Breast Cancer Detection Using Machine Learning

*Project Report of Big Data Mining in Healthcare

1st Akanksha Dewangan
M.Tech, CSE (MT19049)
Indraprastha Institute of Information Technology
New Delhi, India
akanksha19049@iiitd.ac.in

2nd Priyanka Boral
M.Tech, CSE (MT19127)
Indraprastha Institute of Information Technology
New Delhi, India
priyanka19127@iiitd.ac.in

3rd Reecha Kumari Giri
M.Tech, CSE (MT19134)
Indraprastha Institute of Information Technology
New Delhi, India
reecha19134@iiitd.ac.in

4th Shalini Bhardwaj
M.Tech, CSE (MT19045)
Indraprastha Institute of Information Technology
New Delhi, India
shalini19045@iiitd.ac.in

***Abstract*—In the medicinal field, early detection of any disease will make it curable without affecting living organisms. In most cases, people can't detect the disease before it becomes chronic and cause mortality. In women breast cancer is one of the conditions which get cured if it is detected earlier without getting spread all over the body. It is one of the significant reasons for the death of women across the world. It has been seen that practitioners sometimes diagnose it wrongly or misinterpret it due to not having proper technologies. We have classified cells to be malignant or benign using various machine learning models on Wisconsin breast Cancer Dataset from the UCI repository. Some of the major preprocessing steps are removal of class imbalance by SMOTE, outlier detection using ensemble method, and feature selection using a correlation matrix. We have used the 5-fold cross-validation technique in the dataset to select the best model and avoid overfitting. We evaluate all the models based on accuracy, sensitivity, specificity, false discovery rate, false omission rate, and Mattews correlation co-efficient. We have applied data mining models like KNN with k=5 (95% accuracy), SVM(96% accuracy), LightGBM(97% accuracy), Gaussian process with RBF kernel (99% accuracy). Also using neural networks like LSTM (Long short term memory) which is a Recurrent neural network (99% accuracy), sequential Neural Networks (deep learning with Dense layers) which is ANN(Artificial Neural Network) (97%accuracy), MLP(Multi-Layer Perceptron)(97% accuracy). For observing which model will be best then we plot the Receiver Operating Characteristic (ROC) curve between True Positive rate and False Positive Rate then look for AUC (area under the curve) score. Models gave AUC as follows KNN(94.8%), SVM(94.7%), Light GBM(97%), Gaussian process RBF(99.3%), MLP(97%), LSTM (99.3%), Sequential NN such as ANN(97%). By performing all the evaluation we can say that deep learning algorithms like LSTM are best among all because its AUC is 99.3% (Accuracy 99%) as AUC tells which model predicts best and accurate so for us it is LSTM. Next Gaussian RBF is best with accuracy 99% and AUC 99.3%. Among all models, KNN is the worst performer with accuracy 95%and AUC 94.7%. By applying LSTM Neural networks we have increased the AUC to 99.3% and accuracy to be 99% after proper visualization and preprocessing methods. For controlling breast cancer there are two important aspects of early detection and risk reduction. If by using these machine learning techniques we can detect breast cancer in early stages then we can save the lives of many women.**

***Index Terms*—Classification, Neural Network, Accuracy, Models, Outlier detection, Cross-Validation, Smote algorithm, AUC score, ROC Curve.**

I. Introduction

World Health Organization (WHO) reported the breast cancer is the most common cancer amongst women globally [1]. It is also the highest ranked type of cancer cause the death among women in the world .Usually, breast cancer is easily detected if some symptoms appear. However, many women who are suffering from breast cancer have no symptoms. For early detection regular breast cancer screening is very important .Several types of research have been done on early detection of breast cancer to start treatment and increase the chance of survival. Most of the studies concentrated on mammogram images. However, mammogram images sometimes have a risk of false detection that may endanger the patient's health. It is vital to find alternative methods which are easier to implement and can produce a more reliable prediction. Although any lump formed by body cells may be referred to technically as a tumor. Not all tumors are malignant (cancerous). Most breast lumps – 80% of those biopsied – are benign (non-cancerous).Early detection of breast cancer aids for early diagnosis and treatment, because the prognosis is very important for long term survival [5]. Since early detection, diagnosis, and treatment of cancer can reduce the risk of death, it plays a significant role in saving the life of the patient. Any delay in detection of cancer in early stages leads to disease progression and complication of treatment [5], therefore long waiting time prior to diagnosis of breast cancer and starting the treatment process is of high concern.Hence prediction of benign and malignant is worked upon to give close to accurate results.

There are numerous modern techniques have been evolved with the evolution of technology for the prediction of breast cancer. Some of the methods are: Support Vector Machines, Decision Tree, Neural Network, Bayesian networks, k-nearest neighbors, etc. In most of the studies SVM and KNN are applied and observed to be showing great accuracy around 97 or 98%.In studies several distances and different values of the nearest neighbors parameter k, by using different classification rules in the k-nearest neighbors algorithm. The performance is evaluated in term of the classification accuracy rate. However some neural network models are also applied but they are also seen to be showing similar accuracy close to SVM.

