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

Eric Agyemang
(eagyem2@ilstu.edu)

Illinois State University
Department of Mathematics
Normal IL, USA.
Supervisor: Dr. Pei Geng

December 1, 2020

Outline

- Problem Background.
- Objective of the Study.
- Research Questions.
- Methodology.
- Model Results.
- Concluding Remarks.
- Future Research.

Problem Background

- Breast cancer is a type of cancer in which the cells in the breast grow out of control.
- Is the second most common cause of death from cancer in women after lung cancer and over 50 thousand new cases reported every year in the US.
- May be caused by factors such as genetic transition, obesity, abnormal hormone, breast density, menstrual history, heavy drinking, etc.
- Is mostly diagnosed in women between age 55 and 64 and men between 60 and 70 (Cancer Treatment Center of America, 2019).
- Accurate prediction of breast cancer clinical outcomes (Survive/death) is an integral to the successful decision making which can result to better patient care (Bartfay, 2006).

Objective of the Study

- To identify the key factors that affect breast cancer patients' last status within the target population in the US.
- To identify the key factors that affect the last status of male and female breast cancer patients separately within the target population in the US.
- To use machine learning and deep learning techniques to build models which predict breast cancer patient's last status and determine the best model for predicting the patient's last status in contributing to improve clinical decision making in the US.

Research Questions

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

Methodology

Data Collection

- Data on 100,002 breast cancer patients were collected covering three states in the US for the period 1992 - 2019 and were collected from the CDC.
- Data set contains 25 characteristics of breast cancer grouped under demographic risk factors and mammographic descriptors.
- Demographic risk factors: Race, Marital status, Gender, Family history of breast cancer, Prior history of breast surgery, age, etc.
- Mammographic descriptors: Mass size, Mass stability, Mass Density, Amorphous, Eggshell, Milk, Axillary Adenopathy, etc.

Data Collection continued....

- **Summary information of the data set**

State	Gender		Patient Last Stat		Obs.
	Male	Female	Alive	Dead	
San Fran.	201	1501	466	1,236	**1,702**
Connect.	605	45,766	28,733	17,638	**46,371**
New Jersey	14,793	37,136	17,486	34,443	**51,929**
Total	**15,599**	**84,403**	**46,685**	**53,317**	**100,002**

Methodology Cont...

Model Construction

- Five classification models based on sex with the patient's demographic risk factors, and mammographic descriptors are developed using Logistic Regression, Feed Forward Neural Networks, Support Vector Machine, and Random Forest Classification.
- We develop the models to predict breast cancer patients' last status for both sex, and each sex separately.
- Use classification model evaluation techniques to assess the models performance for comparison.
- Python-Jupyter Notebook used for the analysis.

Methodology Cont...

Logistic Regression

- Is a machine learning technique that falls within the class of supervised learning under the classification models.
- 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.
- Probability of a patient's last status (alive or dead) could be estimated with this model using the odd ratios and ranges from 0 to 1.
- With the knowledge of the patient's Age, Gender, Family history of breast cancer, Prior history of breast cancer surgery etc., we can predict the patients last status.

Multi-layer Perceptron (MLP)

- Is a supervised learning algorithm that learns a function by training on a data set with a number of dimensions for the input and a number of dimensions for the output.
- This model can learn a non-linear function approximator for either a classification or regression approach.

