International Journal of Industrial Engineering & Production Research September 2020 Vol. 31, No. 3: 379-386
DOI: 10.22068/ijiepr.31.3.379

RESEARCH PAPER

Breast Cancer Disease Prediction With Recurrent Neural Networks (RNN)

Sangapu Venkata Appaji*[1], R Shiva Shankar[2], K.V.S. Murthy[3] & Chinta Someswara Rao[4]

Received 3 May 2020; Revised 13 July 2020; Accepted 24 August 2020; Published online 30 September 2020

ABSTRACT

Cancer is a collaborative amalgamation of diseases that involves abnormal increase in cell growth with the potential of occupying and attacking the entire body. According to studies, breast cancer most likely occurs in women and it has become the second biggest cause of female death. Due to its widespread penetration and significance, many researchers have analyzed the phenomenon and further studies are still required to reach an optimum outcome. This study applies deep learning technique in conjunction with Recurrent Neural Networks (RNN) to predict the formation of breast cancer disease so that doctors will perform the diagnosis more properly. To assess the efficiency of the proposed method, breast cancer data belonging to UC Irvine repository were used. Precision, recall, accuracy, and f1 score of the proposed method showed good scores and the proposed technique performed well.

KEYWORDS: *Cancer; Breast Cancer; Deep learning; RNN.*

1. Introduction

Breast cancer disease is a malignant growth in breast fleshy tissue [1]. Breast malignancy can be identified by abnormality in breasts, an adjustment fit as a fiddle, dimples on the skin, areola-producing liquid, changes in the areola in recent past, or a red or covered fix on the skin [2]. Among dangerous causes of breast cancer progression in females are obesity, lack of physical exercise, etc. [2, 3].

Results of breast cancer growth are contingent on malignancy type, degree of sickness, and individual's age [4]. Survivability after breast cancer could extend to 5 years among 80-90% of such women in the US and the UK [6,7].

The growth of breast cancer has joined the ranks of top diseases among females in India with a mean age of 25.8 per 100,000 ladies and mortality rate of 12.7 per 100,000 ladies [8].

Information on various recent cases of nation-wide cancer related to breast growth libraries was scrutinized for infection and death rates. The mean age for the frequency rate of cancer because of breast carcinoma was determined at 41 for every 100,000 ladies in Delhi, followed by Chennai (37.9), Bangalore (34.4), and Thiruvananthapuram District (33.7).

It is anticipated that the numerical values of cases for India in 2020 and onwards continue to rise. Wellbeing, accessibility of growing broadcast scientific projects, and treatment offices/clinics may lead to positive clinical outcome in India.

It can be concluded that, with reference to the shortages of statistical data and previous findings, it is quite necessary to dedicate much more time and effort to this area and detect the disease at early stages, which may be the key to preventing further damages.

2. Literature Survey

In the following, some studies conducted in this domain are discussed.

Lavanya and Rani [9] described the ensembles of the decision tree by capturing previous results of the single decision tree model. Of note, the performance of ensembles of classifiers in predicting the formation of breast cancer has not frequently been explored.

* **Corresponding author:** ***Sangapu Venkata Appaji***
venkataappaji.sangapu@gmail.com

1. *Department of CSE, KKR & KSR Institute of Technology and Sciences, Guntur, A.P, India.*
2. *Department of CSE, S.R.K.R Engineering College, Bhimavaram, W.G. District, A.P. India.*
3. *Department of CSE, S.R.K.R Engineering College, Bhimavaram, W.G. District, A.P, India.*
4. *Department of CSE, S.R.K.R Engineering College, Bhimavaram, W.G. District, A.P.India.*

Kashish Goyal et al. [10] proposed a method for identifying the condition of cancer in patients, either in the benign or virulent state. The required data were derived from WISCONSIN dataset in UC Irvine machine-learning warehouse. WISCONSIN dataset comprises the related data of patients at risk of cancer development or those with recurrent cancer.