Proper data preprocessing steps tend to show good quality result for any machine learning problem. In most of the studies preprocessing is not talked about much and accuracy is the only accuracy majorly talked upon. Hybrid models and ensemble techniques are also rarely experimented in the studies which may show better results.

We have tried to implement all kinds of models recently used these days. We have used ensemble technique based models like Light LGBM,most of the variation of neural networks like Ann, Rnn by properly analyzing the loss functions to obtain maximum results. Metric other than accuracy like AUC (area under the curve) which plot ROC curve between 'True Positive rate' vs 'False Positive Rate' is also observed.Preprocessing steps like feature selection, class imbalance and outlier detection are also performed which has helped us to achieve AUC of 99.3 by LSTM.Cross-validation is also done to avoid overfitting. Detailed methodology and results are discussed in further sections.

II. Material and Methods

A. Dataset

Dataset is taken from Wisconsin Breast Cancer Dataset from UCI repository.Dataset having 568 rows and 33 columns.Information about the attributes which are regarding nucleus are 'id'(ID number), 'diagnosis'(classes), 'radius mean'(mean of distances from center to points on the perimeter), 'texture mean' (standard deviation of gray-scale values) , 'perimeter mean', 'area mean', 'smoothness mean' (local variation in radius lengths) , 'compactness mean' (perimeter**2 / area - 1.0) , 'concavity mean' (severity of concave portions of the contour) , 'concave points mean0' (number of concave portions of the contour) , 'symmetry mean', 'fractal dimension mean'("coastline approximation" - 1) , 'radius se', 'texture se', 'perimeter se', 'area se', 'smoothness se', 'compactness se', 'concavity se', 'concave points se', 'symmetry se', 'fractal dimension se', 'radius worst','texture worst','perimeter worst', 'area worst', 'smoothness worst','compactness worst', 'concavity worst', 'concavepoints worst', 'symmetry worst', 'fractal dimension worst', 'Unnamed: 32'. Dataset having binary class Malignant ('M') is 37.7% and Benign ('B') is 63.7 %.

B. Test-Train Split

We split dataset into 20% test data and 80 % train data with random State=0 and them transform them into standard scalar format.

C. Preprocessing

-The dataset consist of 2 classes Malignant ('M') and Benign ('B').This class is encoded in 1 for 'M' and 0 for 'B' so that dataset will contain only integer,float type of values and put this encoded column name 'class'. While checking for null values 'Unnamed: 32' containing all 'NAN' values so we dropped it. -Transformed dataset into standardscaler format so that all the values of the dataset came across in the same range.

