APPLICATION OF LOGISTIC REGRESSION AND NEURAL NETWORK MODELS IN BREAST CANCER RISK ESTIMATION.

Research Questions:

This project is guided by the following research question:

1. Which factors affect the survival and death rate of all breast cancer patients in the USA?
2. Which factors affects male and female breast cancer patients survival and death rate separately in the USA?
3. Which model is the best fit for estimating breast cancer risk factors in contributing to improve clinical decision making?

To answer these questions, data on breast cancer cases in San Francisco, Connecticut, and New Jersey from the years 1992 to 2016 were collected from CDC website. These states are part of the top 10 states in the USA which experience high rate of cancer contraction. A total of 100,002 observations were collected and the variables collected include mammographic descriptors and demographic risk factors. The dependent variable chosen for this study is the Patients Last Status (Alive or dead) and we first analyze all patients together and next split the data gender (Male and Female data) and analyze both sex separately to get the predictions.

Here we evaluate the model performance and compare the prediction of logistic regression with that of Multi-layer network (Deep learning). This is done using the confusion matrix to obtain the accuracy of prediction, sensitivity analysis, specificity analysis, Positive Predicted Value, Negative Predicted Value, and the Receiver Operating Characteristic (ROC) curve.

DEFINITION OF MODEL EVALUATION APPROACHES USED.

1. **Sensitivity Analysis:** This tells us the probability that the model predicts **positive as death** given that the **patient actually died**. Thus **Prob(+ as death / patient actually died).**

2. **Positive Predicted Value (PPV) Analysis:** This tells us that, given the model **predict positive as death** on patients, what is the probability that they actually died? Thus **Prob(patient actually died / + as death).**

3. **Specificity Analysis:** This tells us the probability that the model predicts negative as Alive given that the patients are alive. Thus **Prob(- as Alive / patient actually Alive).**

4. **Negative Predicted Value (NPV) Analysis:** This tells us that given that the patients are alive, what is the probability that the model predicts negative as death? Thus **Prob(- as death / patient actually Alive).**

5. **The Receiver Operating Characteristic (ROC) curve:** 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 a prediction model discriminates between healthy patients and patients who died. The value of an AUC varies between 0.5 (ie, random guess) and 1.0 (perfect accuracy). The higher the value of the AUC, the better the model.

RESULTS AND KEY FINDINGS.

In answering our first question, we find that patients' Race, Marital status, Age diagnosed, Mass stability, Number of visit to the hospital, Density, Nipple retraction, Lymph node, Amorphous, Milk, Axillary adenopathy, and Lucent all positively affect the patients survival and death rate, while gender, cancer grade, patients length of stay in the hospital, Laterality, family history, Prior-history of breast cancer surgery, size of lamb, Eggshell, Dystrophic and skin lesion all contribute negatively affect the dependent variable at 5% level.

In answering our second question, considering the male gender separately, we find that cancer grade, suture, Density, Nipple retraction, Lymph node, Amorphous, size of lamb, Lucent, and Dermal positively affects male patients survival and death rate while Number of visit to the hospital, length of stay in the hospital, Laterality, family history, Prior-history of breast cancer surgery, Milk, Dystrophic and skin lesion all negatively affect male patients survival and death rate.

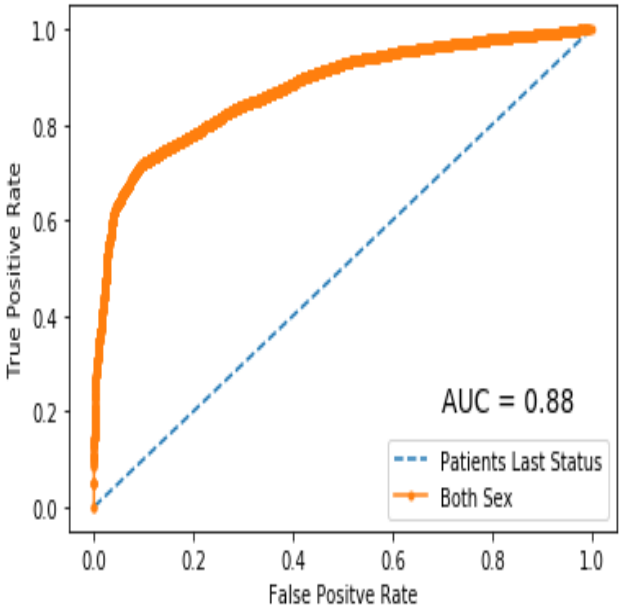
Also, considering the female data separately, we find that female patients' Race, Marital status, Age diagnosed, Mass stability, Number of visit to the hospital, Density, Nipple retraction, Lymph node, Milk, Axillary adenopathy, and Lucent all positively affect the patients survival and death rate, while gender, cancer grade, patients length of stay in the hospital, Laterality, family history, Prior-history of breast cancer surgery, Amorphous, size of lamb, Eggshell, Dystrophic and skin lesion all contribute negatively affect the female patients last status at 5% level.

