Analysis and Prediction of factors affecting Breast and Cervical cancer using Ensemble Techniques

MSc Research Project
MSc in Data Analytics

Yuvaraja Panneerselvam
Student ID: x17150663

School of Computing
National College of Ireland

Supervisor: Swapnil Parashar

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Yuvaraja Panneerselvam

Student ID: X17150663

Programme: MSc in Data Analytics **Year:** 2018

Module: Masters Research Project

Supervisor: Swapnil Parashar

Submission Due Date: 20th December, 2018

Project Title: Analysis and Prediction of factors affecting Breast and Cervical cancer using Ensemble Techniques

Word Count: 6471 **Page Count** 21

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:

Date: 20th December, 2018

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):	

Contents

Analysis and Prediction of factors affecting Breast and Cervical cancer using Ensemble techniques

Yuvaraja Panneerselvam
X17150663

Abstract

Misconception in medical field is that getting older in age makes women vulnerable to easy occurrence of breast and cervical cancer, has prevailed in medical domain for years. This needed to be investigated by considering datasets for both cancers. Several computer aided supports are already providing analysing and prediction of diseases but most of them have bottleneck of more error in prediction and inaccurate results. To fulfil with process of investigation, dataset for breast and cervical cancer taken to analyse whether age is the real complicating root problem among women. Ensemble learner models also implemented to combine functionality of three models in the chase of better accuracy with minimal errors to predict occurrences of cancer accurately.

1 Introduction

Age, diet and activities affect a person's health in drastic fluctuating ways. Carcinoma is the extreme epidemic disease caused around the globe where a person can stay with carcinoma for the rest of their life. Carcinoma termed as benign cancer when cell multiplies at a slower rate compared to malignant cancer that is coined medically after carcinoma cells have developed to final stage where cannot be treated any further.

Cells in human body die faster and regeneration rates becomes slower during cancer. Carcinoma in general is the state at which a person's body cells at a particular part of body starts to rapidly grow absurdly faster than its normal rate of developing and leading to malignant nature. Carcinoma in women prevails in high number of medical cases to which more than half of it seems to be the mortality rates. Carcinoma being that deadly at such certain levels when left unscreened and untreated. Various reasons are behind the cause of a cancer among which majority of causes are due to age, poor diet, activities and exposure to radiation in any form. Women tend to develop cancers in areas that differentiates from the male anatomy. Breast and cervical cancers are high among women when compared to their counterparts. A female's tissue structure and lymph nodes vary from a male form and physique.

Breasts in female form have lymph nodes connecting the cells from body tissues to the mammary glands. Mammary glands are responsible for secretion of milk for a new born from the body of host. Breasts contain layers of branched lobe of lobules and duct that carry blood from lobes to the areola. The cancer develops in these areas that are responsible for milk production. Carcinoma mostly develop in cells that carry information or blood. In breast cancer the carcinoma is formed in lobes in form of solid or lumps filled with fluid, can be formed in the branched lobules and rare cases are near the duct areas.

The causes of breast carcinoma are well associated with production of milk. Breast carcinoma are caused due to when females consume alcohol over the threshold regularly, genetically passed from mother or if runs in the family, radiation exposure during screening tests of any other medical examinations, less knowledge and exposure to screening tests and menopause medications also complicate growth of cancer in breasts.

Symptoms while developing breast carcinoma are pain in lymph areas, change of color in any one areola, feel of lumps when examined closely and lumps may be filled with some lobes secreted fluid or it may be solid.

Cervical cancer originates inside endometrium and on the outer layer of cervix. Women have vagina area acting as pathway for reproduction, cervix on the inner end of vagina, inside of cervix lies the endometrium which is uterus passage to the fallopian tube that contains ovaries on either side. Ovaries are responsible for discharge of old blood and cells that were not required anymore for reproduction purposes. Any complications that causes delay of discharging cells is medically coined as menopause. Menopause is also linked with cervical cancer. Carcinoma hits the uterus and cervix regions, cells develop absurdly at phenomenal rate. If carcinoma gets developed on the cervix which is inner lining of vagina, then the patient will experience enormous pain and complications during monthly discharge of blood from ovaries. Cervical cancer is often mis concluded by patients that it has occurred in vagina but whereas carcinoma grows inside the uterus and also on the outer layer of cervix.

Symptoms while developing cervical carcinoma are pain with irregular discharge of blood from vagina after intercourse or menopause, fluid discharge at random intervals from cervix and increased rate of blood cells discharge from ovaries during monthly cycle of women.

From the existing statements, this research is based on the research question: "***Till what extent does age and gender unquestionably become the reason behind women's breast and cervical cancer occurrences using ensemble learners?"***

Women suffering from cancer has elevated at huge rates in the past few years. This increase in the demise of women was tragic loss to mankind. So, an early detection of carcinoma can be helpful in preventing rapid cancer growth through various screening tests and computer's prediction capabilities. The severity of carcinoma is the reason for this research as it can contribute to existing machine learning models and become a support system for existing computer aided systems.

Prediction errors above five percentage during early examination of breast and cervical cancer has drastic influence over treatment by doctors. If the errors are minimal and close to zero then computer predictions will replace screening tests in near future. So, objective is to develop such a method to predict with more accuracy and less error. To achieve this, ensemble learners are used for increasing prediction score with less errors.

Section 2 talks about related work done and future scope of those research, Section 3 talks about the methodology and section 5 about evaluation of implemented models.

2 Related Work

There are researches done and carried on in breast and cervical cancers. It has to be considered that a wide use of computer aided procedures has already began. Research papers for two different cancers that are breast and cervical cancer were reviewed.

2.1 Breast cancer

2.1.1 Introduction

(Umesh and Ramchandra; 2015) suggests global stats from world health organization throws light over breast cancer patients of 1,80,000 registered in 2008 at united states and in India around 1,10,000 registered. Breast cancer to be one of the severest carcinoma diseases among women according to breast cancer institute.

The problem lies with improper number of screenings tests. (Ahmad and Yusoff; 2013) suggests many patients with identical medical background and diagnosed profiles tend to had different results after screening. (Giger and Huo; 1999) 5 to 30 % who went under mammogram examination were wrongly concluded to be normal whereas they had breast cancer and 10 to 40 % who went under biopsy examination were concluded as having cancer.

(Ahmad and Yusoff; 2013) says fine needle aspiration is a biopsy procedure in detection of breast cancer which had served good results compared to standard mammogram examinations. This biopsy procedure involved penetration of thin long needle into the lesions of malignancy, then that particular masses are observed under microscope for tumor type. (Ahmad and Yusoff; 2013) not only wanted to propose fine needle aspiration as a good and effective screening technique but also utilize random forest prediction technique on the yielded digital data. Random forest gave 72% accuracy and 70 % specificity over dataset containing 450 benign cases and 240 malignant cases of around 700 patients.

(Ahmad and Yusoff; 2013) mammogram are said as standard examining procedure for breast cancer but patients had experienced growth of abnormal cells from benign stage after exposure. Radiologists had knowledge to distinguish the masses of breast images, but computer software still lacked behind with high variance and accuracy error rates. Detection and recognition abilities are different for humans and computers. So, whenever patients sent under mammogram examination they resulted in poor classification from systems as lesions classification was a challenge. Lesions are infected cells grouped together in an area which in turn bulges up and causes malignancy. They are often mistaken for carcinoma. (Giger and Huo; 1999) suggests, ***"The classification method includes three components: 1) automated segmentation of mass regions, 2) automated feature-extraction, and 3) automated classification."*** These methods were used to extract readable information from any digitalized mammogram result. The images moreover had noisy information that cannot be understood by machines. Completed extraction provided enough information to fit and run an artificial neural network model. The model was run over 95 mammogram results of 65 patients where each patient data consisted of more than one mammogram image. Only those images were selected to model as features that had relation with margin of the lesion and density of the masses. Used their own database of patients and ended with 83 % accuracy results from artificial neural networks and 12 % boost from their mammographer results.