Cirkovic BR et al. [11] proposed practical data mining methods for dealing with the information of patients diagnosed with breast cancer in order to estimate the survival rate and disease deterioration. Authors made a comparison between popular machine learning models and concluded that classifiers would help doctors deal with the survivability and recurrence associated with breast cancer.

Kate RJ and Nadig R. [12] investigated the prediction of survivability of patients with breast cancer by applying machine learning techniques. Authors also evaluated models individually and in combination.

Authors [13] collected the necessary data from 100 individuals and the data comprised a combination of cancer and non-cancer cases. Then, they applied the K-means method to categorize the data into relevant and non-relevant. Authors used WEKA to measure the risk factors and raking. Finally, the breast cancer risk level was predicted using Lotus Notes.

Ahmad LG et al. [14] studied the application of data mining techniques and then, developed breast-cancer predictive models for recurrence. They carried out a number of experiments on a group of patients by closely reviewing their status for two years.

Abdelghani Bellaachia and Erhan Guven [15] presented a prediction method for survivability of patients with breast cancer by applying data mining approaches. In this work, authors employed data from SEER Public-Use Data.

Chaurasia et al. [16] presented a model for identifying breast cancer using RepTree, RBF Network, and Simple Logistic. In this study, Simple Logistic was applied to measurement of feature reduction and RepTree and RBF were employed for breast cancer prediction.

3. Methodology

Dataset: In this section, the dataset used for breast cancer was downloaded from UC Irvine repository consisting of 561 instances and 31 attributes, out of which 30 attributes are considered as input attributes and the 1st attribute is considered the target class.

Pre-processing*:* Pre-processing is a significant stage in data mining and is particularly applicable to specific methods for data mining and machine learning. Data collection procedures are often freely controlled which may lead to creation of out-of-range values, wrong combinations of data, null values, missing values, etc. However, the standardization of the data is required for machine learning models. Hence, the data need to be preprocessed.

Model design: Model design comprises of model training and testing.

Model Training: The procedure for training the machine learning model includes learning information from the training data. The training data must contain a correct answer known as a target attribute.

Model testing: A testing dataset is required for simulation model for testing purpose.

Performance measures: In this phase, benchmark model and some other models are evaluated with performance measures including precision, recall, accuracy, and f1 scores.

3.1. Recurrent neural network (RNN)

This study used Recurrent Neural Network (RNN) [17,18,19, 20, 21] as a particular category of artificial neural network in which links among the nodes generate a directed graph along with temporal arrangement.

RNNs were utilized for deep learning and improving the related techniques, which would imitate the movement of neurons in the human brain. RNN nodes are more dominant than other models for predicting the outcomes since these models used back propagation. The architecture of our method is shown in Fig. 2. This method with one input layer consisting of 30 input nodes, three hidden layers consisting of 64, 128, and 256 nodes, and one output layer consisting of one node 0 or 1. ReLU activation function is used in the hidden layer and dropout of about 0.25 is used.