In answering the third question we compare the prediction results of various models used for the analysis.

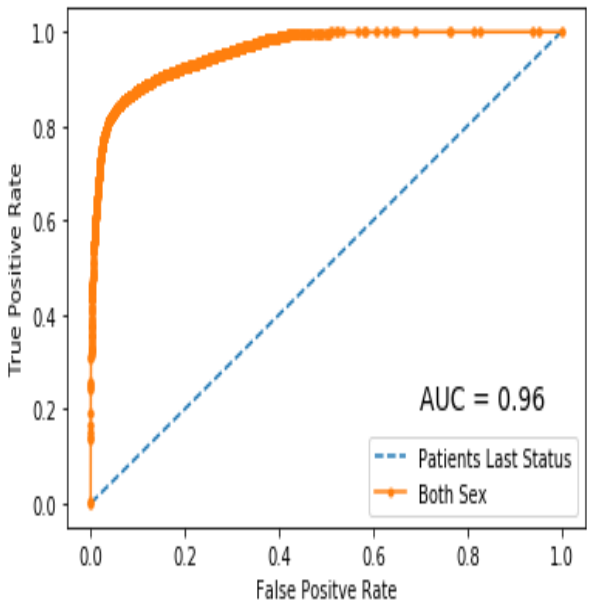
1. COMPARING RESULTS FOR BOTH SEX

EVALUATION METHOD	LOGISTIC REGRESSION	MULTI-LAYER NETWORK	CONVOLUTIONAL NETWORK
Sensitivity Analysis	89.05587%	93.8633%	95.611%
Positive Predicted Value (PPV)	72.140136%	84.71331%	85.8377%
Specificity Analysis	73.43141%	84.01205%	85.2611%
Negative Predicted Value (NPV)	89.67532%	93.54978%	95.41125%
The Receiver Operating Characteristic (ROC) curve, AUC	0.88	0.96	0.90
OVERALL PREDICTION ACCURACY	**80.240988%**	**88.79556%**	**90.26048%**