2.1.2 Methodologies

(Umesh and Ramchandra; 2015) machine learning has shown phenomenal growth in domains that needs prediction. Research done by above people on to achieve a better model that will be close enough to solve detection of malignancy at basic stages to cure breast cancer from features having recurrence information. Dataset worked upon was from surveillance epidemiology and end results to work with association rule mining data mining algorithm since it was widely utilized for bigger datasets with categorical features. Weka toolkit was used for running the model. Considered 17 features from dataset for prediction.

(Ohja and Goel; 2017) Wisconsin Prognostic dataset for breast cancer to apply over classification and clustering algorithms with 32 feature selected variables for prediction. K – means, fuzzy C - means and partitioning around medoids for clustering prediction and support vector machines, K nearest neighbor, decision trees and naïve Bayes for classification prediction. AUTH achieved 63% accuracy from k – means model among clustering models and 81 % accuracy from C5.0 decision trees among classification models.

(Giger and Huo; 1999) the neural network prediction for breast cancer dataset as model emerged as a new comer during publish of journal. While many research papers focused on using models, one researcher emphasized through feature selection intensity of breast cancer detection can be enhanced.

(Kermani et al; 1995) used genetic algorithm to extract features from data that contributed value to cause of occurrence. Then fitted their training and testing data in neural network and genetic algorithm which gained accuracy of 92 % and 94 % respectively. Genetic algorithm was observed to have carefully sort the features that helped in prediction as well.

(Khariwal and Mishra; 2018) artificial neural networks and logistic algorithm as part of classification of biopsy result combined by scatter and violin plot. Authors wanted to prove that artificial neural network and logistic regression are feasible prediction models for breast cancer. Feature selection models include random forest, statistical feature selection of chi square test and single variate selection. Ensemble learners were introduced at basic phase for execution under voting task where algorithms will be sent under execution followed on to validation of predicted results where algorithm with highest accuracy will be considered.

(Fogel et al.; 1999) results from each epoch from artificial neural networks to form a curve that will benefit when and where to adjust weights and that’s not it, the resultant results from artificial neural networks to analyze statically under ROC curve.

(Gupta et al; 2018) another research paper focusing on accuracy and enhance prediction of breast cancer was by making use of K – fold validation technique to train and validate the models from regression and classification applied on his dataset. Among the research papers studied. It was the only one who underwent k-fold technique for training and validation, used r square metric also for validation in contrast to other authors and tested predicted datasets with and without its best features.

2.1.3 Results of developed models

(Gupta et al; 2018) 94 %, 90 %, 93 %, 100 %, 90.9 % accuracies from K nearest neighbor, support vector machines, decision trees and multiple linear regression models respectively using all features from datasets whereas using highly co related 5 features from dataset yielded in this paper accuracy with surprising 10 % decline.

(Khariwal and Mishra; 2018) had 98 % accuracy when put the dataset under voting algorithm with ensemble learners.

2.1.4 Conclusion

(Ohja and Goel; 2017) research motto was to find which algorithm can be utilized for prediction initial phases of breast cancer to lower the intensity of carcinoma. Accuracy needed enhancement was future scope and to consider larger dataset than the one used for more efficient features in breast cancer. (Umesh and Ramchandra; 2015) it was observed by that menopause and size of lymph nodes showed high correlation in causing breast cancer. Patients of above 30 age happened to have high cancer rate than younger people. Tumor size more than 50 were at high time to be coined as malignant breast cancer.

2.2 Cervical cancer

2.2.1 Introduction

Around 200,000 women lost their lives suffering from cervical cancer and out of that sample 90 % of those deaths occurred in countries that had poor to medium developed economy in the year 2012. Cervical cancer has been now the world's highest terrifying death's causing carcinoma disease among women in developed and undeveloped countries. (Kurniawati et al; 2016) suggests 14,500 cases were registered yearly in Indonesia and this estimate has thrown light in the statistics from world health organization quoting Indonesia is the country with highest rates of mortality by cervical cancer.

(Hasan et al.; 2017) In united states, only about 65 % of female patients out of the sample of 16 million went under pap smear screening tests cervical cancer. Probability of women who can develop cervical cancer without pap smear screening test in New Zealand was 0.01 and probability of women who may lose their lives from this dreadful disease was 0.005. Screening tests for cervical cancer has lowered the carcinoma rate by 65 % and mortality rates by 90 %. Reason behind less screening tests are due to less exposure and minimal knowledge in people to cervical cancer. Cervical cancer screening has become a taboo in low economy countries. Treatment for this carcinoma is expensive than any other infections causing agents in medical field. (Nirmal et al; 2013) yearly expenditure for cervical cancer treatments has crossed over 4 million United states dollars by the state. Reason or cause and spread of cancers in general is not detecting them in early stages.

Cervical cancer is caused due to spread or untreated condition of human papillomavirus also known as HPV. This is a virus that has affected patients via sexual contact. HPV conditions can be cured easily as compared to other infections from sexual contact such as herpes and gonorrhea.

(Kurniawati et al; 2016) treatment of cervical cancer involves several screenings from patients before proceeding with biopsy medical procedures and those include Pap test, Pap smear test and human papillomavirus test. Pap smear test procedure is of taking cells right from cervix to examine it under microscope for cancer spread, it also examines non-carcinoma regions like infections for clarity since inflammations and infections both can be mis concluded during focus from concave lens in equipment's. Pap smear test also finds out other carcinoma inside cervical area like ovaries complications and endometrial corpus cancer that originates at fallopian tubes of a woman.

Symptoms when a woman has developed cervical cancer are bleeding after intercourse or after nominal monthly cycles and bleeding followed after menopause. (Shetty and Shah;

2018) suggests symptoms of heavy and irregular menstruation will only be seen as part of cervical cancer.

2.2.2 Methodologies

(Hasan et al.; 2017) used Microsoft excel software to enroll female participants and prepared questionnaires which were distributed to female participants out of which who required help with exposure to cervical cancer or had interest in enrolling for screening process were sorted to followed up procedures. Human papillomavirus being the prominent and root cause behind all cervical cancer complication cases registered.

Use of classifiers and regressors from machine learning world into this medical field can help drastically in lowering carcinoma deaths. (Kurniawati et al.; 2016) used naïve Bayes, support vector machines and random forest data mining model's fore prediction with dataset containing information of patients, their symptoms and biopsy results from pap smear screening tests. Utilization of statistical performance metrics and ROC curves helped during validation.

2.2.3 Results of developed models

(Shetty and Shah; 2018) preferred the usage of support vector machine which yielded him accuracy of 78 % where all features were considered into the training model. (Kurniawati et al.; 2016) used naïve Bayes, support vector machines and random forest data mining model's got accuracy of 78 %, 78.66 %, 80 % and precision metrics of 69 %, 66 % and 75 % respectively. Other authors considered enrollment and gathering of data rather than applying machine learning models to be more prominent.

2.2.4 Conclusion

(Shetty and Shah; 2018) not many infections fade away or cure itself when given time unlike HPV. HPV can be said harmless if treated at early stages and infection prone age group are adults. (Peng and Zhao; 2016) the problem of cervical cancer being age bounded complication among women, only patients of age group 35 to 65 had cervical screenings and cancer and average age among the dataset was 47 years.

3 Research Methodology

KDD approach is the pillar for this research. KDD stands for "Knowledge discovery and databases" and approach is basically a methodology already followed in many industries to have concise data work flow pipeline that helps industries not only in gathering data but throughout the development process till evaluation. KDD is well known for its definite results in a high workflow architecture. In contrast to KDD, there are other methodologies present but their focus is more on the implementation side of work flow and in deploying a product to the market whereas in this research no such methodologies are required as no product being developed or deployed. (Azevedo and Santos; 2008) suggests KDD involves phases that forms the prominent pillars to work with data and they are **selection, preprocessing, transformation, data mining** and **evaluation**. **Selection phase** involves making data accessible for working that is to make retrieved or fetched information from various sources ready for integrating into an environment to start functioning upon.

Information is well studied to understand the base problem, choosing target variables and other features that serves to contribute to the problem. **Preprocessing phase** involves tidying up the data retrieved from sources as different sources have different format of information storage. If worked on image dataset then removing of noise and rescaling are form of data cleaning and if worked with numerical datasets then removal of incomplete values, missing values and outliers are performed. **Transformation phase** involves combining many columns one column without losing information and also selecting proper attributes that contribute to target variable by removable of irrelevant variables. **Data mining phase** is models are applied over the split target and features to perform prediction. **Evaluation phase** involves summarization of all results from executed models and analyzing for patterns to focus on research question.

