National College of Ireland

Machine Learning Applications in Predicting Breast Cancer Survival Using Gene Information

MSc Research Project
Msc Data Analytics

Sharath Kasaraghatta Thimmaraya Gowda
Student ID: x20117507

School of Computing
National College of Ireland

Supervisor: Dr.Rashmi Gupta

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Sharath Kasaraghatta Thimmaraya Gowda
Student ID:	x20117507
Programme:	Msc Data Analytics
Year:	2021
Module:	MSc Research Project
Supervisor:	Dr.Rashmi Gupta
Submission Due Date:	16th August 2021
Project Title:	Machine Learning Applications in Predicting Breast Cancer Survival Using Gene Information
Word Count:	6261
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Sharath Kasaraghatta Thimmaraya Gowda
Date:	16th August 2021

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Machine Learning Applications in Predicting Breast Cancer Survival Using Gene Information

Sharath Kasaraghatta Thimmaraya Gowda
x20117507

Abstract

Gene Expression is a computation concept that replicates human thinking and logic by distributing the information at various levels of precision. Implementing Gene and therapies in the prediction of breast cancer survivability using gene expression through designing of a system enhances the machine learning performance. The suggested method uses data granulation to train and test the classifier at various hierarchies. There are two aspects to the project presented, one where breast cancer histopathological image dataset is used to predict breast cancer with the image dataset then the second dataset, also the main objective of the project, to use the METABRIC breast cancer gene dataset to predict the survival of breast cancer patients using the gene information. To appraise the performance of the planned method, a PCA(Dimensionality Reduction on Gene) system is used based on the preferred technique such as Multilayer Perceptron, K Nearest Neighbours , Support Vector Classifier, and Tensorflow Boosted Estimator's Method. Two different health datasets are applied in the comparative analysis to evaluate the performance of the preferred approach to various classifiers. Multiple Instance learning is the novel approach carried out on Breast cancer prediction using Histopathological images, while Tensorflow Boosted estimators is the novel approach performed on Breast cancer Survivability. Results show that the proposed method improves the performance of the classification and generates improved level of prediction accuracy than the Convolutional Neural Network and Artificial Neural Networks. The accuracy of Attention based multiple instance learning for breast cancer prediction was about 91.04% while the accuracy of survivability of breast cancer patients using gene information is 100%.

1 Introduction

Breast cancer affected 2.3 million people globally in 2020, with over 685000 fatalities. By 2020, 7.8 million women diagnosed with breast cancer in the previous five years are still alive, making breast cancer the most common disease.In the 1980s, survival rates improved in countries where early detection systems were combined with a variety of treatment choices to eliminate invasive disease[1].

Classification of data plays a vital role in the medical trials, diagnoses and decision-making processes.Despite the fact that surgical procedures and medication regimens for treating breast cancer are continually improving, individual patient clinical results remain

[1] https://www.who.int/news-room/fact-sheets/detail/breast-cancer

difficult to predict due to a number of clinically related variables(Clough et al.; 2010).The quest for better prognosis and therapeutic outcome prediction has prompted extensive research on alternative biomarkers based on Bbreast cancer molecular profiling, as well as innovative prediction models and algorithms, which may overcome the inherent limits of earlier methods. High-throughput sequencing technologies, in particular, have been critical to the success of this new strategy.

Breast cancer starts from breast tissue, which is detected by a lump in the breast, and there are certain alterations under normal circumstances. Breast cancer tumors are classified into two types: benign (noncancerous) and malignant (cancerous) (Cancerous).Various practice techniques, such as self-examination, clinical examination, ,screening equipment with mammography, have been used to identify breast cancer. Breast cancer detection may be difficult when it is said that it is a mix of several diseases rather than a single disease(Giri and Saravanakumar; 2017).A biopsy is a diagnostic technique in Histopathology that can identify whether or not a suspicious region is malignant. The pathologist's diagnosis is made by visually inspecting histopathological pictures under a microscope, which is considered as the confirming gold standard for diagnosis(Gurcan et al.; 2009).The combined system will support decision making during the process of diagnosing of breast cancer and controlling the risk of any patient suffering from breast cancer through the predictions.

1.1 Motivation

This paper is motivated by the existence of data for use in the classification, and the availability of several other tests in the past literature for computational analysis and comparison. Intelligent methods have been applied in the medical field for data clustering such as Boosted tree estimators, decision trees and support vector machines (SVM), Convolutional Neural Networks(CNN), Multilayer perceptron etc. If the data is processed efficiently, it may be utilized to improve treatment efficacy and prognosis by leveraging accurate diagnostic procedures(López-García et al.; 2020). These facts motivates us to make build models using the deep learning and machine learning algorithms to predict the patients having breast cancer using the Hstopathological images and then predicting the survival of breast cancer using the gene information.

1.2 Research Question and Objectives

The research project has two research questions to answer:

1. *"How much improvement can be done in predicting the breast cancer using the Histopathology images using Multiple instance learning, Naive Bayes and Support vector Classifiers against the Convolutional Neural Networks?"*

2. *"How much improvement can be done in predicting the survival of breast cancer using Tensorflow boosted tree estimators and Multilayer Perceptron on Artificial Neural Network?"*

1.2.1 Research Objective

- Create a new model based on modifications to the current model that overcomes the limits of the old model.

- Perform visualizations to see the insights on the data.
- Evaluate the performance of the state-of-the-art model, modified model, and deep learning model using metrics such as sensitivity, specificity, accuracy, classification report, and so on.

1.3 Flow of the Paper

Section 2 explains the related work done over time. Section 2.1 debriefs on the work done on predicting breast cancer using histopathological images as the source and Section 2.2. Section 3 briefs the Methodology involved in the performing the analysis in the research.

Data collection, data pre-propressing, and models implemented to investigate the dynamics of breast cancer detection and predicting the survivability are explained in detail under the subsections of Section 3. Design Specification and Implementation are explained in Section 4. In section 5, the outcomes and evaluation to compare the suggested models' results are debriefed and tabulated. Finally, the work is concluded and discussion is suggested in Section 6.