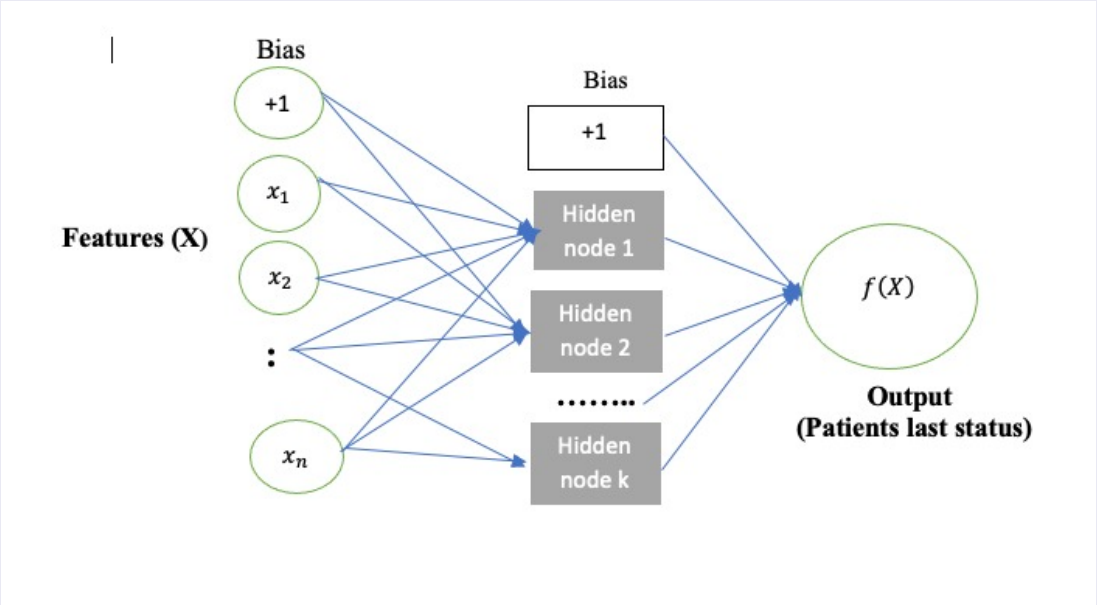

Figure 1: One Hidden Layer MLP

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.
- It is mostly trained using the activation function back propagation.

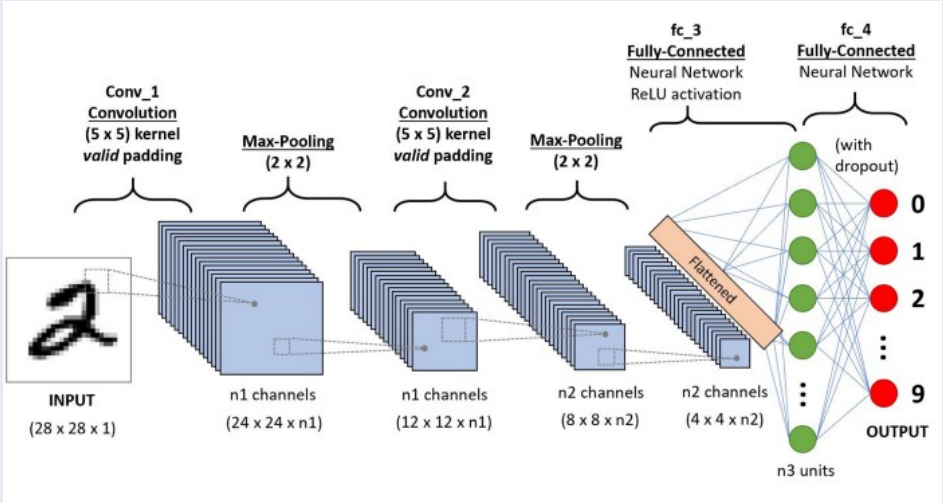

Figure 2: CNN sequence to classify patients last status

- This process trains the CNN and all the weighs and parameters are optimized to correctly classify images from the data set.

Radial Basis Function (RBF) Network

- Is the commonly used type of the artificial neural network for function approximation problems.
- 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.
- Classifications only occur in the second phase where the linear combination of hidden functions are driven to output layer

Radial Basis Function (RBF) Network Continued...

- The second training phase also updates the weighting vectors between the hidden layer and output layer of the model.