Furthermore, use of KDD methodology phases in conjugation with this research is explained below.

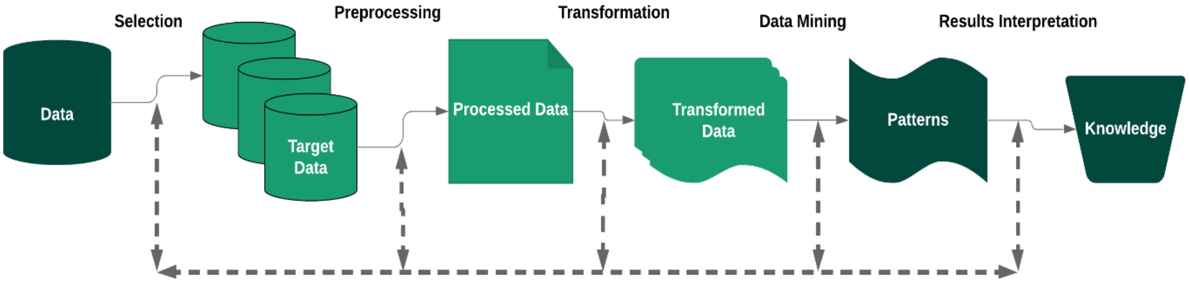


Figure: KDD methodology

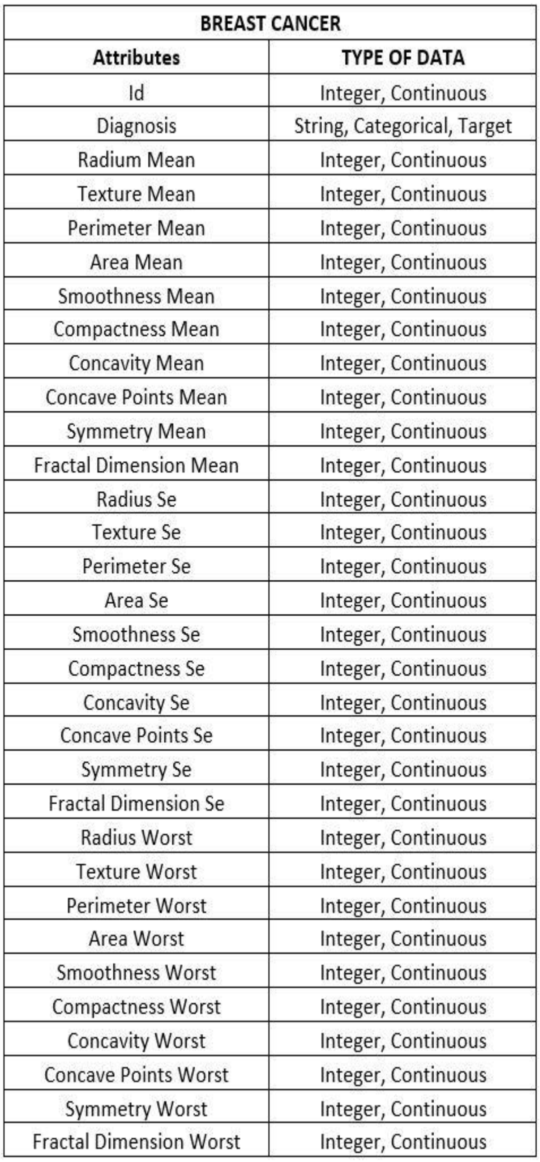
3.1 Data selection

Research needed two datasets from the source since its dealing with two different cancers in women. The motto of the project has to be satisfied by considering dataset that has biopsy results that are the cancer screening results of patients. Any information containing patient details and sensitive information have been avoided while selection.

Breast and cervical cancer datasets were accessed from website Kaggle. Breast and cervical cancer both had biopsy results as the target variable for prediction to split the dataset into features of what are causing the carcinoma spread and what the results of screening are concluded.

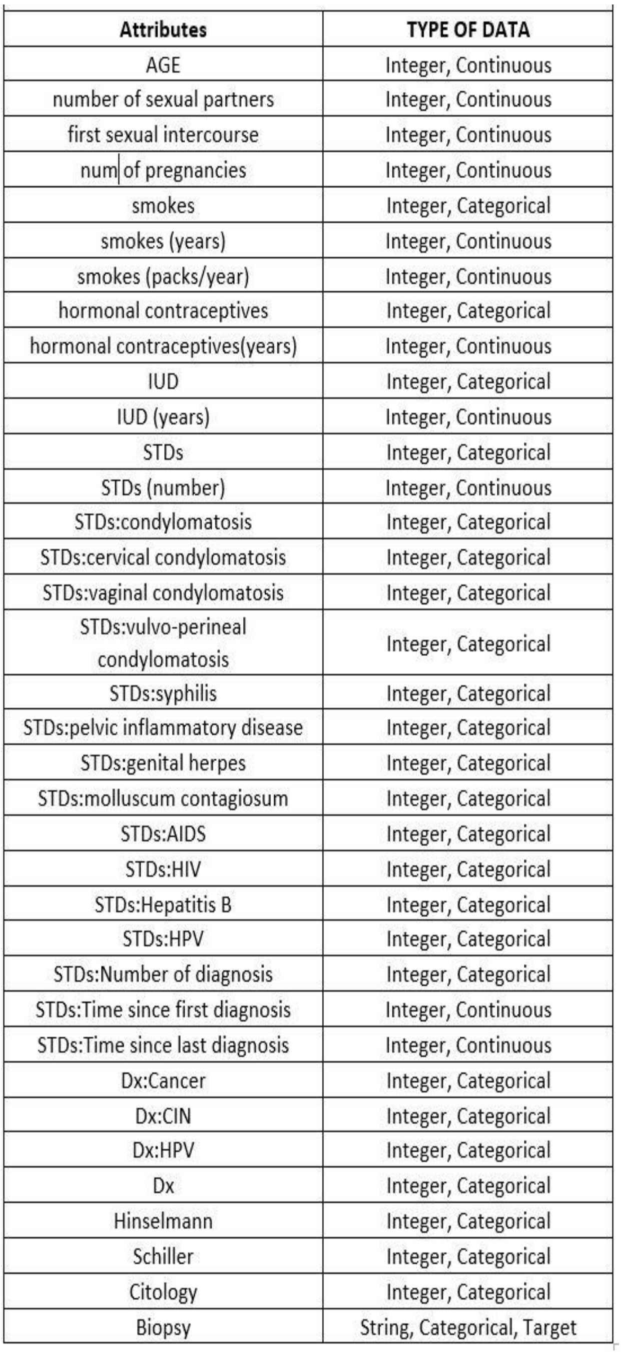
3.2 Data preprocessing

The prominent step in any data project is data preprocessing. Here the data are checked for outliers and undergone cleaning. If any cells have missing values then those observations will be dropped or if any cells are filled with not available values then they are replaced with mean value of that particular column where the cell resides. Some columns having information in string format have to converted to numerical or categorical variables respectively. Irregular symbols and signs to be dropped from datasets.