2 Related Work

2.1 Dataset 1: Breast Cancer prediction using Histopathology Images

2.1.1 Approach

Several techniques have been used to demonstrate the use of data classification in predictive models. For example, Ghasem Ahmad et al. (2013) has applied a combination of algorithms such as Naïve Bayes, Decision Tree, and neural network for analyzing medical datasets. There are many features in the combination and a lot of time is consumed in the prediction. Therefore, the process of reducing the features is necessary through strategic selection. Bazazeh and Shubair (2016) used decision tree as well as neural networks. Additionally, Bektaş and Babur (2016) used a set of algorithms including decision tree, K nearest neighbor, CNN network and Naïve Bayes algorithm to predict the occurrence of Breast Cancer and implemented data mining approaches for detecting the risk rating of Breast Cancer. On the other hand, Nghe et al. (2007) applied Random Forest, ANN, and Genetic algorithm in the predictive model. Associative grouping is another efficient way to integrates association protocol mining and clustering to a predictive model achieving high accuracy.

With the innovations in image processing and deep learning techniques and tools, scientists have continuously developed algorithms to aid in the reduction in the amount of work involved around pathology and maximize the analytical efficiency by computerizing the traditional approaches through the application of CAD systems. For example, nuclei morphological test is done to for the classification of tissues as either benign or malignant. Raczkowski et al. (2019) applied hand-made elements, like morphological characteristics for training and testing a classification model on approximately 500 distinct images and attained an accuracy measure ranging between 83% and 92%. Gupta and Chawla (2020) contributed to the refinement of Histopatology images dataset by designing segmentation with watershed. The accuracy achieved in this case was between 72.85% to 96.17%. On

the other hand, the structure of the tissue configuration was used to perform clustering of histolopatology images dataset. Gupta and Chawla (2020) collected spatial color intensity graphs for separation of the epithelial layers and the resulting classification was trained to read statistical texture characteristics and reported between 75% and 72% accuracy. Belsare and Mushrif (2011) operated on three-group clustering method for breast histology dataset and grouped the histopatology image datasets in a traditional, aggressive, and non- aggressive carcinoma. Gupta and Chawla (2020) applied a linked elements approach for testing and training of an SVM classifier on a set of binary images. This attained a 92% to 97.7%. The clustering technique was further cascaded to train an SVM algorithm on local binarized pictures and attained a 97% accuracy. He also integrated the elimination method to resolve disagreement. Recent research works have used Histopatology image dataset as an advanced technique in the field of image processing as well as artificial intelligence to solve issues concerning image clustering (Komura and Ishikawa; 2018). Before then, features collected by Convolutional Neural Networks (CNN) were used to train the image patches through clustering loss function.

2.1.2 Method

Maity and Das (2017) applied a mix of techniques such as Decision tree, SVM, deep learning algorithm and K nearest neighbor. Since the datasets had noise content, they attempted to clean the datasets prepare them for processing and lessen the dimensionality in the dataset. The result was that high accuracy could be attained through neural networks. Putri et al. (2015) did a detailed discussion about breast cancer and various symptoms of breast cancer and the use of diverse types of categorization and clustering algorithms as well as tools. König (2009) provided an analysis with data mining and found that different approaches had different spread of features and different levels of breast cancer prediction accuracies. Bazazeh and Shubair (2016) shows that the dataset on Breast Cancer had excessive, replicated information, and needed reprocessing and reselection of features on the dataset to attain results that are more accurate.

Umadevi and Marseline (2017) provides the analysis of the existing research works concerning Breast Cancer prediction using data mining as the preferred technique by many researchers. They discuss the test and training on the datasets applied like the Breast Cancer dataset obtained from the UCI data repository. The analysis in the discussion was conducted using a random selection of tools like Weka, Data melt, R Studio, and python for data mining. The conclusion from their analysis is that the application of one algorithm in data mining provides higher accuracy in the prediction of breast cancer survivability as well as other diseases using gene expression. Nevertheless, the application of hybrid system with multiple algorithms could enhance the accuracy of Breast Cancer prediction.

Huang et al. (2017) presents discussion on the machine learning technology and suggests an architecture tool for gathering and processing the dataset to reduce the dimensionality. Ghasem Ahmad et al. (2013) and Bazazeh and Shubair (2016) applied three models of data classification including decision tree, Naïve Bayes, and ANN networks and concluded that the ANN Network algorithm had the highest performance compared to the other two classifiers. In summary on the methods, disease prediction is vital to medical practices as a prevention and therapeutic interventions are proposed on implicit or clear prospects about future results of health results. Umadevi and Marseline (2017) uses the term “Precision medicine” to describe the medical model applied in data classi-

fication, where interventions target individuals and risk classes as opposed to the wider population of breast cancer patients. Precision medicine is essential in the medical practice as it allows physicians to develop strategic plan to aid in reducing the number of patients.

2.1.3 Results

Results of research from various medical practitioners is that Breast cancer is one of the primary agents of cancer-related deaths among women allover the world. From the 2018 studies with data from WHO, approximately 17% of causes of breast cancer cases were from Indonesia (Umadevi and Marseline; 2017). The population of breast cancer patients globally was reaching 42.1 patients in every population of 100,000 people.

König (2009) also carried out the study for predicting the prevalence of breast cancer considering four major parameters, such as patient age, sugar level in blood and Body Mass Index (BMI). In the application of three algorithms, logistic regression, SVM and Random Forest, the analysis found that the SVM algorithm had the most accurate performance compared to the rest of the methods. With regards to the specificity level and sensitivity of the methods used in the past research works, the ANN algorithm had sensitivity of 87% and specificity of 90%. On the same note, Huang et al. (2017) carried out studies to compare the statistical discriminant algorithm and the ANN model to predict the risk of breast cancer. The result of this study is that the ANN algorithm had (93.16%), while the Discriminant analysis had (75.5%) accuracy of prediction. The research conducted by Huang et al. (2017) in early prediction and detection of breast cancer with the ANN and SVM algorithms (Umadevi and Marseline; 2017). The study showed that the SVM algorithm had an accuracy of 95% in predicting breast cancer while the ANN had 94% for ANN Method. The Histopathological Image approach is discussed in detail by Belsare and Mushrif (2011), (Komura and Ishikawa; 2018) and (Raczkowski et al.; 2019). The collection contains 277,524 RGB pictures with a resolution of pixels generated from 162 HE-stained breast histopathology samples (Belsare and Mushrif; 2011). The CNN was trained using a vast volume of image patches obtained from the BC WSI. The picture patches are tiny and extracted from digital tissue samples. The raw data was scaled from 0 to 256. For analyzing various categorization algorithms, the proposed task is scaled from 0 to 1. The data set is split into two sections: test and training. By avoiding overfitting, data splitting helps to improve model accuracy. The proposed model was tested with two different splitting: 70-30 and 80-20. According to the results of the analysis, a handful of the tissues contain malignant cells.The past methods have done better in image clustering, including the histopathological image datasets. Researchers have been successful in modifying the CNN structure for the breast histopatology-connected problems. The amount of data (size) and the level of complexity in the test and training datasets were maximized through random sequences and reflecting or replication of the images. Raczkowski et al. (2019) gathered patches of 100 * 100 from the histopatology images with grid sample for classification of aggressive carcinoma. The extraction of features and the eventual tissue arrangement led to an accuracy level of 77

2.1.4 Critique On Strength and Weakness/Limitation

1. *Strengths:*

   The major strength of this study is the availability of a wide range of literature,

from which detailed comparison is done. The available literature compares several algorithms as well as their levels of prediction accuracy, such that the most accurate algorithms can be selected. The second strength is the availability of the latest datasets, to be used for classification of data and prediction of breast cancer using histopathological images.