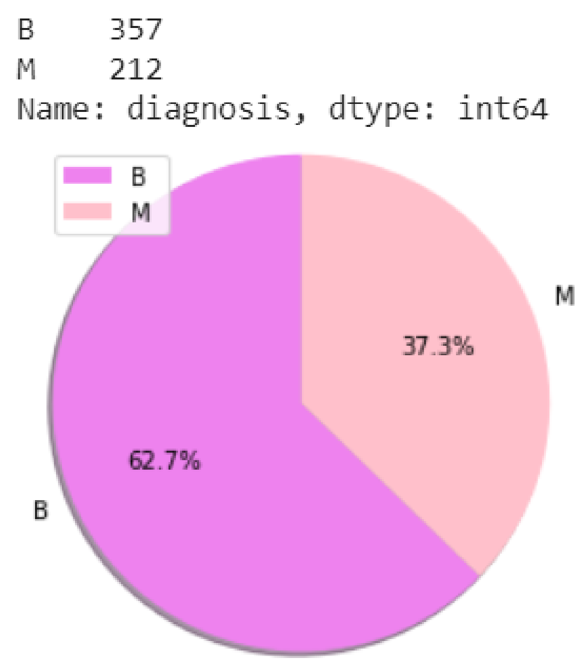

Fig. 1. class percentage distribution in dataset

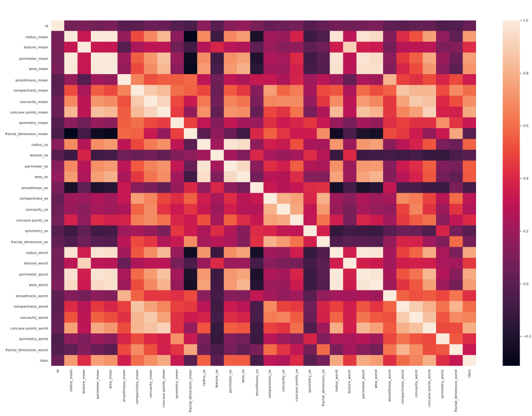

Fig. 2. Correlation Matrix of the dataset

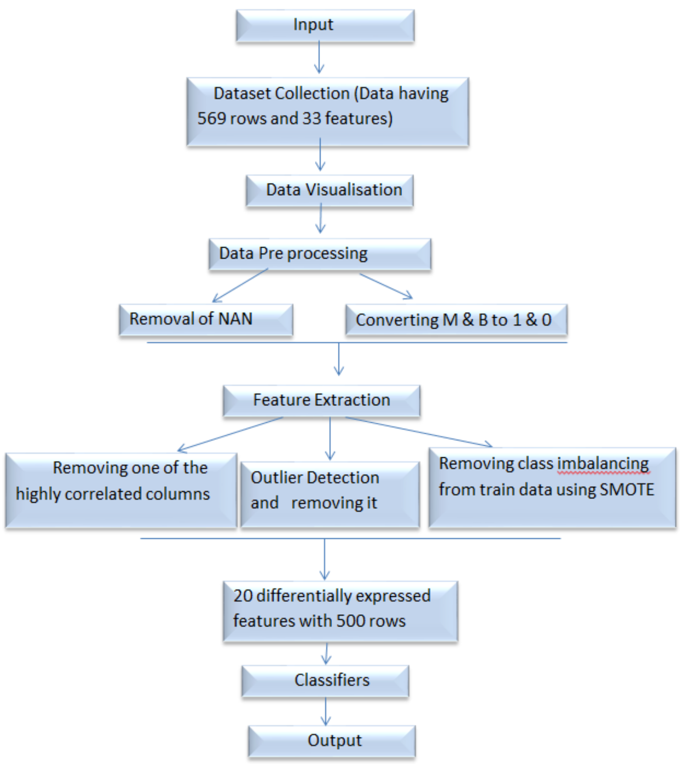

Fig. 3. FLOW CHART FOR METHODOLOGY

D. Methodology

1.Initially data is loaded and unzipped,then read as a data frame which has 569 rows and 33 columns.

2.Visualization for each of the attributes so that we can do feature selection out of already existing attributes. Following visualization are done:
-There is a pie chart in Fig.1 plotted which shows that initially the dataset had 37.3% of class 'M' and 62.7% of class 'B'.
-sns.countplot is plotted for classes so information about the total count of each class is present in the dataset is counted.
-Correlation Matrix is plotted and shown in Fig.2 between all the attributes so that we can remove the correlated columns. If correlation score is 1 corresponding to any two attribute names then they are highly correlated so we can remove any one of the two attribute. After removal of correlated features we have only 20 features left.
-Histogram for every attribute is also plotted.

3. Outliers detection is done by using ensemble method with 'Isolation Forest' model where parameters are- Isolation Forest(n estimators=100,max samples='auto',contamination=float(.12),max features=1.0, bootstrap=False, n jobs=-1, random state=42, verbose=0). In this model it will fit the dataset and predict the anomaly.Here for all rows in dataset starting from 0 to 569, it will return -1 if row is outlier else 1. So in outlier index we contain index of all the outliers which returned as -1 ,and there 69 outliers in data frame so we have to remove them. After outlier removal our dataset have 500 rows and 20 columns.

4. Class balancing
Class balance is done so that all the dataset should not biased toward one class which is maximum.for class balancing SMOTE algorithm is applied which is over sampling algorithm so that it equalize all the classes.In our dataset after preprocessing we left with 500 rows in which one class have count 269 so SMOTE equalise both '0' and '1' classes to 269 by making multiple copies of minimum class. Now dataset having size of 538 rows and 20 columns. Dataset get shuffle after applying class balance because it will make dataset as a sequence in which data set of same class are keep together which will not train well the model.

5. Evaluation parameters functions:
Evaluation of predictions and actual values are done by the help of comprising terms of confusion matrix;
-TP=true positive(correctly identified)
-TN=true negative(incorrectly identified)
-FP=false positive(correctly rejected)
-FN=false negative(incorrectly rejected)
Following are some evaluation parameters-
-Accuracy=(tp+tn)/(tp+tn+fp+fn)
-Sensitivity=tp/(tp+fn)
-Specificity=tn/(tn+fp)
-False Discovery Rate=fp/(fp+tp)
-False Omission Rate=fn/(fn+tn)
-Matthews Correlation coefficient=((tp*tn)-(fp*fn))/((tp+fp)*(tp+fn)*(tn+fp)*(tn+fn))**(0.5)
-ROC (receiver operating characteristic curve) is plotted between true positive rate and false positive rate to see AUC (Area under ROC curve) in percentage to see which model is good for prediction.

6. Cross validation:
We have used the K-fold technique for model selection in our preprocessed dataset. We used 5-fold technique in which we keep 4 chunks in train set and 1 fold for test set and then fit it in model and predict the test label,keep it repeating 5 times and calculate all the evaluation parameters after fitting in 80% train data in model and predict for 20% test data labels.

E. Models

We applied following models:

1) KNN Classifier: As KNN is one of supervised learning we train the model by keeping k values as 3 and 5.It is observed that overall validation accuracy in 5 folds at k=3 is 96.16% and k=5 is 96.16%,but test accuracy for k=3,5 are 95%,95% respectively.So we can say that by predicting by looking at 3 nearest neighbors gives best at both validation and test accuracy hence it is better then k=5.ROC curve