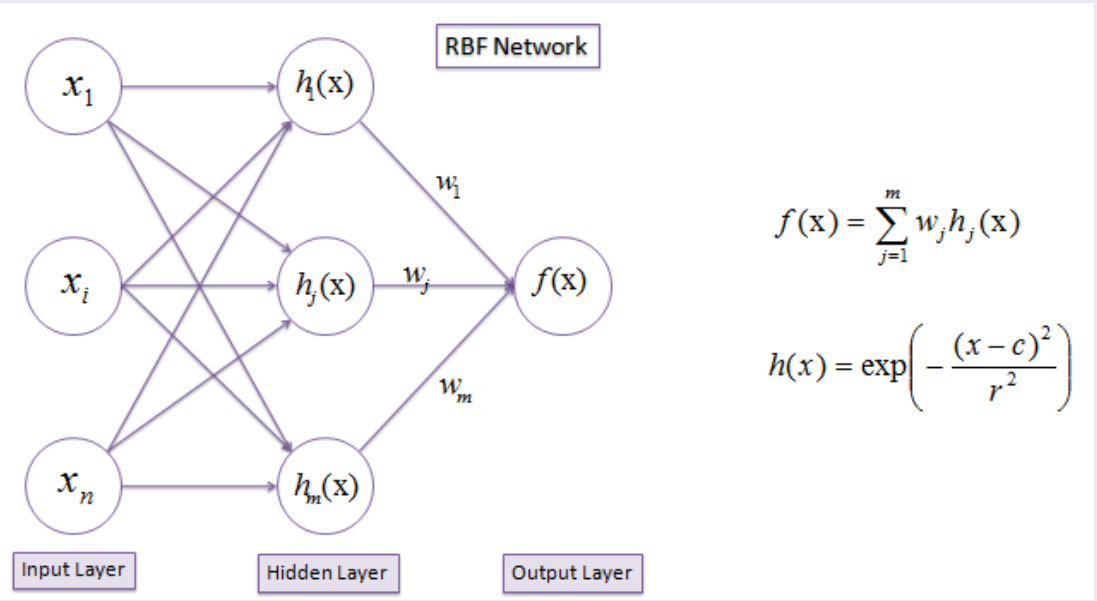

Figure 3: RBF sequence to classify patients' last status

Methodology Cont...

Random Forest Networks

- 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.
- There are n estimators hyper parameter during the training which are the number of trees the algorithm builds before taking the averages of the predictions.

Random Forest Networks Continued...

