it, used labels are Breast Invasive Ductal Carcinoma, Breast Invasive Lobular Carcinoma, Breast Invasive Mixed Mucinous Carcinoma and Breast Invasive Ductal and Lobular Carcinoma.

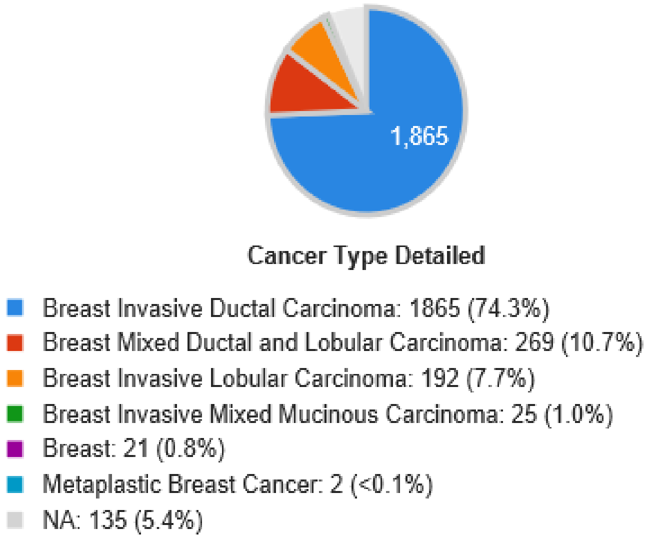


Fig 7. Cancer Type Detailed

B. Data Cleaning & Data Preprocessing

Data cleaning is basically the process of removing errors and anomalies or replacing observed values with true data values in order to gain greater value in analytics.
Data preprocessing is a step when raw data is selected and transformed into a clean dataset which will is feasible for machine learning algorithms. Preprocessing of data is considered as a very crucial step for any data analysis problem and is often an excellent idea to prepare the given data in a way that it best exposes the structure of the problem to the machine learning algorithms in an efficient way.

The following activities are involved while data cleaning and then preprocessing:

i) Features: The best set of genes or features to be selected for our models Missing Values: The number of missing values are calculated for each column of the features(data_expression_median) file and if the values exceed 20% of the whole dataset, that particular feature is dropped as these gene id's won't add value to our models while classifying. There are no null values in features dataset, and maximum null values found for "hugo_symbol" column are only 2.

Transformation: Transpose is applied on the dataset to get values of Patient_Id as column data and that of "hugo_symbol" as row data. This transpose is done by considering "hugo_symbol" as a header.

ii) Labels: Missing Values: As explained in data inspection, we have considered four labels for our model, and labels are selected, which holds medical importance as well.

Transformation: We then applied label encoder to convert categorical labels into unique numerical values. Such as for every unique value like in Chemotherapy if patient has gone for chemo or not

iii) Combined Dataset: The combined dataset is obtained by merging the features' and labels' files to create a dataframe. Inner join is performed by taking "Patient_Id" as a primary key from both the files and "hugo_symbol" as a header from features file. The dataset thus obtained as a result of merging multiple data sources, may contain duplicates or near-duplicate instances. Deduplication is performed and these duplicate values are dropped

C. Feature Selection & Extraction

Feature Selection: With more features included from the dataset, the model becomes more complicated and can overfit the model or data may contain noise and potentially damage the design. The model will generalize better by eliminating such unimportant features. Selection of features is a strategy in which we pick the features in our data that most relate to the target variable. In other words, for the target variable, we choose the best predictors.

Feature Extraction: Extraction of features is a reduction in dimensionality that reduces the initial set of raw data to more manageable classes for processing. The extraction process is useful in reducing the number of resources needed for processing without sacrificing essential or relevant data. Feature Extraction is a process of conveying the given raw data into a set of instance points embedded in a standardized, distinctive and machine-understandable space.

The key difference between feature selection and extraction is that feature selection keeps a subset of the original features while feature extraction creates brand new ones. The following steps are involved in feature selection and extraction:

SelectKBest: This method selects features according to the k highest scores. It can be used for feature selection or dimensionality reduction on sample sets, either to improve estimators' accuracy scores or to boost their performance on a very high-dimensional dataset. It takes a parameter as a score function, which must apply to a pair (X, y). The score function must return an array of scores, one for each feature X[:, i] of X. SelectKBest then retains the first k features of X with the highest scores. We are using Chi2 as a score function and then based on the score assigned to features we are selecting the n_features features with the highest values for the test chi-squared statistic from X, which must contain only non-negative features such as boolean or frequencies (e.g., term counts in document classification), relative to the classes.

Over here, we have selected the top 20 features using chi2 as a score function, and hence, SelectKBest computes the chi2 statistic between each feature. Our features include gene_id and 20 such features are selected for each of the four labels(cellularity, os_status, chemotherapy and cancer_type_detailed). A smaller value predicts that the feature is independent of the label and large value means that the feature is non-randomly related to label, and so likely provides important information.

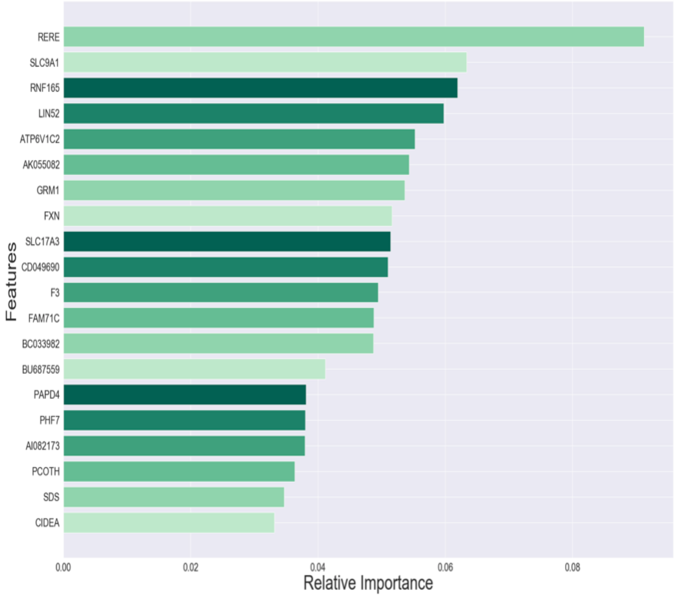
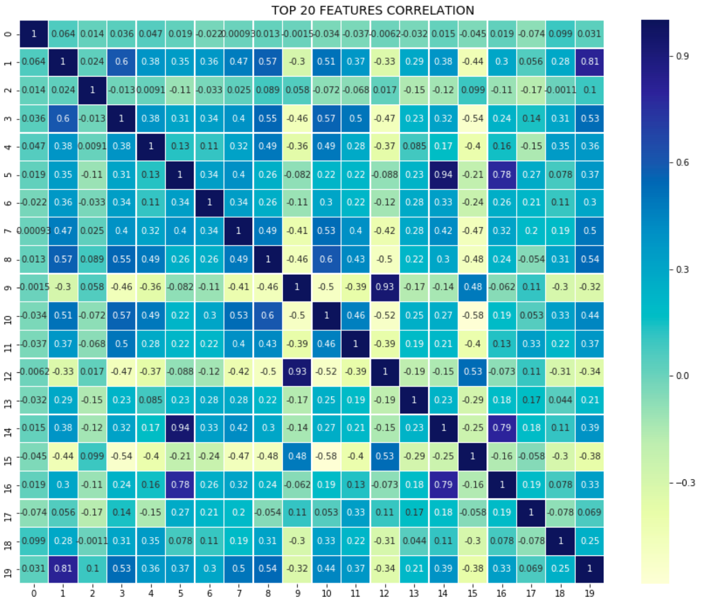


Fig 8. Top 20 Important Features in Dataset

Only k number of features are retained based on the scores and contribution in the model. We then represented the correlation between the top twenty features using a correlation heat map to visualize the relationship between features which are selected by selectKbest.

