APPLICATION OF LOGISTIC REGRESSION AND NEURAL NETWORK MODELS TO BREAST CANCER PATIENTS

A project submitted in partial fulfillment of the requirements for the degree of

Master of Science

by

Eric Agyemang
eagyem2@ilstu.edu

December, 2020

DECLARATION

I Eric Agyemang here by declare that the Project entitled "Application of Logistic Regression and Neural Network Models to Breast Cancer Patients" submitted by me, for the award of the degree of *Master of Science* to Illinois State University is a record of bonafide work carried out by me under the supervision of Dr. Pei Geng, Illinois State University, Faculty of Arts and Sciences, Department of Mathematics, Normal IL, USA.

I further declare that the work reported in this project has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Place: Normal IL

Date:12/08/2020 **Signature of the Candidate**

CERTIFICATE

This is to certify that the Project entitled "Application of Logistic Regression and Neural Network Models to Breast Cancer Patients" submitted by Mr. ERIC AGYEMANG, Illinois State University, Faculty Of Arts And Sciences, Department of Mathematics, Normal IL for the award of the degree of *Master of Science*, is a record of bonafide work carried out by him/her under my supervision.

The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university. The project fulfills the requirements and regulations of the University and in my opinion meets the necessary standards for submission.

Place: Normal IL

Date: 12/08/2020 **Project Supervisor**

(Dr. Pei Geng)

ABSTRACT

Breast cancer is a disease in which the cells in the breast grows out of control. It is the second most common cause of death from cancer in women after lung cancer and over 50 thousand new cases are reported every year in the United States. The survival rates of breast cancer depend on many factors. Using a sample of 100,002 breast cancer patients from three states in the United States covering the period 1992 – 2019, this study focuses on developing predictive models based on patient's demographic risk factors, and mammographic descriptors to predict breast cancer patient's last status (Alive or dead) and examine the influence of these factors on the patient's last status. Using machine learning and deep learning techniques, five models were developed using logistic regression, feed forward neural networks, and Random Forest Classification to predict the breast cancer patient's last status for both sex and each sex separately. We compared the output of all models and found that the Convolutional Neural Network outperformed all models in both sex and each sex separately. However, in assessing the influence of the risk factors on patient's last status, the Random Forest is proposed to be the best fit compared with all the other model. This work provides insight into increasing the effectiveness of machine learning in contributing to improve clinical decision making.

Keywords: *Model Evaluation*, *Prediction Accuracy*, *Breast Cancer*, *Abnormal Cells* , *Clinical Outcomes*.

ACKNOWLEDGEMENT

With immense pleasure and deep sense of gratitude, I wish to express my sincere appreciation to my project supervisor Dr. Pei Geng, at the Department of Mathematics, Illinois State University, for her relentless effort to make this project come to a successful end.

I am most grateful to Dr. Krzysztof Ostaszewsk, Dr. Michael Plantholt, and Dr. Ranee Thiagarajah at the Department of Mathematics, Illinois State University, for their advice and directions throughout my course of study in the mathematics department.

I wish to also express my deepest appreciation to the entire staff of Illinois State University Mathematics department and my colleagues for their tremendous contribution they have made in shaping my perspective towards life.

Last but not the least, I would like to thank my wife Gifty Afrakomah and my daughter Kristie Agyemang for their constant encouragement and moral support along with patience and understanding.

Above all, my sincere gratitude goes to the Lord God Almighty for bringing me to this far.

Place: Normal IL

Date: 12/08/2020 **Eric Agyemang**

CONTENTS

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

1.1 Background of Study

Cancer is a disease characterized by an uncontrolled growth and spread of abnormal cells in the parts of the body. An uncontrolled growth and spread of these abnormal cells in a patient's body could result in death. Cancer has become a global health problem causing deep suffering to the individual patients, the patient's family, and the communities at large. Cancer is the second leading cause of death throughout the world (Xu, 2018) and according to the world Health Organization, 58 million people died in 2008 out of which 7.6 million people died out of cancer which was 13% of the total deaths in that year. In 2015, cancer accounted for death of about 8.8 million people all over the globe (WHO, 2017).

Breast cancer is a type of cancer in which the cells in the breast grow out of control. It is the second most common cause of death from cancer in women after lung cancer and over 50 thousand new cases are reported every year in the United States. Breast cancer is a non-skin cancer genetic transition disease caused by factors such as obesity, abnormal hormone, breast density, menstrual history, a sedentary lifestyle, heavy drinking, etc. The risk of getting breast cancer develops in women with the age 62, while an average age of 68 dies from breast cancer. It is mostly diagnosed in women age 55 and 64 while with men between 60 and 70 (see Cancer Treatment Center of America, 2019). There are many types of breast cancer determined based on where it develops in the breast; invasive or non-invasive, driven by hormones or proteins. They include Adenocystic Carcinoma, Angiosarcoma, Ductal carcinoma, Inflammatory breast cancer, lobular Carcinoma, Metaplastic Carcinoma, and Pyllodes Tumor. The survival rate of breast cancer depends on many factors. Surgery is the common treatment option to breast cancer. The other ways of treatment of patients diagnosed of breast cancer include chemotherapy, hormone therapy, radiation therapy, and the targeted therapy (American Cancer Society, 2019).

Accurate prediction of clinical outcomes is an integral to the successful decision making which can result to better patient care (Bartfay, 2006). A person diagnosed of breast cancer who undergo accurate prediction under biopsy on the basis of mammographic findings may help prevent missing a breast cancer or performing biopsy of a noncancerous lesion. Physicians use probabilities calculated using heuristic methods based on training and experience to predict the outcome of a disease (death or alive) although these methods may be very necessary, they may be subject to bias leading to systematic errors (see Kahneman, Slovic and Tversky, 1982).

Computer models could provide some assistance in processing a large number of variables and in bridging the gap between risk factors and risk estimation. A variety of computer models have been developed in the area of machine learning and statistics which can be used to predict clinical outcomes. Some of such models include logistic regression, neural networks (NNs), Bayesian networks, decision trees, etc. Previous studies have shown both logistic regression and NNs to be useful tools in medical diagnosis.

This study is one of the preeminent studies focusing on understanding what factors contribute to males and females breast cancer risk estimation using sampled data on three States in the United States and examining empirically the influence of demographic risk factors and mammographic descriptors on the breast cancer patient's outcome (last status) after being diagnosed and determining comparatively which model best predict patient's outcomes. The accuracy of a mammographic and ultrasound interpretations depends on various factors such as characteristic of the population, patient's age, the experience of the radiologist, and previous knowledge of the BI-RADS guideline (see Liberman, 1998; Hong, 2005).

1.2 Research Objectives

The present study aims to examine the key factors that affect the last status (survival or death) of breast cancer patients within the target population in the United States of America (USA). Here, we aim at assessing the key factors' influence on both male and female patients' last status together, and separately for comparison of the results. We will achieve these using the logistic regression, and neural network models, then evaluate these predictive models, compare the prediction accuracy to determine which model is best for estimating breast cancer risk factors in contributing to improve clinical decision making in the USA.

1.3 Research Questions

This study was guided by the following research question:

1. What key factors affect breast cancer patients' last status within the target population in the USA?
2. What factors affects the last status of male and female breast cancer patients separately within the target population in the USA?
3. Which model is the best fit for predicting breast cancer patient's last status in contributing to improve clinical decision making?

1.4 Significance of the Study

This study does not only contributes to extent literature, it also contribute to creating the awareness, bringing to knowledge and understanding of the factors that affect breast cancer patients, serving as a guide to both patients and physicians which intends to open the doors to a more appropriate means for prevention, accurate method of diagnosis, better prediction of clinical outcomes, and treating breast cancer in the USA. These will help ensure an improved quality of life of breast cancer patients which has a commanding effect of mitigating the mortality rate of the breast cancer patient.

1.5 Scope of the Study

The study is conducted using data collected on the three states San Francisco, Connecticut, and New Jersey in the USA. It seeks to know the factors that affect the last status of all breast cancer patients together within these three states. It also has an objective to examine which factors affects the last status of male and female breast cancer patients separately for the three states and ends by evaluating predictive model to determine the best fitted model for estimating breast cancer risk factors in contributing to improve clinical decision making.

1.6 Organization of the Study

This study is composed of five chapters. The first chapter comprises an introduction to the study, objectives to the study, research questions, significance of the study, scope of the study, and the organization of the study. Chapter two reviews relevant literature and examines the gaps that have made this research necessary. Chapter three is on the study's methodology and discusses the study area, the methods and approaches for the study, the analysis of the data which includes the procedures to be followed to reach the conclusion. Chapter four presents the results generated using Python-Jupiter Notebook tabular formats for easy interpretation. The final chapter five is on the study's key findings and conclusions.

Next, we turn to a brief discussion of the literature review and set our argument for the discussions of findings of study.

CHAPTER 2

LITERATURE REVIEW

There are a number of research works that relate with this study area. One of them was conducted by Turgay et al., 2010 where these researchers compared two of most frequently used computer models in clinical risk estimation, the logistic regression and artificial neural network model in breast cancer risk estimation (see also Chang and Hsu, 2009; Yusuff et al., 2012; Ghamdi, 2002). In this work, the researcher compared these models based on mammographic descriptors and demographic factors to prove complementary in contributing to improve clinical decision making. They found that the main advantage of ANNs over logistic regression model lies in the hidden layers of nodes of the ANNs and are useful when there are implicit interactions and complex relationships in the data, whereas logistic regression model is a better choice when one needs to draw statistical inference from the required output. They concluded that neither models can replace the other but the two may be used complementary to aid decision making.

A similar work was done by Burivong and Amornvithayacharn, 2011 where the researchers tried to determine the accuracy of BI-RADS 4 subcategories using mammographic and ultrasound images of patients who were sub-categorized into 4A, 4B, and 4C (See Balleyguier et al., 2007). These researchers used the Cohen's kappa (k) test to calculate the inter-observer agreement for the description of mammographic, ultrasound lesions, and the subcategory assessment. They concluded that malignancy rate in subcategories 4B and 4C were significantly higher than in 4A. They proposed the improvement of the accuracy of mammographic and breast ultrasound interpretation by means of census evaluation to reduce underestimated malignancy.

Another work which compared the multinomial logistic regression analysis and the discriminant analysis in predicting the stage of breast cancer was conducted by Maiprasert and Krieng, 2012. These researchers also developed and tested predictive models to determine the probability that a patient is detected at any stage of breast cancer or non-breast cancer based on tumor cells characterized by abnormal growth of breast cancer (See also Yuri, Rochadi and Danarto, 2016; Concato, Feinstein Holford, 1993; Bandhita and Noparat, 2006). This work concluded that the ordinal logistic regression model can use few variables in a prediction stage of breast cancer and is 1.4% accurate in prediction that discriminant model based on classification.

Zangmo and Tiensuwan, 2018 employed the logistic regression to identify the factors that affect the survival of all cancer patients as well as male and female patients separately (see also Roy and Guria, 2008). This was done using patients last status as the dependent variable which

included alive, death or unknown. The researcher used the binomial logistic regression for the two cases of patients' last status alive or death (see Stephenson, 2008), and a multinomial logistic regression for the cases alive, death or unknown cases (see Madhu et al., 2014). The results from the analysis of this work show that the last status of the patients and the variable of personal clinical data are significant with factors age, length of stay and the cancer sites affecting the patients' last status.

The present study provides additional insight into assessing factors affecting the survival or death of breast cancer patients by using a comparative approach incorporating more models. These include the logistic regression, deep learning (MLP network), Convolutional Neural Networks (CNN), and Radial Basis Function network (RBF) not included in the literature discussed above.

Next, we turn to the methodology with a briefly discussion of the models used which will help us better understand and answer the research questions posed earlier in this paper.

CHAPTER 3

METHODOLOGY

3.1 Data and Model Construction

3.1.1 Data

The data used for this study consist of a sample of 100,002 breast cancer patients who have been treated at various hospitals in the USA covering the period 1992 - 2019 and were collected from the USA Center for Disease Control (CDC) under the Surveillance, Epidemiology, and End results program. The states in the US used for this study from which the sampled data were collected are San Francisco, Connecticut, and New Jersey with each state uniquely identified by a registry ID number. The data set contains 25 characteristics of the breast cancer found in Table A.2 in the Appendix. The total number of male patients were 15599, while female patients were 84403. The number of patients detected with breast cancer in grade 1 was 16358, grade 2 was 36123 and grade 3 was 47521. The number of patients who are married were 48698, while the unmarried patients were 51304. The total number of patients alive after the treatment at the various hospitals which the data covered were 46685 and the total death cases equal 53317. San Francisco had 1702 patients, Connecticut had 46371, and New Jersey had 51929. Table A.1 in Appendix shows the summary information of the data set. The three states were chosen based on availability of data for the chosen years considered for the study and there were no missing values for the entries in the data set.[1]

3.1.2 Model Construction

Machine learning and deep learning techniques have proven useful in biomedical science when classifying disease types based on genetic information. In this study, five models based on sex with the patient's demographic characteristics, and mammographic descriptors are developed using logistic regression, feed forward neural networks, Support Vector Machine, and Random Forest Classification to predict breast cancer patients' last status for both sex, and each sex separately. We then compare the output of each model to propose the best model for predicting breast cancer risk patients' last status, contributing to improve clinical decision making. These were achieved by first splitting the data int eighty percent training set and twenty percent test

[1]See Appendix Table A.2 for complete list of the variables used for the study with the variable definition and python codes.

set. The models were developed using the training data set while the testing data set was used for prediction. Python-Jupyter Notebook was used for the analysis.