ROC curve for Logistic regression



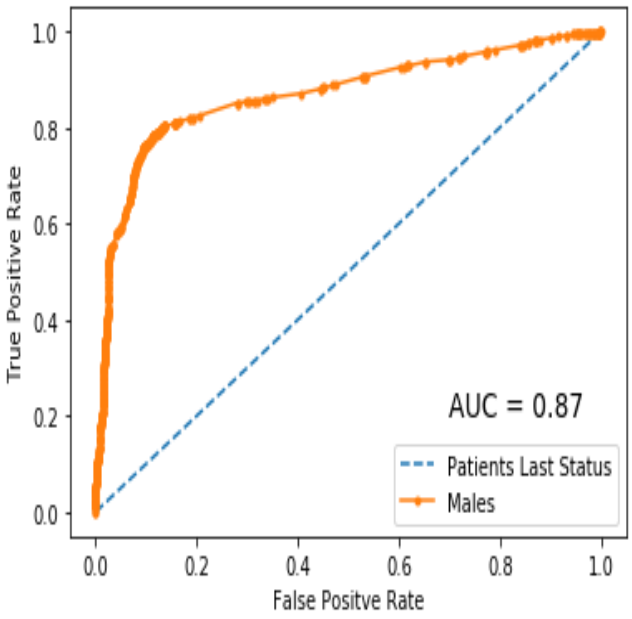
ROC Curve for Neural network



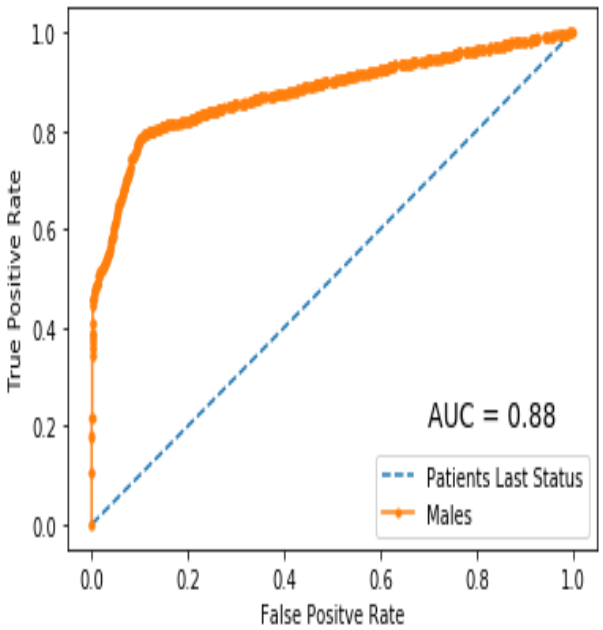
2. COMPARING RESULTS FOR MALE GENDER

EVALUATION METHOD	LOGISTIC REGRESSION	MULTI-LAYER NETWORK	CONVOLUTIONAL NETWORK
Sensitivity Analysis	91.95980%	92.4114%	93.5094%
Positive Predicted Value (PPV)	80.1850%	80.03894%	79.5699%
Specificity Analysis	69.375470%	69.42580%	69.6881%
Negative Predicted Value (NPV)	86.491557%	87.3358%	89.4785%
The Receiver Operating Characteristic (ROC) curve, AUC	0.87	0.88	0.83
OVERALL PREDICTION ACCURACY	**82.33974%**	**82.5320%**	**82.9807%**

ROC curve for logistic regression male gender gender



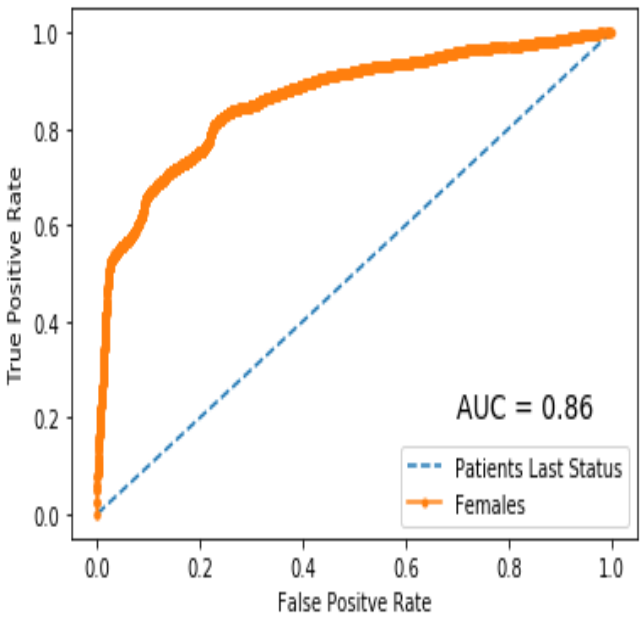
ROC curve for ML Neural network male



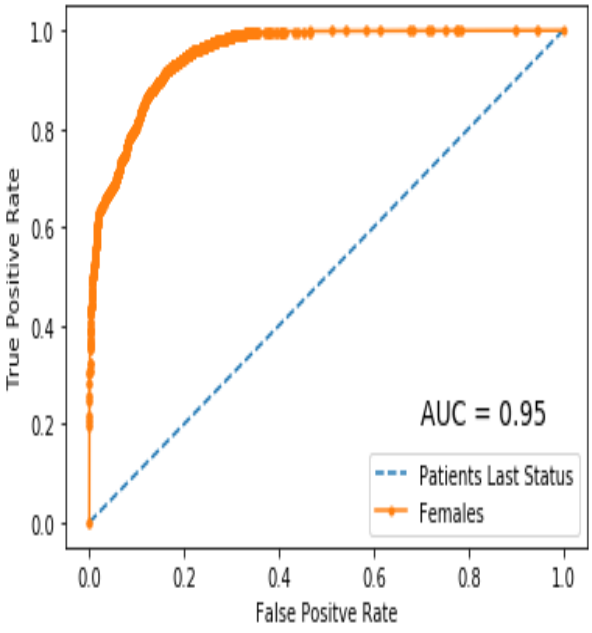
3. COMPARING RESULTS FOR FEMALE GENDER

EVALUATION METHOD	LOGISTIC REGRESSION	MULTI-LAYER NETWORK	CONVOLUTIONAL NETWORK
Sensitivity Analysis	78.90153%	94.3277%	0.95094%
Positive Predicted Value (PPV)	78.746088	75.9300%	89.71179%
Specificity Analysis	77.82077%	79.0941%	89.6595%
Negative Predicted Value (NPV)	77.981095%	95.2259%	95.0685
The Receiver Operating Characteristic (ROC) curve, AUC	0.86	0.95	0.92
OVERALL PREDICTION ACCURACY	**78.37213**	**85.3622%**	**92.30495%**

ROC curve for logistic regression Female gender



ROC curve for ML Neural network Female Gen.



Comparing all results of the predictions for various models we can conclude that the Multi-layer network model is better to predict breast cancer risk estimation.

SUMMARY STATISTICS FOR THE POPULATION:

	PATIENTS STATUS		
GENDER	**ALIVE**	**DEATH**	**TOTAL PATIENTS**
MALE	5,354	10,245	**15,599**
FEMALE	41,331	43,072	**84,403**
TOTAL PATIENTS	**46,685**	**53,317**	**100,002**