D. Normalization

Normalization is a technique that is often used as part of data training. The goal of normalization is to adjust numeric column values in the dataset to a standard level, without distorting variations in value ranges. It is needed when there are different ranges of features.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
0	1	0.064	0.014	0.036	0.047	0.019	-0.022	0.00093	0.013	-0.0015	-0.034	-0.037	-0.0062	-0.032	0.015	-0.045	0.019	-0.074	0.099	0.031
1	0.064	1	0.024	0.6	0.38	0.35	0.36	0.47	0.57	-0.3	0.51	0.37	-0.33	0.29	0.38	-0.44	0.3	0.056	0.28	0.81
2	0.014	0.024	1	-0.013	0.0091	-0.11	-0.033	0.025	0.089	0.058	-0.072	-0.068	0.017	-0.15	-0.12	0.099	-0.11	-0.17	-0.0011	0.1
3	0.036	0.6	-0.013	1	0.38	0.31	0.34	0.4	0.55	-0.46	0.57	0.5	-0.47	0.23	0.32	-0.54	0.24	0.14	0.31	0.53
4	0.047	0.38	0.0091	0.38	1	0.13	0.11	0.32	0.49	-0.36	0.49	0.28	-0.37	0.085	0.17	-0.4	0.16	-0.15	0.35	0.36
5	0.019	0.35	-0.11	0.31	0.13	1	0.34	0.4	0.26	-0.082	0.22	0.22	-0.088	0.23	0.94	-0.21	0.78	0.27	0.078	0.37
6	-0.022	0.36	-0.033	0.34	0.11	0.34	1	0.34	0.26	-0.11	0.3	0.22	-0.12	0.28	0.33	-0.24	0.26	0.21	0.11	0.3
7	0.00093	0.47	0.025	0.4	0.32	0.4	0.34	1	0.49	-0.41	0.53	0.4	-0.42	0.28	0.42	-0.47	0.32	0.2	0.19	0.5
8	0.013	0.57	0.089	0.55	0.49	0.26	0.26	0.49	1	-0.46	0.6	0.43	-0.5	0.22	0.3	-0.48	0.24	-0.054	0.31	0.54
9	-0.0015	-0.3	0.058	-0.46	-0.36	-0.082	-0.11	-0.41	-0.46	1	-0.5	-0.39	0.93	-0.17	-0.14	0.48	-0.062	0.11	-0.3	-0.32
10	-0.034	0.51	-0.072	0.57	0.49	0.22	0.3	0.53	0.6	-0.5	1	0.46	-0.52	0.25	0.27	-0.58	0.19	0.053	0.33	0.44
11	-0.037	0.37	-0.068	0.5	0.28	0.22	0.22	0.4	0.43	-0.39	0.46	1	-0.39	0.19	0.21	-0.4	0.13	0.33	0.22	0.37
12	-0.0062	-0.33	0.017	-0.47	-0.37	-0.088	-0.12	-0.42	-0.5	0.93	-0.52	-0.39	1	-0.19	-0.15	0.53	-0.073	0.11	-0.31	-0.34
13	-0.032	0.29	-0.15	0.23	0.085	0.23	0.28	0.28	0.22	-0.17	0.25	0.19	-0.19	1	0.23	-0.29	0.18	0.17	0.044	0.21
14	0.015	0.38	-0.12	0.32	0.17	0.94	0.33	0.42	0.3	-0.14	0.27	0.21	-0.15	0.23	1	-0.25	0.79	0.18	0.11	0.39
15	-0.045	-0.44	0.099	-0.54	-0.4	-0.21	-0.24	-0.47	-0.48	0.48	-0.58	-0.4	0.53	-0.29	-0.25	1	-0.16	-0.058	-0.3	-0.38
16	0.019	0.3	-0.11	0.24	0.16	0.78	0.26	0.32	0.24	-0.062	0.19	0.13	-0.073	0.18	0.79	-0.16	1	0.19	0.078	0.33
17	-0.074	0.056	-0.17	0.14	-0.15	0.27	0.21	0.2	-0.054	0.11	0.053	0.33	0.11	0.17	0.18	-0.058	0.19	1	-0.078	0.069
18	0.099	0.28	-0.0011	0.31	0.35	0.078	0.11	0.19	0.31	-0.3	0.33	0.22	-0.31	0.044	0.11	-0.3	0.078	-0.078	1	0.25
19	0.031	0.81	0.1	0.53	0.36	0.37	0.3	0.5	0.54	-0.32	0.44	0.37	-0.34	0.21	0.39	-0.38	0.33	0.069	0.25	1

Fig 9. Correlation Heatmap

Scaling: We have used Standard Scaler, which standardizes features by removing the mean and scaling to unit variance. Centering and scaling happen independently on each feature by computing the relevant statistics on the samples in the training set. Standard Scaler is normalizing the top twenty features which are extracted after the feature selection and extraction process using SelectKBest method. These features are scaled for the next step, where the model is trained for classification.

E. Data Splitting Architecture for Model classification

The model is prepared after normalizing the features extracted and the labels selected. The dataset is divided into training and testing phase for classification. The different algorithms are then applied on the training dataset to fit the model.

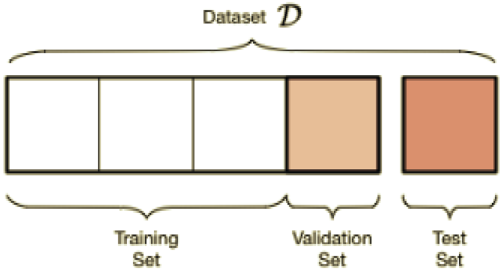


Fig 10. Data splitting

Training & Testing: We are using "train_test_split" method of scikit-learn to randomly pick the data to create the train and test split, which is desirable in real-world applications to avoid artifacts existing in the data preparation process. We are splitting the dataset in the ratio of 70% for training and 30% for testing, testing data will be kept unseen before applying oversampling on training data split for balancing the classes in the next step.

Validation set: In this phase, we are using a 10-fold cross-validation score to validate the training results for overfitting and underfitting.

F. Oversampling the Dataset

Oversampling and undersampling in data analysis are techniques used to adjust the class distribution of a dataset (i.e. the ratio between the different classes/categories represented).

If a class of data is the over-represented majority class, under-sampling may be used to balance it with the minority class. Under-sampling is used when the amount of collected data is sufficient. Conversely, when one class of data is the underrepresented minority class in the data sample, oversampling techniques may be used to duplicate these results for a more balanced amount of favourable results in training. Oversampling is used when the amount of data collected is insufficient.

Over here, we have applied oversampling to the dataset because we are reducing the features in the data set, thus giving the model fewer data to classify. Let's say, we have a data set of 10000 data and there are only 100 data points for class 1 while others are class 0. Now after performing under-sampling, we are reducing the data set to 1100 samples where 1000 fall under class 0 and 100 under class 1. So we are getting rid of almost 9000 samples and feeding it to the model. Hence, the model is more prone to error.

NOTE: Though different models use different methods to do under-sampling, the result is that we have less number of samples in the data set, and that is the reason we are using oversampling. We considered SMOTE and ADASYN techniques to oversample and chose the latter one for our model.

SMOTE first finds the n-nearest neighbours in the minority class for each of the samples in the class, and then it draws a line between the neighbours a generates random points on the lines which are not suitable for real-world applications. ADASYN is an improved version of SMOTE, in ADASYN after creating the samples it adds small random values to the points, thus making it more realistic. In other words, instead of all the samples which are linearly correlated to the parent, they have a little more variance in them, i.e. they are a bit scattered.

G. Dimensionality Reduction for Representation Learning

We have 24360 features in our dataset which is hard to visualize to get the insights and on top of that we have multi-class labels, in order to do so we are selecting top 20 features first then applying Principal Component Analysis (PCA) and Kernel-PCA (KPCA) to visualize the data in 1D and 2D form and also by rotating the axes to view data points from different perspective.

1). Cancer Type Detailed: This label is to predict the location of cancer in the breast, we have four classes in this label to predict the area affected by cancer.

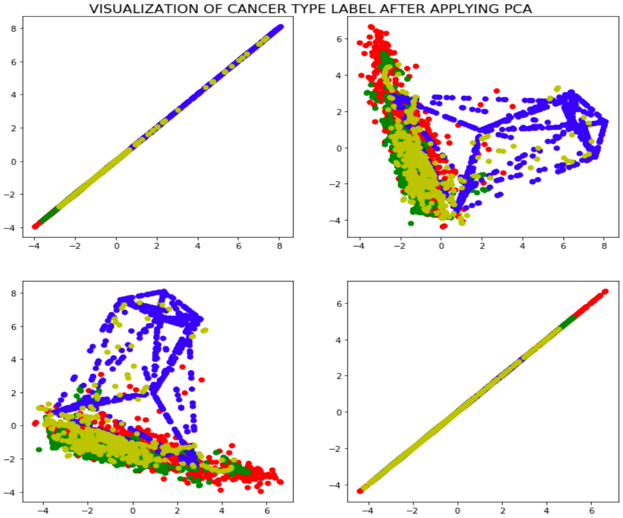


Fig 11. Visualization of Cancer Type using PCA

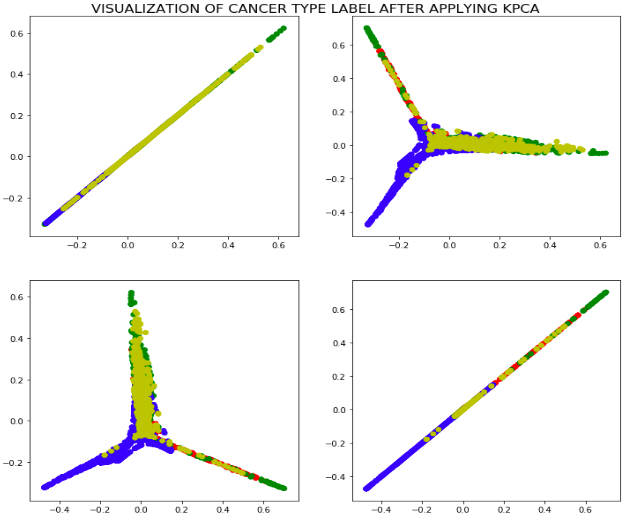


Fig 12. Visualization of Cancer Type using KPCA

2). Cellularity: Estimation of cancer cellularity is a critical task, which is conventionally achieved by manually reviewing the hematoxylin and eosin- (H&E-) stained microscopic slides of cancer sections. In this project, we develop an automatic and direct method to estimate cellularity and we have three classes in this label which explains the amount of area affected such as low, moderate and high.