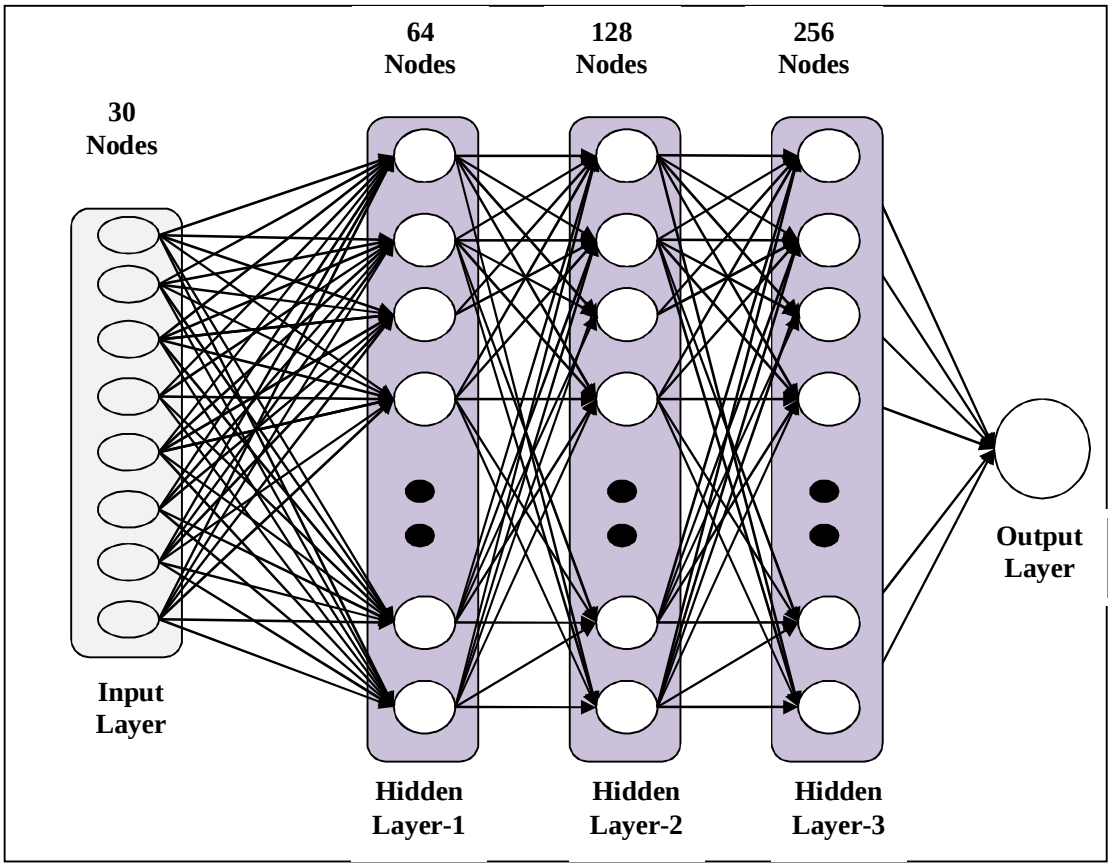

Fig. 2. RNN architecture

4. Dataset

In this research, the dataset used for breast cancer was downloaded from UC Irvine repository. This dataset comprise 569 instances and 31 attributes, among which the first attribute diagnosis is chosen as the target class variable and 2 to 31 attributes are considered as input variables.

The dataset includes data from 569 instances with 31 characteristics: "diagnosis", "radius_mean", "texture_mean", "perimeter_mean", "area_mean", "smoothness_mean", "compactness_mean", "concavity_mean", "concave points_mean", "symmetry_mean", "fractal_dimension_mean", "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", "concave points_worst", "symmetry_worst", and "fractal_dimension_worst".

Fig. 3 shows the histogram of 2 to 31 features and the significance of every feature. Fig. 4 shows the Bar plot of Attribute 1; Fig. 3 lists Class 0 with 357 instances and Class 1 with 212 instances. Fig. 5 shows the diagonal correlation matrix of 2 to 31 input attributes.