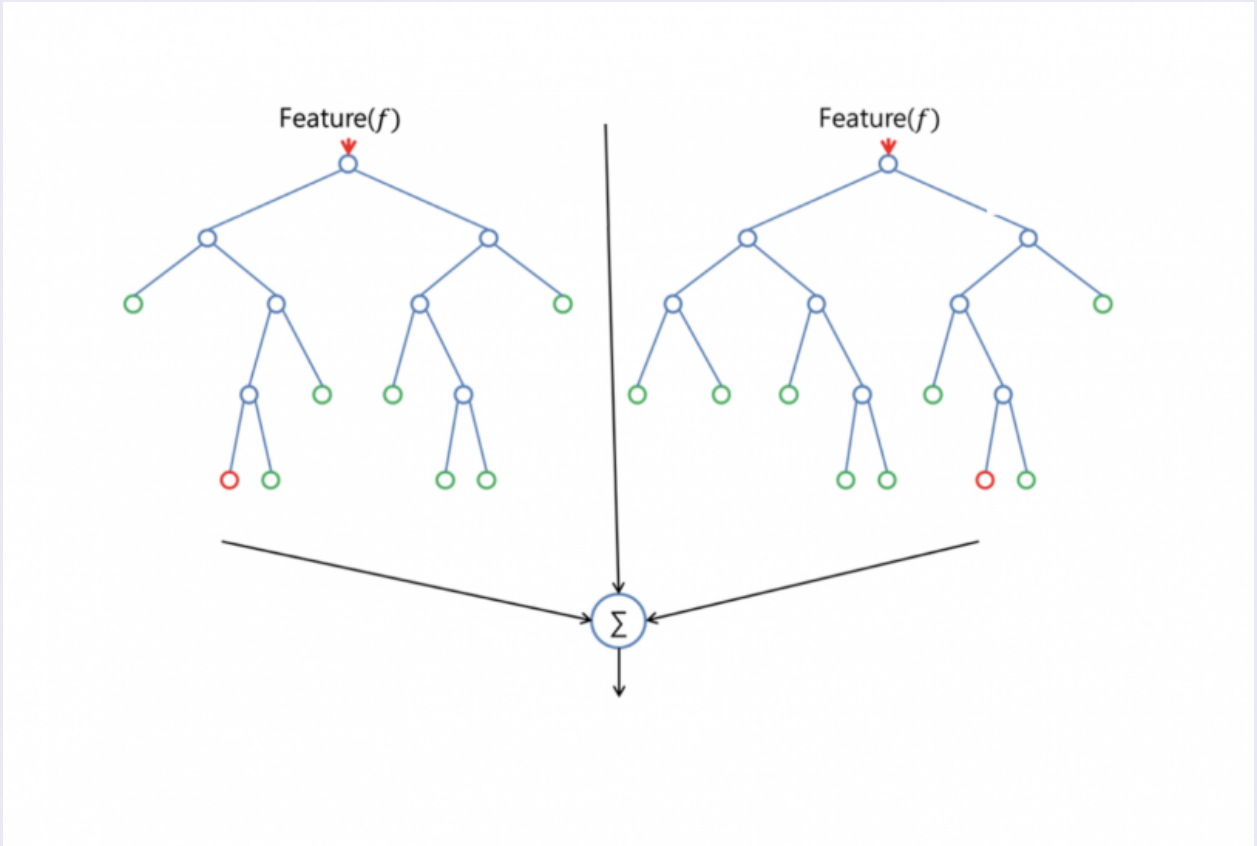

Figure 4: Random Forest Networks to classify patients' last status

Model Evaluation

- **Accuracy (ACC) :** The classification accuracy is the number of correct predictions on the total number of inputs samples. It is given by: $ACC = \frac{TP+TN}{TP+TN+FP+FN}$
- **Sensitivity or True Positive Rate (TPR):** The probability that the model predicts positive as death given that the patient actually died. It is given by: $TPR = \frac{TP}{TP+FN}$
- **Positive Predictive Value (PPV):** Given that the model predict positive as death, what is the probability that the patients actually died ? This is given by: $PPV = \frac{TP}{TP+FP}$
- **Specificity or True Negative Rate (TNR):** The probability that the model predicts negative as Alive given that the patients are actually alive. This is given by: $TNR = \frac{TN}{TN+FP}$

Model Evaluation Continued...

- **Negative Predictive Value (NPV):** Given that the patients are alive, what is the probability that the model predicts negative as death? This is given by: $NPV = \frac{TN}{TN+FN}$
- **The Receiver Operating Characteristic (ROC) curve:** This graph summarizes the performance the classifier model. It is generated by plotting the TPR against the FPR as the threshold is varied for assigning observations to the given class.
- **The area under a ROC curve (AUC):** indicates how well the predictive model discriminates between healthy patients and patients who died. value of an AUC varies between 0.5 and 1.0.

Model Results.

Logistic Regression for Both Male and Female Patients

- Considering the first research question "What key factors affect breast cancer patients" last status within the target population in the US ?"

Table 1: Logistic Regression Results for Male and Female Patients.

Variable	**Estimate**	**P**$>|z|$
Patient's Race	0.0056	0.000
Marital Status	0.0538	0.004
Patient's Gender	-0.0592	0.024
Age at Diagnosed of Breast Cancer	0.0066	0.000
Mass Stability	0.2178	0.000
Number of Visits to Hospital	0.0335	0.000
Length of stay in Hospital	-0.1740	0.000
Laterality	-0.0810	0.000

Logistic Regression for Both Male and Female Continued...

Table 1 Continued

Variable	**Estimate**	**P**$>\|z\|$
Family History of Breast Cancer	-3.3407	0.000
Prior History of Breast cancer Surgery	-0.3006	0.000
Mass Density of the cancer	0.2365	0.000
Nipple Retraction Status	2.1587	0.000
Presence of Lymph Node	0.3620	0.000
Presence of Calcification Amorphous	0.1636	0.000
Mass Size	-0.1563	0.000
Presence of Calcification Eggshell	-0.4962	0.000
Presence of Calcification Milk	0.8963	0.000
Presence of Axillary Adenopathy	1.0621	0.000
Presence of Dystrophic	-0.7367	0.000
Presence of Calcification Lucent	1.9973	0.000
Presence of Skin Lesion	-1.6329	0.000

Logistic Regression for Male and Female Separately

Logistic Regression For Male Patients

- Considering the first part if the second research question " What factors affects the last status of male breast cancer patients separately within the target population in the USA?"

Table 2: Logistic Regression Results for Male Breast Cancer Patients.