3). Chemotherapy: Chemotherapy for breast cancer is used in addition to other treatments, such as surgery, radiation or hormone therapy. Receiving chemotherapy for breast cancer may increase the chance of a cure, decrease the risk of the cancer returning, alleviate symptoms from the cancer or help people with cancer live longer with a better quality of life [https://www.mayoclinic.org/tests-procedures/chemotherapy-for-breast-cancer/about/pac-20384931]. We have two classes in this label which explains if a patient has gone for Chemotherapy or not. If the cancer has recurred or spread, chemotherapy may control the breast cancer to help you live

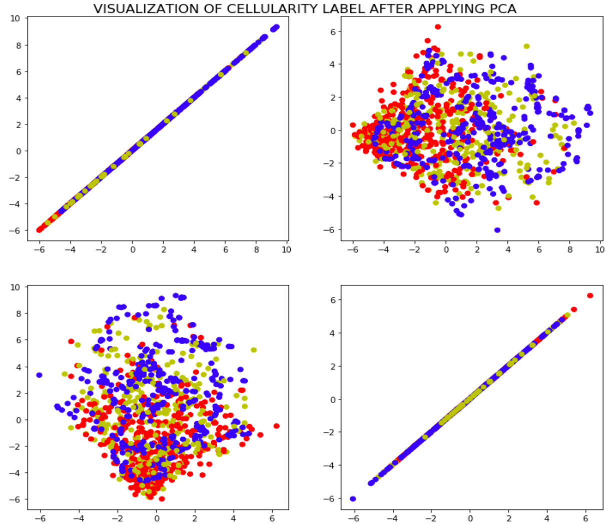


Fig 13. Visualization of Cellularity using PCA

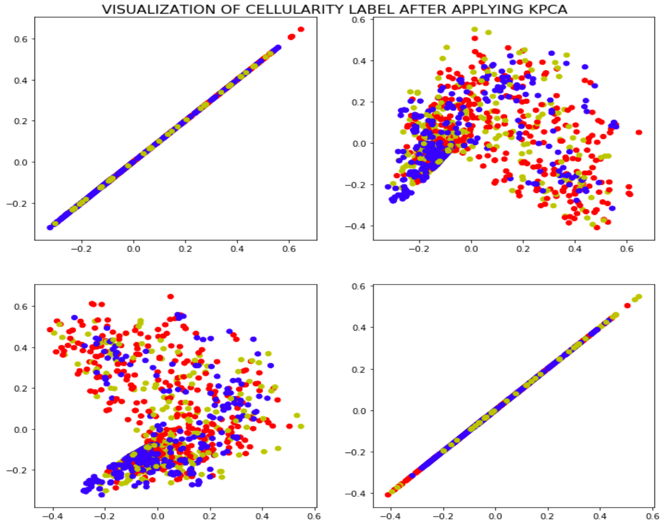


Fig 14.Visualization of Cellularity using KPCA

longer. Or it can help ease symptoms the cancer is causing.

VI. Algorithms & Techniques

1) **Gaussian Naive Bayes Classifier (GaussianNB):** The Bayesian Classification represents a supervised learning method as well as a statistical method for classification. The equation of Bayes' theorem can be stated as Gaussian Naive Bayes is an algorithm having a probabilistic approach which involves prior and posterior probability calculation of the classes in the dataset and the test data given a class respectively. Using the Bayes rule, the prior probability of belonging to each class can be learnt and estimated using the training data with ignoring the marginal probabilities based on the conditional probability of each variable X given the class label C. Likelihood: It is probability of a feature within a class. For example, if we want to calculate P(Chemotherapy — "Yes"), where Chemo is a feature, and "Yes" is a

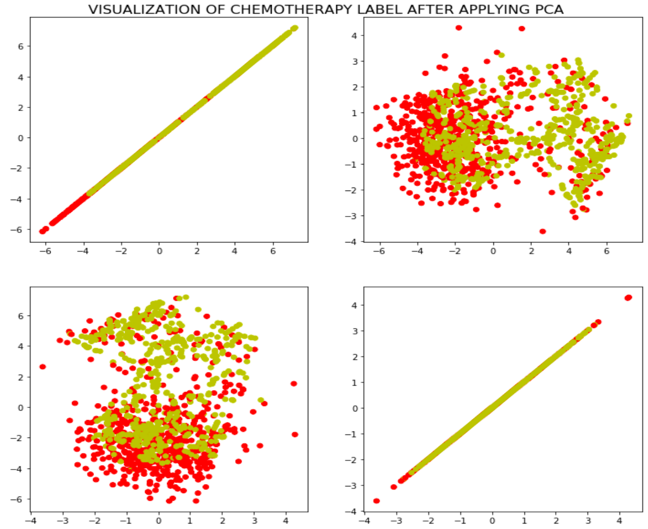


Fig 15. Visualization of Chemotherapy using PCA

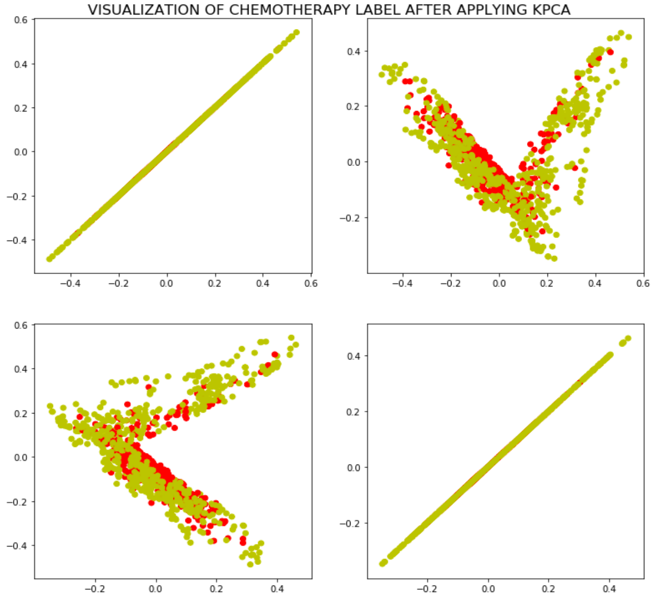


Fig 16. Visualization of Chemotherapy using KPCA

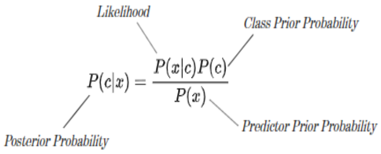

Likelihood
Class Prior Probability
$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$
Posterior Probability
Predictor Prior Probability

Fig 17. Bayes Theorem

class, we will count all ”Yes”es, or all times we went to suggest a patient for Chemo, (and ignore ”No”s) when Chemo is not recommended, divided by the overall observed days in our data set. Prior Likelihood or Class

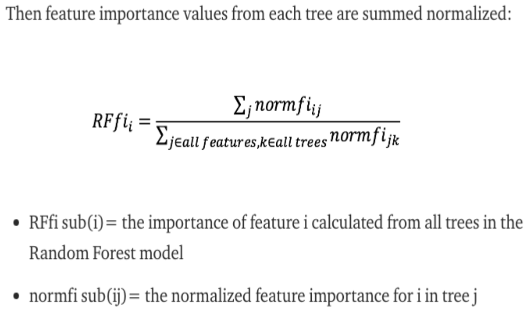
Prior Probability is a probability of a class. For example, if we have to calculate P(”No”), we will count number of all ”No”s, or, the when Chemo is not recommended, divided by the overall times Chemo is recommended in our data set.
Posterior probability is the probability of an event occurring after taking into consideration new samples.