BREAST CANCER	
Attributes	**TYPE OF DATA**
Id	Integer, Continuous
Diagnosis	String, Categorical, Target
Radium Mean	Integer, Continuous
Texture Mean	Integer, Continuous
Perimeter Mean	Integer, Continuous
Area Mean	Integer, Continuous
Smoothness Mean	Integer, Continuous
Compactness Mean	Integer, Continuous
Concavity Mean	Integer, Continuous
Concave Points Mean	Integer, Continuous
Symmetry Mean	Integer, Continuous
Fractal Dimension Mean	Integer, Continuous
Radius Se	Integer, Continuous
Texture Se	Integer, Continuous
Perimeter Se	Integer, Continuous
Area Se	Integer, Continuous
Smoothness Se	Integer, Continuous
Compactness Se	Integer, Continuous
Concavity Se	Integer, Continuous
Concave Points Se	Integer, Continuous
Symmetry Se	Integer, Continuous
Fractal Dimension Se	Integer, Continuous
Radius Worst	Integer, Continuous
Texture Worst	Integer, Continuous
Perimeter Worst	Integer, Continuous
Area Worst	Integer, Continuous
Smoothness Worst	Integer, Continuous
Compactness Worst	Integer, Continuous
Concavity Worst	Integer, Continuous
Concave Points Worst	Integer, Continuous
Symmetry Worst	Integer, Continuous
Fractal Dimension Worst	Integer, Continuous

Figure: Breast cancer dataset attributes



Attributes	TYPE OF DATA
AGE	Integer, Continuous
number of sexual partners	Integer, Continuous
first sexual intercourse	Integer, Continuous
num of pregnancies	Integer, Continuous
smokes	Integer, Categorical
smokes (years)	Integer, Continuous
smokes (packs/year)	Integer, Continuous
hormonal contraceptives	Integer, Categorical
hormonal contraceptives(years)	Integer, Continuous
IUD	Integer, Categorical
IUD (years)	Integer, Continuous
STDs	Integer, Categorical
STDs (number)	Integer, Continuous
STDs:condylomatosis	Integer, Categorical
STDs:cervical condylomatosis	Integer, Categorical
STDs:vaginal condylomatosis	Integer, Categorical
STDs:vulvo-perineal condylomatosis	Integer, Categorical
STDs:syphilis	Integer, Categorical
STDs:pelvic inflammatory disease	Integer, Categorical
STDs:genital herpes	Integer, Categorical
STDs:molluscum contagiosum	Integer, Categorical
STDs:AIDS	Integer, Categorical
STDs:HIV	Integer, Categorical
STDs:Hepatitis B	Integer, Categorical
STDs:HPV	Integer, Categorical
STDs:Number of diagnosis	Integer, Categorical
STDs:Time since first diagnosis	Integer, Continuous
STDs:Time since last diagnosis	Integer, Continuous
Dx:Cancer	Integer, Categorical
Dx:CIN	Integer, Categorical
Dx:HPV	Integer, Categorical
Dx	Integer, Categorical
Hinselmann	Integer, Categorical
Schiller	Integer, Categorical
Citology	Integer, Categorical
Biopsy	String, Categorical, Target

Figure: Cervical cancer dataset attributes

3.3 Data transformation

After datasets goes through thorough cleaning, time for feature selection to select proper attributes that contribute target prediction and avoids over fitting. Feature selection step is mandatory if bigger datasets are considered as they will reduce the training time significantly compared to normal execution with all features.

3.4 Data mining

For this classification-based research, it was appropriate to utilize the power of decision trees, support vector machines, artificial neural networks, logistic regression, k nearest neighbor, random forest and ensemble learners by combining models already executed for increased accuracy, precision and recall.

3.5 Data evaluation

As classification models were applied over datasets, accuracy of model performance, precision of predicted true values and recall of actual true values from the models were evaluated.

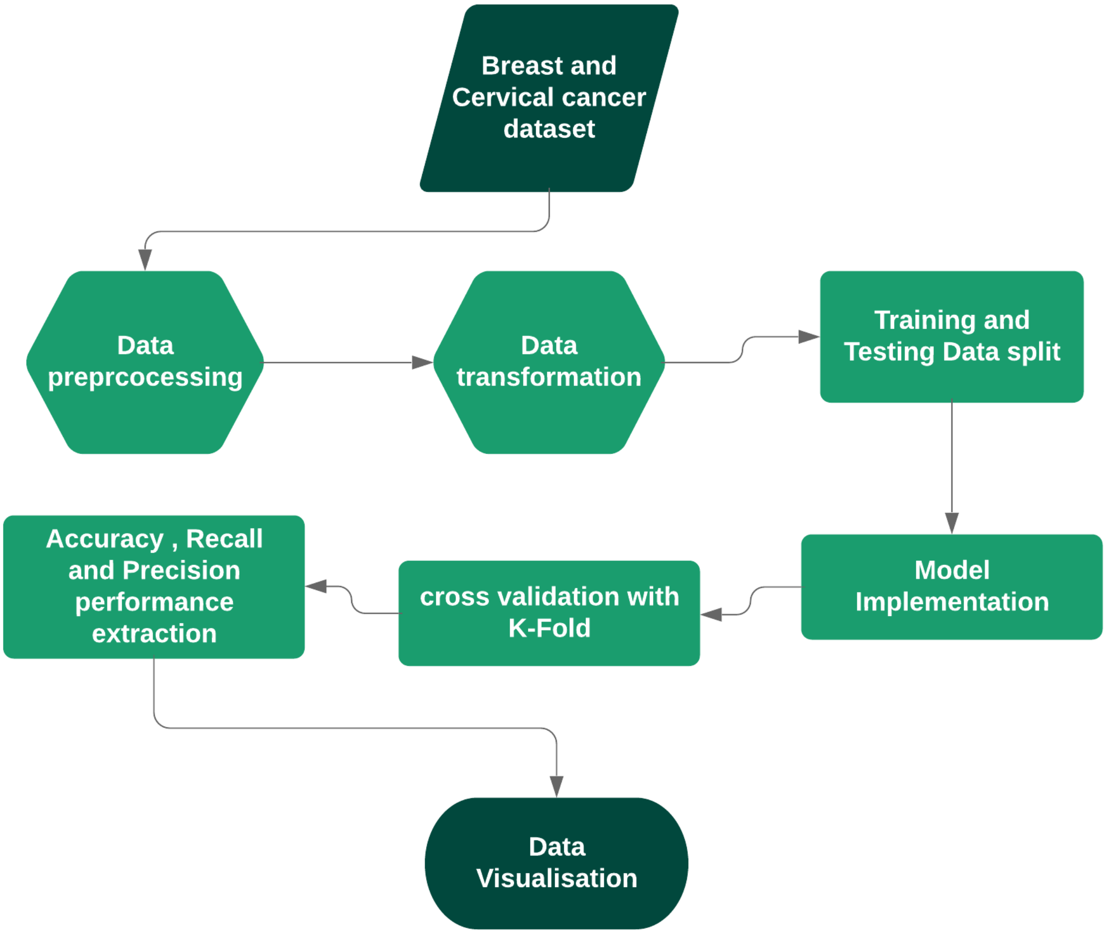

Breast and Cervical cancer dataset
Data preprcocessing
Data transformation
Training and Testing Data split
Model Implementation
cross validation with K-Fold
Accuracy , Recall and Precision performance extraction
Data Visualisation

Figure: Process flow diagram of cancer datasets

4 Implementation

First and second phases from KDD are basic implementation as in laying the groundwork for core third phase which is implementation of data mining models over dataset.

4.1 Exploration of datasets

Understanding of dataset was essential and exploring through the details of information it was observed that for breast cancer, area, perimeter, smoothness, radius, concavity, compactness of lymph nodules and texture of breast regions can act as features. For cervical cancer, smoking, pregnancies, intercourse encounters, consumption of hormonal contraceptives, statuses of sexually transmitted diseases can act as features for prediction.