3.1.2.1 Logistic Regression

This is a type of machine learning technique that falls within the class of supervised learning under the classification models. It is used to model medical problems due to the fact that the methodology is well established, and coefficients can have intuitive clinical interpretation. It is used to examine the relationship between a dependent variable eg. Alive or death, present or absent; and a given set of explanatory variables such as mammographic lesions, demographic variables etc. The practical application of this model is our study, where the breast cancer patient's last status at a specific time period is predicted with the knowledge of the patient's Age, Gender, Family history of breast cancer, Prior history of breast cancer surgery etc. Let X_i; where $i = 1, 2, 3, \ldots n$ represents the explanatory variables and Y represent the patient's last status. Then $Y = 0$ if patient is alive and $Y = 1$ if death. The logistic regression describing the relationship above is $\log(\frac{p(x)}{1-p(x)})$ = α + ($\sum_{k=1}^{n} \beta_i x_i$); where β_i represent the coefficients of the explanatory variables estimated from the data and α is the intercept of the model (see Ayer et al., 2010). The probability that a patient is dead or alive could be estimated with this model using the odd ratios.

3.1.2.2 Multi-layer Perceptron (MLP)

This is a supervised learning algorithm that learns a function $f(.) : R^n \rightarrow R^o$ by training on a data set where n is the number of dimensions for the input and o is the number of dimensions for the output. Having a number of features, $X = x_i$; where $i = 1, 2, 3, \ldots n$ reprinting patients' breast cancer specifications such as family history, Mass Size, Milk, Lymph Node, etc. and a target y which represent patient's last status in our case. It can learn a non-linear function approximator for either a classification or regression approach. Figure 3.1 is the MLP network process. Here, each neuron in the hidden layer transforms the value from the previous layer with a weighted linear summation $\sum_{i=1}^{n} w_i x_i$ followed by a non-linear activation function $g(.) : R \rightarrow R$ which is like the hyperbolic $tanfunction$. The hidden layer receives the values from the hidden layer and then transforms them into output values with the model continuing the coefficients and intercepts in a list of weighted matrices.

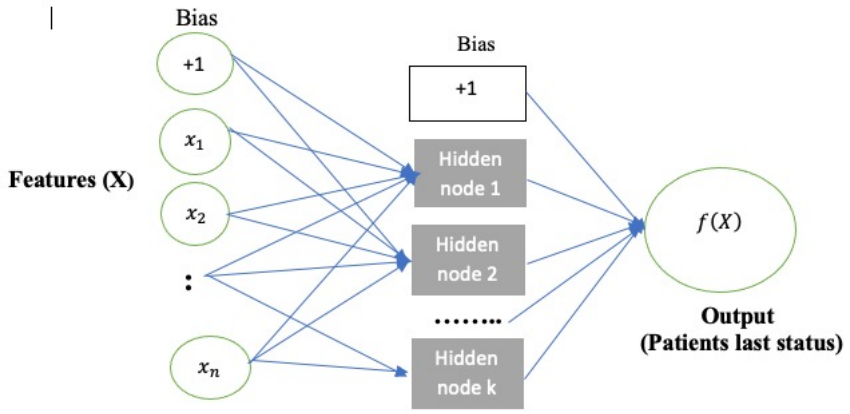

Fig. 3.1 One Hidden Layer MLP

3.1.2.3 Convolutional Neural Network (CNN)

This is a deep learning algorithm which can take an input image, assign importance to the various aspect of the image and is able to differentiate one aspect from the other. The pre-processing required in a CNN is much lower as compared to the other classification algorithms. It is mostly trained using the activation function back propagation. Below is a figure describing the structure and how a CNN operates. This process trains the CNN and all the weighs and parameters are optimized to correctly classify images from the data set. The total error at the output layer is computed by: Total Error = $\sum_{i=1}^{n} \frac{1}{2}(TargetProb. - OutputProb.)^2$

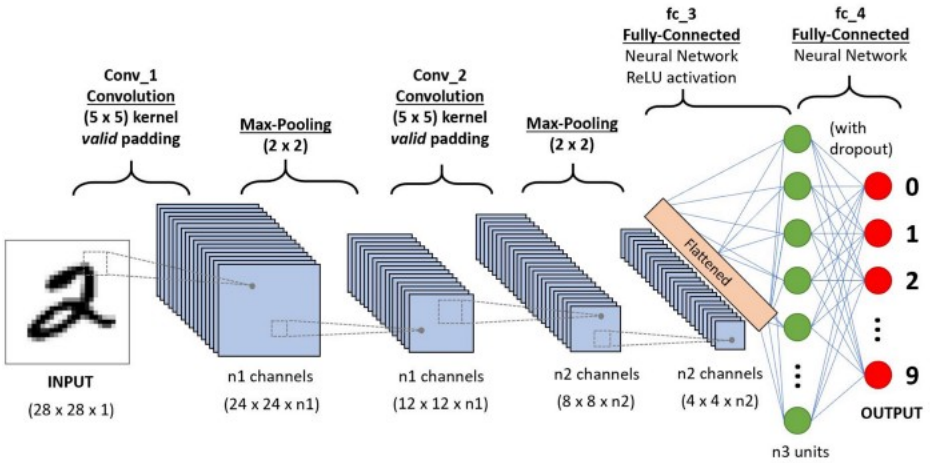

Fig. 3.2 CNN sequence to classify patients last status
(See, Toward Data science. Retrieved from *https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53*)

3.1.2.4 Radial Basis Function (RBF) Network

The commonly used type of the artificial neural network for function approximation problems is the Radial Basis Function (RBF) network. It is composed of three layers which includes the input layer, hidden layer, and output layer. The RBF network in its simplest form is a three-layer feed forward neural network which is strictly limited to one hidden layer called the feature vector (Ahmadian et al., 2018). Increasing the dimension of the feature vector increases the linear separability vector. The Gaussian function are generally used for RBF and is equal to $\phi(r) = exp(\frac{-r^2}{2\sigma^2})$; where the radial distance $r = \|x - t\|$ and $\sigma > 0$. Below is the figure describing the RBF neural network. The first stage of training the model is done by clustering algorithm where the number of cluster centers needed are defined and computed which then assigned as the receptors for the hidden neurons. Classifications only occur in the second phase where the linear combination of hidden functions are driven to output layer. The second training phase also updates the weighting vectors between the hidden layer and output layer of the model.

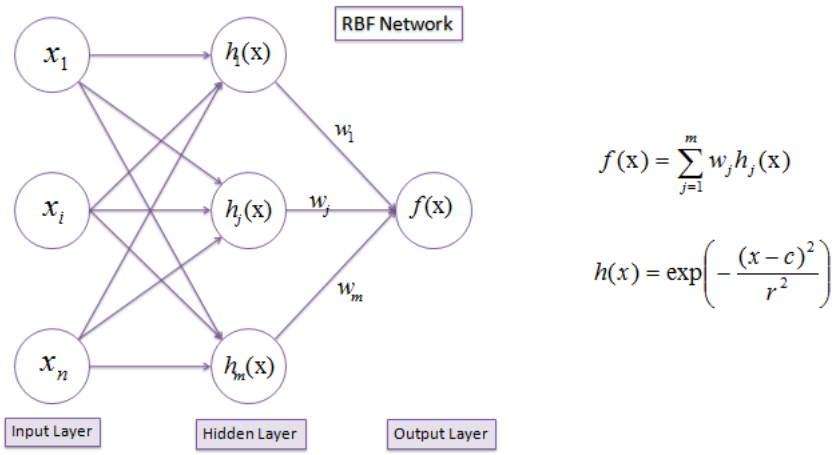

Fig. 3.3 RBF sequence to classify patients' last status
(see Toward Data science Retrieved from *https://towardsdatascience.com/radial-basis-functions-neural-networks-all-we-need-to-know-9a88cc053448*)

3.1.2.5 Random Forest Networks

A random forest which is also known as random decision forest is a supervised learning algorithm which is an ensemble learning method for classification, regression problems and other tasks which forms the majority of the current machine learning system. It operates by the construction of multitude of decision trees at a training time and outputting the classes which are

the mode of classification or average prediction of the individual trees.[2] It is a meta estimator that fits the number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and over-fitting. There are n estimators hyper parameter during the training of the random forest networks which are the number of trees the algorithm builds before taking the averages of the predictions. Figure 3.4 shows the random forest description with two trees.

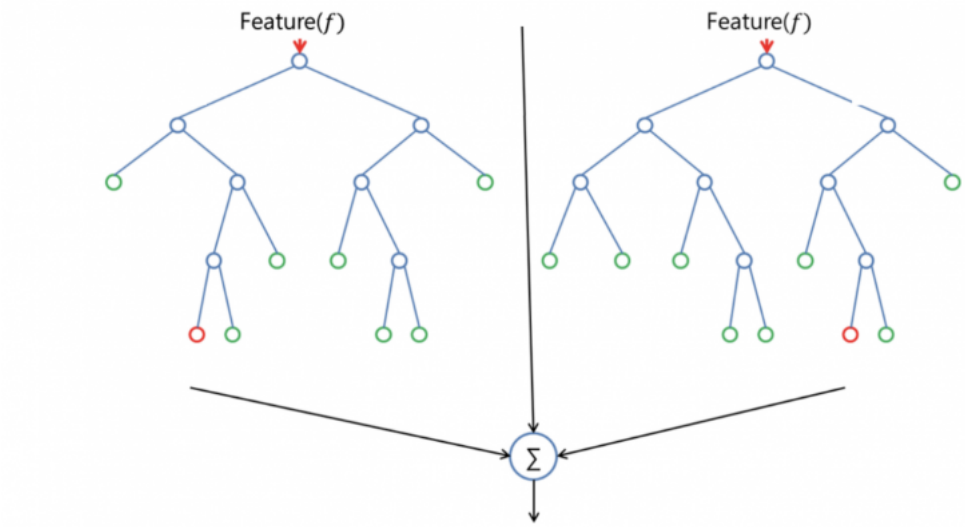

Fig. 3.4 The Random Forest with Two Trees to classify patients' last status (Retrieved from *https://builtin.com/data-science/random-forest-algorithm*)

Next, we turn to a brief description of the model evaluation techniques and set our arguments for the discussion of the choice of the best model for this study.

3.1.3 Model Evaluation

After training and fitting the models above, we evaluate their performance and compare the accuracy of prediction of each model to make decision in order to answer the research questions posed earlier in this paper. This is done using the confusion matrix to obtain the Overall accuracy of prediction, Sensitivity, Specificity, Positive Predictive Value (PPV), Negative Predictive Value (NPV), and the Receiver Operating Characteristic (ROC) curve showing the Area Under the ROC curve (AUC). These are model evaluation techniques used for assessing the perfor-

[2]See *https://www.google.com/search?q=rANDOM+fORESRTie=utf-8oe=utf-8client=firefox-b-1*

mance of classification models and for the selection of the best model for decision making.

3.1.3.1 Model Evaluation Approaches Used for the Study.

Below are the description of the criteria for evaluating the fitted classification models described from above.