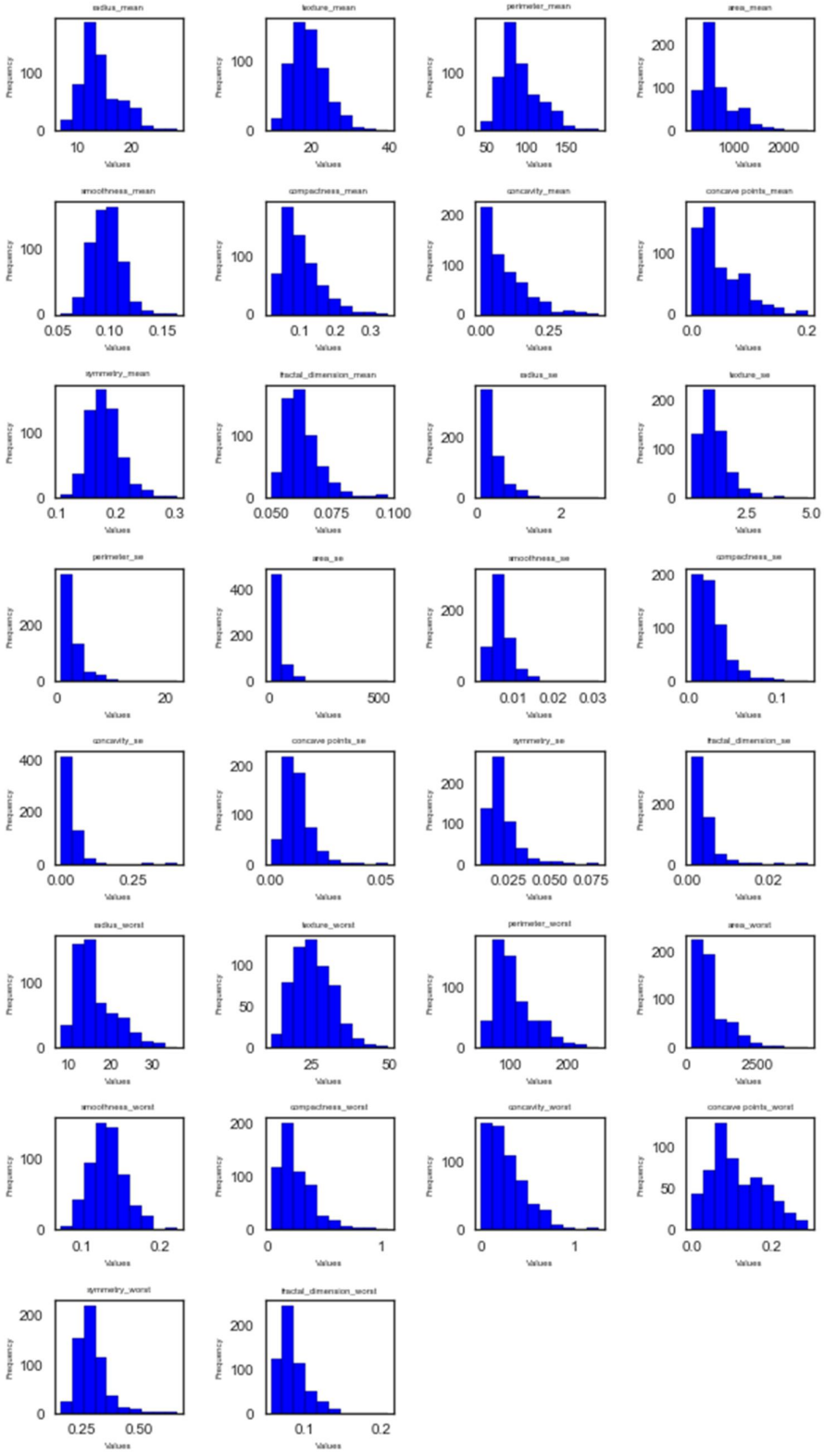

Fig. 3. Histogram of 2 to 31 attributes

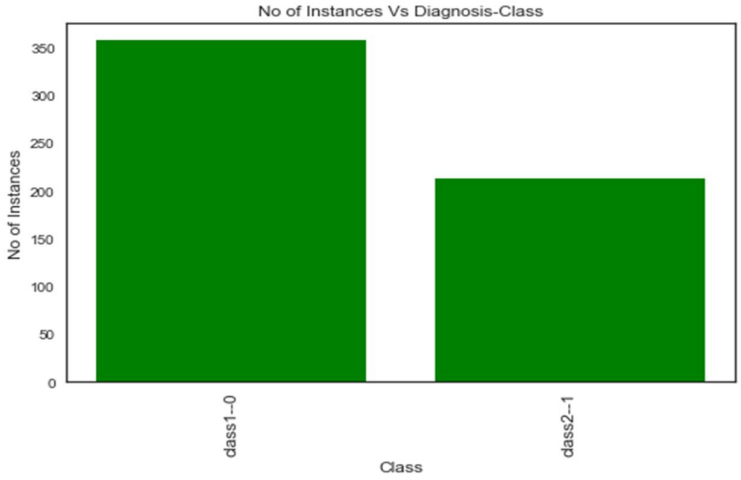

Fig. 4. Bar plot diagnosis attribute