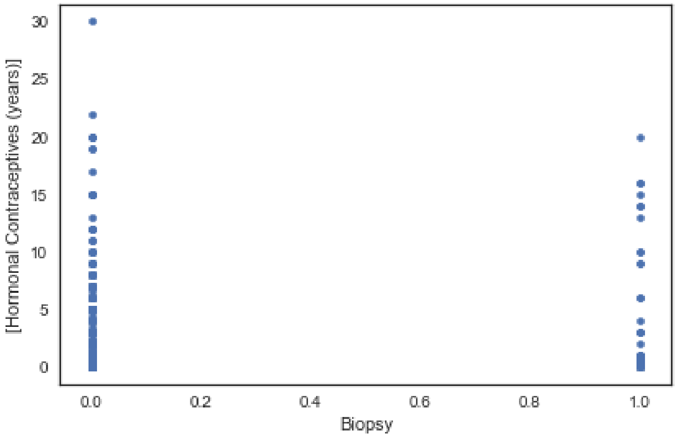


Figure: Cervical patients and consumption of hormonal contraceptives

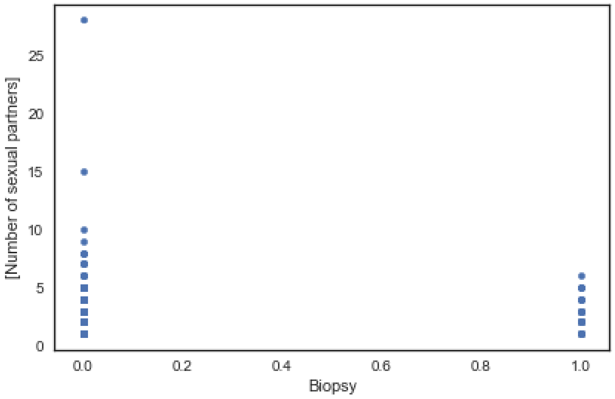


Figure: Cervical patients and number of sexual partners

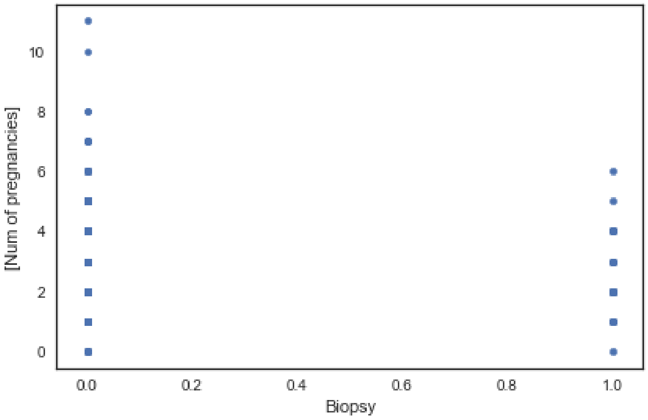


Figure: Cervical patients and pregnancy counts

4.2 Cleaning

In breast cancer dataset, last column was removed as it had no value but only empty cells occupying the computer resource and also id column was removed as it could mislead while fitting into models. Several columns had numbers captured in string format which had to be turned to proper integers to further avoid errors while fitting into data mining models. Biopsy results column was converted to dichotomous column from string variable using for loop in python as it was the target. After these steps final check for not available values were done to ensure there were no missing values.

In cervical cancer dataset, some cells with question marks were replaced with not available values. Since presence of not available values were more than expected, all those cells were replaced column wise with their mean value corresponding to the variable. Continuous and categorical were properly converted from strings to integers and dichotomous variables.

Both datasets had outliers but were not removed as a smaller number of observations will affect the quality of model training since both were around only 500 rows and 13 columns in dimension.

4.3 Model input preparation

Datasets were reasonable in length so feature selection was not necessary. Only correlation of all columns in both datasets were observed and attributes were selected accordingly. Data was split into training and testing using K-fold validation. K fold validation is better way of training data before running data mining models as in traditional splitting testing data never gets used while in k-fold validation testing data is used after training and also avoids overfitting complications while training the model.

4.4 Implementation of Artificial neural networks

Artificial neural networks were used from the library Keras. In **breast** and **cervical cancer** datasets, four hidden layers were implemented in the network for training the data and applying cost function, first layer with 24 neurons that are the nodes for processing the incoming 35 inputs attributes with relu activation function, second layer with 12 neurons with relu, third with 6 neurons with relu and final fourth with single neuron with sigmoid activation function.

Relu function handled multiple inputs well, optimized 0 to 1 flat value and sigmoid at last layer was for output that needed to be classified as binary value. 50 Iterations were run over dataset to ensure cost function met its saturation point. Selected attributes were sent to the network in form of arrays. Model performed significantly faster as the load was put on graphics processing unit. Keras library has functionality in adjusting weights after every iteration through automation where cost function work got accomplished.

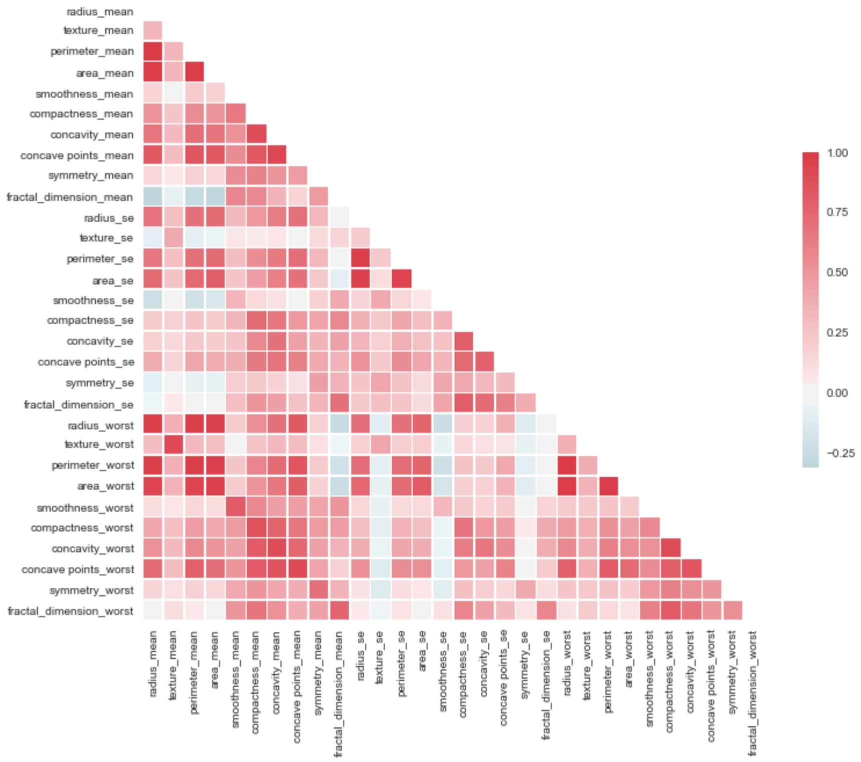


Figure: Correlation matrix of breast cancer variables

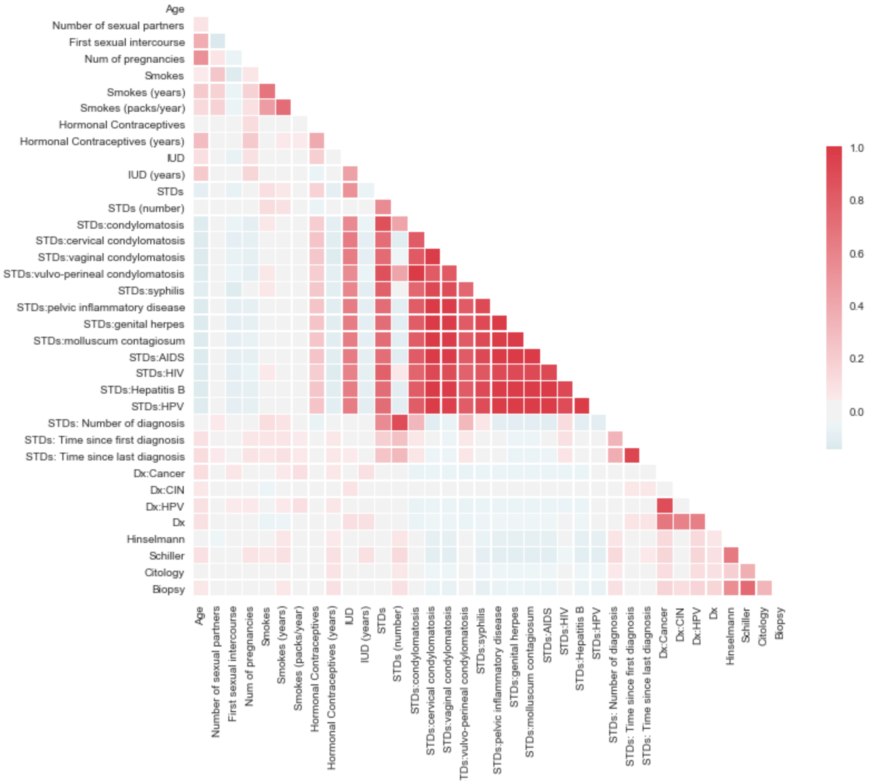


Figure : Correlation matrix of cervical cancer variables

4.5 Implementation of Logistic regression

As the dependent attribute is dichotomous, there was a hurdle in dataset where the dependent variable was in string format and underwent data type conversion to integer 32 bytes to run over the model. Model ran quicker than expected and cross validation using K-fold validation. Random state while training model was set to false as data split and validation handled by k-fold validation. The inputs were continuous variable and output were categorical variable.