- Accuracy (ACC): This is the number of predictions the given classification model got right. The classification accuracy is the number of correct predictions on the total number of inputs samples. This works better only if we have equal number of samples in each class. It is given by: $ACC = \frac{TP+TN}{TP+TN+FP+FN}$
- Sensitivity or True Positive Rate (TPR): This tells us the probability that the model predicts positive as death given that the patient actually died. That is the Prob (+ as death / patient actually died). It is given by: $TPR = \frac{TP}{TP+FN}$
- Positive Predictive Value (PPV) or Precision: This tells us that, given that the model predict positive as death on patients, what is the probability that they actually died? That is the Prob (patient actually died / + as death). This is given by: $PPV = \frac{TP}{TP+FP}$
- Specificity or True Negative Rate (TNR): This tells us the probability that the model predicts negative as Alive given that the patients are alive. That is the Prob (- as Alive / patient actually Alive). This is given by: $TNR = \frac{TN}{TN+FP}$
- Negative Predictive Value (NPV): This tells us that given that the patients are alive, what is the probability that the model predicts negative as death? That is the Prob (- as death / patient actually Alive). This is given by: $NPV = \frac{TN}{TN+FN}$
- The Receiver Operating Characteristic (ROC) curve: This is a commonly used graph that summarizes the performance of a classifier model over all possible thresholds. It is generated by plotting the True Positive Rate against the False Positive Rates as the threshold is varied for assigning observations to the given class. In classifying breast cancer patients as alive or dead with use of receiver operating characteristic (ROC) curves, the area under a ROC curve (AUC) indicates how well the predictive model discriminates between healthy patients and patients who died. The value of an AUC varies between 0.5 (That is random guess) and 1.0 (perfect accuracy). The higher the value of the AUC, the better the model. All analysis was completed using Python-Jupyter Notebook.

Next, we turn to the model results and the discussion of the key findings obtained

CHAPTER 4

MODEL RESULTS

The given Table 4.1 to Table 4.6; Table A.3, and Figure 4.1 to Figure 4.20 in these sections shows the summary results of this project followed by the discussion. Here, we present the results for all breast cancer patients together and later present that of both genders separately which aid us to better understand the subject being researched and give the responses to the research questions posed earlier in this paper.

First, we turn to the results of both male and female patients together.

4.1 MALE AND FEMALE BREAST CANCER PATIENTS.

This section of the study emphasis on both male and female patients together. Below are the summary results of key finding of each model and their performance evaluation.

4.1.1 Logistic Regression Model for Male and Female Breast Cancer Patients.

Table 4.1 shows that all variables tested statistically significantly affect all breast cancer patients' survival and death except for the patient's Cancer Grade, Suture, and Dermal with the p-values of respectively 0.198, 0.153, 0.210 that are statistically insignificant. All other variables have the p-values statistically significant at 1% level except Gender with p-value statistically significant at 0.024.

While a breast cancer patient's Race, Marital Status, Age at Diagnosed, Mass Stability, Number of Visits to the hospital, Mass Density, Nipple Retraction, Lymph Node, Amorphous, Eggshell, Milk, Axillary Adenopathy, and Lucent have positive influence on patient's last status, Gender, Length of stay in the hospital after admission, Laterality, Family history of breast cancer, Prior history of breast cancer surgy, Mass Size, Eggshell, Dystrophic, and Skin Lesion negatively influence patient's last status. These results are interesting to be used for clinical decision making by patients and physicians.

Table 4.1 Logistic Regression Model Results for Male and Female Breast Cancer Patients.

Variable	**Estimate**	**Std. Error**	**z**	**P**$>|z|$	**95% C.I**
Race	0.0056	0.001	8.874	0.000	(0.004, 0.007)
MarST	0.0538	0.019	2.895	0.004	(0.017, 0.090)
Gender	-0.0592	0.026	-2.255	0.024	(-0.111, -0.008)
AgeDiag	0.0066	0.001	10.397	0.000	(0.005, 0.008)
Grade	-0.0159	0.012	-1.286	0.198	(-0.040, 0.008)
Stability	0.2178	0.035	6.280	0.000	(0.150, 0.286)
No.Visits	0.0335	0.003	12.467	0.000	(0.028, 0.039)
Lstay	-0.1740	0.003	-51.229	0.000	(-0.181, -0.167)
Laterality	-0.0810	0.004	-22.634	0.000	(-0.088, -0.074)
FamHist	-3.3407	0.063	-53.148	0.000	(-3.464, -3.217)
PrioBSurgy	-0.3006	0.025	-12.034	0.000	(-0.350, -0.252)
Suture	0.0427	0.022	1.932	0.053	(-0.001, 0.086)
Density	0.2365	0.015	16.194	0.000	(0.208, 0.265)
NipRet	2.1587	0.033	64.839	0.000	(2.093, 2.224)
LyNode	0.3620	0.029	12.287	0.000	(0.304 , 0.420)
Amorph	0.1636	0.030	5.422	0.000	(0.104, 0.223)
Size	-0.1563	0.015	-10.446	0.000	(-0.186, -0.127)
Eggshell	-0.4962	0.024	-20.468	0.000	(-0.544, -0.449)
Milk	0.8963	0.025	35.627	0.000	(0.847, 0.946)
AxiAden	1.0621	0.021	50.061	0.000	(1.021, 1.104)
Dystroph	-0.7367	0.038	-19.439	0.000	(-0.811, -0.662)
Lucent	1.9973	0.037	53.432	0.000	(1.924, 2.071)
Dermal	0.0292	0.023	1.254	0.210	(-0.016, 0.075)
SkinLesion	-1.6329	0.020	-80.003	0.000	(-1.673, -1.593)

4.1.1.1 Model Evaluation for Male and Female Patients' Logistic Regression

From the confusion matrix in Appendix Figure A.1, we get the overall accuracy of prediction of 80.24% with a Sensitivity of 0.734, Specificity of 0.891, Positive Predictive Value of 0.897, Negative Predictive Value of 0.721 and AUC of 0.88.

This model's performance is good since all model evaluation techniques give high values as required. Figure 4.1 is the graph showing the ROC curve with the AUC for this logistic regression model for the Male and Female patients considered.

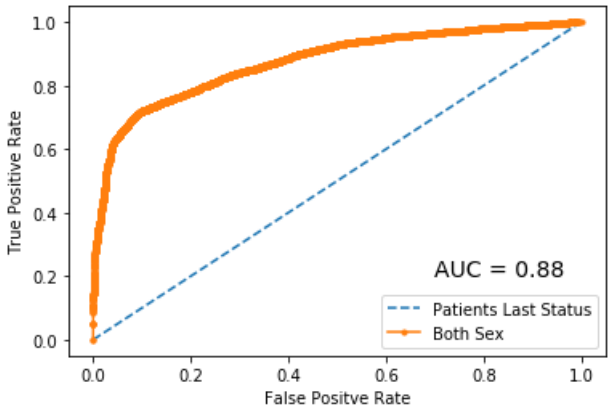

Fig. 4.1 The ROC curve showing the AUC for Male and Female logistic regression model.

4.1.2 Multi-Layer Perceptron (MLP) for Male and Female Breast Cancer Patients

The fitted Multi-Layer Perceptron model gives the confusion matrix results indicated in Appendix Figure A.4. From this figure, the overall accuracy of prediction of breast cancer patient's last status is 86.181% with Sensitivity of 0.791, Specificity of 0.953, Positive Predictive Value of 0.956, Negative Predictive Value of 0.78 and AUC of 0.90. This model also performs very great because all model evaluation techniques used here give good values. Figure 4.2 is the graph of the ROC curve showing the AUC for the Multi-Layer Perceptron model for these patients considered.

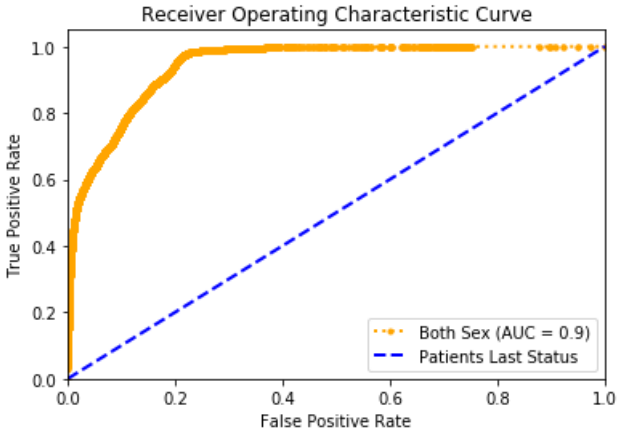

Fig. 4.2 The ROC curve showing the AUC for Male and Female Multi-Layer Perceptron.

4.1.3 Convolutional Neural Network (CNN) for Male and Female Breast Cancer Patients

The fitted CNN give an overall accuracy of prediction of 90.750% with Figure 4.3 showing the training and validation accuracy of the model. From this figure, the training and validation lines flow in a sink manner indicating that there is a good fit and no over fitting occurs here.

The confusion matrix in Appendix Figure A.7 gives a Sensitivity of 0.904, Specificity of 0.911, Positive Predictive Value of 0.897, Negative Predictive Value of 0.917 and AUC of 0.96. This model also performs very great because all model evaluation techniques give higher values. Figure 4.4 is the graphs of the ROC curve showing the AUC for the Convolutional Neural Network model for these patients considered.

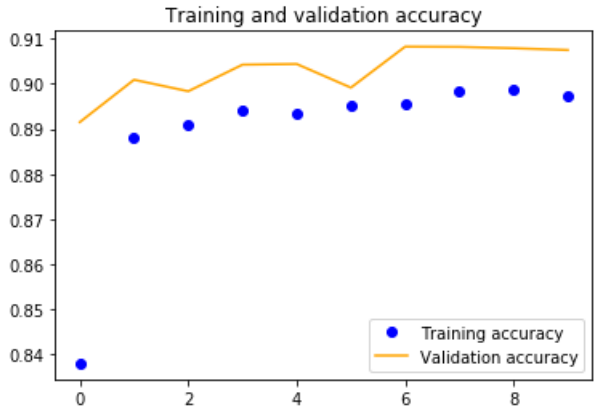

Fig. 4.3 Training and validation accuracy for Male and Female CNN

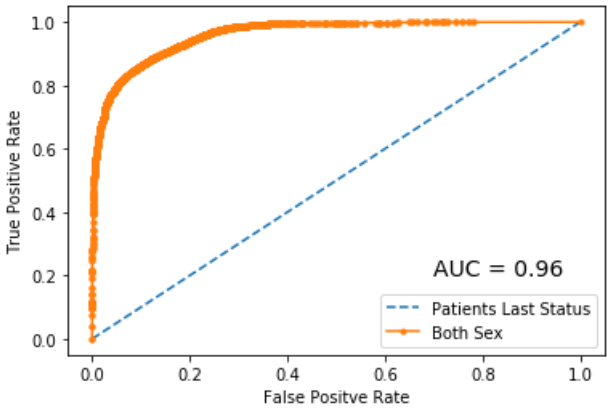

Fig. 4.4 The ROC curve showing the AUC for Male and Female CNN

4.1.4 Radial Basis Function (RBF) Network for Male and Female Breast Cancer Patients

The fitted Radial Basis Function Networks give the confusion matrix indicated in Appendix Figure A.10. From this figure, the overall accuracy of prediction of breast cancer patient's last status is 90.171% with Sensitivity of 0.864, Specificity of 0.940, Positive Predictive Value of 0.935, Negative Predictive Value of 0.873 and AUC of 0.90. This model performs very great because all model evaluation techniques give higher values. Figure 4.5 is the graphs of the ROC curve showing the AUC for the Radial Basis Function Network for these patients considered.

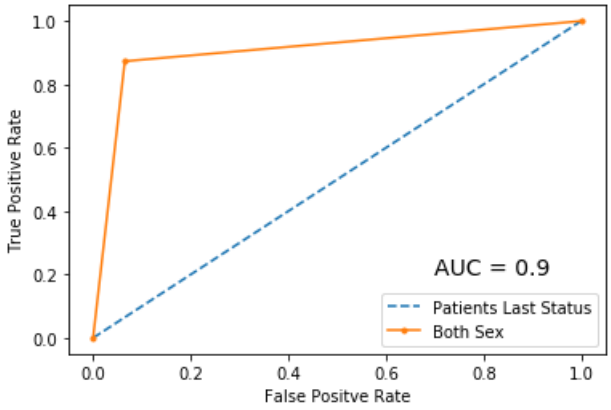

Fig. 4.5 The ROC curve showing the AUC for Male and Female RBF model.

4.1.5 Random Forest Network for Male and Female Breast Cancer Patients

The fitted Random Forest Network give the confusion matrix indicated in Appendix Figure A.13. From this figure, the accuracy of prediction of breast cancer patient's last status is 90.311% with Sensitivity of 0.859 , Specificity of 0.950, Positive Predictive Value of 0.947, Negative Predictive Value of 0.865 and AUC of 0.91. This model performs very great because all model evaluation techniques give higher values. Figure 4.6 is the graphs of the ROC curve showing the AUC for the Random Forest Network for these patients considered.