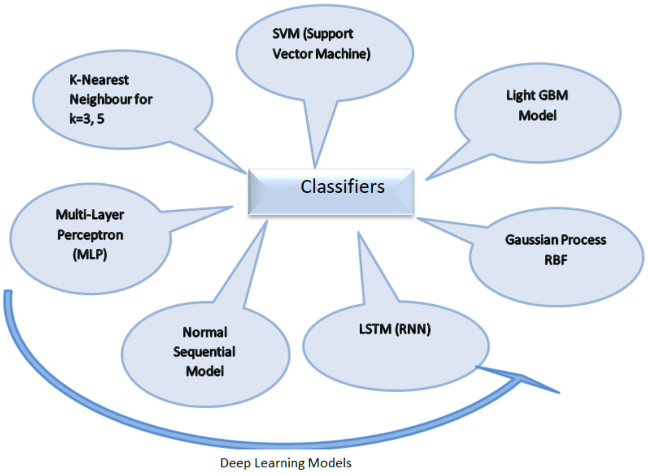

Fig. 4. MODELS USED

is plotted for k=3 ,5 having area under curve (AUC) is 94.7%,94.7% respectively so model with k=3 and 5 both were best.

2) Light GBM Classifier: Light GBM is a gradient boosting framework that uses tree based algorithm. It will grows vertically i.e. leaf wise growth, while other tree based algorithms grow horizontally i.e. level wise. It overall validation accuracy is 97.07% and test accuracy on 20 % test set is 99%.its AUC while plotting ROC is 98.5 %.

3) Gaussian process with RBF: GPR is a non-parametric CLASS that uses the probability distribution to predict.it used Radial basis function (RBF) as a kernel popularly. In our model we used kernel as 1.0 * RBF(1.0). It is giving validation accuracy 97.1% with test accuracy 99%.While plotting ROC it will give AUC 99.3%.

4) SVM Classifier: Support vector machine(SVM) is a type of deep learning algorithm that can be used as supervised learning for classification in a dataset.Here default SVC function is used as a model where kernel is ‘RBF’.It will give over all validation accuracy 96.74% and test accuracy is 96%,ROC curve is also plotted which will give AUC as 94.7%.

5) LSTM (RNN): It is a neural network LSTM (Long Short Term Memory) is know for RNN (Recurrent Neural Network). It setup is like initially sequential() is used as model where 2 layer LSTM is applied with 32,16 neurons in 1st and 2nd layer respectively, activation function ‘relu’ is used then at output layer activation function is used is softmax.Then compile it with ‘sparse categorical cross entropy’ as a loss and optimizer is ’adam’ as it is adaptive according to dataset. Also early stopping is used with patience 10 so that once its accuracy start decreasing it will run for 10 more step and look for best model and save model in mcp save variable.After that train the model for 100 epochs with batch size as 5 and also checking for validation accuracy.It will result in 98.85% as validation accuracy and test accuracy is 99%.There are certain graphs like actual data vs predicted data,accuracy vs epochs and loss vs epochs to observe the rates within each parameters.When its ROC is plotted its AUC is 99.3%.

6) Dense Sequential Model: It is deep neural network know for ANN(Artificial Neural Network).In this model sequential where 2 dense layer is applied with 32,16 neurons respectively along with relu as an activation function,and one output layer with activation function softmax. other then layer everything is same as LSTM(RNN) as explained before.But while fitting the model it will apply 120 epochs with batch size 5.It’s validation accuracy is 98.28% and test accuracy is 97%.There are certain graphs like actual data vs predicted data,accuracy vs epochs and loss vs epochs to observe the rates within each parameters.When its ROC is plotted its AUC is 97%.

7) MLP: It is multi layer perception and belongs to the class of feed forward artificial neural network and use back propagation for training will consider it for supervised learning technique. It is a good learner can learn fastly even in small dataset. In our model we used the activation function relu ,batch size 50,learning rate is adaptive,solver is ”adam” and run it till maximum iteration of 500 because it will not converge in small iteration and give warning in sklearn library of python.It gives validation accuracy 99.9%,test accuracy is 97%.When we draw the ROC curve it will resulted AUC as 97%.

III. RESULTS

The result of Fig.5 Table shows that Gaussian process classifier and LSTM(RNN) perform best over all other models applied in the dataset if judge on the basis of accuracy, sensitivity, specificity for them is 99%,100% and 98.5% respectively for both models.
The parameter Mathews correlation coefficient value indicate purity of classifier. Higher the value or equal to 1 then it will be pure binary classifier,and in our dataset purity is highest in GPC(97%) and LSTM (RNN)(97.78%) are best models.Hence purity in decreasing order is: GPC = LSTM, LGBM = MLP = Dense sequential (ANN), SVM, KNN. Among all KNN worst in while looking for purity.
False omission rate(FOR) are 0% in GPC and LSTM. For measuring the false negatives which are incorrectly rejected while predicting the classes.Among all SVM FDR is 4.34% which is highest among all hence not good in terms of FOR. Now if we talk about False discovery rate(FDR) which return total false discoveries among total experiments which should be less.And in our observation FDR is 2.94% in both GP classifier and LSTM which is least hence best but if we look for other models SVM (3.2%) better then KNN(8.8%),MLP(5.8%),Dense sequential(5.88%),Hence we can say conclude that KNN is worst and return lots of false