2) **Logistic Regression:** Logistic regression is a classification algorithm used by a discrete set of classes to assign observations. Logistic regression transforms its output using the logistic sigmoid function to return a probability value. Logistic Regression uses a complex cost function, which can be defined as the ‘Sigmoid function’ or also known as the ‘logistic function’. In order to map predicted values to probabilities, Sigmoid function is used. The function maps any real value into another value between 0 and 1. The formula for Sigmoid function can be gives as: Logistic regression uses an equation as the representation. Input values (x) are combined linearly using weights or coefficient values (referred to as the Greek capital letter Beta) to predict an output value (y). Output value of the model is a binary value (0 or 1) rather than a numeric value.

$$y = e\hat{(}b0 + b1*x) / (1 + e\hat{(}b0 + b1*x))$$

Where y is the predicted output, b0 is the bias or intercept term and b1 is the coefficient for the single input value (x). Each column in your input data has an associated b coefficient (a constant real value) that must be learned from your training data.

3) **Random Forest Classifier:** Decision tree is a non-parametric model that can classify data based on a tree of decision rules. Random forests is an ensemble learning method that operates by constructing a multitude of decision trees at training time and outputs the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random forest classifier creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object. It can handle both categorical and numerical variables without much preprocessing. The low correlation between models is the key. Uncorrelated models can generate more accurate ensemble predictions than any of the individual predictions. The reason for this effect is that trees protect each other from their individual errors (as long as not all of them are constantly mistaken in the same direction) While some trees may be wrong, many other trees will be right, so the trees can move in the right direction as a group. The feature importance for each decision tree is normalized in random forest and then average out to remove the bias. The equation for it is as follows:



Then feature importance values from each tree are summed normalized:

$$RFfi_i = \frac{\sum_j normfi_{ij}}{\sum_{j\in all\ features, k\in all\ trees} normfi_{jk}}$$

- RFfi sub(i) = the importance of feature i calculated from all trees in the Random Forest model
- normfi sub(ij) = the normalized feature importance for i in tree j

Fig 18. Equation for Random Forest Classifier

4) **Gradient Boosting:** Gradient Boosting is a machine learning technique for regression and classification issues that generates a predictive model in the form of a set of weak predictive models, typically decision trees. It builds the model in a stage-wise fashion and it generalizes them by allowing optimization of an arbitrary differentiable loss function. So, the idea behind gradient boosting algorithm is to repeatedly leverage patterns and enhance a model with weak predictions and make it better. When we reach a stage where data doesn’t have any pattern that could be modeled, we can avoid modeling (otherwise it could lead to over-fitting). Algorithmically, we are minimizing our loss function, such that test loss reach its minima. The loss function used in gradient boosting is as follows:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

Fig 19. Loss Function for Gradient Boosting

5) **Extreme Gradient Boosting:** Extreme Gradient Boosting(XGBoost) is one of the implementations of Gradient Boost decision trees designed for speed and performance, but what makes XGBoost unique is that it uses a more regularized model formalization to control overfitting, which gives it better performance. XGBoost assigns positive and negative values to every decision made. XGBoost dominates structured or tabular datasets on classification and regression predictive modeling problems. Considering the potential loss to create a new branch for all possible splits (especially if we consider the case where there are thousands of features and hence thousands of possible splits), XGBoost solves this inefficiency by looking at the distribution of features across all data points in a leaf and by using this information, it reduces the search space for possible splits of features. XGBoost is used by IBM Watson to classify tumors.

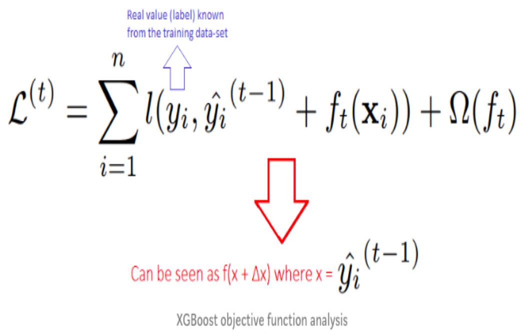

Real value (label) known from the training data-set
$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)) + \Omega(f_t)$
Can be seen as f(x + Δx) where x = $\hat{y}_i^{(t-1)}$
XGBoost objective function analysis

Fig 20. Objective Function for XgBoost Classifier

6) **Support Vector Machine:** Support Vector Machine (SVM) is a supervised machine learning algorithm which is used for classification problems. The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N is number of features) that distinctly classifies the data points. The main objective is to find a plane that has the maximum margin, i.e. the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence. Hyperplanes are decision boundaries that help classify the data points falling on either side of the hyperplane can be attributed to different classes. Learning of hyperplane is done by using linear algebra i.e. kernel, to linearly separate each variable. For high dimensional data other kernels are used and are of four types: Linear, Polynomial, RBF (Radial Basis function), Sigmoid. SVM minimizes the misclassification errors by maximizing the margin. We have used RBF which non-linearly separate the variables. For distance metric, squared Euclidean distance is used.

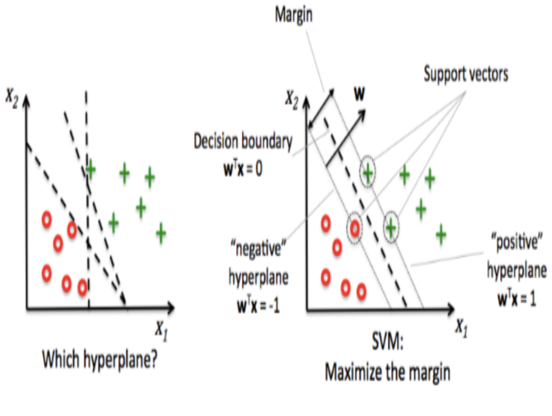


Fig 21. Support Vector Machine

VII. Experimental Setup

The Breast Cancer (BC) dataset is used in this project for carrying out the experiments based on our proposed methodologies. This dataset has been introduced, described, and analyzed in the paper by Pereira et al. 2016 in the journal Nature Communications. Additionally, the BC data set can also be downloaded from the Cancer Genomics Bio Portal by searching the name "Pereira et al. Commun 2016" and also from this link (http://www.cbioportal.org/data_sets.jsp). We have explained earlier in detail how we have combined the gene ids and labels from the dataset, which results in the final dataset. It has total of 24360 attributes (features) that first aids in predicting the precise location of the cancer in the breast from Cancer_Type_Detailed label, second in predicting the cellularity or number of tissues affected by the cancer, the total cancer cell number revealed a positive relationship with survival (Overall survival status), third in predicting if a patient should be recommended chemotherapy or not. Additionally, we have also predicted the overall survival status of a patient. Firstly, we are analyzing the dataset at depth for insights about the genes and labels which are of medical importance. We are using pandas profiling package to create profile reports of all the features in the Data Frame which calculates the count, mean, standard deviation and seventy-fifth percentile. Next, data cleaning and data processing is conducted while data cleaning the missing values are found, and it is observed that column unnamed has missing values for all entries and hence those field are removed. Also, the first field which has patient id is dropped as it has no relevant contribution to the prediction. After cleaning and combining the gene records with labels still contains duplicate records which are then also removed from the data frame.

After the cleaning of the dataset, first, we are applying SelectKBest with chi2 score function to select top twenty features (Gene ids or Hugo symbols) out of 24360 features. Secondly, features are then scaled using Standard Scaler library which standardizes features (Gene id's) because learning algorithm like RBF kernel of SVM assume that all features are centered around 0 and have variance in the same order. Third, for analyzing the correlation between twenty selected features, we are creating heat maps using seaborn, and for correlation, we are using the Pearson method to demonstrate the relation between all the attributes. Next, the dataset is split into training (70%) and testing (30%) sets. Our dataset contains imbalanced classes to overcome the partiality we are applying ADASYN method for oversampling on training split and to keep testing data split unseen. Then, all the classification algorithms, including logistic regression, random forest, SVM, Naive Bayes, XGBoost and Gradient boosting, is applied to the data set with all the twenty attributes. We then measure the performance of models based on their Accuracy, Precision, Sensitivity, Specificity, Negative Predictive value, number of misclassified samples and F1 score with respective four labels used in this project to find out the relevancy and accuracy of the models for a particular problem.

Initially, we were able to predict the location of cancer in the breast with an accuracy of 80%, to improve the accuracy we then combined scaling features, classification models and GridSearchCV together into pipeline, pipeline will scale the data first then run classifiers by selecting the best parameters from grid search to optimize the accuracy. In GridSearchCV we are passing all the parameters a model take as a input to estimate the best combination of parameters to achieve higher accuracy.