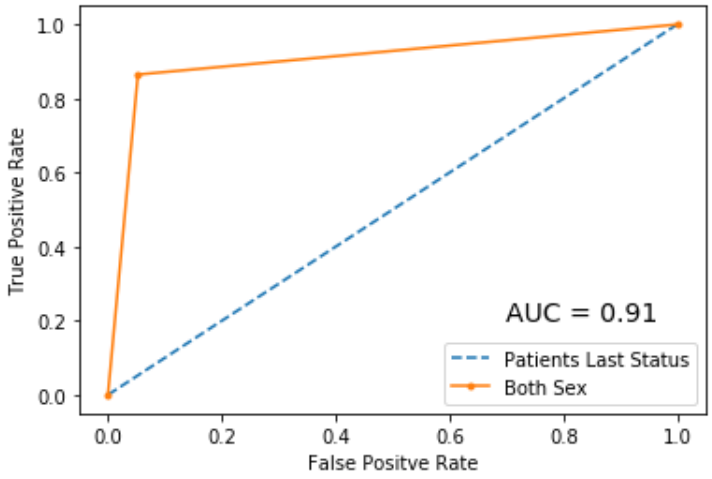

Fig. 4.6 The ROC curve showing the AUC for Male and Female Random Forest Network.

The table 4.2 below is the summary results of all fitted models' evaluation for the male and female breast cancer patients together. This table will help us to make comparison among all models studied in this project for the discussions section.

From this table, none of the models studied performed poorly. All models' evaluation techniques gave a good score. The Convolutional Neural Network gave a higher overall prediction accuracy, high precision with higher AUC compared with those of other model and these give an indication of a better fitted model. The Logistic regression performs lower compared with all other models' overall prediction accuracy, AUC, and all other evaluation techniques.

Table 4.2 Summary of All Model Results for Male and Female Breast Cancer Patients

Evaluation Method	**Logistic Regression**	**Multi-Layer Network**	**Convolutionary Network**	**RBF**	**Random Forest**
Sensitivity Analysis	0.73431	0.79059	0.90361	0.86355	0.858801
Positive Predictive Value (PPV)	0.89675	0.95590	0.89663	0.93496	0.947295
Specificity Analysis	0.89056	0.95320	0.91082	0.93988	0.949723
Negative Predictive Value (NPV)	0.72140	0.78031	0.91695	0.87315	0.864723
The Receiver Operating Characteristic (ROC) curve, AUC	0.88	0.90	0.96	0.90	0.91
Overal Prediction Accuracy	**80.241%**	**86.181%**	**90.750%**	**90.171%**	**90.3105%**

Next, we turn to the male breast cancer patients' case and look at the results obtained from the analysis of this study.

4.2 MALE BREAST CANCER PATIENTS.

This section of the study emphasizes on the male breast cancer patients only. Below are the summary results of the key findings for each model and their performance evaluation.

4.2.1 Logistic Regression Model for Male Breast Cancer Patients.

Table 4.3 is the results for the logistic regression model which shows that, all variables tested statistically significantly affect male breast cancer patients' survival and death at 5% level except the patient 's Race, Marital Status, Age at Diagnosed, Mass Stability, Laterality, Eggshell, Axillary Adenopathy and Dystrophic that are statistically insignificant at 5% level. While a Male patients Grade, Suture, Mass Density, Nipple Retraction, Lymph Node, Amorphous, Mass Size, Lucent, and Dermal have positive influence on Male patients' last status, Number of Visits to the hospital, Length of stay in the hospital after diagnosed, Family history of breast cancer, Prior history of breast cancer surgery, Milk, and Skin Lesion negatively influence male patients last status. These results are surprising since one may expect these variable to have positive effect on the Male patients' last status.

Table 4.3 Logistic Regression Model Results for Male Breast Cancer Patients.

Variable	**Estimate**	**Std. Error**	**z**	**P**$>|z|$	**95% C.I**
Race	-0.0008	0.002	-0.353	0.724	(-0.005, 0.004)
MarST	-0.0004	0.050	-0.009	0.993	(-0.099, 0.098)
AgeDiag	0.0026	0.002	1.357	0.175	(-0.001, 0.006)
Grade	0.1192	0.035	3.425	0.001	(0.051, 0.187)
Stability	0.0759	0.086	0.883	0.377	(-0.093, 0.245)
No.Visits	-0.0218	0.010	-2.180	0.029	(-0.041, -0.002)
Lstay	-0.1344	0.012	-10.926	0.000	(-0.158, -0.110)
Laterality	-0.0121	0.009	-1.288	0.198	(-0.030, 0.006)
FamHist	-2.6038	0.405	-6.423	0.000	(-3.398, -1.809)
PrioBSurgy	-0.3056	0.131	-2.325	0.020	(-0.563, -0.048)
Suture	0.1467	0.062	2.373	0.018	(0.026, 0.268)
Density	0.2654	0.063	4.196	0.000	(0.141, 0.389)
NipRet	0.8825	0.147	6.009	0.000	(0.595, 1.170)
LyNode	2.6183	0.231	11.352	0.000	(2.166, 3.070)
Amorph	1.4480	0.173	8.354	0.000	(1.108, 1.788)
Size	0.5909	0.068	8.684	0.000	(0.458, 0.724)
Eggshell	0.1048	0.067	1.561	0.118	(-0.027, 0.236)
Milk	-0.8173	0.180	-4.553	0.000	(-1.169, -0.465)
AxiAden	0.0887	0.063	1.400	0.161	(-0.035, 0.213)
Dystroph	-0.1351	0.162	-0.835	0.404	(-0.452, 0.182)
Lucent	0.9937	0.144	6.921	0.000	(0.712, 1.275)
Dermal	0.1767	0.063	2.796	0.005	(0.053, 0.300)
SkinLesion	-2.7320	0.068	-40.352	0.000	(-2.865, -2.599)

4.2.1.1 Model Evaluation for Male Logistic Regression

From the confusion matrix in Appendix Figure A.2, we get the overall accuracy of prediction of 82.339% with a Sensitivity of 0.694, Positive Predictive Value 0.865, Specificity of 0.920, Negative Predictive Value of 0.802 and AUC of 0.87. This model performs well because all model evaluation techniques give good values. Figure 4.7 is the graph of the ROC curve showing the AUC for the Male logistic regression model for these patients considered.

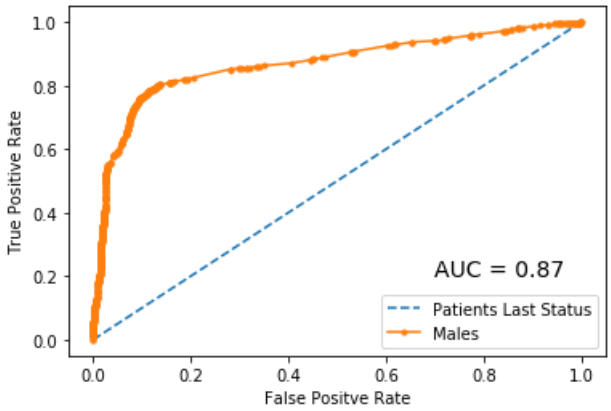

Fig. 4.7 The ROC curve showing the AUC for Male logistic regression model.

4.2.2 Multi-Layer Perceptron (MLP) for Male Breast Cancer Patients

The fitted Multi-Layer Perceptron model of the male patients give the confusion matrix indicated in Appendix Figure A.5. From this figure, the overall accuracy of prediction of male breast cancer patients' last status is 83.0% with Sensitivity of 0.692, Specificity of 0.935, Positive Predictive Value of 0.894, Negative Predictive Value of 0.794 and AUC of 0.88. This model also performs very well because all model evaluation techniques give good values. Figure 4.8 is the graph of the ROC curve showing the AUC for the Multi-Layer Perceptron model for these patients considered.

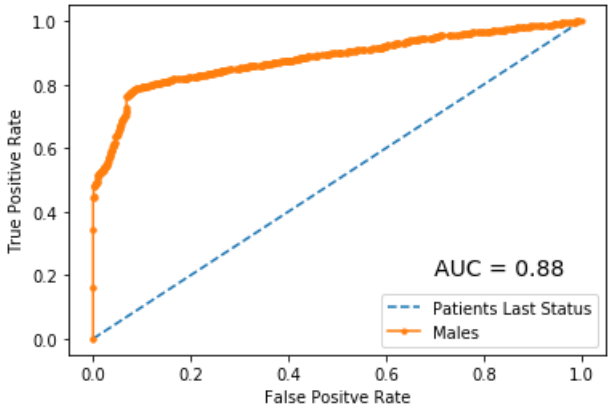

Fig. 4.8 The ROC curve showing the AUC for Male Multi-Layer Perceptron model.

4.2.3 Convolutional Neural Network (CNN) for Male Breast Cancer Patients

The fitted CNN for the male patients give an overall accuracy of prediction of 95.994% with Figure 4.9 showing the training and validation accuracy of the model. From this figure, the training and validation lines also flow in a sink manner indicating that there is a good fit and no over fitting occurs. The confusion matrix in Appendix Figure A.8 gives a Sensitivity of 0.997,

Specificity of 0.942, Positive Predictive Value of 0.889, Negative Predictive Value of 0.999 and AUC of 0.98. This model also performs very great because all model evaluation techniques used give higher values. Figure 4.10 is the graphs of the ROC curve showing the AUC for the Convolutional Neural Network for these patients considered.

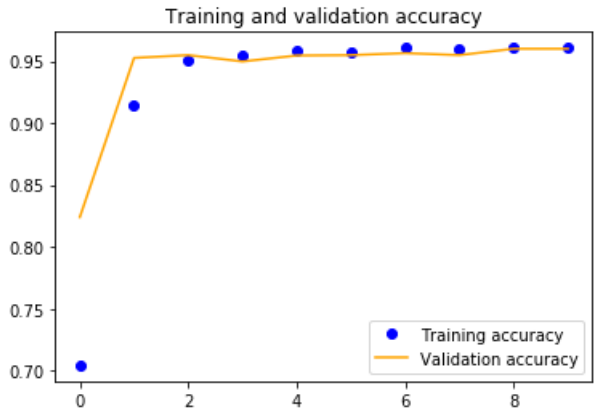

Fig. 4.9 Training and validation accuracy for Male CNN

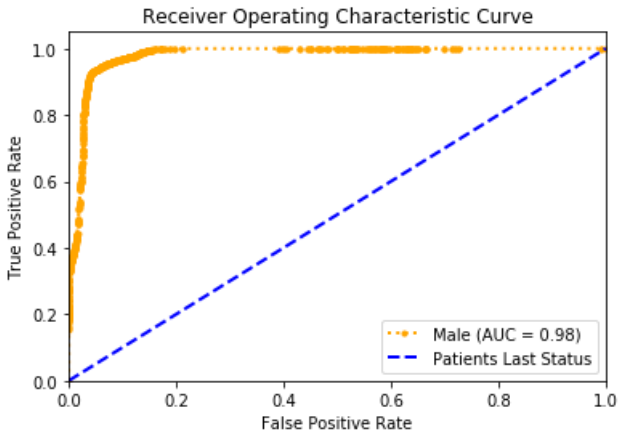

Fig. 4.10 The ROC curve showing the AUC for Male CNN

4.2.4 Radial Basis Function (RBF) Network for Male Breast Cancer Patients

The fitted Radial Basis Function Networks give the confusion matrix indicated in Appendix Figure A.11. From this figure, the accuracy of prediction of breast cancer patients' last status is 83.558% with Sensitivity of 0.697, Specificity of 0.947, Positive Predictive Value of 0.914, Negative Predictive Value of 0.915 and AUC of 0.85. This model performs very well because all model evaluation techniques give high values. Figure 4.11 is the graphs of the ROC curve showing the AUC for the Radial Basis Function Network for these patients considered.

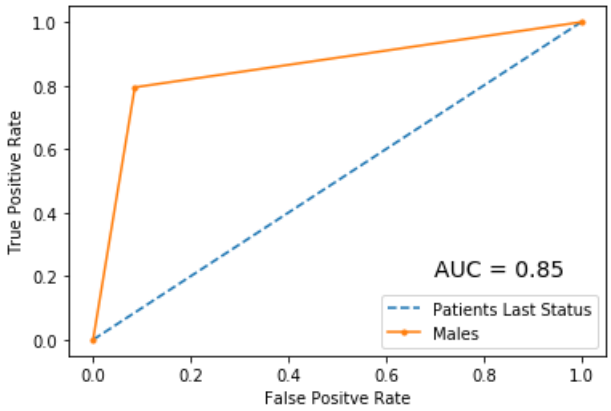

Fig. 4.11 The ROC curve showing the AUC for Male RBF model.