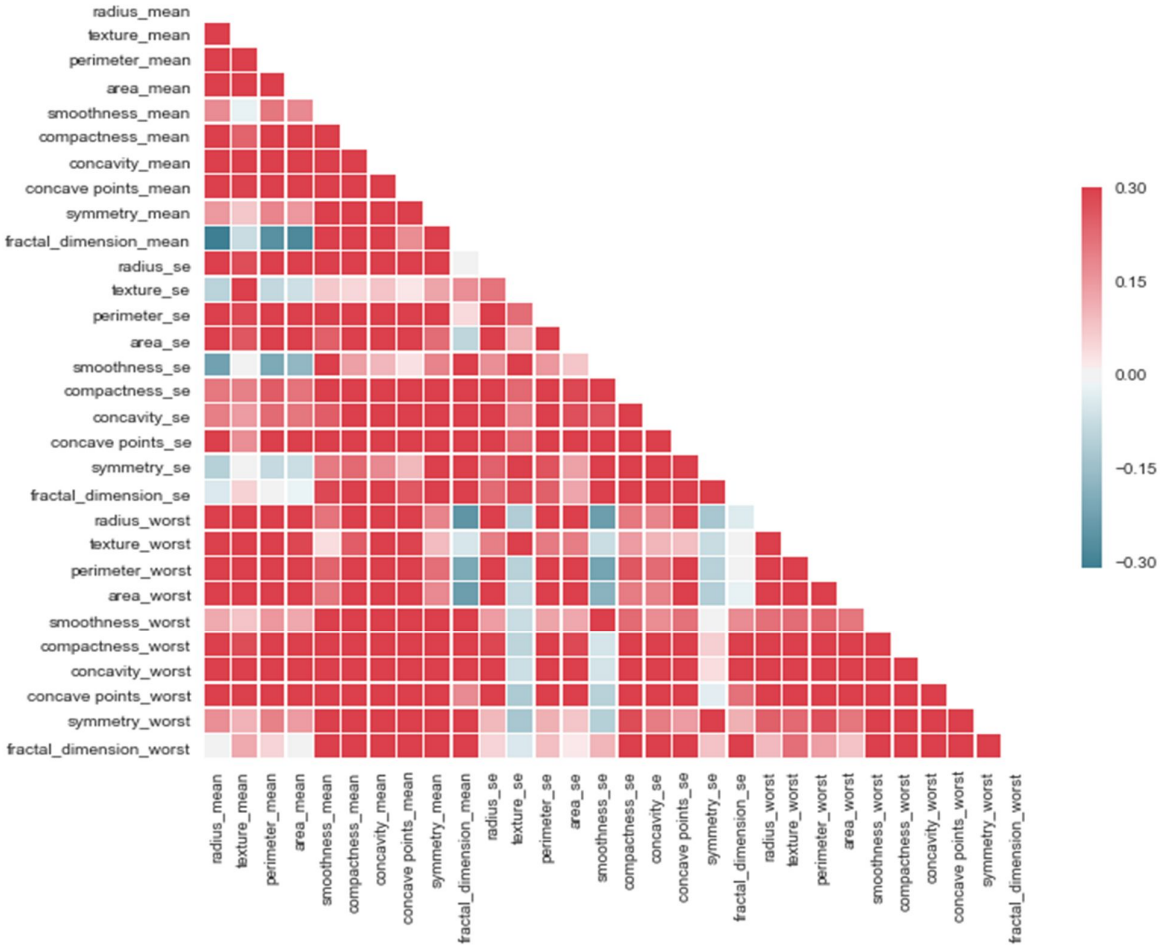

Fig. 5. Diagonal correlation matrix of 2 to 31 input attributes

5. Results

This section discusses the outcomes of the proposed method.