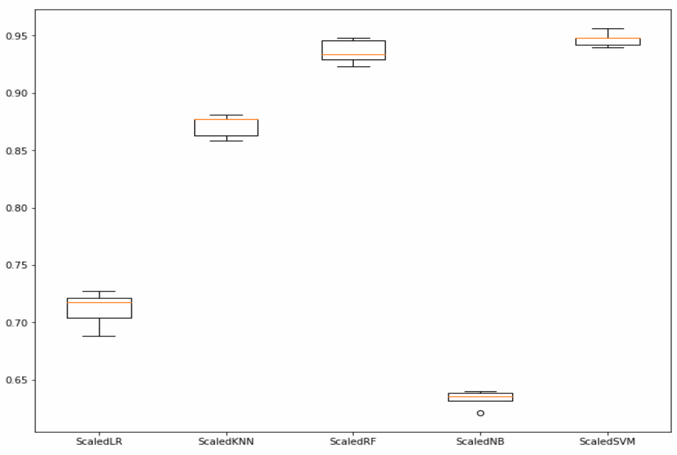


Fig 22. Performance after scaling

In the figure above we have tested which models are performing better once scaling is performed and from this observation it can be concluded that SVM, Random forest works well after scaling and then we further applied this comparison on XGBoost and optimized selected classifiers more to increase the efficiency. we are using various algorithms due to their specific advantages. For example, we are using random forest algorithm since it enables easier feature selection by processing just the top few nodes. Similarly, SVM is a learning algorithm that uses a linear function space hypothesis in a high-dimensional feature space which enables to classify non-linear data easily, and, Gradient Boosting (GB) is a combination of boosting method with gradient descent. GB is built by making a new model to predict errors from the previous model. Iteratively, a new model is added to fix the error from the previous model until no more fixes conducted. XGBoost is a more efficient and scalable GB version. XGBoost assigns positive and negative values to every decision made. XGBoost is used by IBM Watson to classify tumors. We have considered LR and Naive Bayes as a baseline methods to check how they perform before selecting typical models like XG and others. Finally, after predicting the type of the breast cancer with high accuracy we are also predicting the tissues affected by cancer i.e High, Low or Moderate, predicting whether patient should go for chemo or not (from the data analysis we found that 78% patients died who didn't go for chemo and there is 25% increase in survival rate if went for chemo) and lastly overall survival status of a patient to predict cancer is malignant or benign. From the results below we can conclude that XGBoost, Random forest and SVM is able to predict with high accuracy.

VIII. RESULTS

In this section, we have discussed the results obtained by applying different classifier algorithms on the feature column ”hugo_symbol” and multi-class labels like Cancer Type Detailed, Cellularity, Chemotherapy and OS Status. We have calculated the performance metrics like accuracy, sensitivity, specificity, precision, negative predictive value, F-1 Score and misclassified samples for all the labels. We have shown the confusion matrix of all the four labels for three classifiers that have top 3 performances(XGBoost, SVM and RF). The features have been classified into the number of labels accordingly where the diagonal in the confusion matrix represents the number of samples classified correctly for each class. For all the multi-class labels XGBoost classifier has performed the best because it has an immensely high predictive power which makes it the best choice for accuracy in events as it possesses both linear model and the tree learning algorithm, making the algorithm almost 10x faster than existing gradient booster techniques. Moreover, it does parameter tuning and deals with the irregularities of the data. It implements parallel processing through parallel gradient boosting decision trees. Also, XGBoost has built-in routines to handle missing values. XGBoost tries different things as it encounters a missing value on each node and learns which path to take for missing values in future. Naive Bayes Classifier has performed the worst for all the four multi-class labels because it considers Gaussian (normal) distribution of data as Naive Bayes classifier makes a strong assumption on the shape of data distribution, i.e. any two features are independent given the output class. Also, for any possible value of a feature, it estimates a likelihood value by a frequent approach this results in probabilities going towards 0 or 1, which in turn leads to numerical instabilities and worse results. Also, since the classes are imbalanced, it results in skew probabilities.

A. *CANCER TYPE DETAILED*

For this label, we have obtained cancer type detailed classified as Breast Invasive Ductal Carcinoma(class 0), Breast Invasive Lobular Carcinoma(class 1), Breast Mixed Ductal and Lobular Carcinoma(class 2) and Breast Invasive Mixed Mucinous Carcinoma(class 3). In Table, I, different performance metrics is calculated. We have got accuracy values like 99.10%, 96.13% and 95.83% for XGBoost, SVM and Random Forest classifiers respectively. Naive Bayes has performed the worst amongst the six of them because it has got the highest number of misclassified samples. XGBoost has performed the best and has given the highest values for all the performance metrics and hence can classify all the samples accurately with least misclassified samples. Sensitivity(Recall), which is a measure of the proportion of actual positive cases that got predicted as positive is the highest for XGBoost classifier, and the value is 95.0%. Precision, which is the positive predictive value or the fraction of the positive predictions that are positive, is obtained 100% for XGBoost. The other two classifiers like Gradient Boost and Logistic Regression have given moderate results while Naive Bayes has given the worst results for all the performance metrics calculated.

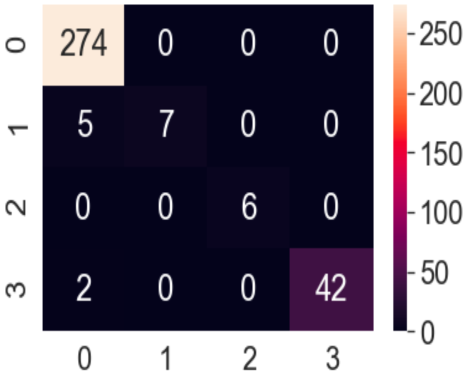


Fig 23. Confusion Matrix For Cancer Type Xgboost

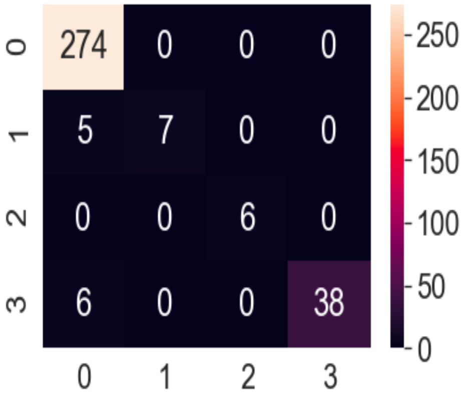


Fig 24. Confusion Matrix For Cancer Type Random

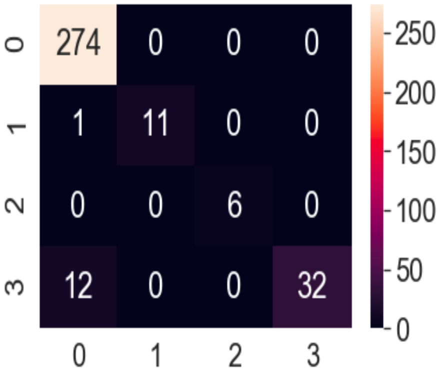


Fig 25. Confusion Matrix For Cancer Type SVM

We are displaying Confusion Matrix for three classifiers based on performance metrics that have classified the samples accurately with least misclassified samples. From the matrix of XGBoost, it is observed that it has classified 274 samples for class 0, 7 for class 1, 6 for class 2 and 42 for class 3. We have got 5 misclassified samples for class 1 and 2 for class 3. Similarly, we can see the confusion matrix for the same label for classifiers like Random Forest, and SVM respectively arranged in order of the number of misclassified samples. The confusion matrix of these three classifiers have been shown where XGBoost is performing the best and shown at first (Fig), Random Forest being the second(Fig) and SVM being the third(Fig).

B. CELLULARITY

For this label, we have obtained cellularity classified as high(class 0), low(class 1) and moderate(class 2). In Table II, all the values calculated for different performance metrics have been shown. We have got accuracy values like 99.70%, 98.52% and 86.68% for XGBoost, Random Forest and SVM classifiers respectively. Naive Bayes has performed the worst amongst the six of them because it has got 158 misclassified samples. XGBoost has performed the best and has given the highest values for all the performance metrics and hence can classify all the samples accurately with least misclassified samples. Sensitivity(Recall) for XGBoost classifier is 99.0%. Precision is obtained 100% for XGBoost and 48% for Naive Bayes. Over here, the performance metrics values of Random Forest and Gradient Boost are almost similar. The other classifier, Logistic Regression have given moderate results while Naive Bayes has given the worst results for all the performance metrics calculated.