2. *Weaknesses / Limitations:* The main limitation of the use of assumption of stability in the state of the society's factors affecting the patients' conditions. The selection of algorithm is done at a different time, after the collection of the datasets. The situation of patients' health during the analysis time could differ from the situation during the data analysis. There is need to the formulate the model for problem solving to take care of the uncertainties and changes in the health conditions and maintain the level of accuracy and sensitivity in the prediction results

2.2 Dataset 2: Breast cancer survivability prediction using gene information

2.2.1 Approach

Huang et al. (2017) suggested "An automatic system for diagnosing of medical conditions and enhancing medical care and minimizing costs. In this paper, the predictive model is designed efficiently to decide on the rules to use for predicting the risk rating in patients considering the parameter representing the patient's health condition. The priority of the rules depends on the needs and expectations of the user. The evaluation of the model performance is done in view of grouping accuracy and the to show the potential of the system in accurate prediction of breast cancer survivability using gene expression.

Li et al. (2021), applied a combination of decision tree, Naïve Bayes, and ANN to construct Intelligent model for predicting Breast Cancer. To improve data visualization and simplify the data interpretation, it shows the outcome in graphical and tabulated form. By delivering efficient therapies, it also facilitates the reduction of cost of therapies.

2.2.2 Method

Umadevi and Marseline (2017) presents a list of causes of Breast Cancer including physical dormancy, poor diets, obesity, and irregular blood pressure. Siegel et al. (2021) also presents comparison of prevalence of breast cancer across various classes of target population.

Singhal and Pareek (2018) on the other hand did an experiment and the outcome showed the repeat of neural network providing high prediction accuracy relative to other algorithms like support vector machine, CNN, and Naïve Bayes. In that case, neural networks had better performance in Breast Cancer prediction. They also attained a model that predicted silent existence of breast cancers and presented early warning to the user of the model.

Siegel et al. (2021) used a large size of data in data mining, a daily-collected data that could not be interpreted manually. The outcome indicates that data mining and deep learning can be used in the effective prediction of breast cancer survivability using gene expression with using the datasets. In this study, two separate datasets are analyzed on the prediction of Breast Cancer risk and survival ability. It does analysis and comparison

of the different classification techniques and algorithms, operating on the Breast Cancer dataset.

Vanneschi et al. (2011) applied many algorithms, including K -Nearest Neighbor, decision tree, random forest, and Naïve Bayes, SVM, and logistic model tree. In the comparative analysis, Naïve Bayes had the best results. In the application of UCI Breast Cancer repository for dataset, the J48 technique took the shortest duration to construct and provided the best outcome.

Siegel et al. (2021) did a comparative analysis on many algorithms, including ANN, KNN, Naïve Bayes and SVM to predict Breast Cancer, and KNN produced the best results in accuracy. Montazeri et al. (2015) applied Logistic regression, ANN, and gradient boosting algorithms in machine learning. The result was that random forest algorithms had the most accurate performance in the prediction of the breast cancer survivability using gene expression. They say that this is the first experimentation using machine learning techniques to routine patient data in electronic records. The source of the dataset is the Clinical Practice Research Datalink (CPRD) (Kourou et al.; 2015). These are the electronic medical records which contains all the medical related information including the statistics of human population, medical biodata, medical appointments with specialists. It also covers the information about the medicine consumption, results, and information about inpatient hospital services. Since the various predictive genes precisely represent the same basic breast cancer survival and risk, then it is necessary to perform an evaluation of genes as components of gene expression and apply the information to as a guiding principle for selecting the predictive genes (Iwamoto and Pusztai; 2010). Ideally, one would like to have detailed gene expression dataset, which will then be integrated in the selection model for genes with the possible causative link to the breast cancer. This integration has been largely impractical due to the inability to find adequate size of data sample. Another problem has been the complex nature of gene interactions. Consequently, the difficulty of retrieving the gene expression information must be confronted in other techniques. One approach is to use the assumption that genes with related expression are in one cluster in an expression (or otherwise have a relationship even when there is no direct interaction between them). Abraham Abraham et al. (2010) used unsupervised gene module discovery approach, computed the discrete rating of module actions. They applied the ratings as vital feature in the Naïve Bayes classification. The outcome from research by Adnan et al. (2020)4 was that the classification with gene expression was more accurate in predicting the survivability of breast cancer results compared the classification with individual genes.

2.2.3 Results

Results of research from various medical practitioners is that breast cancer-related deaths were among the highest in women all over the world. From the analysis by Tapak et al. (2019), approximately 14% of causes of breast cancer cases were from South Korea, and the population of breast cancer patients globally was reaching 39.7 patients in each population of 100,000 people. This is echoed by Kourou et al. (2015), who also carried out the study to predict the prevalence of breast cancer considering 4 main parameters, like the patient age, blood sugar level and the pressure in the body. In the use of the various algorithms, the logistic regression, the SVM and Random Forest, the analysis, the outcome was that the SVM algorithm had the highest accuracy of classification of the data compared to the rest of the techniques (Maity and Das; 2017). With regards to

the specificity level as well as the sensitivity of the methods used in the past literature, Li et al. (2021) shows that the sensitivity of ANN algorithm was 89% and its specificity was 87%. On the other hand, Moncada-Torres et al. (2021) carried out research works to evaluate the statistical discriminant algorithm and the ANN model to forecast on the risk of breast cancer. The findings were that the ANN algorithm had (91%), while the Discriminant analysis had (77.7%) degree of accuracy in predicting breast cancer risk and survival rate. The research conducted by Kourou et al. (2015) in early prediction and detection of breast cancer with the ANN and SVM algorithms. The study showed that the SVM algorithm had an accuracy of 91% in predicting breast cancer while the ANN had 94% for ANN Method.