5.1. Evolution metrics

Accuracy: It is calculated as "the total number of two right predictions, True Positive (TP) + True Negative (TN), divided by the overall number of a dataset Positive (P) + Negative (N)".

$Accuracy=TP+TN/P+N$

Precision: It is calculated as "the count of right positive predictions (TP) divided by the total count of positive predictions (TP + FP)". Precision is likely called a positive predictive number.

$Precision=TP/TP+FP$

Recall: It is calculated as "the count of right positive predictions (TP) divided by the total count of positives (P)". Recall is likely called the the true positive rate or sensitivity.

$Recall=TP/P$

Confusion matrix: It shows the number of correct and incorrect predictions with count values and is broken down by each class. The matrix provides us with a better understanding of mistakes and errors in classifiers and more importantly, of the categories of mistakes being made.

Testing dataset's Confusion matrix for RNN is shown in Figure 6, and precision, recall, f1-score, and support of RNN are shown in Table 1. By considering the data of precision-recall shown in Table 1, a graph is drawn in Figure 7. ROC curve is shown in Figure 8 and ROC curve with a threshold is shown in Figure 9.

Tab. 1. Evolution metrics of RNN for diabetes dataset

	precision	Recall	f1-score	Support
0	0.98	0.97	0.98	66
1	0.96	0.98	0.97	48
avg/total	0.97	0.97	0.97	114