discovery results.
ROC curve for all the models between their true positive rate and false negative rate hence we got the area under curve(AUC) whose percentage if greatest then model will predict accurate.In our observations GP classifier and LSTM having AUC of 99.3% which is best among all, LGBM, MLP and Dense sequential having same AUC (97%) and other models like KNN,SVM having AUC 94.7% which is worst among all models.
Fig. 6 shows ROC curve of KNN for K=5
Fig. 7 shows ROC curve of SVM classifier
Fig. 8 shows ROC curve of Gaussian Process classifier with RBF
Fig. 9 shows ROC curve of LGBM
Fig. 10 shows ROC curve of LSTM which is RNN
Fig. 11 shows ROC curve of MLP classifier
Fig. 12 shows ROC curve of Dense Neural Network which is ANN
Fig. 13 shows which is accuracy graph for Gaussian process classifier with RBF
Fig. 14 shows accuracy for LSTM which is RNN model

Model	Accuracy	sensitivity	specificity	False discovery rate	False omission rate	Mattews correlation coefficient	AUC
KNN	95%	93.93%	95.5%	8.8%	3.03%	88.8%	94.7%
SVM	96%	90.9%	98.5%	3.2%	4.34%	90.9%	94.7%
LGBM	97%	96.9%	97%	5.8%	1.515%	93.29%	97%
GP(RBF)	99%	100%	98.5%	2.94%	0%	97.7%	99.3%
MLP	97%	96.96%	97.01%	5.88%	1.515%	93.29%	97%
LSTM(RNN)	99%	100%	98.5%	2.94%	0%	97.78%	99.3%
Dense Sequential (ANN)	97%	96.96%	97.01%	5.88%	1.515%	93.29%	97%