Variable	**Estimate**	**P**$>|z|$
Patient's Race	-0.0008	0.724
Marital Status	-0.0004	0.993
Age at Diagnosed of Breast Cancer	0.0026	0.175
Cancer Grade	0.1192	0.001
Mass Stability	0.0759	0.377
Number of Visits to Hospital	-0.0218	0.029

Logistic Regression for Male Patients Continued

Table 2 Continued

Variable	**Estimate**	**P**$>|z|$
Length of stay in Hospital	-0.1344	0.000
Family History of Breast Cancer	-2.6038	0.000
Prior History of Breast cancer Surgery	-0.3056	0.020
Presence of Suture	0.1467	0.018
Mass Density of the cancer	0.2654	0.000
Nipple Retraction Status	0.8825	0.000
Presence of Lymph Node	2.6183	0.000
Presence of Calcification Amorphous	1.4480	0.000
Mass Size	0.5909	0.000
Presence of Calcification Milk	-0.8173	0.000
Presence of Calcification Lucent	0.9937	0.000
Presence of Calcification Dermal	0.1767	0.005
Presence of Skin Lesion	-2.7320	0.000

Logistic Regression For Female Patients

- Considering the second part of the second question " What factors affects the last status of female breast cancer patients separately?"

Table 3: Logistic Regression Results for female Breast Cancer Patients.

Variable	**Estimate**	**P**$>|z|$
Patient's Race	0.0059	0.000
Marital Status	0.0730	0.000
Age at Diagnosed of Breast Cancer	0.0078	0.000
Age at Diagnosed of Breast Cancer	-0.0393	0.004
Mass Stability	0.1699	0.000
Number of Visits to Hospital	0.0507	0.000
Length of stay in Hospital	-0.1656	0.000
Laterality	-0.0942	0.000

Logistic Regression For Female Patients

table 3 Continued

Variable	**Estimate**	**P**$>|z|$
Family History of Breast Cancer	-3.2622	0.000
Prior History of Breast cancer Surgery	-0.2662	0.000
Mass Density	0.1879	0.000
Nipple Retraction Status	2.2790	0.000
Presence of Lymph Node	0.1978	0.000
Presence of Calcification Amorphous	-0.1152	0.001
Mass Size	-0.2425	0.000
Presence of Calcification Eggshell	-0.5637	0.000
Presence of Calcification Milk	0.9166	0.000
Presence of Axillary Adenopathy	1.2250	0.000
Presence of Dystrophic	-0.6857	0.000
Presence of Calcification Lucent	2.0236	0.000
Presence of Skin Lesion	-1.4800	0.000

Model Results Cont...