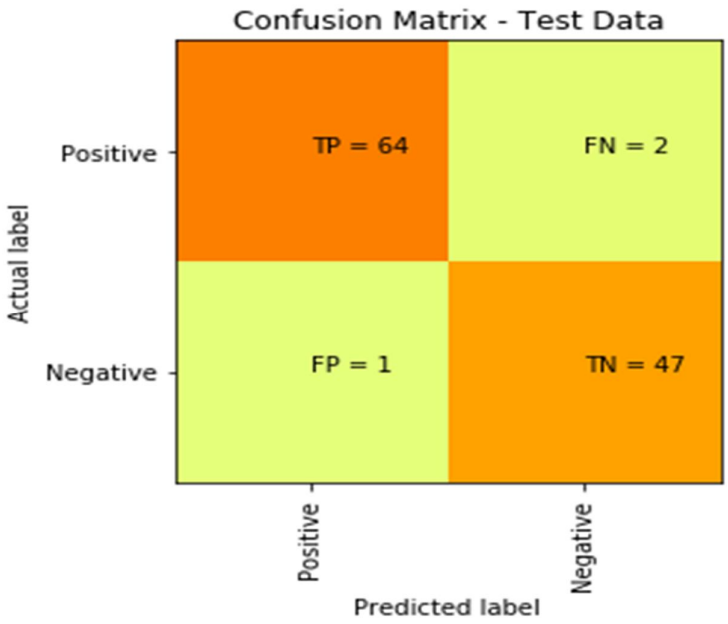

Fig. 6. Confusion matrix

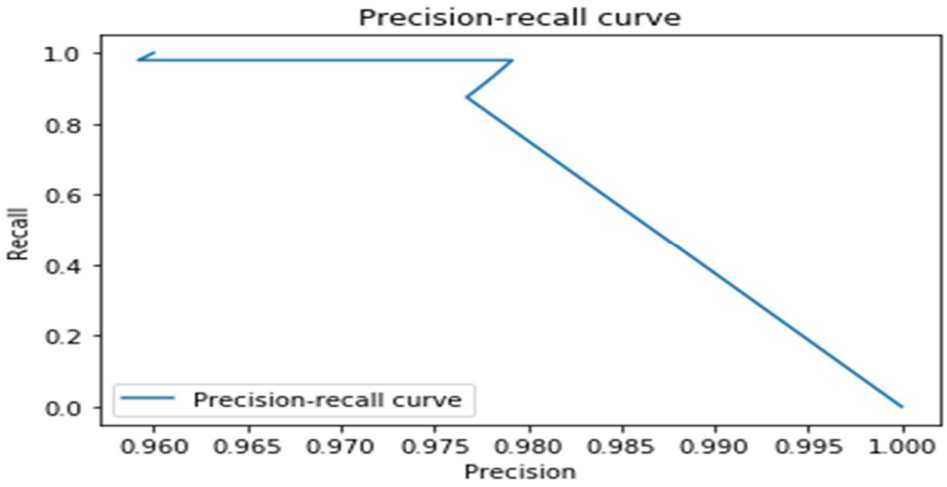

Fig. 7. Precision-recall plot

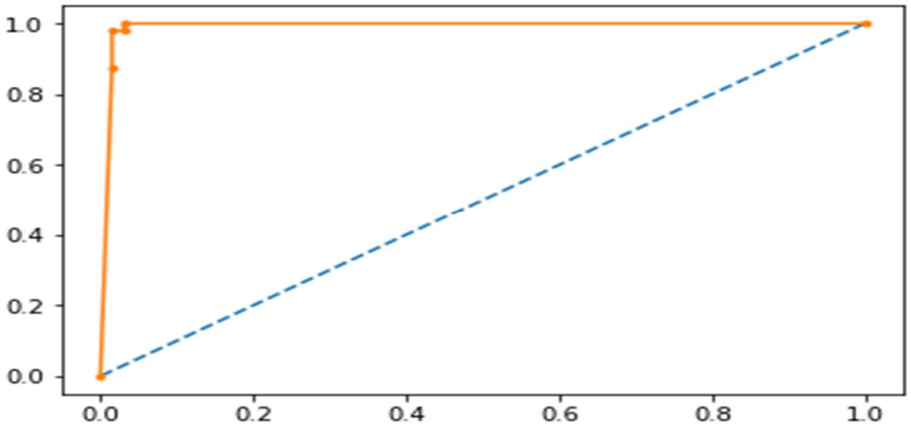

Fig. 8. ROC plot

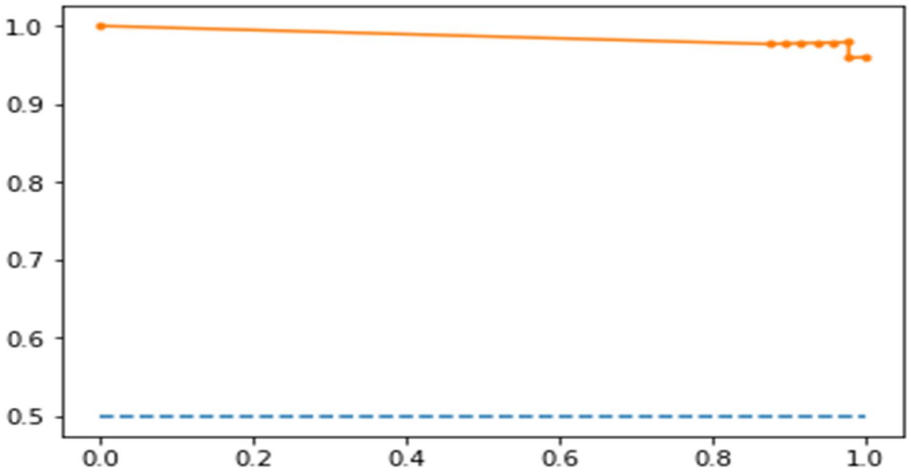

Fig. 9. ROC plot with a threshold

6. Conclusions

Recent statistics concerning cancer illustrated that breast cancer growth had already joined the ranks of top diseases among females in India with a mean age of 25.8 per 100,000 ladies and death rate of 12.7 per 100,000 ladies. Given the significance of the cancer, this study derived the required breast cancer dataset from UCI repository with 569 instances out of which 357 instances with Class 0 and 212 instances with Class 1. This study discussed the diagnosis of breast cancer disease by using deep learning methods called RNN, which will help doctors handle the issue and prescribe proper medicine. Based on experimental results, the RNN model exhibited the 97% of f1 score.