Fig. 5. Meausre at different parameters

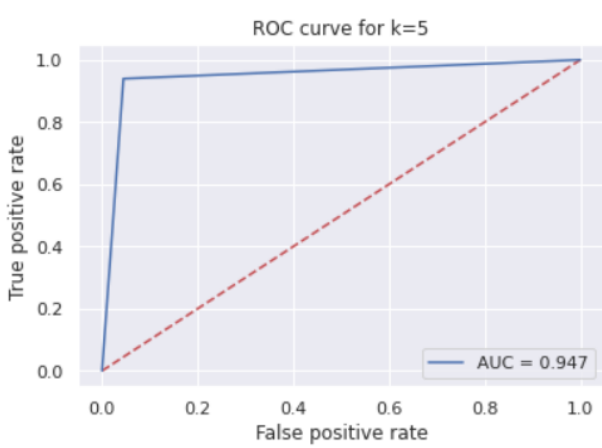

Fig. 6. ROC curve of knn at k=5

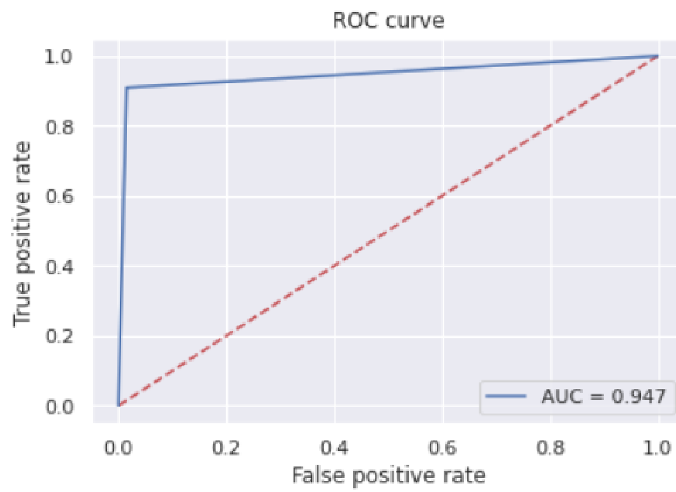

Fig. 7. ROC curve of SVM classifier

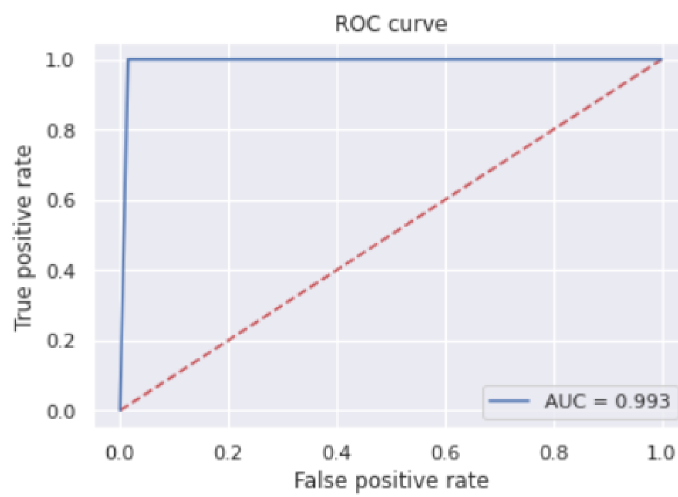

Fig. 8. ROC curve of GP classifier with RBF

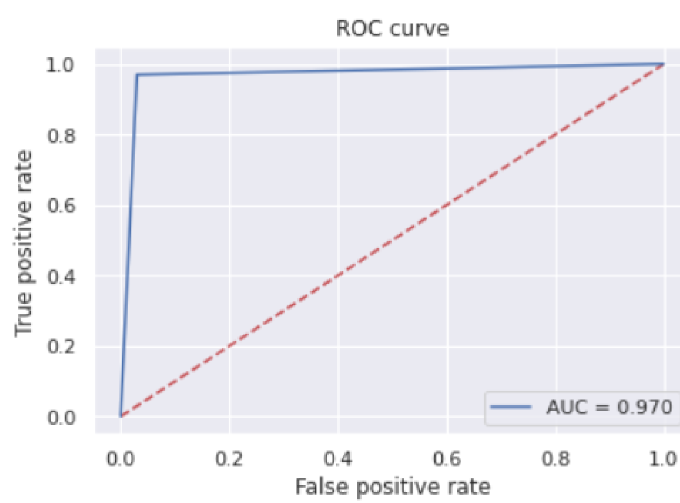

Fig. 9. ROC curve of Light GBM classifier

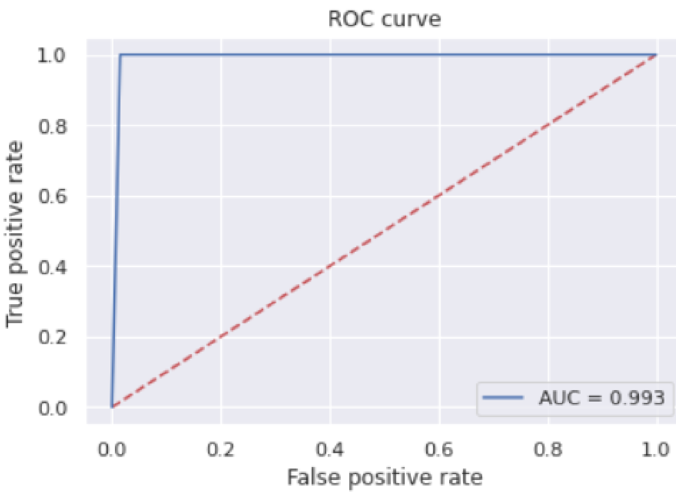

Fig. 10. ROC curve of LSTM model

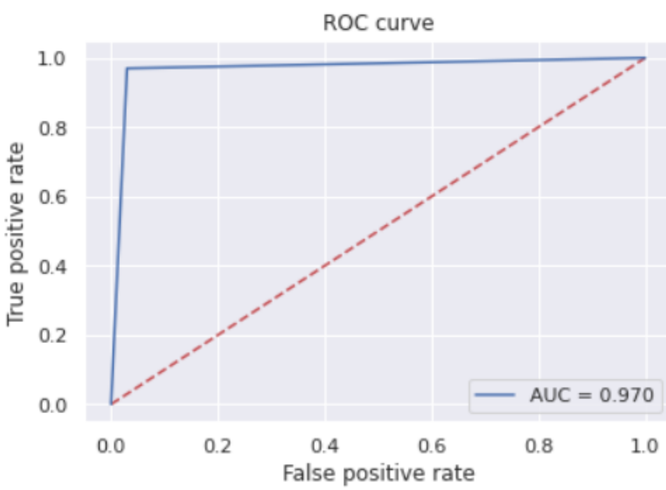

Fig. 11. ROC curve of MLP Classifier

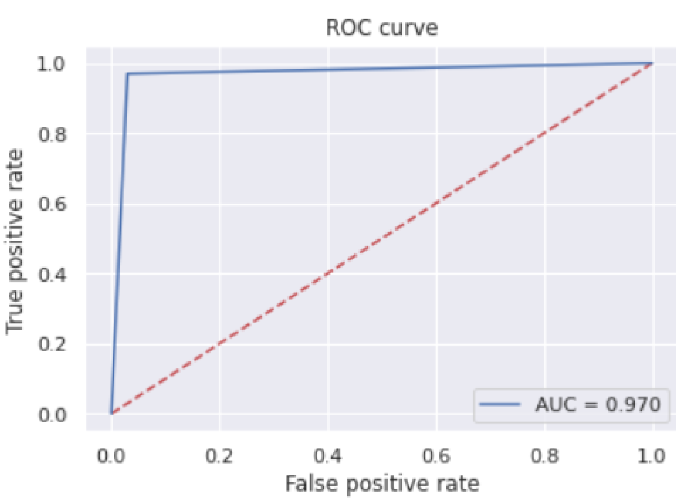

Fig. 12. ROC curve of Dense Neural Networks

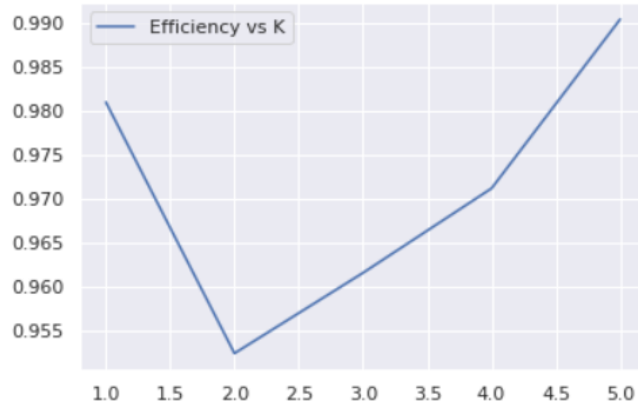

Fig. 13. Accuracy graph for Gaussian Process classifier for 5-folds

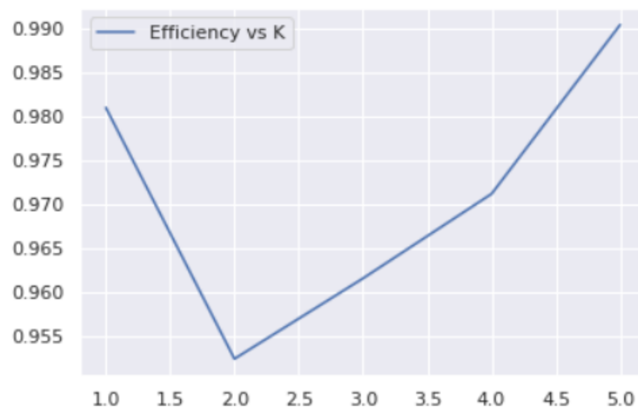

Fig. 14. Accuracy graph for LSTM (RNN) for 5-folds

IV. DISCUSSION

Breast cancer prediction is major area in Medical field. It is one of the chronic and dangerous diseases which cause major reason of death among all the women. So if we detect it earlier by using various supervised classifiers and neural networks models results in save precious life's of women worldwide.In this paper we applied various models and among them Gaussian process classifier with RBF and LSTM (long short term memory) which is RNN (Recurrent Neural Network) is best among all by observing all the parameters like accuracy,sensitivity,specificity,False Discovery rate,False Omission rate,Matthew's correlation coefficient.And when ROC graph for all it will results in best area under curve in graph which meant for best prediction model.
Supervised classifiers like K-nearest neighbor classifier results worst among all the solution with least accuracy 95% compare to others and majorly due to having 8.8% highest False discovery rate which is incorrectly reject null hypothesis also gave among all. SVM classifier should not used because it is giving highest false omission rate(4.34%) which is not good

in terms of false negative rate it should be less and having accuracy (96%) with AUC (94.5%). LGBM boosting classifier and Dense neural network both were giving good accuracy of 97% both but should not used because of having more false discovery rate of 5.8%.Hence we should focus as a best solution to both Gaussian process classifier and LSTM (Recurrence neural network) which is best not just because highest accuracy and AUC, but having false omission rate is almost negligible (0%) in our dataset.