4.2.5 Random Forest Network for Male Breast Cancer Patients

The fitted Random Forest Network for the male patients give the confusion matrix indicated in Appendix Figure A.14. From this figure, the accuracy of prediction of breast cancer patients last status is 94.94% with Sensitivity of 0.997, Specificity of 0.906, Positive Predictive Values of 0.907, Negative Predictive Value of 0.997 and AUC of 0.95. This model performs very great because all model evaluation techniques give higher values. Figure 4.12 is the graphs of the ROC curve showing the AUC for the Random Forest Network for these patients considered.

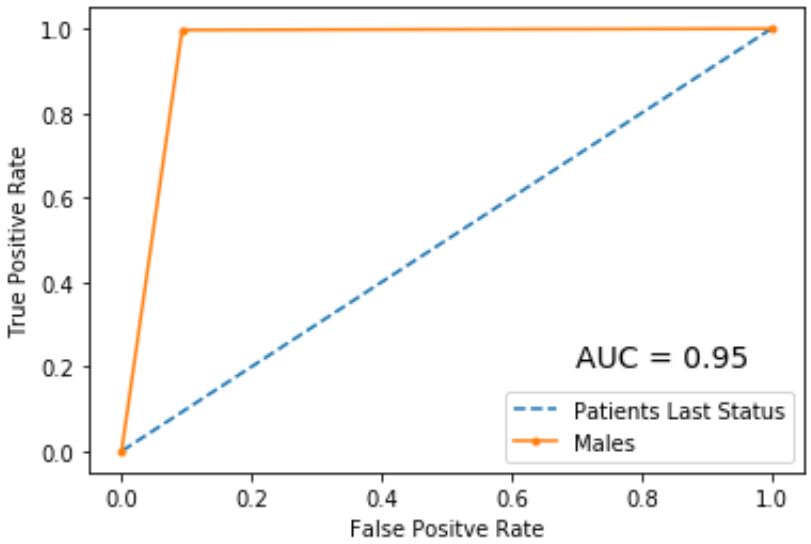

Fig. 4.12 The ROC curve showing the AUC for Male Random Forest Network.

4.2.6 Male Breast Cancer Patients Variables Importance to the Random Forest.

This is an aspect of the Random Forest network that is advantageous over all other neural networks. Here, we determine the contribution of each variable to the model and dropping the least contributing variable helps improve the model. Appendix Table A.3 is the table showing the variables and their importance to the model to Random Forest models. For the Male patients, skin lesion highly contributes to the model with over 34.94% followed by mass size with over 22.79% , followed by the length of stay in the hospital after diagnosis with over 10.32%. Family history of male breast cancer is the least with 0.0021% which implies that the male breast cancer survival or death is not much dependent on the patient's family history. Figure 4.13 shows the distribution of the variables and their importance to the Random Forest model.

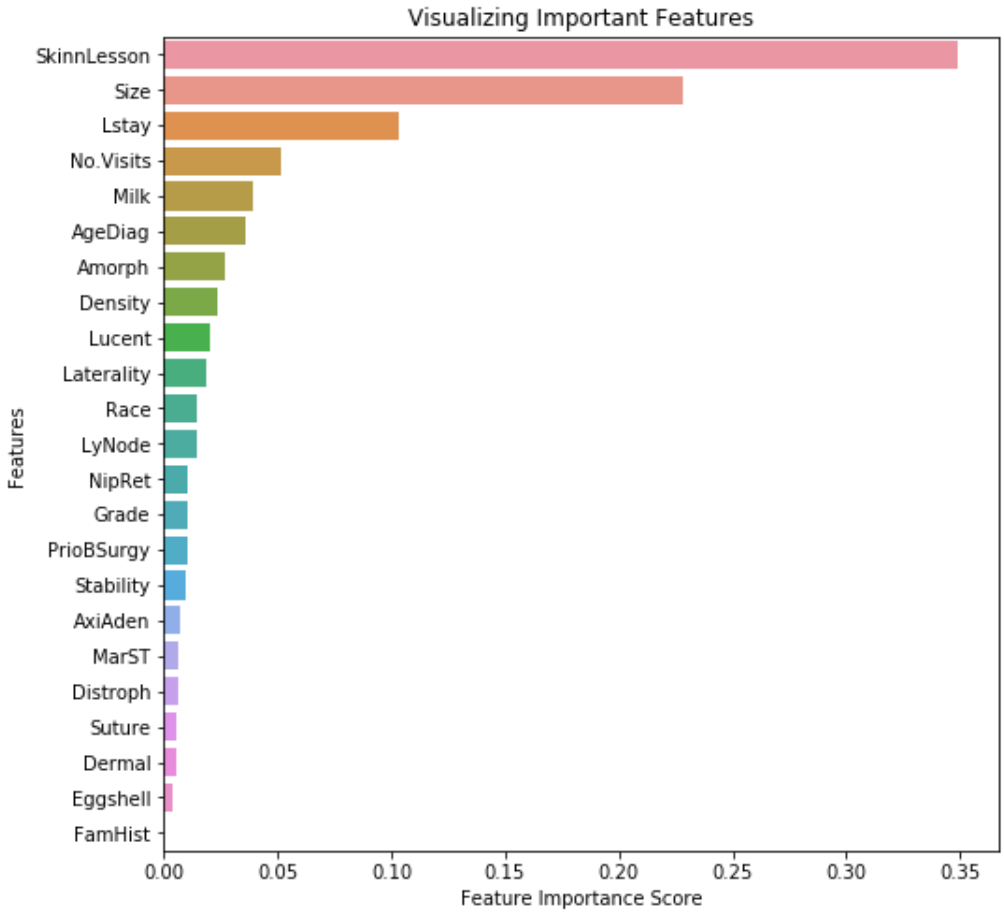

Fig. 4.13 Distribution of Male patient's variable importance to the Random Forest.

The table 4 below is the summary results of all fitted models' evaluation for the male breast cancer patients. This will help us to make comparison among all models studied in this project for discussions.

From this table, all the models performed well because considering the model evaluation techniques used, we obtained a good score for each. The Convolutional Neural Network in this case gave a higher overall prediction accuracy, high precision with higher AUC compared with that of other model and these give an indication of a better fitted model.

The Logistic regression performs lower compared with all other models' overall prediction accuracy, and precision. The RBF is not performing very well with the AUC compared with that of other models.

Table 4.4 Summary of All Models Results for Male Breast Cancer Patients

Evaluation Method	**Logistic Regression**	**Multi-Layer Network**	**Convolutionary Network**	**RBF**	**Random Forest**
Sensitivity Analysis	0.69376	0.69214	0.99796	0.69749	0.99696
Positive Predictive Value (PPV)	0.86492	0.89380	0.88849	0.91447	0.90683
Specificity Analysis	0.91960	0.93528	0.94245	0.94725	0.90561
Negative Predictive Value (NPV)	0.80185	0.79426	0.99901	0.915	0.99691
The Receiver Operating Characteristic (ROC) curve, AUC	0.87	0.88	0.98	0.85	0.95
Overal Prediction Accuracy	**82.33974%**	**83.00%**	**95.994%**	**83.5577%**	**94.942%**

4.3 FEMALE BREAST CANCER PATIENTS

This section of the study emphasizes on the female patients only. Below are the summary results of key finding of each model and their performance evaluation.

4.3.1 Logistic Regression Model for Female Breast Cancer Patients.

Table 4.5 below shows that all variables tested statistically significantly affect female breast cancer patients' survival and death at 1% level except the patient 's Suture and Dermal that are statistically insignificant with p-value of 0.187, 0.438 respectively. While a patients' Race, Marital Status, Age at Diagnosed, Mass Stability, Number of Visits to the hospital, Mass Density, Nipple Retraction, Lymph Node, Milk, Axillary Adenopathy, and Lucent positively influence the female patient's last status, Cancer Grade, Length of stay in the hospital after admission, laterality, Family history of breast cancer, Prior history of breast cancer surgy, Amorphous, Mass Size, Eggshell, Dystrophic and Skin Lesion negatively influence female patient's last status.

Table 4.5 Logistic Regression Model Results for Female Breast Cancer Patients.

Variable	**Estimate**	**Std. Error**	**z**	**P**$>|z|$	**95% C.I**
Race	0.0059	0.001	9.034	0.000	(0.005,0.007)
MarST	0.0730	0.021	3.523	0.000	(0.032 ,0.114)
AgeDiag	0.0078	0.001	11.084	0.000	(0.006 ,0.009)
Grade	-0.0393	0.014	-2.863	0.004	(-0.066, -0.012)
Stability	0.1699	0.039	4.314	0.000	(0.093, 0.247)
No.Visits	0.0507	0.003	17.406	0.000	(0.045, 0.056)
Lstay	-0.1656	0.004	-45.392	0.000	(-0.173, -0.158)
Laterality	-0.0942	0.004	-23.309	0.000	(-0.102, -0.086)
FamHist	-3.2622	0.064	-50.609	0.000	(-3.388, -3.136)
PrioBSurgy	-0.2662	0.026	-10.256	0.000	(-0.317, -0.215)
Suture	-0.0323	0.024	-1.321	0.187	(-0.080, 0.016)
Density	0.1879	0.015	12.304	0.000	(0.158, 0.218)
NipRet	2.2790	0.035	64.601	0.000	(2.210, 2.348)
LyNode	0.1978	0.031	6.426	0.000	(0.137, 0.258)
Amorph	-0.1152	0.033	-3.479	0.001	(-0.180, -0.050)
Size	-0.2425	0.016	-14.893	0.000	(-0.274, -0.211)
Eggshell	-0.5637	0.027	-20.885	0.000	(-0.617, -0.511)
Milk	0.9166	0.026	35.257	0.000	(0.866, 0.968)
AxiAden	1.2250	0.023	52.419	0.000	(1.179, 1.271)
Distroph	-0.6857	0.040	-17.217	0.000	(-0.764, -0.608)
Lucent	2.0236	0.040	51.090	0.000	(1.946, 2.101)
Dermal	-0.0202	0.026	-0.775	0.438	(-0.071, 0.031)
SkinnLesson	-1.4800	0.022	-66.078	0.000	(-1.524, -1.436)

4.3.1.1 Model Evaluation for Female Logistic Regression

From the confusion matrix in Appendix Figure A.3, we get the overall accuracy of prediction of 78.37% with a Sensitivity of 0.778, Specificity of 0.789, Positive Predictive Value 0.780, Negative Predictive Value of 0.787 and AUC of 0.86. This model performs quite well because all model evaluation techniques used give considerable good values. Figure 4.14 is the graph of the ROC curve showing the AUC for the logistic regression model for these patients considered.

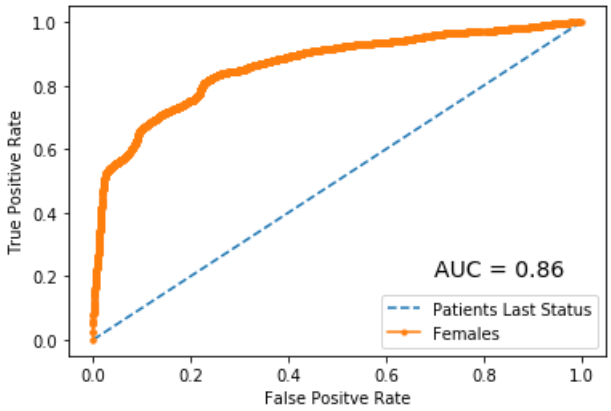

Fig. 4.14 The ROC curve showing the AUC for Female logistic regression model.

4.3.2 Multi-Layer Perceptron (MLP) for Female Breast Cancer Patients

The fitted Multi-Layer Perceptron model for the female patients give the confusion matrix indicated in Appendix Figure A.6. From this figure, the overall accuracy of prediction of female breast cancer patient's last status is 86.25% with Sensitivity of 0.797, Specificity of 0.958, Positive Predictive Value of 0.965, Negative Predictive Value of 0.765 and AUC of 0.80. This model performs well because all model evaluation techniques give good scores. Figure 4.14 is the graph of the ROC curve showing the AUC for the Multi-Layer Perceptron model for these patients considered.

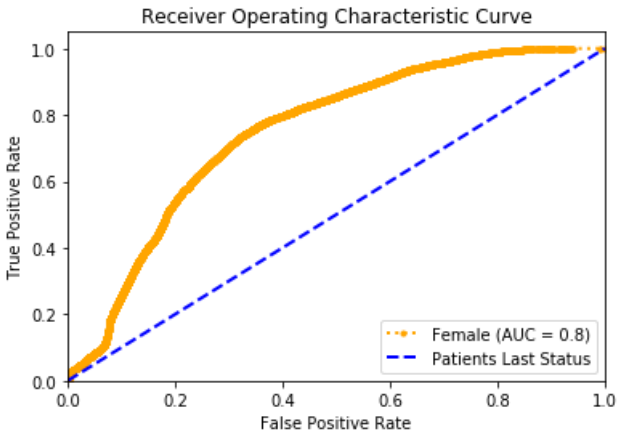

Fig. 4.15 The ROC curve showing the AUC for Female Multi-Layer Perceptron.