2.2.4 Critique On Strength and Weakness/Limitation

1. *Strengths* Concerning the second dataset, the major strength of this study is the extended period of the dataset coverage, spreading to more than 10 years. This is a reasonable period of data coverage for a realistic comparative analysis. Additionally, the available sources of information did compare several algorithms across the wide span of time and maintained consistent analysis of the levels of accuracy and sensitivity. The data eliminates bias due to the random distribution of the personal data of the target groups over the period.

2. *Weaknesses / Limitations* The main limitation of the overconcentration on one algorithm, the K-Nearest Neighbor, using the assumption that it has the highest performance efficiency inherently. The experimental analysis may not present the same scenario as the dynamic algorithm apply at different time, after the collection of the datasets

Table 1: Summary details of the related work

Article	**Author(s) and Year Published**	**Algorithms**	**Accuracy**
Analysis of Histopathological Images for Prediction of Breast Cancer Using Traditional Classifiers with Pre-Trained CNN	Gupta and Chawla (2020)	SVM CNN	97% 95%
Machine learning for improved diagnosis and prognosis in healthcare	Maity and Das (2017)	Decision Tree SVM KNN ANN	89% 91% 91% 94%
Predictive Techniques And Methods For Decision Support In Situation With Poor Data Quality	König (2009)	SVM, Random Forest ANN	93.67% 93% 87%
Comparative study of machine learning algorithms for breast cancer detection and diagnosis	Bazazeh and Shubair (2016)	KNN CNN Naive Bayes	94.98% 84.23% 83%
A Comparative Analysis of Techniques for Predicting Academic Performance	Nghe et al. (2007)	Random Forest ANN	78.92% 92%

A comparison of machine learning techniques for survival prediction in breast cancer	Vanneschi et al. (2011)	Naïve Bayes KNN Decision Tree Random forest SVM Logistic model tree	75.34% 91.5% 76% 84% 87.3% 78.5%
Predicting breast cancer 5-year survival using machine learning: A systematic review	Li et al. (2021)	Decision Tree Naïve Bayes ANN	84.38% 92% 97%
Cancer statistics, Ca-a Cancer Journal for Clinicians	Singhal and Pareek (2018)	CNN Naive Bayes	84.33% 93.2%
Machine learning models in breast cancer survival prediction	Montazeri et al. (2015)	Gradient Boosting Logistic regression ANN	85.89% 84.7% 93.45%
ARA: accurate, reliable and active histopathological image classification framework with Bayesian deep learning	Raczkowski et al. (2019)	Bayesian Deep Learning SVM	86% 93%

3 Methodology

The Methodology follows four stages: first the data collection, second Data Pre-processing, Third data transformation and lastly the model selection.

3.1 Data Collection

Dataset to Breast cancer prediction using Histopataogical Images: The dataset is retrived from Kaggle[2] The initial dataset included 162 whole mount slide images of Breast Cancer (BCa) specimens scanned at 40x. 277,524 patches of 50 × 50 size were taken from this (198,738 IDC negative and 78,786 IDC positive). The file name of each patch is in the following format: uxXyYclassC.png, for example 10253idx5x1351y1101class0.png .Where u is the patient ID (10253idx5), X is the x-coordinate of where this patch was cropped from, Y is the y-coordinate of where this patch was cropped from, and C is the class, with 0 indicating non-IDC and 1 indicating IDC.The original files are located in [3] and the original paper is cited in [4].

Dataset to Breast cancer survivability using gene information: The dataset was retrived from Kaggle [5].The genetics part of the dataset contains m-RNA levels z-score for 331 genes, and mutation for 175 genes.The dataset was collected by Professor Carlos Caldas from Cambridge Research Institute and Professor Sam Aparicio from the British Columbia Cancer Centre in Canada and published on Nature Communications (Pereira et al.; 2016).

[2] https://www.kaggle.com/paultimothymooney/breast-histopathology-images

[3] http://gleason.case.edu/webdata/jpi-dl-tutorial/IDC_regular_ps50_idx5.zip

[4] https://www.ncbi.nlm.nih.gov/pubmed/27563488andhttp://spie.org/Publications/Proceedings/Paper/10.1117/12.2043872

[5] https://www.kaggle.com/raghadalharbi/breast-cancer-gene-expression-profiles-metabric/code

3.2 Data-Preprocessing:

3.2.1 Dataset to Breast cancer prediction using Histopataogical Images

1. **Step 1:** *Data Pre-processing*

   (a) *Data Separation:*The data set is separated into IDC- and IDC+ classes. fn match. lter () is a Python class that is utilized in this procedure. This function was used to return a Boolean value based on a pattern match.

   (b) *Image Resizing:* Resizing an image in terms of dimensions, height, and width using the Python class cv2.resize () while maintaining the image's uniqueness in the resized image.

   (c) *Image Labeling:* If an IDC is present, the value is "1," else it is "0." For constructing data frames, use the pd.DataFrame ( ).

2. **Step 2:** *Data Processing*

   (a) *Data Splitting:* The python class sklearn.model selection. Train test split( ) is used to split data into test and train subsets, which helps to offer random partitioning of subsets and improve model accuracy by reducing overfitting.

   (b) *Data Normalization:* Data from 0 to 256 is scaled to 0 to 1, allowing a wide range of classification methods to work with it.

   (c) *Data Flattening:* The 2D array is flattened into a 1D array using the NumPy python class. ndarray.atten ( ). This class is used to make a replica of the input array. np.ravel() and np.reshape() are two alternative data attenuation methods.

   (d) *Data Rebalancing:* This method resamples the dataset for imbalanced classes by employing two methods: random oversampling and random undersampling. Imblearn.under sampling () is the Python class that was used. The minority class examples are duplicated during random oversampling, resu lting in over-stating. In the case of Random Undersampling, minority class examples are excluded.

3. **Step 3:***Data Transformation using Incremental PCA:* When the dataset to be decomposed is too big to fit in memory, incremental principal component analysis (IPCA) is commonly used as a substitute for principal component analysis (PCA). IPCA uses an amount of memory that is independent of the number of input data samples to create a low-rank approximation for the input data.