Our proposed solution with statistical Gaussian process classifier with RBF and deep learning neural networks LSTM which is Recurrent neural network will be very much helpful in early detection of cancer with giving best accuracy 99%,99% respectively with negligible False omission rate 0% for both hence helpful in medical field to save life of women.

V. Contribution of Each Author

Akanksha Dewangan (MT19049) : Removing class imbalance, light GBM, Gaussian process(varying kernel), analysis implementation.

Priyanka Boral (MT19127) : KNN model(with varying k), Dense neural network(ANN), correlation feature selection implementation.

Reecha Kumari Giri (MT19134) : Visualization of dataset, LSTM (RNN), MLP neural network implementation.

Shalini Bharadwaj (MT19045) : Preprocessing of dataset, outlier detection, SVM model, Gaussian process (varying kernel) implementation.

References

[1] https://archive.ics.uci.edu/ml/datasets/ Breast+Cancer+Wisconsin+(Diagnostic)

[2] M. M. Islam, H. Iqbal, M. R. Haque and M. K. Hasan, "Prediction of breast cancer using support vector machine and K-Nearest neighbors," 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), Dhaka, 2017, pp. 226-229, doi: 10.1109/R10-HTC.2017.8288944.

[3] S. Karthik, R. Srinivasa Perumal and P. V. S. S. R. Chandra Mouli,"Breast Cancer Classification Using Deep Neural Networks"· February 2018,Research Gate.

[4] Tahmooresi, M., Afshar, A., Rad, B.B., Nowshath, K.B. and Bamiah, M.A., 2018. Early Detection of Breast Cancer Using Machine Learning Techniques. Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 10(3-2), pp.21-27.

[5] Medjahed, S.A., Saadi, T.A. and Benyettou, A., 2013. Breast cancer diagnosis by using k-nearest neighbor with different distances and classification rules. International Journal of Computer Applications, 62(1).

[6] Padmavati, J., 2011. A comparative study on breast cancer prediction using RBF and MLP. International Journal of Scientific Engineering Research, 2(1), pp.1-5.

[7] S. Sharma, A. Aggarwal and T. Choudhury, "Breast Cancer Detection Using Machine Learning Algorithms," 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), Belgaum, India, 2018, pp. 114-118, doi: 10.1109/CTEMS.2018.8769187.

[8] M. R. Al-Hadidi, A. Alarabeyyat and M. Alhanahnah, "Breast Cancer Detection Using K-Nearest Neighbor Machine Learning Algorithm," 2016 9th International Conference on Developments in eSystems Engineering (DeSE), Liverpool, 2016, pp. 35-39, doi: 10.1109/DeSE.2016.8.