Male Patients Variables Importance to the Random Forest

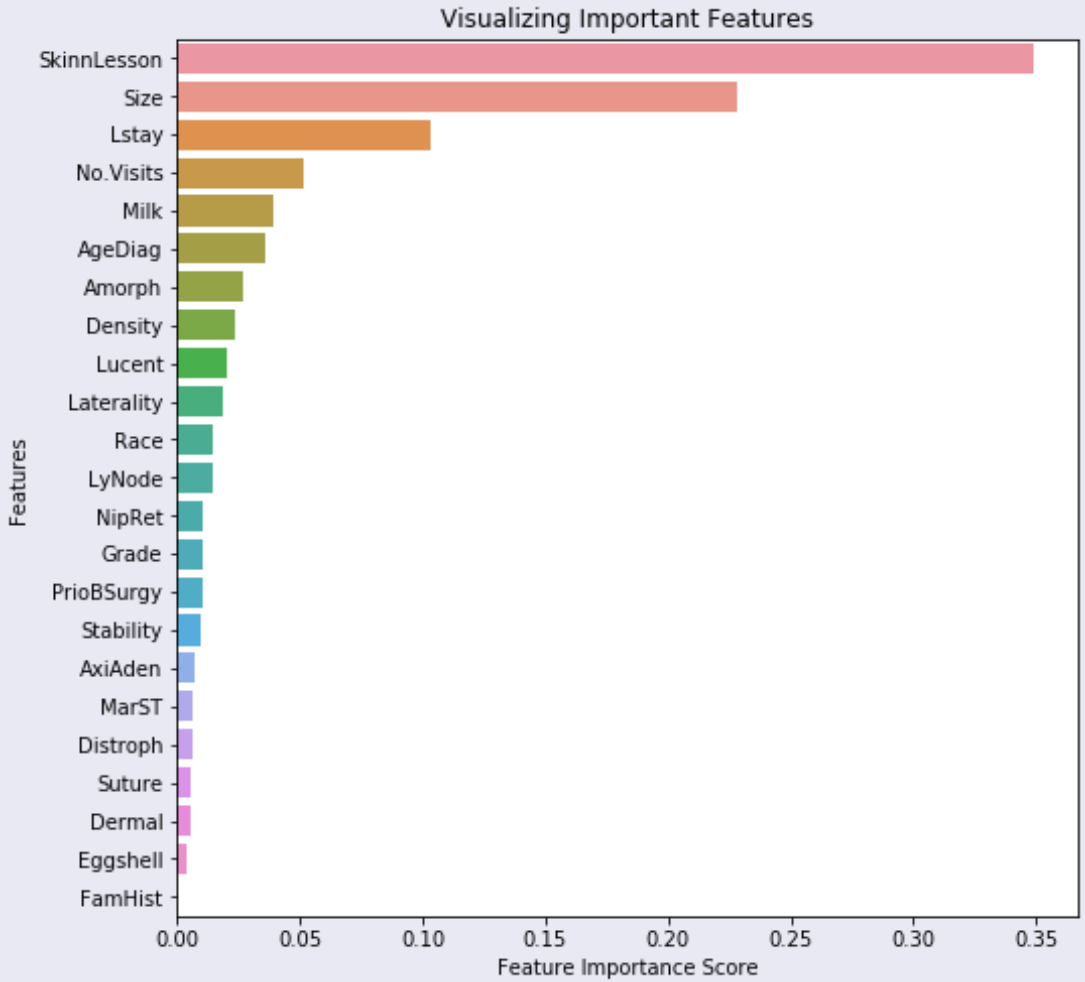

Figure 5: Distribution of Male patient's variable importance to Random Forest

Model Results Cont...