4. **Step 4:** *Data Modeling* Data Model Fitting: The model is tested for accuracy using K-cross validation, several classification models, and CNN.

   (a) *Data Plotting:* To plot data, use the Python method matplotlib.pyplot.gure ( ). Matplotlib is a Numpy mathematical extension. The charting model's accuracy and loss are calculated using the epoch's number. IDC- is used to define all class-zero images, while IDC+ is used to define class-one images.

3.2.2 Dataset to Breast cancer survivability using gene information

1. **Step 1:***Data Preprocessing:* Observing the data reveals that some columns have missing data. This will affect the model by adding noise, degrading the performance. Hence, we drop such data points, which is a valid action considering these points constitute a small percentage of the overall data.

   The data set contains a good number of categorical variables, which are denoted using strings. This must be rectified as learning models cannot handle string data directly. The solution is to use label encoders. In our case, we use the LabelEncoder provided by Python library Scikit-Learn. This assigns a number to each of the categories and replaces the the columns with the respective label numbers.

   We also observed that the columns involved are not of the same scale. This might adversely the affect the model. For that reason, we need to scale each of the columns such that they have comparable values while maintaining the distribution of the data. To implement this, we use the MinMaxScaler provided by a scale on which scales the given column between 0 and 1 using the minimum and maximum values of that particular column does the final data set has all columns with values 0 to 1.

   Another issue to observe while running the model was errors with respect to 2 column names. It turns out the library expects the column names to follow certain rules and it seem like some of these columns where not adhering to the rules hence we rename all column as numbers representing their respective positions in the data set.

2. **Step 2:** *Data Transformation using PCA:* The data set being used has over 600 columns of which not all of them might contribute much to the actual distribution of data. Hence, it makes sense to you choose only e those representations of data that actually contribute to the distribution this helps remove noise from the data set and helps improve the accuracy of the model at the same time ensuring computer resources are used efficiently.

   To reduce the dimensions of the data set we use the concept of principal component analysis, which is a tried and tested method. In order to apply this method we need to specify the number of components required for the target number of columns. To decide this we have a look at the curve of the below graph.

   This graph indicates the amount of variance contributed by the different number of components. We see that out of the 693 columns 600 columns explain almost 100% of the variance in the data set. Further observation shows that 100 components explain about 80% of the variance; this shows that using columns greater than hundred might contribute only 20% of the remaining variance. Hence, it is a fair assumption that 100 components can explain the majority of variance; in other words, we need to extract 100 components from the principal component analysis model.

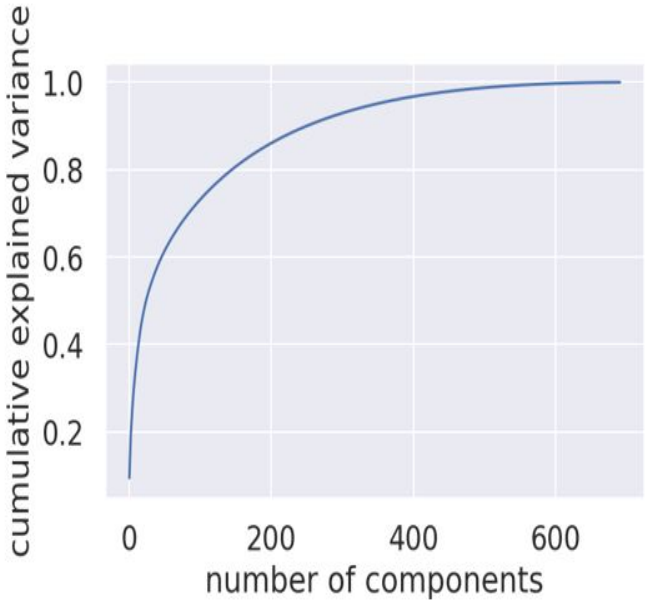

Figure 1: Dimensionality Reduction

3.3 Model Selection

The type of data set being used and the problem statement define the model selection process. The chosen problem clearly falls under the classification category. We looked at models that exclusively perform classification. The goal here is to provide a comparative analysis of different approaches to classification on the given data set.

Dataset to Breast cancer prediction using Histopataogical Images

After much deliberation and study the following models were chosen to be applied on the data set:

- Support vector classifier: A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems.
- Naive Bayes Classifier: Naive Bayes is a probabilistic machine learning method that may be used to solve a wide range of classification problems.
- Convoluted Neural Networks (CNN): The name "convolutional neural network" refers to the network's use of the convolutional mathematical procedure. Convolutional neural networks are a form of neural network that uses convolution rather than ordinary matrix multiplication in at least one layer.
- Attention Based Neural Networks: Multiple instance learning (MIL) is a supervised learning technique in which a single class label is applied to a group of examples.

Each of these models have drastically drift different approaches to solving the same problem. Hence we shall see which methods worked really well with the given data set

Dataset 2

After much deliberation and study the following models were chosen to be applied on the data set:

- Support vector classifier
- K nearest neighbours: K-Nearest Neighbors (KNN) is a basic machine learning method for regression and classification problems. KNN algorithms takes data and apply similarity metrics to categorize fresh data points. AClassification is done by a majority vote to its neighbors. The information is allocated to the class with the most neighbors.As the number of nearest neighbors grows, so does the value of k, and so does the accuracy.
- Multilayer Perceptron Classifier: A multilayer perceptron (MLP) is a class of feed-forward artificial neural network (ANN).here are at least three levels of nodes in an MLP: an input layer, a hidden layer, and an output layer. Backpropagation is a supervised learning technique used by MLP during training.
- Boosted tree estimator classifier: Gradient boosting is a machine learning approach for regression, classification, and other problems that generates a prediction model from an ensemble of weak prediction models, most often decision trees.

4 Design Specification and Implementation

The implementation of both stages of this study was facilitated on popular cloud based Virtual Machine service - Google Colaboratory, which specializes in providing Python-based computing resources, including specialized hardware options to include Graphics Processing Units (GPUs) which speed up the implementation stage significantly.

The design specification is converted to runnable code using Python as the operating language. This is because Python provides a very strong support for machine learning and deep learning model development. Some of the libraries used in this implementation include:

- Keras (ML model development)
- Tensorflow (ML model development - assistance)
- Pandas (Data handling)
- Scipy (Statistics)

As proposed the there are two different health datasets considered for the project.