4.6 Implementation of K nearest neighbor

Categorical independent variable enabled k nearest neighbor to form classes for all input attributes and form circles around each class. Iterated for loop over classifier to evaluate an efficient k value that was later used to again train the classifier for maximum efficiency. The optimum value of k by breast cancer classifier was 24 and by cervical cancer classifier was 15.

4.7 Implementation of Decision trees

Decision trees were initiated with base parameters for number of levels and branches in every node. Training and testing the model was processed through k-fold cross validation as it not only makes use of every subset of data, accurately validates and avoided overfitting.

4.8 Implementation of Random Forest

Random forest is the ensemble version of decision trees as it has voting after several decision trees are executed and from the results the tree that has maximum votes gets considered. Implemented random forest with 50 trees that are counts of trees used to run the model in the forest and random state set to 5 that corresponds to seed value of randomizing data. Random forest is better a classifier than decision trees as it avoids all overfitting complications from datasets. Default values were taken for number of features at every split and for number of levels up to which tree can build.

4.9 Implementation of Ensemble learner model

This phase is the core of this research, as it deals with the objective of this research to develop a model and attain accuracy that will benefit in predicting the future biopsy results of patients before the convergence of malignant stage from benign occur. Ensemble learning is the innovation in machine learning that enables to concatenate the computing potential of various models and get the best from all the worlds.

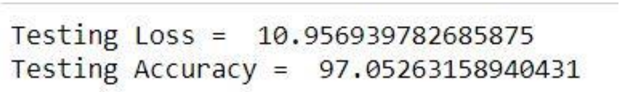
Here, voting algorithm was utilized with Logistic regression, k nearest neighbor and random forest acting as input models whose results were selected by the base classifier. The other method used was bagging algorithm that ran varied subset of original information from training data and averaged out the final results from all the subsets.

5 Evaluation

Data mining models for classification datasets are evaluated on basis of recall, precision and accuracy of the model. Overfitting and underfitting of data mining models are also considered from evaluation point of view. Every model underwent two ways of validation which is solo run and one with k-fold validation run except Artificial neural networks and ensemble learners as results from these two high functional models are already averaged out results.

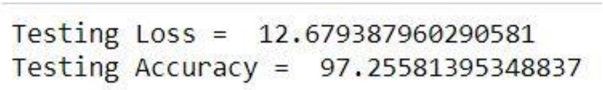
5.1 Evaluation of Artificial neural networks

Every hidden layer produces varied outputs and cost function at end of every epoch was checked which then got weight adjusted to secrete final output. Neural network's accuracy corresponds to correct predictions of the number of patients those who had benign cancer and who suffered from malignant cancer. Expected overfitting did not occur as epoch iteration count given met saturation point.



```
Testing Loss =  10.956939782685875
Testing Accuracy =  97.05263158940431
```

Figure : Neural networks in Breast cancer

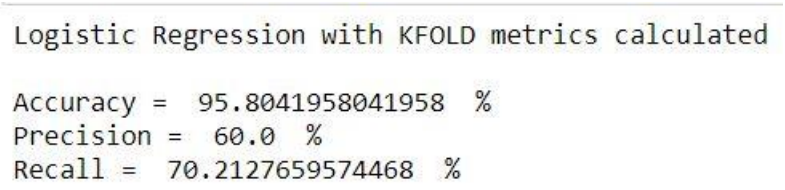


```
Testing Loss =  12.679387960290581
Testing Accuracy =  97.25581395348837
```

Figure : Neural networks in cervical cancer

5.2 Evaluation of logistic regression

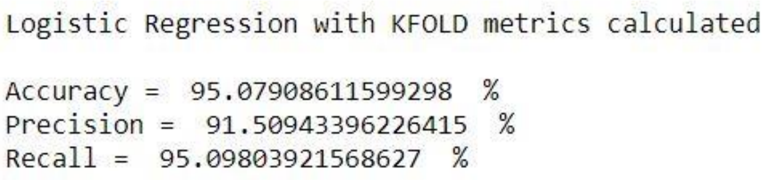
Both breast and cervical cancer models yielded same accuracy of 95%. The percentage of patients who suffered from carcinoma and were predicted accurately by breast cancer model with 60% and by cervical cancer model with 91%. Drastic differences in the precision rate even though trained by exact same model.



```
Logistic Regression with KFOLD metrics calculated

Accuracy =  95.8041958041958  %
Precision =  60.0  %
Recall =  70.2127659574468  %
```

Figure : Logistic regression in Breast cancer



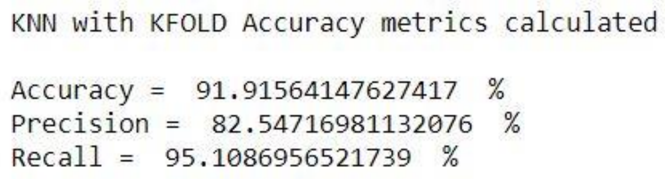
```
Logistic Regression with KFOLD metrics calculated

Accuracy =  95.07908611599298  %
Precision =  91.50943396226415  %
Recall =  95.09803921568627  %
```

Figure : Logistic regression in cervical cancer

5.3 Evaluation of K nearest neighbour

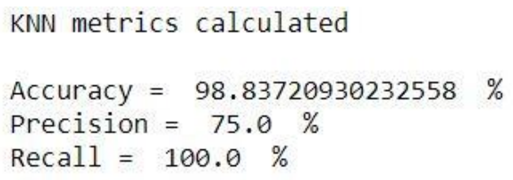
The percentage of patients who were predicted as having carcinoma and when compared by actual dataset had carcinoma was 95% recall in breast cancer model and 100% recall in cervical cancer model. The cervical model experienced overfitting during k-fold validation. This model experienced low accuracy and precision in k-fold scores compared to all other models.



```
KNN with KFOLD Accuracy metrics calculated

Accuracy =  91.91564147627417  %
Precision =  82.54716981132076  %
Recall =  95.1086956521739  %
```

Figure : K nearest neighbor in Breast cancer



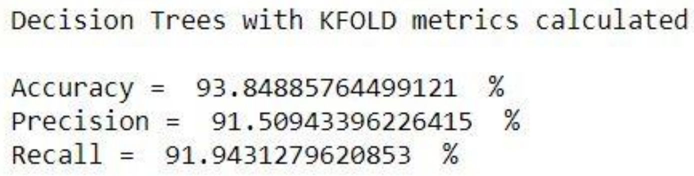
```
KNN metrics calculated

Accuracy =  98.83720930232558  %
Precision =  75.0  %
Recall =  100.0  %
```

Figure : K nearest neighbor in cervical cancer

5.4 Evaluation of decision trees

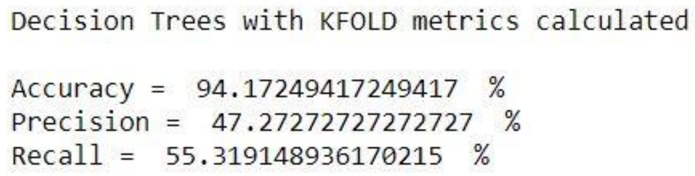
Overfitting was experienced when deployed with 100 number of trees and random state set to 30 so reduction of parameters were found to be necessary. Decision trees experienced low accuracy score out of all models in breast cancer and more accuracy than logistic regression during k-fold procedure.



```
Decision Trees with KFOLD metrics calculated

Accuracy =  93.84885764499121  %
Precision =  91.50943396226415  %
Recall =  91.9431279620853  %
```

Figure : Decision trees in Breast cancer



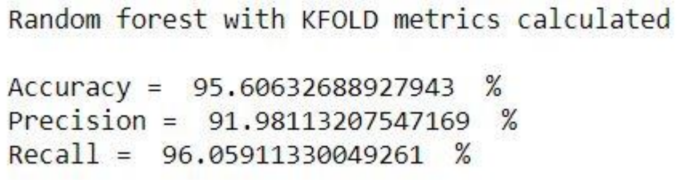
```
Decision Trees with KFOLD metrics calculated

Accuracy =  94.17249417249417  %
Precision =  47.27272727272727  %
Recall =  55.319148936170215  %
```