References

[1] "Breast Cancer", NCI, Jan, (1980).

[2] "Breast Cancer Treatment", NCI, (2014).

[3] "World Cancer Report 2014", WHO, (2014).

[4] Collignon J, Lousberg L, Schroeder H, Jerusalem G., "Triple-negative breast cancer: treatment challenges and solutions. Breast Cancer: Targets and Therapy", (2016).

[5] "World Cancer Report", IARC, (2008).

[6] "SEER Stat Fact Sheets: Breast Cancer", NCI, (2014).

[7] Solomon T, Rachet B, Whitehead S, Coleman MP., "Cancer survival in England: patients diagnosed 2007–2011 and followed up to 2012", National Statistics, (2013).

[8] Malvia S, Bagadi SA, Dubey US, Saxena S., "Epidemiology of breast cancer in Indian women", APJCO, (2017), pp. 289-295.

[9] Lavanya D., Rani K.U., "Ensemble decision tree classifier for breast cancer data", IJITCS, (2012), pp. 17-24.

[10] Goyal K, Sodhi P, Aggarwal P, Kumar M., "Comparative Analysis of Machine Learning Algorithms for Breast Cancer Prognosis", IEEE Proceedings on CCN, (2019), pp. 727-734.

[11] Cirkovic BR, Cvetkovic AM, Ninkovic SM, Filipovic ND., "Prediction models for estimation of survival rate and relapse for breast cancer patients", IEEE Proceedings on BIBE , (2015), pp. 1-6.

[11] Kate RJ, Nadig R., "Stage-specific predictive models for breast cancer survivability", IJMI, (2017), pp. 304-311.

[12] Ahmed K, Habib MA, Jesmin T, Rahman MZ, Miah MB., "Prediction of breast cancer risk level with risk factors in perspective to bangladeshi women using data Mining", IJCA, (2013).

[13] Ahmad LG, Eshlaghy AT, Poorebrahimi A, Ebrahimi M, Razavi AR., "Using three machine learning techniques for predicting breast cancer recurrence", JHMI, (2013).

[14] Abdelghani B and Erhan G, "Predicting Breast Cancer Survivability Using Data Mining Techniques", GWU, (2012).

[15] Chaurasia, Vikas, Pal, Saurabh, "Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability" IJCSMC, (2014), pp. 10-22.

[16] Mikolov T, Karafiát M, Burget L, Cernocky J, Khudanpur S, "Recurrent neural network based language model", Proceedings on SCA, (2010).

[17] Hansen LK, Salamon P, "Neural network ensembles", IEEE Transactions on PAMI, (1990), pp. 993-1001.

[18] Che Z, Purushotham S, Cho K, Sontag D, Liu Y., "Recurrent neural networks for multivariate time series with missing values", Scientific reports, (2018).

[19] Miikkulainen R, Liang J, Meyerson E, Rawal A, Fink D, Francon O, Raju B, Shahrzad H, Navruzyan A, Duffy N, Hodjat B., "Evolving deep neural networks", Artificial Intelligence in the Age of Neural Networks and Brain Computing, (2019), pp. 293-312.

[20] Du KL, Swamy MN. , "Recurrent neural networks", Neural networks and statistical learning, (2019), pp. 351-371.

Follow This Article at The Following Site:

Venkata Appaji S, Shiva Shankar R, Murthy K, Someswara Rao C. Breast Cancer Disease Prediction With Recurrent Neural Networks (RNN). IJIEPR. 2020; 31 (3) :379-386
URL: http://ijiepr.iust.ac.ir/article-1-1069-en.html