Dataset to predict the breast cancer using Histopathology Images:

The design specification for (*Breast Cancer Prediction using Histopathological Images*) is shown in the figure Figure 2

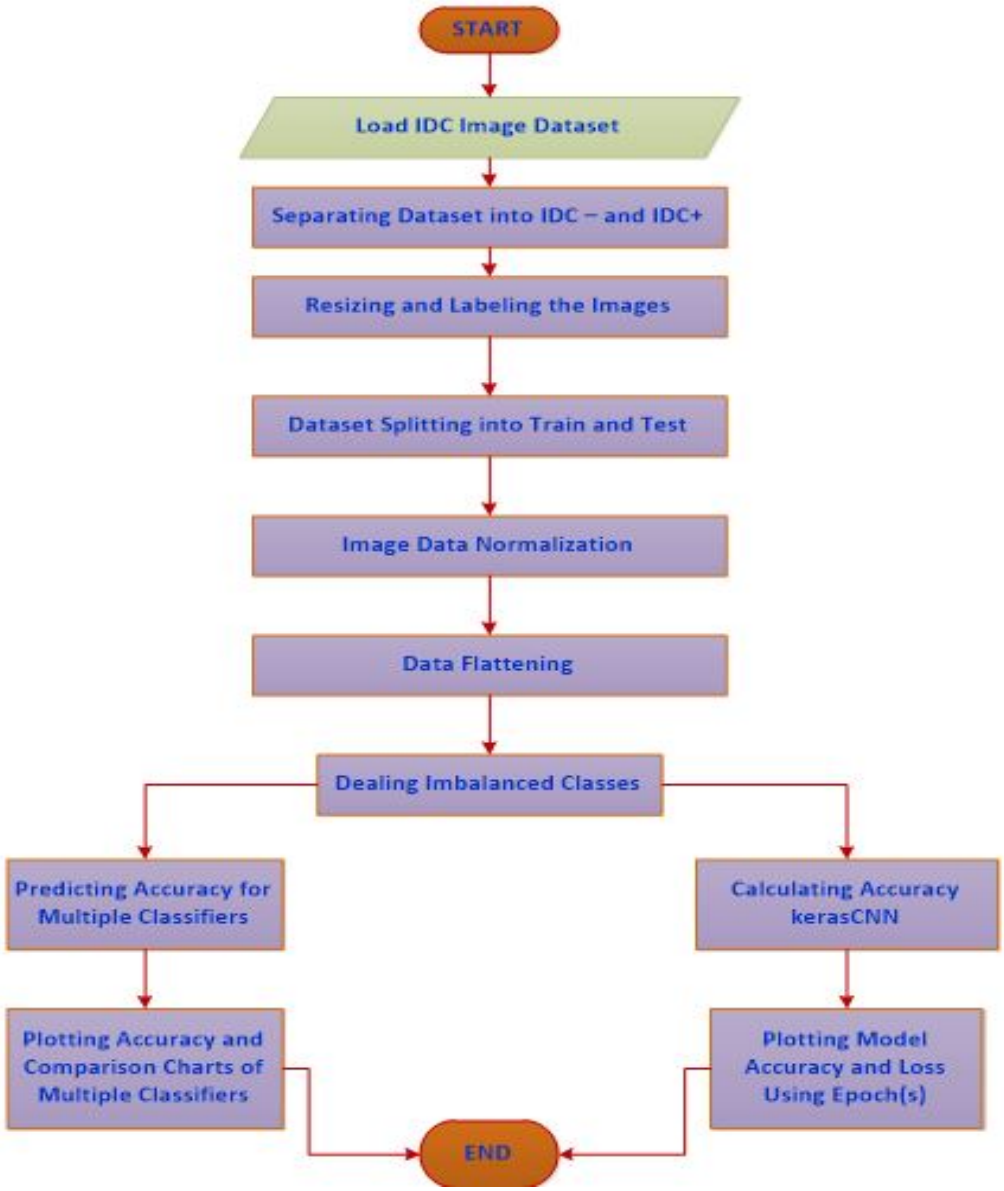

Figure 2: Flow Diagram of the prediction of Breast cancer

A precise IDC BC prediction model with CNN and multiple classifiers are proposed in this paper. The accuracy level of these classifiers is estimated with the help of CNN and diagnosis of cancer is achieved with the help of WSI (Whole Slide Image). The process of loading the image dataset is the first step. Following that, the dataset was split into two sections: IDC+ and IDC-. The IDC- symbol denotes a class zero image, while the IDC+ symbol denotes a class one image. IDC+ is represented by one, while IDC- is represented by zero. The image has been scaled as well as labeled. These operations are part of the preprocessing stage.

The suggested model's next stage is data processing, which includes data splitting, data normalization, data flattening, and data rebalancing. At the data splitting stage, the dataset was separated into two halves (testing and training), which helped to improve

the model's accuracy. The data is then structured to ensure compatibility when using classifiers. The data-attending step is used to turn a two-dimensional array into a one-dimensional array. After then, data were sampled for balancing out unbalanced datasets. The two methods for resampling are random under sampling and random oversampling. IDC+ and IDC- are balanced in this proposed effort to cope with the imbalanced classes by reshaping data. After balancing the data, K-fold cross-validation was used to test the accuracy of multiple classifiers. Multiple classifiers are supplied with different-sized inputs, and the assistance function is disabled. Finally, the accuracy value and comparison chart for the various classifiers are plotted. Additionally, epochs were used to depict the model accuracy and loss of CNN.

Dataset to predict the surviavabiliy of Breast Cancer Patients

The design Specification for the dataset 2(Breast Cancer Survivability prediction using the Gene Information). The Process follows KDD methodology (*Knowledge Discovery and Database*). The flow diagram is as shown in the below Figure 3