4.3.3 Convolutional Neural Network (CNN) for Female Breast Cancer Patients

The fitted CNN for the female patients give an overall accuracy of prediction of 88.78% with Figure 4.16 showing the training and validation accuracy of the model. From this figure, the training and validation line flow in a sink manner indicating that there is a good fit and there

is no over fitting. The confusion matric in Appendix Figure A.9 gives a Sensitivity of 0.948, Specificity of 0.845, Positive Predictive Value of 0.958, Negative Predicted Value of 0.813 and AUC of 0.95. This model performs very great because all model evaluation techniques give higher values. Figure 4.17 is the graphs of the ROC curve showing the AUC for the Convolutional Neural Network model for these patients considered.

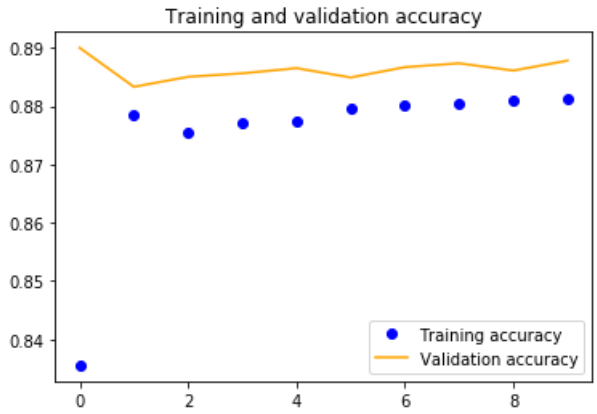

Fig. 4.16 Training and validation Accuracy for Female CNN

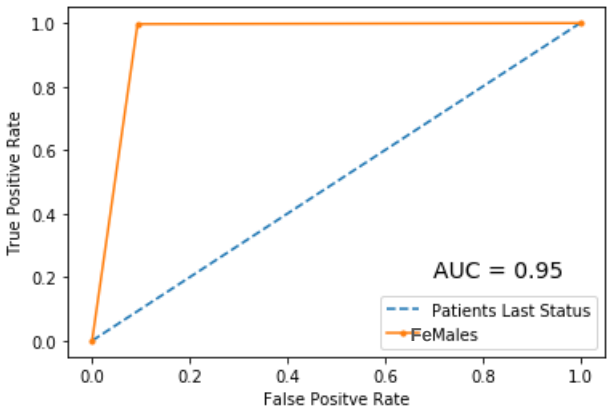

Fig. 4.17 The ROC curve showing the AUC for Female CNN

4.3.4 Radial Basis Function (RBF) Network for Female Breast Cancer Patients

The fitted Radial Basis Function Networks for the female patients give the confusion matrix indicated in Appendix Figure A.12. From this figure, the accuracy of prediction of female breast cancer patient's last status is 85.93% with Sensitivity of 0.791 , Specificity of 0.960, Positive Predictive Value of 0.967, Negative Predicted Value of 0.756 and AUC of 0.86. This model performs very good because all model evaluation techniques give higher values. Figure 4.18 is the graphs of the ROC curve showing the AUC for the Radial Basis Function Network for these patients considered.

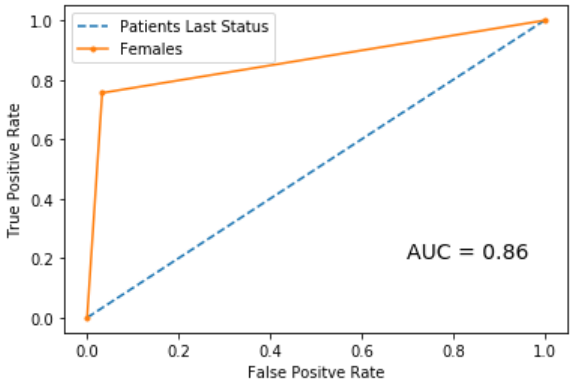

Fig. 4.18 The ROC curve showing the AUC for Female RBF model.

4.3.5 Random Forest Network for Female Breast Cancer Patients

The fitted Random Forest Network for the female patients give the confusion matrix indicated in Appendix Figure A.15. From this figure, the overall accuracy of prediction of female breast cancer patient's last status is 88.640% with sensitivity of 0.854, specificity of 0.930, Positive Predictive Value of 0.933, Negative Predicted Value of 0.847 and AUC of 0.89. This model performs very well because all model evaluation techniques give higher values. Figure 4.19 is the graphs of the ROC curve showing the AUC for the Random Forest Network for these patients considered.

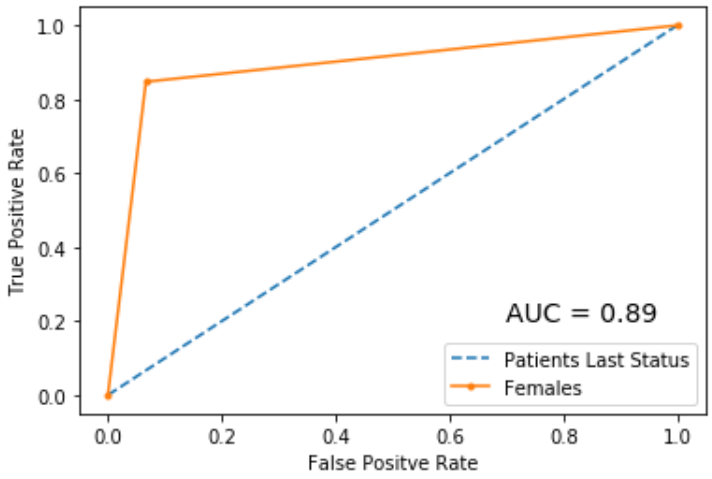

Fig. 4.19 The ROC curve showing the AUC for Female Random Forest Network.

4.3.6 Female Patients Variables Importance to the Random Forest.

In this aspect of the Random Forest network we determine the contribution of each variable to the female breast cancer patients' model. The female patients' variables contribution to the model as seen from Appendix Table A.3 shows that, Length of stay in the hospital after admission contributes the most to the model with over 13.22% followed by Mass Density with over 10.16% , followed by the Nipple Retraction with over 10.02%. Mass Stability is the least with 0.46% which implies that the female breast cancer survival or death is not much dependent on the patient's Mass Stability. Figure 4.20 is the distribution of variables in the Random Forest model.

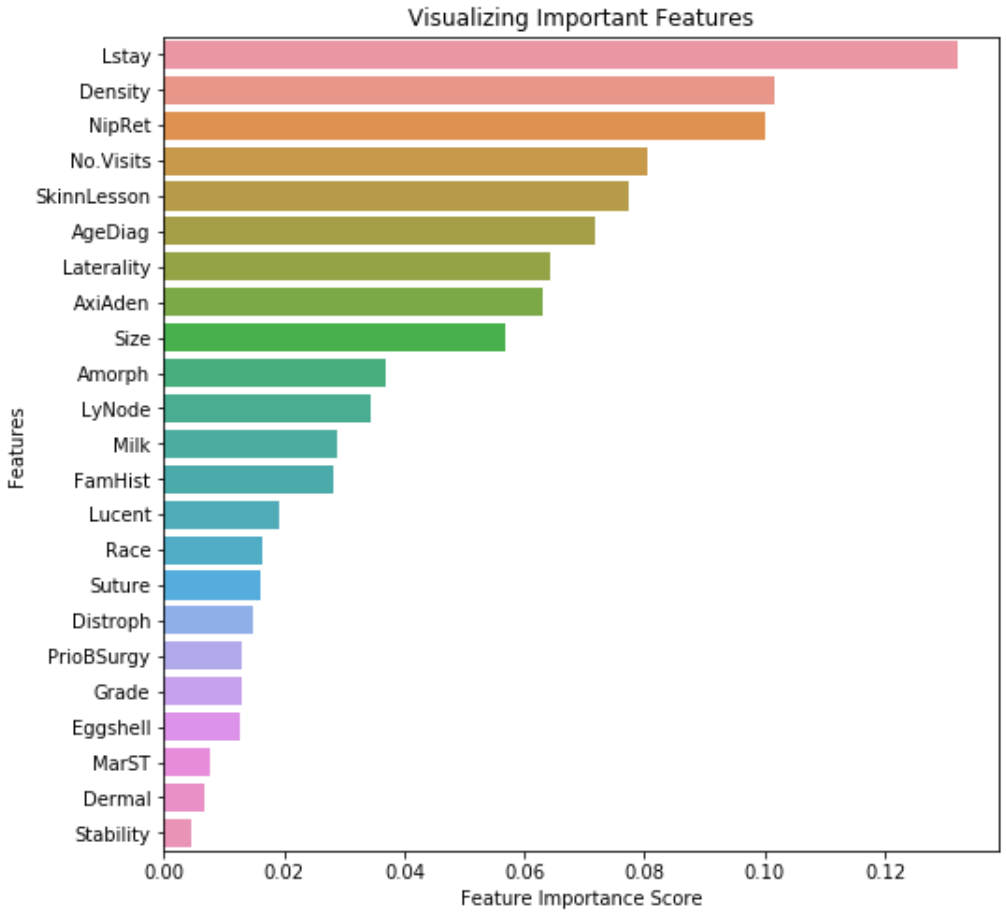

Fig. 4.20 Distribution of Female patient's variable to Random Forest model.

The Table 4.6 below is the summary results of all fitted models' evaluation for the female breast cancer patients. This will help us to make comparison among all models studied in this project for discussions.

From this table, all models perform well because all the evaluation techniques employed give us higher scores. The Convolutional Neural Network in this case gives the higher overall prediction accuracy, a good precision with a good AUC. This gives an indication of a better fitted model compared with all other models. The Logistic regression performs lower compared with all other model's overall percentage of prediction, and precision. The RBF has the same AUC as that of the logistic regression but has a higher overall prediction accuracy than the logistic regression.

Table 4.6 Summary of All Model Results for Female Breast Cancer Patients.

Evaluation Method	**Logistic Regression**	**Multi-Layer Network**	**Convolutionary Network**	**RBF**	**Random Forest**
Sensitivity Analysis	0.77821	0.79688	0.94805	0.79135	0.85405
Positive Predictive Value (PPV)	0.77981	0.96462	0.95820	0.96704	0.93323
Specificity Analysis	0.78902	0.95763	0.84506	0.95998	0.92993
Negative Predictive Value (NPV)	0.78746	0.76486	0.81278	0.75617	0.84745
The Receiver Operating Characteristic (ROC) curve, AUC	0.86	0.80	0.95	0.86	0.89
Overal Prediction Accuracy	78.372%	86.251%	88.780%	85.9250%	88.640%

Next, we turn to the discussions of our results obtained which sets the platform for answering the research questions posed earlier in this paper.

CHAPTER 5

DISCUSSION OF RESULTS

The tables and figures presented in the above sections show the procedures and results obtained from the analysis of this study which will help us to give responses to the research questions posed earlier in this paper. In this section, we look at the discussion of the key findings in the three cases considered in this study.[1] These will help us answer the research questions posed earlier.

5.1 Male and Female Breast Cancer Patients.

Considering the first research question "What factors affect the last status of all breast cancer patients with the target population in the USA ?" we use the key findings from the logistic regression for both the male and female patents to answer this question. Here, we found that a patient's Race, Marital Status, Age at Diagnosed, Mass Stability, Number of Visits to the hospital, Mass Density, Nipple Retraction, Lymph Node, Amorphous, Eggshell, Milk, Axillary Adenopathy, Lucent each is statistically significant at 1% level; Gender is statistically significant at 5% level, Length of stay in the hospital after admission, Laterality, Family history of breast cancer, Prior history of breast cancer surgy, Mass Size, Eggshell, Dystrophic, and Skin Lesion each is statistically significantly at 1% level affect breast cancer patient's last status. This answers the first research question.

The key findings of Gender, and Length of stay in the hospital after admission are consistent with that of Zangmo and Tiensuwan (2018), While Age at diagnosed, and Number of Visits to the hospital opposes their key findings. Also, the findings of Family history of breast cancer coincide with those of Colditz (1993); Gui (2001); and Raw (1998), while that of Milk coincides with the findings of Gatta (2006). Further, the findings obtained for Milk, Mass Size and Skin Lesion are consistent with those of Balleyguier (2007); Eltoukhy (2010); Maiprasert and Krieng, 2012; and Moezzi (1996).