Figure : Decision trees in cervical cancer

5.5 Evaluation of Random forest

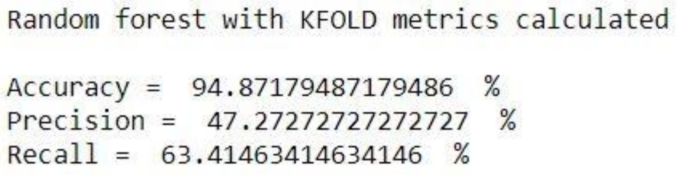
Highest non-ensemble k-fold accuracy was attained with random forest reaching up to 95% in single run in breast cancer and 94% in cervical cancer and also highest accuracy in non-k-fold validation.



```
Random forest with KFOLD metrics calculated

Accuracy =  95.60632688927943  %
Precision =  91.98113207547169  %
Recall =  96.05911330049261  %
```

Figure : Random forest in Breast cancer



```
Random forest with KFOLD metrics calculated

Accuracy =  94.87179487179486  %
Precision =  47.27272727272727  %
Recall =  63.41463414634146  %
```

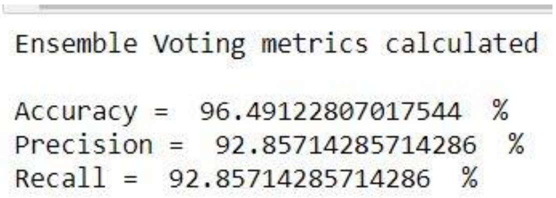
Figure : Random forest in cervical cancer

5.6 Evaluation of Ensemble learner model

Training data were successfully inserted into ensemble voting and bagging algorithms that yielded high accuracy under voting approach for both breast and cervical cancer datasets.

5.6.1 Voting

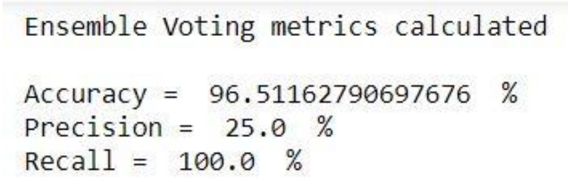
Voting was the combination of selecting best votes from logistic regression, random forest and k nearest neighbor models. As observed from single model predictions above, highest votes were registered from logistic regression and random forest.



```
Ensemble Voting metrics calculated

Accuracy =  96.49122807017544  %
Precision =  92.85714285714286  %
Recall =  92.85714285714286  %
```

Figure : Ensemble voting in breast cancer model



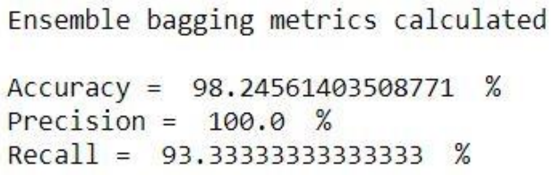
```
Ensemble Voting metrics calculated

Accuracy =  96.51162790697676  %
Precision =  25.0  %
Recall =  100.0  %
```

Figure : Ensemble voting in cervical cancer model

5.6.2 Bagging

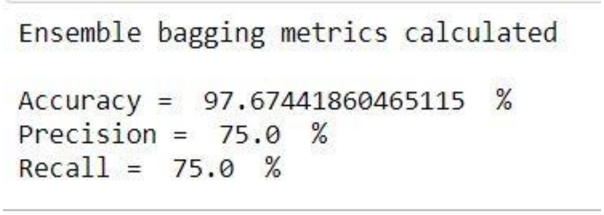
Randomly selection of training data to train data in such a way that all data gets utilized is the bagging approach and they yielded high accuracy, precision and recall in breast and cervical cancer datasets than voting methods executed.



```
Ensemble bagging metrics calculated

Accuracy =  98.24561403508771  %
Precision =  100.0  %
Recall =  93.33333333333333  %
```

Figure : Ensemble bagging in breast cancer model



```
Ensemble bagging metrics calculated

Accuracy =  97.67441860465115  %
Precision =  75.0  %
Recall =  75.0  %
```

Figure : Ensemble bagging in cervical cancer model

5.7 Discussion

To sum up, K-fold cross validation avoided all the over fitting that were caused in single prediction run by logistic regression, random forest, k nearest neighbor and decision trees. To the surprise, random forest not only outperformed artificial neural networks in these two short datasets, also had highest accuracies from single training runs.

Thus, the objective of this research was accomplished by running ensemble voting and bagging approaches over breast and cervical cancer datasets which attained high accuracies by evading under fitting and overfitting of short dataset problems.

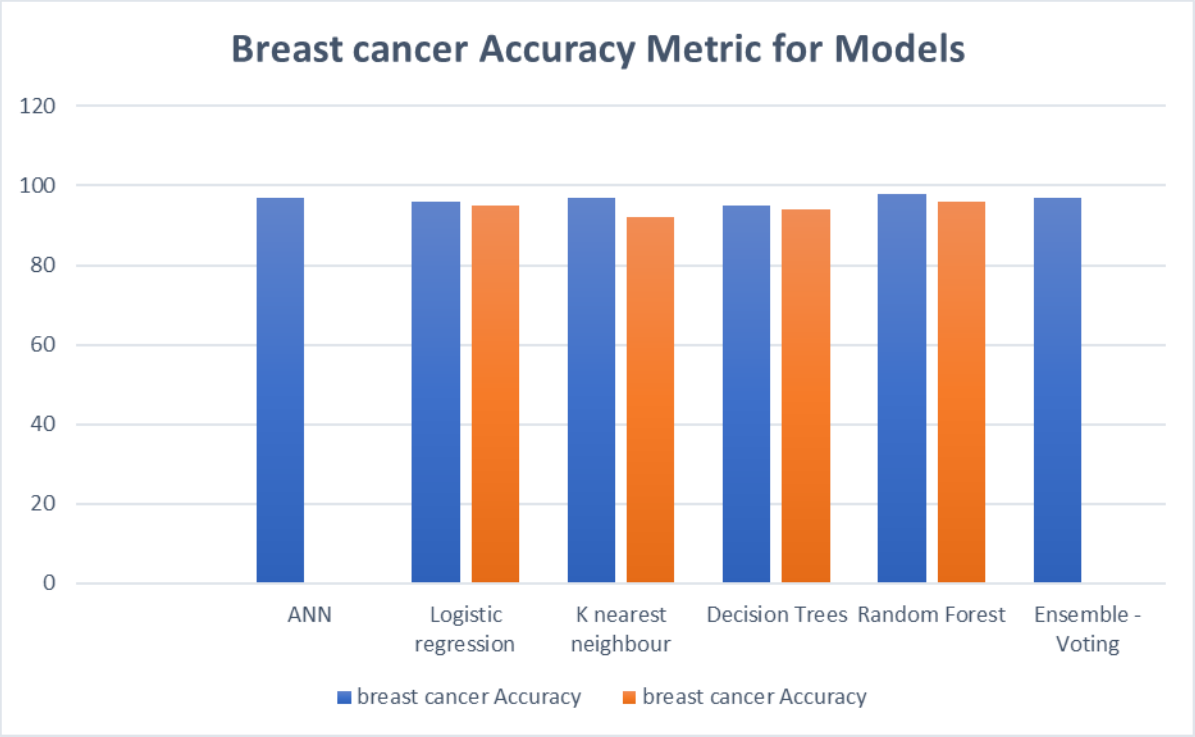


Figure : Breast cancer Accuracy of models with and without k-fold cross validation