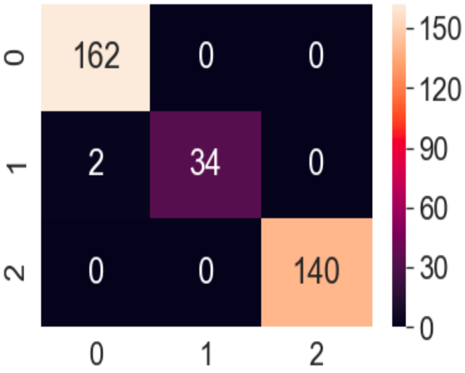


Fig 26. Confusion Matrix For Cellularity Xgboost

Confusion Matrix for 3 classifiers based on performance metrics have been shown that have classified the samples accurately with least misclassified samples. Here, we have shown the confusion matrix for XGBoost, Random Forest and SVM. Since the metrics for Gradient Boost and Random Forest is similar, we haven't considered confusion matrix for Gradient Boost because the working of gradient boost is almost similar to XGBoost with few changes. From the matrix of XGBoost, it can be observed that it has classified 162 samples for class 0, 34 for class 1 and 140 for class 3. We have got 2 misclassified samples for class 1. Similarly, we can see the confusion matrix for the same label for classifiers like Random Forest, and SVM respectively arranged in order of the number of misclassified samples. The confusion matrix of these 3 classifiers have been

Performance Metrics	SVM	Naive Bayes	Random Forest	Xgboost	Gradient	Logistic
Accuracy	96.13%	64.88%	96.72%	97.91%	95.53%	82.44%
Sensitivity(Recall)	91.0%	67.0%	81.0%	95.0%	81.0%	47.0%
Specificity(True Negative Rate)	79.03%	76.47%	77.41%	95.16%	76.19%	14.28%
Precision(Positive Predictive Value)	99.0%	52.0%	99.0%	100.0%	98.0%	52.0%
Negative Predictive Value	100.0%	35.86%	100.0%	100.0%	100.0%	69.23%
F-1 Score(Average of Precision and Sensitivity)	94.0%	56.0%	87.0%	97.0%	88.0%	45.0%
Miss-classified Samples	13	118	11	7	15	59

TABLE I
RESULTS OF PREDICTING BREAST CANCER TYPE

Performance Metrics	SVM	Naive Bayes	Random Forest	Xgboost	Gradient	Logistic
Accuracy	86.68%	53.25%	98.52%	99.70%	98.52%	65.68%
Sensitivity(Recall)	83.0%	55.0%	97.0%	99.0%	99.0%	55.0%
Specificity(True Negative Rate)	82.38%	61.36%	97.15%	99.43%	98.88%	65.90%
Precision(Positive Predictive Value)	91.0%	48%	99.0%	100.0%	99.0%	70.0%
Negative Predictive Value	93.54%	72.97%	100.0%	100.0%	98.30%	78.37%
F-1 Score(Average of Precision and Sensitivity)	86.0%	47.0%	98.0%	99.0%	99.0%	57.0%
Miss-classified Samples	45	158	5	2	5	116

TABLE II
RESULTS OF PREDICTING CELLULARITY TYPE

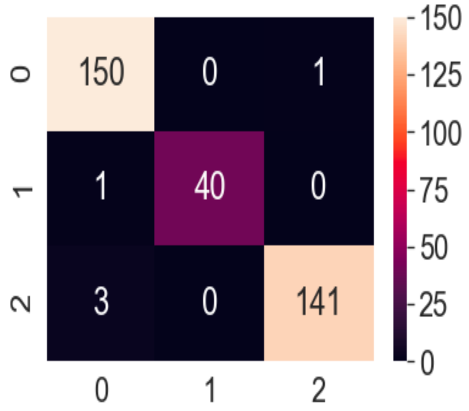


Fig 27. Confusion Matrix For Cellularity Random

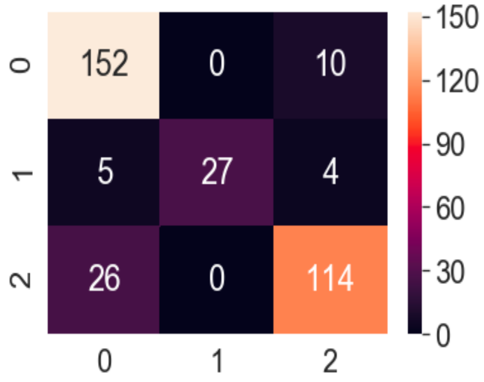


Fig 28. Confusion Matrix For Cellularity SVM

shown where XGBoost is performing the best and shown at first (Fig), Random Forest being the second(Fig) and SVM being the third(Fig).

C. CHEMOTHERAPY

For this label, we have obtained chemotherapy classified as yes(class 1) that represent that the patient has gone through the treatment and no(class 0) representing that the patient has not gone through the treatment. In Table III, all the values calculated for different performance metrics have been shown. We have got accuracy values like 99.5%, 99% and 98.21% for XGBoost, Random Forest and SVM classifiers respectively. Naive Bayes has performed the worst amongst the six of them because it has got 72 misclassified samples. XGBoost has performed the best and has given 0 misclassified samples, and it has the highest values for all the performance metrics and hence can classify all the samples accurately. Sensitivity(Recall) is 100% for XGBoost, Gradient Boost and Random Forest. Precision is obtained 100% for XGBoost. In this section, XGBoost can correctly classify that if the patient has undergone chemotherapy treatment or not.

Confusion Matrix for 3 classifiers based on performance metrics have been shown that have classified the samples accurately with least misclassified samples. From the matrix of XGBoost, it can be observed that it has classified 277 samples for class 0, 61 for class 1. We have got 0 misclassified samples. Similarly, we can see the confusion matrix for the same label for classifiers like Random Forest, and SVM respectively arranged in order of the number of misclassified samples. The confusion matrix of only these 3 classifiers have been shown where XGBoost is performing the best and shown at first (Fig), Random Forest being the second(Fig) and SVM being the third(Fig).

Performance Metrics	SVM	Naive Bayes	Random Forest	Xgboost	Gradient	Logistic
Accuracy	98.21%	78.69%	99%	99.5%	98.52%	84.61%
Sensitivity(Recall)	99.0%	82.0%	100.0%	100.0%	100.0%	94.0%
Specificity(True Negative Rate)	92%	62%	92%	99.10%	100.0%	94.0%
Precision(Positive Predictive Value)	98.0%	91.0%	98.0%	100.0%	98.0%	88.0%
Negative Predictive Value	95.0%	44.0%	100.0%	100.0%	99.40%	61.0%
F-1 Score(Average of Precision and Sensitivity)	99.0%	86.0%	98.0%	99.0%	99.0%	91.0%
Miss-classified Samples	6	72	4	0	5	52

TABLE III
RESULTS OF PREDICTING CHEMOTHERAPY

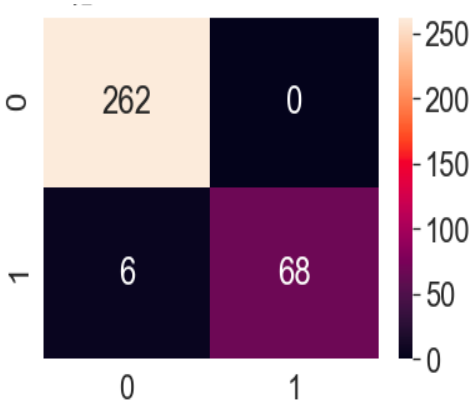


Fig 29. Confusion Matrix For Chemotherapy SVM

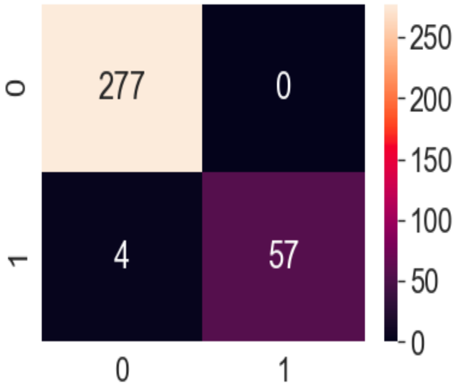


Fig 30. Confusion Matrix For Chemotherapy Random

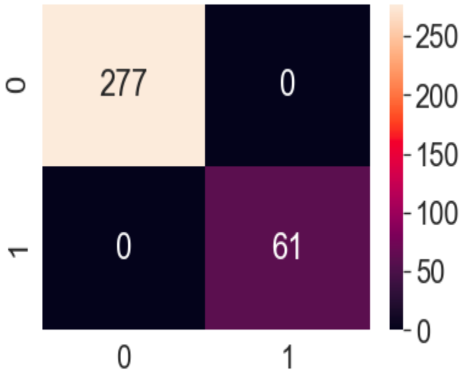


Fig 31. Confusion Matrix For Chemotherapy Xgboost

D. Overall Survival Status

For this label, we have obtained an overall survival status classified as deceased(class 0) and living(class 1). In Table IV, all the values calculated for different performance metrics have been shown. We have got accuracy values like 99.70%, 99% and 94% for XGBoost, Random Forest classifiers and SVM respectively. Naive Bayes has performed the worst amongst the six of them because it has given only 65.6% accuracy while XGBoost has performed the best and has given the highest values for all the performance metrics and hence can classify all the samples accurately with 0 misclassified samples. Sensitivity(Recall) is 100% for XGBoost, and precision is obtained 99.20% for the same. The other classifiers like Gradient Boost and has given moderate results while Logistic Regression has given poor results in comparison to others but a little better in comparison to Naive Bayes that has given worst results for all the performance metrics calculated.