Female Patients Variables Importance to the Random Forest

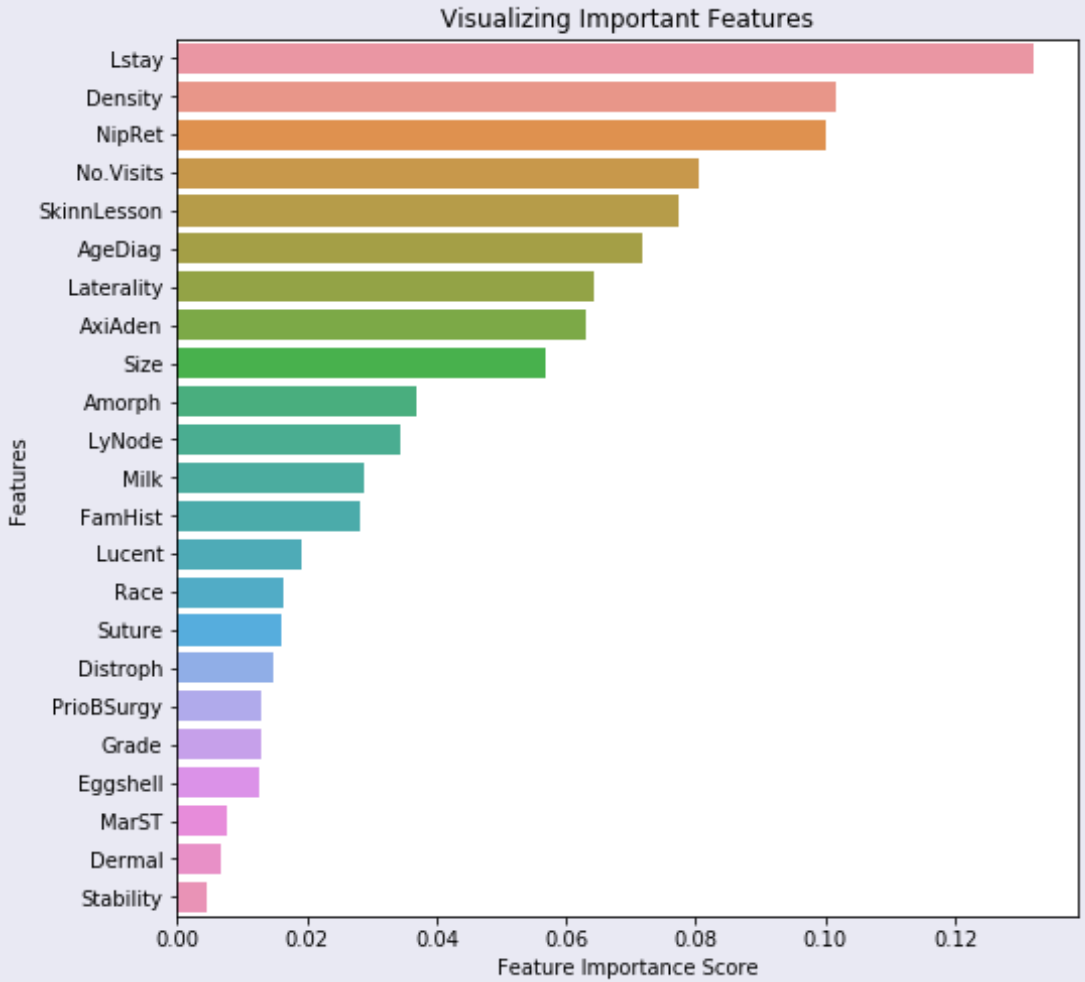

Figure 6: Distribution of Female patient's variable importance to Random Forest

Model Results Cont.

Model Evaluation, Comparison, and Model Selection

- Considering the third research question " Which model is the best fit for predicting breast cancer patient's last status in contributing to improve clinical decision making?"

Table 4: Summary of All Model Results for Male and Female Patients

Method	**Logistic.**	**MLP**	**CNN**	**RBF**	**Rand. F.**
Sensit.	0.734	0.791	0.904	0.864	0.859
PPV	0.897	0.956	0.897	0.935	0.947
Specif.	0.891	0.953	0.911	0.939	0.949
NPV	0.721	0.780	0.917	0.873	0.865
AUC	0.88	0.90	0.96	0.90	0.91
ACC	**80.24%**	**86.18%**	**90.75%**	**90.17%**	**90.31%**

Summary of All Models Results for Male Breast Cancer Patients

- We consider male patients models' evaluation for comparison.

Table 5: Summary of All Model Results for Male Patients

Method	**Logistic.**	**MLP**	**CNN**	**RBF**	**Rand. F.**
Sensit.	0.694	0.692	0.998	0.697	0.997
PPV	0.864	0.894	0.888	0.914	0.907
Specif.	0.919	0.935	0.942	0.947	0.906
NPV	0.802	0.794	0.999	0.915	0.997
AUC	0.87	0.88	0.98	0.85	0.95
ACC	**82.34%**	**83.00%**	**95.99%**	**83.56%**	**94.94%**

Summary of All Models Results for Female Breast Cancer Patients

- We consider female patients models' evaluation for comparison.

Table 6: Summary of All Model Results for Female Patients

Method	**Logistic.**	**MLP**	**CNN**	**RBF**	**Rand. F.**
Sensit.	0.778	0.797	0.948	0.791	0.854
PPV	0.780	0.964	0.958	0.967	0.933
Specif.	0.789	0.958	0.845	0.959	0.929
NPV	0.787	0.764	0.813	0.756	0.847
AUC	0.86	0.80	0.95	0.86	0.89
ACC	**78.37%**	**86.25%**	**88.78%**	**85.92%**	**88.64%**

Concluding Remarks

- Convolutional Neural Network outperforms all other models tested in all three cases.
- The Random Forest is seen here as the second best predictive model next to Convolutional Neural Network.
- In assessing the importance of each risk factor to the model and determining the distribution of the factors, the Random Forest is proposed as best in this work.
- With the Logistic regression, we are able to determine the significance of each variable to the model however the accuracy of prediction was the lowest in all three cases.

Concluding Remarks Cont...

- The Logistic regression, Multi-layer Network, and the RBF do not perform better compared with the Convolutional Neural Network and the Random Forest.
- The significant factors discovered, affecting the survival and death of the patients are very key and the proposed models are useful for clinical decision making by physicians and breast cancer patients for accurate prediction of death or survival chances of such patients which will result to a better patients care.

Future Research

- 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.
- Future study should 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.