[9] M. Gupta and B. Gupta, "A Comparative Study of Breast Cancer Diagnosis Using Supervised Machine Learning Techniques," 2018 Second International Conference on Computing Methodologies and Communication (ICCMC), Erode, 2018, pp. 997-1002, doi: 10.1109/ICCMC.2018.8487537.

[10] S. Lee, M. Amgad, M. Masoud, R. Subramanian, D. Gutman and L. Cooper, "An Ensemble-based Active Learning for Breast Cancer Classification," 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), San Diego, CA, USA, 2019, pp. 2549-2553, doi: 10.1109/BIBM47256.2019.8983317.

[11] Y. Khourdifi and M. Bahaj, "Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification," 2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS), Kenitra, 2018, pp. 1-5, doi: 10.1109/ICECOCS.2018.8610632.

[12] F. Y. A'la, A. E. Permanasari and N. A. Setiawan, "A Comparative Analysis of Tree-based Machine Learning Algorithms for Breast Cancer Detection," 2019 12th International Conference on Information Communication Technology and System (ICTS), Surabaya, Indonesia, 2019, pp. 55-59, doi: 10.1109/ICTS.2019.8850975.

[13] H. Jouni, M. Issa, A. Harb, G. Jacquemod and Y. Leduc, "Neural Network architecture for breast cancer detection and classification," 2016 IEEE International Multidisciplinary Conference on Engineering Technology (IMCET), Beirut, 2016, pp. 37-41, doi: 10.1109/IMCET.2016.7777423.

[14] D. Selvathi and A. AarthyPoornila, "Performance analysis of various classifiers on deep learning network for breast cancer detection," 2017 International Conference on Signal Processing and Communication (ICSPC), Coimbatore, 2017, pp. 359-363, doi: 10.1109/CSPC.2017.8305869.

[15] S. Murugan, B. M. Kumar and S. Amudha, "Classification and Prediction of Breast Cancer using Linear Regression, Decision Tree and Random Forest," 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC), Mysore, 2017, pp. 763-766, doi: 10.1109/CTCEEC.2017.8455058.

[16] T. I. Rohan, Awan-Ur-Rahman, A. B. Siddik, M. Islam and M. S. U. Yusuf, "A Precise Breast Cancer Detection Approach Using Ensemble of Random Forest with AdaBoost," 2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2), Rajshahi, Bangladesh, 2019, pp. 1-4, doi: 10.1109/IC4ME247184.2019.9036697.

[17] O. V. Singh and P. Choudhary, "A Study on Convolution Neural Network for Breast Cancer Detection," 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP), Gangtok, India, 2019, pp. 1-7, doi: 10.1109/ICACCP.2019.8882993.

[18] J. Zheng, D. Lin, Z. Gao, S. Wang, M. He and J. Fan, "Deep Learning assisted Efficient AdaBoost Algorithm for Breast Cancer Detection and Early Diagnosis," in IEEE Access, doi: 10.1109/ACCESS.2020.2993536.

[19] T. Mu and A. K. Nandi, "Detection of breast cancer using v-SVM and RBF networks with self organized selection of centers," The 3rd IEE International Seminar on Medical Applications of Signal Processing 2005 (Ref. No. 2005-1119), London, UK, 2005, pp. 47-52, doi: 10.1049/ic:20050330.

[20] R. D'mello, V. Pardeshi, H. Mehta and S. Dhage, "Comparative Study of Breast Cancer Detection Techniques," 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), Bombay, India, 2019, pp. 1-6, doi: 10.1109/I2CT45611.2019.9033894.

[21] J. Dheeba and S. Tamil Selvi, "Classification of malignant and benign microcalcification using SVM classifier," 2011 International Conference on Emerging Trends in Electrical and Computer Technology, Nagercoil, 2011, pp. 686-690, doi: 10.1109/ICETECT.2011.5760205.

[22] S. Sayed, S. Ahmed and R. Poonia, "Holo entropy enabled decision tree classifier for breast cancer diagnosis using wisconsin (prognostic) data set," 2017 7th International Conference on Communication Systems and Network Technologies (CSNT), Nagpur, 2017, pp. 172-176, doi: 10.1109/CSNT.2017.8418532.

[23] N. Khuriwal and N. Mishra, "Breast cancer diagnosis using adaptive voting ensemble machine learning algorithm," 2018 IEEMA Engineer Infinite Conference (eTechNxT), New Delhi, 2018, pp. 1-5, doi: 10.1109/ETECHNXT.2018.8385355.

[24] R. Radha and P. Rajendiran, "Using K-Means Clustering Technique to Study of Breast Cancer," 2014 World Congress on Computing and Communication Technologies, Trichirappalli, 2014, pp. 211-214, doi: 10.1109/WCCCT.2014.64.

[25] X. Yang, H. Peng and M. Shi, "SVM with multiple kernels based on manifold learning for Breast Cancer diagnosis," 2013 IEEE International Conference on Information and Automation (ICIA), Yinchuan, 2013, pp. 396-399, doi: 10.1109/ICInfA.2013.6720330.