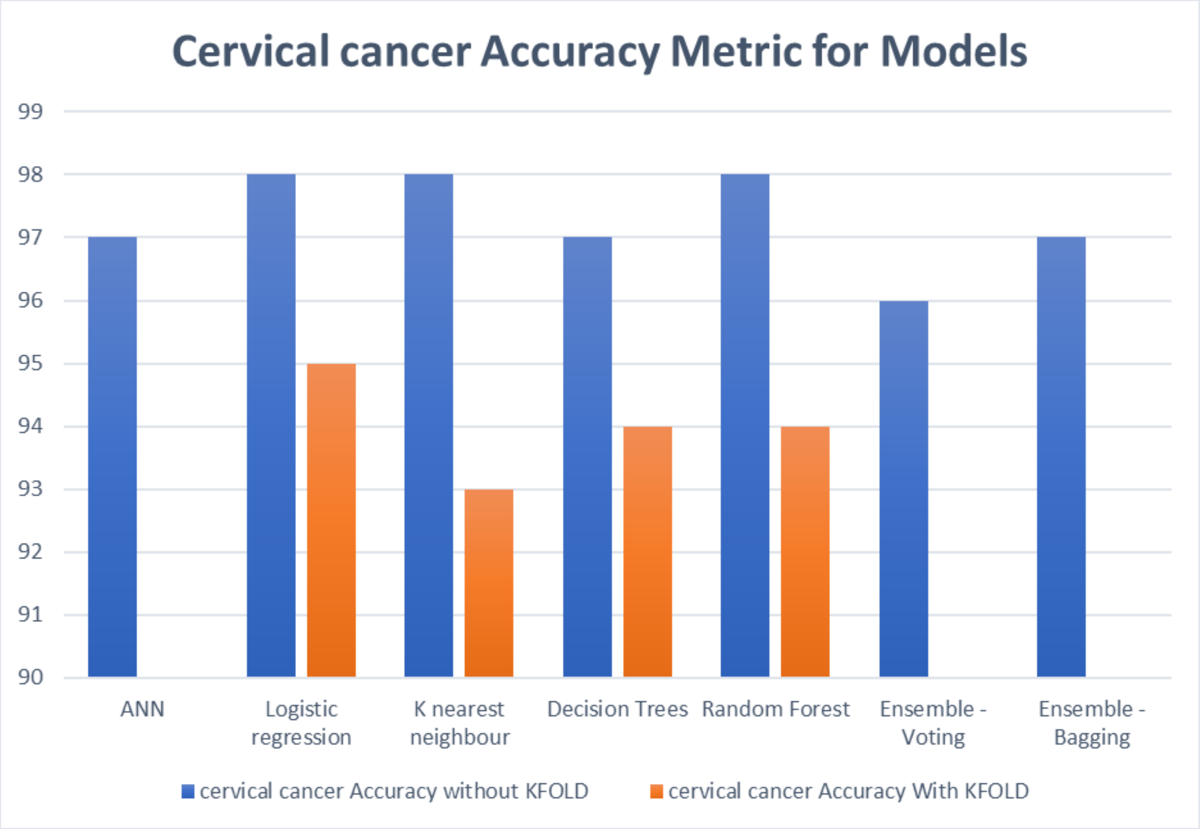


Figure : Breast cancer Accuracy of models with and without k-fold cross validation

6 Conclusion and Future Work

To perform implementation of Ensemble learner model that will predict with accuracies, minimal errors and outperform single prediction models were successfully accomplished. Less to zero errors in analyzing and prediction will help save trillions of lives if fit to present computer aided systems for screening procedures. (Khariwal and Mishra; 2018) also performed the same ensemble learning with voting algorithm and attained 98.50 % accuracy but did not perform bagging or stacking over his research problem and out of which bagging was performed in this research. Random forest and ensemble voting approach performed rich by predicting the exact patients who had cancer and who did not have cancer that is the accuracy, less error ratio from precision and recall which are patients for whom it was wrongly classified.

It was analyzed that in **breast cancer,** patients even with identical medical knowledge, history and records had different types stages of carcinoma. Area, perimeter and circumference size of lymph nodules in breasts above 0.3 had high rates of spreading from benign to malignant stages in no time. Age group who had high breast cancers in the dataset were above 20 and below 35.

It was analyzed that in **cervical cancer,** surprisingly patients with no signs of HIV (Human immunodeficiency virus) and STD (sexually transmitted diseases) suffered from malignancies. Age group when analyzed was seen to be above 20 and below 30 in the dataset. Patients who held pregnancy phase were medium to low likely inclined to have cervical cancer.

Research has to be done with larger dataset of at least a million patients records comprising of all symptoms from breast and cervical cancer for accurate results as it is dealing with patient’s life. Incorrect classification rates should be always lower as this research domain deals with valuable life of a human being.

References

Giger, M.L. and Huo, Z., 1999. Artificial neural networks in breast cancer diagnosis: Merging of computer-extracted features from breast images. In *Evolutionary Computation, 1999. CEC 99. Proceedings of the 1999 Congress on* (Vol. 3, pp. 1768-1769). IEEE.

Ahmad, F.K. and Yusoff, N., 2013, December. Classifying breast cancer types based on fine needle aspiration biopsy data using random forest classifier. In *Intelligent Systems Design and Applications (ISDA), 2013 13th International Conference on* (pp. 121-125). IEEE.

Umesh, D.R. and Ramachandra, B., 2015, December. Association rule mining based predicting breast cancer recurrence on SEER breast cancer data. In *Emerging Research in Electronics, Computer Science and Technology (ICERECT), 2015 International Conference on* (pp. 376-380). IEEE.

Ojha, U. and Goel, S., 2017, January. A study on prediction of breast cancer recurrence using data mining techniques. In *Cloud Computing, Data Science & Engineering-Confluence, 2017 7th International Conference on* (pp. 527-530). IEEE.

Kermani, B.G., White, M.W. and Nagle, H.T., 1995, September. Feature extraction by genetic algorithms for neural networks in breast cancer classification. In *Engineering in Medicine and Biology Society, 1995., IEEE 17th Annual Conference* (Vol. 1, pp. 831-832). IEEE.

Khuriwal, N. and Mishra, N., 2018, March. Breast cancer diagnosis using adaptive voting ensemble machine learning algorithm. In *2018 IEEMA Engineer Infinite Conference (eTechNxT)* (pp. 1-5). IEEE.

Gupta, M. and Gupta, B., 2018, February. A Comparative Study of Breast Cancer Diagnosis Using Supervised Machine Learning Techniques. In *2018 Second International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 997-1002). IEEE.

Fogel, D.B., Angeline, P.J., Porto, V.W., Wasson, E.C. and Boughton, E.M., 1999. Using evolutionary computation to learn about detecting breast cancer. In *Proceedings of the 1999 Congress on Evolutionary Computation-CEC99 (Cat. No. 99TH8406)* (Vol. 3, pp. 1749-1754). IEEE.

Shetty, A. and Shah, V., 2018, July. Survey of Cervical Cancer Prediction Using Machine Learning: A Comparative Approach. In *2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1-6). IEEE.

Hasan, M.R., Gholamhosseini, H., Sarkar, N.I. and Safiuzzaman, S.M., 2017, December. Intrinsic motivated cervical cancer screening intervention framework. In *Humanitarian Technology Conference (R10-HTC), 2017 IEEE Region 10* (pp. 506-509). IEEE.

Nirmal, R., Yun, C., Le, M., Paripoonnanonda, P. and Yi, J., 2013, September. Digital health game on cervical health and its effect on american women's cervical cancer knowledge. In *Games Innovation Conference (IGIC), 2013 IEEE International* (pp. 191-198). IEEE.

Kurniawati, Y.E., Permanasari, A.E. and Fauziati, S., 2016, October. Comparative study on data mining classification methods for cervical cancer prediction using pap smear results. In *Biomedical Engineering (IBIOMED), International Conference on* (pp. 1-5). IEEE.

Peng, W. and Zhao, Y., 2016, October. Comparative of patients with cervical cancer using 3D-CRT and IMRT. In *Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), International Congress on* (pp. 1849-1853). IEEE.

Azevedo, A.I.R.L. and Santos, M.F., 2008. KDD, SEMMA and CRISP-DM: a parallel overview. *IADS-DM*.

Kai, L. and Ping, Z.Z., 2012, August. Using an Ensemble Classifier on Learning Evaluation for E-learning System. In *Computer Science & Service System (CSSS), 2012 International Conference on* (pp. 538-541). IEEE.