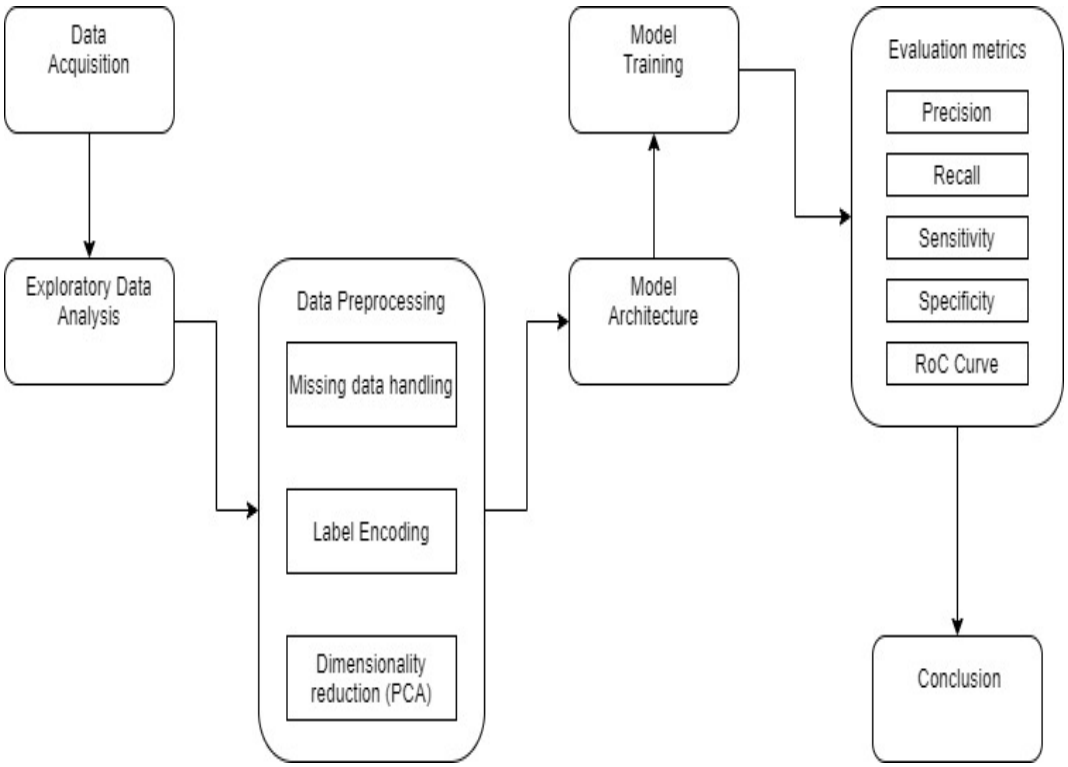

Figure 3: Flow Diagram for the breast cancer survivability prediction

First we load the dataset into Google Colab from Kaggle. Exploratory data analysis was then performed on the dataset to get more insights onto the data. Data pre-processing is the next step in the process flow. Handelling missing data in the dataset is the first priority, else the results will be over fitting or under fitting. Usually data normalization or max min actions are performed to handle the missing data. Label encoding is then performed on the data since the column headers have unnecessary special symbols, which misleads the models during train and evaluation of the model. Principal Component Analysis is conducted to perform dimensionality reduction as the dataset have huge number of columns and not all columns are necessary for prediction/modelling. Further the dataset is split into train and test before application of the model. The data is then

structured to ensure compatibility when using classifier. Data modelling was then applied and finally evaluated using Precision, Recall, Accuracy and ROC curve.

5 Evaluation

The purpose of this section is to provide a comprehensive analysis of the results and main findings of the study as well as the implications of these finding both from academic and practitioner perspective are presented. Only the most relevant results that support your research question and objectives shall be presented. Provide an in-depth and rigorous analysis of the results. Statistical tools should be used to critically evaluate and assess the experimental research outputs and levels of significance.

Given two different implementations inplace, we need two establish two different evaluation reports, one for each part of the study.

Breast Cancer Prediction using Histopathological Images

The suggested model is run with 70-30 splitting (training and testing). For different data sizes, the model accuracy and loss are projected based on epoch numbers. The epoch was used to count how many times the training vector was utilized before updating the weights.

Table 2 shows the accuracy of numerous classifiers on different sized datasets with a 70-30 split.

Table 2: Multiple classifier accuracy comparisons for a 70-30 split

Classifiers	**Accuracy in percentage**	**Precision**	**Recall**	**F1-score**	**Specificity**
Naive Bayes	57.78%	48.57%	94.44%	64.12	33.34%
SVC	91.12%	100%	89.74	94.59%	100%
CNN	78.00%	77%	80%	78%	97.5%
MIL	91.05%				

These scores represent the accuracy of the models. The accuracy is calculated internally using the formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Where, TP $\rightarrow TruePositive, TN \rightarrow TrueNegative, FP \rightarrow FalsePositive, FN \rightarrow FalseNegative$

Here, the Positive class represents the presence of cancer cells, and Negative represents its absence. What the score essentially means is that if accuracy is x%, the model correctly identifies the class the image belongs to x% of the time. Hence, higher the accuracy, the better the model classifies.

One interesting part of this study is the comparison of models of a varied complexity. We observe results from models as simple as Support Vector Classifiers, to extremely complex models like Attention based neural networks. What we observe straight up is increasing the complexity of the model does not necessarily help better the score.

Breast Cancer Survivability using Gene Information

The model is split into 70:30 for train and test accordingly, different evaluation metrices are used for evaluation of the model such as precision, recall, F1-Score and Accuracy.

Table 3: Evaluation Comparison between the models

Classifiers	Accuracy in percentage	Precision	Recall	F1-score
KNN	80%	81%	83%	82%
SVC	100%	100%	100%	100%
MLP	99%	100%	99%	99%
Boosted tree Estimators	100%	100%	100%	

The Model is promising and has achieved much more accuracy than the state of the art model(ANN with 97%) presented in Li et al. (2021).

Decision trees are the one of the best [6] classification models that are not too complex. It is relatively simple, yet extremely reliable, especially when it comes to binary classification. Boosted trees follow the same underlying working of the basic decision trees, but are an enhanced form of these trees, hence we can expect really high scores from these boosted tree estimators, which also explains the result.

6 Conclusion and Discussion

This study has provided insight into how different problems of the same nature can be dealt with. We addressed how both two main aspects of breast cancer: detection and survivability prediction can be used. We also trained and compared machine learning models of varying degrees of complexity, and the conclusion was using extremely complex models like attention based neural networks does not significantly increase its performance, compared to more simpler models like SVC.

There are many ways to improve upon the outcome of the model. Although we used unpaired, independent datasets for this, future scope could be to collect both these data from the same patients, enabling more streamlined treatment of patients. But this is definitely challenging as such data is collected across different departments in the medical field and might be difficult to get collaborative results.

References

Abraham, G., Kowalczyk, A., Loi, S., Haviv, I. and Zobel, J. (2010). Prediction of breast cancer prognosis using gene set statistics provides signature stability and biological context, *BMC Bioinformatics* **11**: 277 – 277.