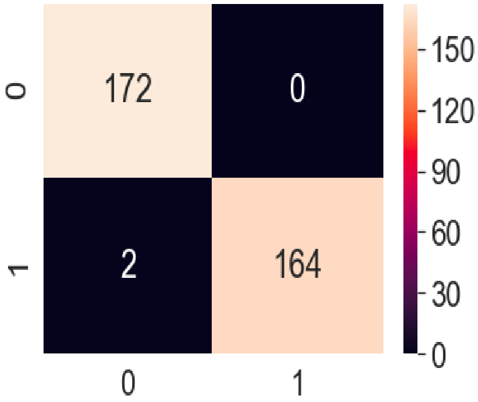


Fig 32. Confusion Matrix For Chemotherapy SVM

Confusion Matrix for 3 classifiers based on performance metrics have been shown that have classified the samples accurately with least misclassified samples. From the matrix of XGBoost, it can be observed that it has classified 172 samples for class 0, 166 for class 1. We have got 0 misclassified samples. Similarly, we can see the confusion matrix for the same label for classifiers like Random Forest, and SVM respectively arranged in order of the number of misclassified samples. The confusion matrix of only top 3 classifiers has been shown. Where XGBoost is performing the best and shown at first (Fig), Random Forest being the second(Fig) and SVM being the third(Fig).

Performance Metrics	SVM	Naive Bayes	Random Forest	Xgboost	Gradient	Logistic
Accuracy	94%	65.6%	99%	99.70%	90.53%	67.75%
Sensitivity(Recall)	98.0%	70.0%	98.0%	100.0%	92.0%	72.0%
Specificity(True Negative Rate)	91%	61%	99%	100.0%	89.0%	64.0%
Precision(Positive Predictive Value)	92.0%	65.0%	99.0%	99.20%	89.0%	67.0%
Negative Predictive Value	97.0%	66.0%	100.0%	100.0%	92.0%	68.0%
F-1 Score(Average of Precision and Sensitivity)	95.0%	68.0%	99.0%	99.34%	91.0%	69.0%
Miss-classified Samples	2	116	1	0	32	109

TABLE IV
RESULTS OF PREDICTING OS STATUS

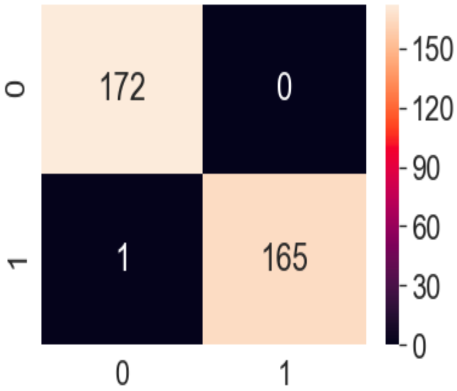


Fig 33. Confusion Matrix For Chemotherapy Random

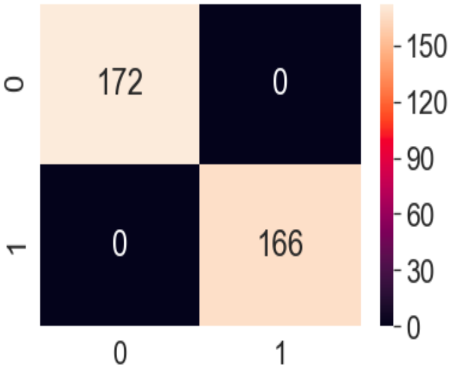


Fig 34. Confusion Matrix For Chemotherapy XGBoost

IX. Ethical, Legal and Societal Aspects of Machine Learning in Breast Cancer Analysis

Machine Learning is an emerging and a broad area and the data used by the machine learning algorithms is also huge. Machine learning algorithms can be trained to give accurate results and humans do not have to invest much time in making decisions. We have considered the principles encoded in algorithms, the need to interpret performance, bias and demonstrability problems, data ownership, privacy and consent, and legal, ethical and professional responsibility. We have also considered potential consequences for patients, including confidence in healthcare, and provide reasons for the apparent rush to introduce machine learning solutions for some social science. Breast cancer treatment is a leading research problem in the technology field, with applications including screening and diagnosis, risk assessment, prognosis and clinical decision-making support, management planning, and precision medicine. We study these innovations' ethical, legal and social consequences. The following figure summarises drivers, risks, solutions and desired outcomes.

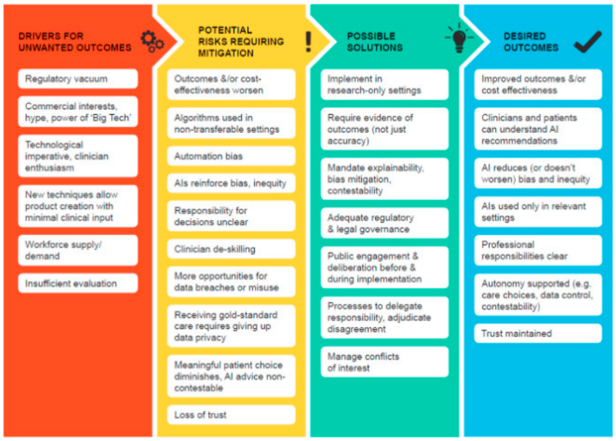

DRIVERS FOR UNWANTED OUTCOMES
Regulatory vacuum
Commercial interests, hype, power of 'Big Tech'
Technological imperative, clinician enthusiasm
New techniques allow product creation with minimal clinical input
Workforce supply/ demand
Insufficient evaluation
POTENTIAL RISKS REQUIRING MITIGATION
Outcomes &/or cost-effectiveness worsen
Algorithms used in non-transferable settings
Automation bias
AIs reinforce bias, inequity
Responsibility for decisions unclear
Clinician de-skilling
More opportunities for data breaches or misuse
Receiving gold-standard care requires giving up data privacy
Meaningful patient choice diminishes, AI advice non-contestable
Loss of trust
POSSIBLE SOLUTIONS
Implement in research-only settings
Require evidence of outcomes (not just accuracy)
Mandate explainability, bias mitigation, contestability
Adequate regulatory & legal governance
Public engagement & deliberation before & during implementation
Processes to delegate responsibility, adjudicate disagreement
Manage conflicts of interest
DESIRED OUTCOMES
Improved outcomes &/or cost effectiveness
Clinicians and patients can understand AI recommendations
AI reduces (or doesn't worsen) bias and inequity
AIs used only in relevant settings
Professional responsibilities clear
Autonomy supported (e.g. care choices, data control, contestability)
Trust maintained

Fig 35. Summarises Drivers, Risks, Solutions and Desired Outcomes

A. Ethical and Moral Aspects

In AI for Good Global Summit in Geneva, 2019 by Wendell Wallach, there are four primary values in every AI/ML ethics list: 1) Privacy, 2) Accountability, 3) Fairness (minimizing bias), and 4) Transparency. The debate about medical Machine learning (ML) often features concern that human clinicians (particularly diagnosticians such as radiologists) will become redundant and future decision making is likely to involve both clinicians and ML systems in some way, requiring management of machine-human disagreement and delegation of responsibility for decisions and errors. ML potentially disrupts traditional conceptions of professional medical responsibility. In the traditional approach, If doctors are directly involved in patient care, but decisions depend on non-explainable AI/ML recommendations, doctors will face challenges regarding their moral and legal responsibility. Doctors may be expected to take responsibility for decisions which they cannot control or explain; if they decline, it is not clear to whom responsibility should be delegated. In responsibilities, unlike healthcare professionals, developers are not required to put the interest of patients first.

In our project, to collect and to ensure the reliability of the dataset, we have downloaded data from cBioportal, which is Cancer Genomics repository. The patient id is anonymized by cBioportal already before we used the dataset for our project, and we cannot use patient id to track people without their consent. Before putting data into models, we have removed the patient ID from the data frame to ensure privacy. The classifiers are open-sourced and available in scikit in python this means that everybody has access to it and some even allow us to modify it and make a profit. We have a complete dataset before putting into a classifier (i.e., a dataset that accurately consolidates a good number of all possible cases) so that our resulting ML system is not biased and it cannot discriminate. To ensure fairness in accuracy and results, we have not applied oversampling on test data to keep it unseen and tested the accuracy on test dataset to make an accountable classifier.