The demographic risk factors such as Race, Marital Status, and Age at Diagnosed positively affect patient's last status. This implies that while brown patients, married patients, and the aged patients are more likely to die of breast cancer, patients who are females, patients with no history of breast surgery, and with no Family history of breast cancer have chance of survival since these negatively affect patients last status. One plausible reason could be that patients without breast

[1]The three cases considered were Both sex, Male sex separately, and Female sex separately.

surgery or patients with no family history of breast cancer takes the disease seriously, takes doctors medications, and follows the direction given by the physicians and as such cancer cells do not growths out of control leading to their high chance of survival. Such kind of patients in the society should be encouraged to give advice and motivation to their co-patients to help them manage the disease as well.

Also, while the presence of the mammographic descriptor such as high: Mass Density, Nipple Retraction, Lymph Node, Amorphous, Milk, Axillary Adenopathy, and Lucent in a patient makes that patient more likely to die, higher Length of stay in the hospital after admission, Laterality, Mass Size, Eggshell, Dystrophic, and Skin Lesion in a patient do not make that patient more likely to die. However, a breast cancer patient survival or death is not affected by the Cancer Grade, Suture, and Dermal. These results oppose the key findings of Colditz (1993); Gui (2001); Rawal (2006); and Yusuff et al. (2012).

5.2 Male Breast Cancer Patients.

Next, considering the second question "What factors affects the last status of male and female breast cancer patients separately within the target population in the USA?", we use the male patients' logistic regression to answer the first part of this question. For the male patients only, we found that the cancer Grade is statistically significant at 1% level, Suture is statistically significant at 5% level, Mass Density, Nipple Retraction, Lymph Node, Amorphous, Mass Size, Lucent, and Dermal each is statistically significant at 1% level, Number of visit to the hospital is statistically significant at 5% level, Length of stay in the hospital after admission, Milk, Family history of breast cancer, Prior-history of breast cancer surgery, and Skin Lesion are statistically significant each at 1% level and affects the male patients' last status. These answer that first part of the second research question. These key findings coincide with those of Ahmed et. al (2012); Borgen et al. (1997), and Kornegoor et al. (2012).

Here, the more men patients visit hospital to access health care, the more the Length of stay in the hospital after a man is admitted, the more likely they survive with breast cancer. Also, the higher the cancer Grade, the more likely males die, and men with Family history of breast cancer has a more likelihood of surviving. One plausible reason for this maybe that, males with family history of breast cancer, and Prior History of breast surgery, presents of Milk, and Skin Lesion are able to manage the disease very well by being serious to cure the disease which helps them to survive. Also, men with such factors might be taking their medication as prescribed, visiting the hospital as required etc., since they all influence negatively on patients last status. However, Race, Marital Status, Age at diagnose, Mass Stability, Laterality, Axillary Adenopathy, and Dystrophic do not affect the male patients' survival or death.

5.3 Female Breast Cancer Patients.

Also, considering the second part of the second research question, we use the female patients' logistic regression to answer this question. we found that female patients' Race, Marital Status, Age diagnosed, Mass Stability, Number of visit to the hospital, Density, Nipple Retraction, Lymph Node, Milk, Axillary Adenopathy, Lucent, cancer Grade, female patients' Length of stay in the hospital after admission, Laterality, Family history of breast cancer, Prior-history of breast cancer surgery, Amorphous, size of lamb, Eggshell, Dystrophic, and Skin lesion all statistically significantly affect the female patients last status at 1% level. These answer the second part of the second research question. These results of coincide with that of Zangmo and Tiensuwan (2018); and Nahleh et al. (2007).

Here, female white patients, female married patients, aged female patients, female patients with Mass Stability and presence of Axillary Adenopathy, and Lucent are more likely to die of breast cancer since they positively affect the female patients' last status, while female patients with less cancer Grade, less Length of stay in the hospital after admission, laterality, Family history of breast cancer, prior history of breast surgery, presence of Amorphous, less Mass Size, presence of Dystrophic, and skin Lesion are more likely to survive since they negatively affect the patients last status. Most of the risk factors in the females contribute negatively to the patients' last status.

5.4 Model Comparison and Model Selection.

In this section, we compare all fitted models' prediction performance using all the model evaluation techniques used in all three cases considered in this study to come up with the best model to predict the survival or death of breast cancer patients. These will help us better understand to answer the third research question posed earlier in this study.[2] Considering the male and female patients together, comparing all five model in this study in Table 4.2, we note that the Convolutional Network gave an overall prediction of 90.750% which is the greatest among the five models with an AUC of 0.96 which is also the greatest. We also note that the Convolutional Network gave a sensitivity of 0.90361, specificity of 0.9108 and Negative Predictive Value of 0.917 which are all the greatest among the five models' evaluation techniques. These give us enough reasons to conclude that the convolutional network performs better than all other models tested in this study for the target population, though the it gave a smaller Positive Predictive Value of 0.897 compared with all other models.

[2]The three cases considered were male and female patients together, Male patients only, and Female patients only.

Using the same criteria for model selection, the Random Forest and the RBF also perform well with an overall prediction of 90.311%, 90.171% respectively and a high percentages of all other model evaluation techniques as compared with the Logistic regression, and the Multi-Layer network.[3]

Also, from Table 4.4, we compared the results of all models for the male patients only. Here, the Convolutional Neural Network performs best compared with all other models. We note an overall prediction accuracy of 95.994% with a very high AUC of 0.98 with all other model evaluation techniques having very high values. These also give us enough evidence to conclude that the CNN is the best model for predicting male breast cancer patients' last status. The Random forest is seen here as the second best predictive model next to CNN with an overall prediction accuracy of 94.942% and an AUC of 0.95. For this model, we are able to determine each variable contribution to the model and the distribution of the entire variables[4] which help us to better understand which factor to focus on to make further proposal to a male patient as to how to manage the disease. All other models do not give us these details. The Logistic regression, Multi-layer Network, and the RBF do not perform better as the other two model.[5] These results are not supported by that obtained by Ayer et al. (2010).

Further, considering the female patients models from Table 4.6, the Convolutional Neural Network gave an overall prediction accuracy of 88.780% which is the greatest among all the five models and a very high AUC of 0.95. All other evaluation techniques of the Convolutional Neural Network provides greater probabilities compared with all other fitted models.[6] These also give use sufficient evidence to conclude that the CNN is the best model for predicting Female breast cancer patients' last status. Here, the Random Forest is performing closer with the Convolutional Neural Network, with an overall prediction accuracy of 88.640% and an AUC of 0.89 which is higher that that of the CNN. The logistic regression performs poorly here while the Multi-layer network and the RBF performs well compared with that of the logistic regression.

In answering the third research question posed earlier in this study, we consider all the three cases analyzed above. Here, we conclude that the Convolutional Neural Network is best for predicting breast cancer patients' survival or death in the target population of this study viewing it from the angels of all patients, and male and female patients separately. These conclusions are supported by the figures and table provided in the discussions in previous paragraphs.

[3]Refer to Table 4.2 for the details of obtained values of all the model techniques noting the obtained values for the Random Forest Network and the RBF compared with that of logistic regression, and the Multi-layer network.

[4]See Appendix Table A.3 and Figure 4.13 for the details of the male variables contribution to the male patients Random Forest model.

[5]The Convolutional Neural Network, and the Random Forest Network.

[6]Refer to Table 4.6 for the details of all models predicted values.

CONCLUDING REMARKS

This study examines the factors that affect breast cancer patients last status in the United States and proposes the best model to predict the patients' last status using data on breast cancer patients treated at various hospitals, and were collected from the CDC for the states of San Francisco, Connecticut, and New Jersey. The collect data covered the period 1992 – 2019 and contained 25 breast cancer patients demographic risk factors and mammographic descriptors.

We found that both male and female patients' Race, Marital status, Age diagnosed, Mass Stability, Number of visit to the hospital, Mass Density, Nipple Retraction, Lymph Node, Amorphous, Milk, Axillary Adenopathy, and Lucent all positively affect the patients' last status, while Gender, cancer Grade, patients Length of stay in the hospital after diagnosed, Laterality, Family history of breast cancer, Prior-history of breast cancer surgery, Mass Size, Eggshell, Dystrophic and Skin Lesion all negatively affect patients' last status.

Considering the male patients separately, we found that cancer Grade, Suture, Density, Nipple Retraction, Lymph Node, Amorphous, Mass Size, Lucent, and Dermal positively affects male patients last status, while Number of visit to the hospital, Length of stay in the hospital after diagnosed, Laterality, Family history of breast cancer, Prior-history of breast cancer surgery, Milk, Dystrophic, and Skin Lesion all negatively affect male patients' last status.

Also, considering the female patients separately, we found that patients' Race, Marital status, Age diagnosed, Mass Stability, Number of visit to the hospital, Mass Density, Nipple Retraction, Lymph Node, Milk, Axillary Adenopathy, and Lucent all positively affect the female patients' last status, while cancer Grade, patients' Length of stay in the hospital after diagnosed, Laterality, Family history of breast cancer, Prior-history of breast cancer surgery, Amorphous, Mass Size, Eggshell, Dystrophic, and Skin Lesion all negatively affect the female patients' last status.

Further, the Convolutional Neural Network was found to be the best model for predicting breast cancer patients last status with higher degree of accuracy and overall prediction performance. Though the Convolutional Neural Network performs best, we are unable to determine each variable contribution to the model and their significance which highlights the disadvantage of using the Convolutional Neural Network. Although the Random Forest is not seen as the overall best model for predicting breast cancer patient's last status, it performs closely well as the Convolutional Neural Network and presents the contribution of each variable to the model, and the distribution of the variables in the model which provides a better fit when a physician, radiotherapist, and a breast cancer patient what to consider making analysis for decision with the factors. This is in fact easier to use model in this case.

The logistic regression is also easier to use and presents the contribution of each variable to the model (Sargent, 2001), as the Random Forest but prediction performances were very

low. The RBF and the Multi-Layer Network did not perform well as the Convolutional Neural Network and Random Forest, and do not also present the contribution of each variable to the model. These highlight the advantages and disadvantages for using each model.

Future research should consider collecting data covering more states in the US on breast cancer patients and possibly extending the scope to other cancer types, and make comparison of the models' prediction performance and their evaluation. It could also consider a deep dive into the comparison of Logistic regression and Random Forest network variable importance which will aid in a better decision making.

REFERENCES

1. Xu, J. Q., Murphy, S. L., Kochanek, K. D., Bastian, B., and Arias, E.(2018). Deaths: final data for 2016. *Centers for Disease Control and Prevention*, 67(5), 2018–1120
2. World Health Organization (2017). Retrieved from *http://www.who.int/mediacentre/factsheetfs297/en/*
3. Ahmadian, A., (2016). Numerical Models for Submerged Breakwaters. Coastal Hydrodynamics and Morphodynamics.
4. Cancer Treatment Center of America (2019). Retrieved from *https://www.cancercenter.com/cancer-types/breast-cancer*
5. Bartfay, E., Mackillop, WJ., and Pater, JL. (2006). Comparing the predictive value of neural network models to logistic regression models on the risk of death for small-cell lung cancer patients. *European Journal of cancer care*, 15(2), 115–124.
6. Kahneman, D., Slovic, P., and Tversky A. (1982). Judgment under uncertainty: heuristics and biases.*Cambridge, England: Cambridge University Press.*
7. Liberman, L., Abramson, A. F., Squires, F. B., Glassman, J. R., Morris, E. A., and Dershaw, D. D. (1998). The breast imaging reporting and data system: positive predictive value of mammographic features and final assessment categories. *American Journal of Roentgenology*, 171(1), 35-40.
8. Hong, A. S., Rosen, E. L., Soo, M. S., and Baker, J. A. (2005). BI-RADS for sonography: positive and negative predictive values of sonographic features. *American Journal of Roentgenology*, 184(4), 1260-1265.
9. American Cancer Society, 2019. Retrieved from *https://www.cancer.org/cancer/breast-cancer/understanding-a-breast-cancer-diagnosis/types-of-breast-cancer.html*

10. Ayer, T., Chhatwal, J., Alagoz, O., Kahn Jr, C. E., Woods, R. W., and Burnside, E. S. (2010). Comparison of logistic regression and artificial neural network models in breast cancer risk estimation. *Radiographics*, 30(1), 13-22.

11. Chang, C. L., and Hsu, M. Y. (2009). The study that applies artificial intelligence and logistic regression for assistance in differential diagnostic of pancreatic cancer. *Expert Systems with Applications*, 36(7), 10663-10672.

12. Yusuff, H., Mohamad, N., Ngah, U. K., and Yahaya, A. S. (2012). Breast cancer analysis using logistic regression. *International Journal of Research and Reviews in Applied Sciences*, 10(1), 14-22.