Adnan, N., Lei, C. and Ruan, J. (2020). Robust edge-based biomarker discovery improves prediction of breast cancer metastasis.
URL: *https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-03692-2*

Bazazeh, D. and Shubair, R. (2016). Comparative study of machine learning algorithms for breast cancer detection and diagnosis, *2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA)*, pp. 1–4.

[6] https://www.analyticsvidhya.com/blog/2021/05/5-classification-algorithms-you-sh ould-know-introductory-guide/: :text=5.-,Decisionntuitions%20and%20interpretations

Bektaş, B. and Babur, S. (2016). Machine learning based performance development for diagnosis of breast cancer, *2016 Medical Technologies National Congress (TIPTEKNO)*, pp. 1–4.

Belsare, A. and Mushrif, M. (2011). Histopathological image analysis using image processing techniques: An overview, *Signal Image Process Int J* **3**.

Clough, K., Kaufman, G., Nos, C., Buccimazza, I. and Sarfati, I. (2010). Reply to comments on: Improving breast cancer surgery: A classification and quadrant per quadrant atlas for oncoplastic surgery.
URL: *https://link.springer.com/article/10.1245/s10434-010-1302-yciteas*

Ghasem Ahmad, L., Eshlaghy, A., Pourebrahimi, A., Ebrahimi, M. and Razavi, A. (2013). Using three machine learning techniques for predicting breast cancer recurrence, *Journal of Health Medical Informatics* **4**: 124–130.

Giri, P. and Saravanakumar, K. (2017). Breast cancer detection using image processing techniques, *Oriental journal of computer science and technology* **10**: 391–399.

Gupta, K. and Chawla, N. (2020). Analysis of histopathological images for prediction of breast cancer using traditional classifiers with pre-trained cnn, *Procedia Computer Science* **167**: 878–889.

Gurcan, M. N., Boucheron, L. E., Can, A., Madabhushi, A., Rajpoot, N. M. and Yener, B. (2009). Histopathological image analysis: A review, *IEEE Reviews in Biomedical Engineering* **2**: 147–171.

Huang, M. W., Chen, C. W., Lin, W. C., Ke, S. W. and Tsai, C. F. (2017). Svm and svm ensembles in breast cancer prediction, *PLOS ONE* **12**: 1–14.
URL: *https://doi.org/10.1371/journal.pone.0161501*

Iwamoto, T. and Pusztai, L. (2010). Predicting prognosis of breast cancer with gene signatures: Are we lost in a sea of data?
URL: *https://genomemedicine.biomedcentral.com/articles/10.1186/gm202citeas*

Komura, D. and Ishikawa, S. (2018). Machine learning methods for histopathological image analysis, *Computational and Structural Biotechnology Journal* **16**: 34–42.
URL: *https://www.sciencedirect.com/science/article/pii/S2001037017300867*

König, R. (2009). Predictive techniques and methods for decision support in situations with poor data quality.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V. and Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction, *Computational and Structural Biotechnology Journal* **13**: 8–17.
URL: *https://www.sciencedirect.com/science/article/pii/S2001037014000464*

Li, J., Zhou, Z., Dong, J., Fu, Y., Li, Y., Luan, Z. and Peng, X. (2021). Predicting breast cancer 5-year survival using machine learning: A systematic review, *PLOS ONE* **16**: 1–23.
URL: *https://doi.org/10.1371/journal.pone.0250370*

López-García, G., Jerez, J. M., Franco, L. and Veredas, F. J. (2020). Transfer learning with convolutional neural networks for cancer survival prediction using gene-expression data, *PLOS ONE* **15**(3): 1–24.
URL: *https://doi.org/10.1371/journal.pone.0230536*

Maity, N. G. and Das, S. (2017). Machine learning for improved diagnosis and prognosis in healthcare, *2017 IEEE Aerospace Conference*, pp. 1–9.

Moncada-Torres, A., van Maaren, M. C., Hendriks, M. P., Siesling, S. and Geleijnse, G. (2021). Explainable machine learning can outperform cox regression predictions and provide insights in breast cancer survival.
URL: *https://www.nature.com/articles/s41598-021-86327-7citeas*

Montazeri, M., Montazeri, M., Montazeri, M. and Beigzadeh, A. (2015). Machine learning models in breast cancer survival prediction, *Technology and health care : official journal of the European Society for Engineering and Medicine* **24**.

Nghe, N. T., Janecek, P. and Haddawy, P. (2007). A comparative analysis of techniques for predicting academic performance, *2007 37th Annual Frontiers In Education Conference - Global Engineering: Knowledge Without Borders, Opportunities Without Passports*, pp. T2G–7–T2G–12.

Pereira, B., Chin, S.-F., Rueda, O. M., Vollan, H.-K. M., Provenzano, E., Bardwell, H. A., Pugh, M., Jones, L., Russell, R., Sammut, S.-J. and et al. (2016). The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes.
URL: *https://www.nature.com/articles/ncomms11479*

Putri, I., Cholissodin, I. and Setiawan, B. D. (2015). Optimasi metode adaptive fuzzy k-nearest neighbor dengan particle swarm optimization untuk klasifikasi status sosial ekonomi keluarga, **5**.

Raczkowski, , Możejko, M., Zambonelli, J. and Szczurek, E. (2019). Ara: accurate, reliable and active histopathological image classification framework with bayesian deep learning, *Scientific Reports* **9**.

Siegel, R. L., Miller, K. D., Fuchs, H. E. and Jemal, A. (2021). Cancer statistics, 2021, *CA: A Cancer Journal for Clinicians* **71**(1): 7–33.
URL: *https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21654*

Singhal, P. and Pareek, S. (2018). Artificial neural network for prediction of breast cancer, *2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2018 2nd International Conference on*, pp. 464–468.

Tapak, L., Shirmohammadi-Khorram, N., Amini, P., Alafchi, B., Hamidi, O. and Poorolajal, J. (2019). Prediction of survival and metastasis in breast cancer patients using machine learning classifiers, *Clinical Epidemiology and Global Health* **7**(3): 293–299.
URL: *https://www.sciencedirect.com/science/article/pii/S2213398418301829*

Umadevi, S. and Marseline, K. S. J. (2017). A survey on data mining classification algorithms, *2017 International Conference on Signal Processing and Communication (ICSPC)*, pp. 264–268.

Vanneschi, L., Farinaccio, A., Mauri, G., Antoniotti, M., Provero, P. and Giacobini, M. (2011). A comparison of machine learning techniques for survival prediction in breast cancer, *BioData Mining* **4**.