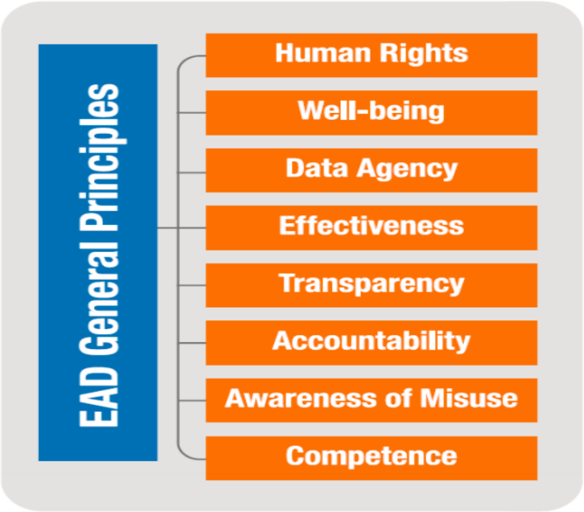


Fig 36. EAD Principles

B. Legal Aspects

Due to the rapid implementation of machine learning into healthcare, we actually have no clear regulator, no clear trial process and no clear accountability trail. Virtually, no courts have established guidelines relevant to those who should be held legally liable if any machine learning algorithm causes harm, and the usually voluble world of legal scholarship has been especially silent. Since ML algorithms require large amounts of good quality data to train and test, ownership and consent to the use and security of that data are critical issues. This is compounded by the rapid acceleration of large technology firms like IBM and Google into healthcare and the proliferation of breast cancer-related start-ups. Regardless of the specifics behind these ideas, it is important to maintain a system of transparency and comprehension surrounding machine learning models, during which the data is scrutinized throughout all stages of the machine learning process to ensure fair practices that do not perpetuate biases

In our work, for the Breast Cancer Analysis, the methods that are used to build models give the results based on the data fed. The data classification and the results obtained are according to the features. The predictions and the outcomes by the algorithms are solely concerned to provide the results for the ease of the hard work of the doctors and not determine the entire decision without an expertise advice. Let's say, if a patient is not required to have a chemotherapy treatment but the algorithm predicts an outcome for the treatment to be done, it should still be carried out with a doctor's guidance. Our work is not creating any legal harm to the patients as the approaches taken into account are dependent on the given patients' gene data from cBioPortal and all the classifiers are open-sourced and we did not use any of the patented softwares in our project.

C. Societal Aspects

In the coming years, ML will take over the entire manual work of humans and the society through all the automation techniques that will be implemented. Using algorithms to evaluate and cross-reference symptoms to databases containing millions of other cases and diseases has resulted in quicker identification of infection and illness, saving lives through faster care, and reducing the time doctors spends in the health system. Hospitals are currently using ML algorithms to identify tumours more accurately in different scans and to analyse various cancer types, and machine learning is being applied to accelerate research into cancer cure. The ease of ML algorithms might have people to lose their jobs because of the automation techniques and doctors will become lazy because all of the results are predicted by machines. The researchers might not be having any unique ideas and might lose their ability to think.

In our project, the methodologies used do not have any impact on the jobs because we are making the tasks of the doctors easy by automating the prediction of type of cancer in breast and this will help the doctors to work efficiently and give appropriate treatment to the patients without wasting any further time. We are also automating the prediction of number of tissues affected called cellularity in breast cancer which can help doctors to suggest patients for radio therapy or chemotherapy when level of cellularity is high to save the patients with early treatment. Instead of worrying about people losing jobs we can train AL/ML to hospitals, doctors and staff to treat and cure millions of women efficiently.

X. CONCLUSION & FUTURE WORKS

This project attempts to solve the problem of automatic detection of the type of breast cancer, amount of tissues affected by cancer called cellularity, the requirement for chemotherapy and overall survival status of the patient using machine learning algorithms. We are using five different algorithms on breast cancer dataset. This work makes an effort to predict the type of breast cancer in its early stages and cellularity which can be further used to recommend chemotherapy if the number of tissues affected is high or moderate in order to increase the survival of patients. In the first phase, we proved that the three most popular algorithms RF, XGBoost and SVM can achieve high performance after effective scaling. The second phase conducted focus on combining feature selection method to select the best set of genes as a feature to improve the accuracy performance and remove redundant features. It is advantageous to exclude less relevant features as it is beneficial in the face of computational cost. Finally, in the last, we deduced how to combine scaling automatically, sampling of classes and optimized parameter selection of models of the machine learning supervised classifiers. The proposed algorithms selected the best parameters among the various configurations. With 20 features, SVM, Random Forest, and XGBoost yield the best results with a precision score of more than 96%, while XGBoost performs the best with 20 features with a precision score of 98% for predicting cancer type, 99% for cellularity, 99.5% for chemotherapy and 99.7% for overall survival status. This project shows that prediction algorithms with feature selection outperform those without feature selection in all cases. All the experiments are performed using Python libraries. The high level of reported cases in breast cancer care in AI and ML development creates both opportunities and responsibilities. In our project we identified that if we can predict the type of cancer first and then on the basis of that we can further classify the amount of cellularity and further on the basis of that if we can suggest patient for chemotherapy or not. In simple words, if we can automate the multiple label prediction it will be very beneficial for doctors.

XI. CONTRIBUTION

Mayank Semwal The documentation of project report w.r.t (abstract, oversampling, dimensionality reduction for representation learning, experimental setup, conclusion and ethics in machine learning), dataset exploration and deciding labels for classification and cleaning, research on classifiers, feature selection using chi2, evaluation metrics (Validation score, misclassified samples, sensitivity, F1-score), ADASYN, implementation of classifiers using pipeline (XGBoost, SVM), scaling results, parameter tuning and optimization for better accuracy.

Priyanka Motwani The documentation of report w.r.t (abstract, introduction, data inspection, feature selection and extraction, normalization, theory about algorithms and techniques and legal in machine learning), Study of related works based on their proposed approaches and comparison, dataset exploration and splitting, research on papers on breast cancer in medicine and machine learning, feature extraction from dataset, evaluation metrics (correlation matrix, NPV, PPV), comparison of smote and adasyn for accuracy, classifiers (Logistic regression, Random forest), parameter tuning.

Vipul Malhotra The documentation of project report w.r.t (abstract, results of performance by our classifiers like accuracy, sensitivity, specificity, NPV, PPV and misclassified samples, confusion matrix, formulas of classifiers), dataset exploration, cleaning and combining, research on classifiers, feature selection using f_classification, evaluation metrics (accuracy, confusion matrix, precision, Recall), SMOTE, implementation of classifiers (gradient boosting, naive Bayes), parameter tuning for better accuracy.

REFERENCES

[1] Abreu, P. H., Santos, M. S., Abreu, M. H., Andrade, B., Silva, D. C. (2016). Predicting Breast Cancer Recurrence Using Machine Learning Techniques. ACM Computing Surveys, 49(3), 1–40.

[2] Al-Shargabi, B., Al-Shami, F. (2019). An experimental study for breast cancer prediction algorithms. Proceedings of the Second International Conference on Data Science, E-Learning and Information Systems - DATA 19.

[3] Breast Cancer(METABRIC, Nature 2012 & NAT Commum 2016) - https://www.cbioportal.org/study/summary?id=brca$_metabric$

[4] Gayathri, B. M., Sumathi, C. P. (2016). Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer. 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC).

[5] Goel, V. (2018, October 12). Building a Simple Machine Learning Model on Breast Cancer Data. Retrieved from https://towardsdatascience.com/building-a-simple-machine-learning-model-on-breast-cancer-data-eca4b3b99fa3.

[6] Leventis, D. (2019, July 11). XGBoost Mathematics Explained. Retrieved from https://towardsdatascience.com/xgboost-mathematics-explained-58262530904a.

[7] Lg, A., At, E. (2013). Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence. Journal of Health Medical Informatics, 04(02).

[8] Lindqvist, N., Price, T. (2018). Evaluation of Feature Selection Methods for Machine Learning Classification of Breast Cancer (Dissertation)

[9] Mushtaq, Z., Yaqub, A., Hassan, A., Su, S. F. (2019). Performance Analysis of Supervised Classifiers Using PCA Based Techniques on Breast Cancer. 2019 International Conference on Engineering and Emerging Technologies (ICEET).

[10] Prateek. (2019). Breast Cancer Prediction: Importance of Feature Selection. Advances in Intelligent Systems and Computing Advances in Computer Communication and Computational Sciences, 733–742.

[11] What is breast cancer? - Canadian Cancer Society. (n.d.). Retrieved from https://www.cancer.ca/en/cancer-information/cancer-type/breast/breast-cancer/?region=on.

[12] Jamal, A., Handayani, A., Septiandri, A. A., Ripmiatin, E., Effendi, Y. (2018). Dimensionality Reduction using PCA and K-Means Clustering for Breast Cancer Prediction. Lontar Komputer : Jurnal Ilmiah Teknologi Informasi, 192–201.

[13] https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/