13. Al-Ghamdi, A. S. (2002). Using logistic regression to estimate the influence of accident factors on accident severity. *Accident Analysis and Prevention*, 34(6), 729-741.

14. Ahmed, A., Ukwenya, Y., Abdullahi, A., and Muhammad, I. (2012). Management and outcomes of male breast cancer in Zaria, Nigeria. *International journal of breast cancer*. Retrieved from *https://doi.org/10.1155/2012/845143*

15. Burivong, W., and Amornvithayacharn, O. (2011). Accuracy of subcategories A, B, C in BI-RADS 4 lesions by combined mammography and breast ultrasound findings. *Journal of Medicine and Medical Sciences*, 2(3), 728-733.

16. Balleyguier, C., Ayadi, S., Van Nguyen, K., Vanel, D., Dromain, C., and Sigal, R. (2007). BIRADS™ classification in mammography.*European journal of radiology*, 61(2), 192-194.

17. Maiprasert, D., and Kitbumrungrat, K. (2012). Comparison multinomial logistic regression and discriminant analysis in predicting the stage of breast cancer.*International Journal of Computer Science and Network Security (IJCSNS)*, 12(4), 44.

18. Yuri, P., Rochadi, S., and Danarto, R. (2016). A Device for Predicting Prostate Cancer Risk: A Logistic Regression. *J Pros Canc*, 1(111), 1-5.

19. Cancato, J., Feinstein, A. R., and Holford, T. R. (1993). The risk of determining risk with multivariate models. Ann Intern Med, 118(3), 201-210.

20. Bandhita, P., Noparat, T. (2006). Ordinal regression analysis in factors related to sensorial hearing loss of the employee in industrial factory in Lampang Thailand. Mathematic. *Statistics and Their Application, Penang.*

21. Zangmo, C., and Tiensuwan, M. (2018). Application of logistic regression models to cancer patients: a case study of data from jigme dorji wangchuck national refrerral hospital (jdwnrh) in bhutan. In *Journal of Physics: Conference Series*, 1039(012031), 1-10.

22. Roy, S. S., and Guri, S. (2008). Diagnostics in logistic regression models. Retrieved from *https://doi.org/10.1016/j.jkss.2007.03.001*

23. Stephenson, B., Cook, D., Dixon, P., Duckworth, W., Kaiser, M., Koehler, K., and Meeker, W. (2008). Binary response and logistic regression analysis. Retrieved from *http://www.stat.wisc.edu/mchung/teaching/MIA/reading/GLM.logistic.Rpackage. Pdf*

24. Madhu, B., Ashok, N. C., and Balasubramanian, S. (2014). A multinomial logistic regression analysis to study the influence of residence and socio-economic status on breast cancer incidences in southern Karnataka. *International Journal of Mathematics and Statistical Invention*, 2(5), 01-8.

25. Ahmadian, A. S., Simons, R. R. (2018). Estimation of nearshore wave transmission for submerged breakwaters using a data-driven predictive model. *Neural Computing and Applications*, 29(10), 705-719.

26. Colditz, G. A., Willett, W. C., Hunter, D. J., Stampfer, M. J., Manson, J. E., Hennekens, C. H., and Speizer, F. E. (1993). Family history, age, and risk of breast cancer: prospective data from the Nurses' Health Study. *Jama Health Forum*, 270(3), 338-343.

27. Gui, G. P. H., Hogben, R. K. F., Walsh, G., A'Hern, R., and Eeles, R. (2001). The incidence of breast cancer from screening women according to predicted family history risk: does annual clinical examination add to mammography? *European Journal of Cancer*, 37(13), 1668-1673.

28. Rawal, R., Bertelsen, L., and Olsen, J. H. (2006). Cancer incidence in first-degree relatives of a population-based set of cases of early-onset breast cancer. *European Journal of Cancer*, 42(17), 3034-3040.

29. Gatta, G., Pinto, A., Romano, S., Ancona, A., Scaglione, M., and Volterrani, L. (2006). Clinical, mammographic and ultrasonographic features of blunt breast trauma. *European journal of radiology*, 59(3), 327-330.

30. Eltoukhy, M. M., Faye, I., and Samir, B. B. (2010).Breast cancer diagnosis in digital mammogram using multiscale curvelet transform. *Computerized medical imaging and graphics*, 34(4), 269-276.

31. Moezzi, M., Melamed, J., Vamvakas, E., Inghirami, G., Mitnick, J., Quish, A., and Feiner, H. (1996). Morphological and biological characteristics of mammogram-detected invasive breast cancer. *Human pathology*, 27(9), 944-948.

32. Borgen, P. I., Senie, R. T., McKinnon, W. M., and Rosen, P. P. (1997). Carcinoma of the male breast: analysis of prognosis compared with matched female patients. *Annals of Surgical Oncology*, 4(5), 385-388.

33. Kornegoor, R., Moelans, C. B., Verschuur-Maes, A. H., Hogenes, M. C., De Bruin, P. C., Oudejans, J. J., and Van Diest, P. J. (2012). Oncogene amplification in male breast cancer: analysis by multiplex ligation-dependent probe amplification. *Breast cancer research and treatment*, 135(1), 49-58.

34. Nahleh, Z. A., Srikantiah, R., Safa, M., Jazieh, A. R., Muhleman, A., Komrokji, R. (2007). Male breast cancer in the veterans affairs population: a comparative analysis. Cancer: *Interdisciplinary International Journal of the American Cancer Society*, 109(8), 1471-1477.

35. Sargent DJ. (2001). Comparison of artificial neural networks with other statistical approaches: results from medical data sets. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, 91(58), 1636–1642.

36. Toward Data science (2020). Retrieve from *https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53*)

37. Toward Data science (2020). Retrieve from *https://towardsdatascience.com/radial-basis-functions-neural-networks-all-we-need-to-know-9a88cc053448*)

APPENDIX A

Table A.1 Summary information of the data set

State	**Registry ID**	**Gender**		**Patient Last Status**		**Observation**
		Male	**Female**	**Alive**	**Dead**	
San Francisco	1501	201	1501	466	1,236	**1,702**
Connecticut	1502	605	45,766	28,733	17,638	**46,371**
New Jersey	1544	14,793	37,136	17,486	34,443	**51,929**
Total		**15,599**	**84,403**	**46,685**	**53,317**	**100,002**

Table A.2 Data variable description with python codes.

VARIABLE	PYTHON VARIABLE	DEFINITION AND CODE
Patient's last status	PatStatus	The patients last status after treatment (Alive = 0 or death = 1)
Patient's race	Race	The race/ethnicity of the patients (Black = 0 or Brown = 1)
Marital status	MarST	(Married = 1 or Not married = 0)
Gender	Gender	(Male = 1 or Female = 0)
Age at diagnose	AgeDiag	The age at which the patients was diagnosed of breast cancer, (Young = 0 or Aged = 1)
Family history of cancer	FamHist	The family history of breast cancer, (Yes =1 or No = 0)
Prior history of breast cancer surgery	PrioBSurgy	History of surgery of breast cancer (Yes = 1 or No = 0)
Mass size	Size	Description of mass size (small $<$ 3 cm or large $\geq$ 3cm)
Grade	Grade	Patients grade of breast cancer (grade 1 = 0, grade 2 = 1 or grade 3 = 2)
Length of stay	Lstay	Length of admission in the hospital after diagnosed $<$ 5 days or $\geq$ 5 days.
Laterality	Laterality	Describes the side of paired organ/body on which reportable tumor originated (Right = 1 or Left = 0)
Number of visit	No.Visits	Number of visits to the hospital ($<$ 5 times or $\geq$ 5 times)
Mass stability	Stability	Describes the Mass of cancer (Unstable = 1 or stable = 0)
Suture	Suture	Describes the calcification (Present = 1 or Not present = 0)
Mass Density	Density	Mass density of the cancer (High = 3, Equal = 2, Low = 1, Fat containing = 0)
Nipple Retraction	NipRet	The status of nipple retraction (Present = 1 or Not present = 0)
Lymph Node	LyNode	A special case of the presence of lymph node in the patients (Present = 1 or Not present = 0)
Amorphous	Amorph	The calcification with the presence of Amorphous (Present = 1 or Not present = 0)
Eggshell	Eggshell	The calcification with the presence of eggshell (Present =1 or Not present = 0)

Table A.2 Continued

VARIABLE	PYTHON VARIABLE	DEFINITION OF VARIABLE AND PYTHON CODE
Milk	Milk	The calcification with the presence of milk (Present = 1 or Not present = 0)
Axillary Adenopathy	AxiAden	The associated finding of Axillary Adenopathy in the cells (Yes = 1 or No = 0)
Dystrophic	Dystroph	The calcification with presence of Dystrophic in the cell (Present = 1 or Not present = 0).
Lucent	Lucent	The calcification with presence of Lucent (Present = 1 or Not present = 0).
Dermal	Dermal	The calcification with presence of Dermal in cells (Present = 1 or Not present = 0).
Skin lesion	Skin Lesson	An associated mammography finding with skin lesion (Present = 1 or Not present = 0).

Table A.3 Random Forest Variable importance for the Male and Female patients.

MALE		**FEMALE**	
Variable	**Variable Importance**	**Variable**	**Variable Importance**
SkinnLesson	0.349386	Lstay	0.132272
Size	0.227863	Density	0.101575
Lstay	0.103242	NipRet	0.100182
No.Visits	0.051164	No.Visits	0.080408
Milk	0.039392	SkinnLesson	0.077362
AgeDiag	0.035818	AgeDiag	0.071854
Amorph	0.026599	Laterality	0.064168
Density	0.023298	AxiAden	0.063047
Lucent	0.020689	Size	0.056779
Laterality	0.018861	Amorph	0.036994
Race	0.014629	LyNode	0.034484
LyNode	0.014476	Milk	0.028956
NipRet	0.010778	FamHist	0.028135
Grade	0.010599	Lucent	0.019203
PrioBSurgy	0.010189	Race	0.016249
Stability	0.009298	Suture	0.01595
AxiAden	0.006768	Distroph	0.014919
MarST	0.006625	PrioBSurgy	0.013072
Distroph	0.005998	Grade	0.012817
Suture	0.005569	Eggshell	0.01259
Dermal	0.00527	MarST	0.007673
Eggshell	0.003468	Dermal	0.00671
FamHist	0.000021	Stability	0.004602

	Act 1s	Act 0s
Pred 1s	8286	954
Pred 0s	2998	7763

Figure A.1: Confusion Matrix for Logistic Regression on Box Sex

	Act 1s	Act 0s
Pred 1s	922	144
Pred 0s	407	1647

Figure A.2: Confusion Matrix for Logistic Regression on Male Sex

	Act 1s	Act 0s
Pred 1s	6435	1817
Pred 0s	1834	6795

Figure A.3: Confusion Matrix for Logistic Regression on Female Sex

	Act 1s	Act 0s
Pred 1s	8887	410
Pred 0s	2354	8350

Figure A.4: Confusion Matrix for Multi-Layer Network on Box Sex

	Act 1s	Act 0s
Pred 1s	951	113
Pred 0s	423	1633

Figure A.5: Confusion Matrix for Multi-Layer Network on Male Sex

	Act 1s	Act 0s
Pred 1s	7960	292
Pred 0s	2029	6600

Figure A.6: Confusion Matrix for Multi-Layer Network on Female Sex

	Act 1s	Act 0s
Pred 1s	8336	961
Pred 0s	889	9815

Figure A.7: Confusion Matrix for CNN on Both Sex

	Act 1s	Act 0s
Pred 1s	980	123
Pred 0s	2	2015

Figure A.8: Confusion Matrix for CNN on Male Sex

	Act 1s	Act 0s
Pred 1s	6642	1530
Pred 0s	364	8345

Figure A.9: Confusion Matrix for CNN on Female Sex

	Act 1s	Act 0s
Pred 1s	8639	601
Pred 0s	1365	9396

Figure A.10: Confusion Matrix for RBF on Box Sex

	Act 1s	Act 0s
Pred 1s	973	91
Pred 0s	422	1634

Figure A.11: Confusion Matrix for RBF on Male Sex

	Act 1s	Act 0s
Pred 1s	7980	272
Pred 0s	2104	6525

Figure A.12: Confusion Matrix for RBF on Female Sex

	Act 1s	Act 0s
Pred 1s	8807	490
Pred 0s	1448	9256

Figure A.13: Confusion Matrix for Random Forest on Both Sex

	Act 1s	Act 0s
Pred 1s	983	101
Pred 0s	3	969

Figure A.14: Confusion Matrix for Confusion Matrix for Random Forest on Male Sex

	Act 1s	Act 0s
Pred 1s	7701	551
Pred 0s	1316	7313

Figure A.15: Confusion Matrix for Random Forest on Female Sex