BREAST CANCER DISEASE PREDICTION WITH RECURRENT NEURAL NETWORKS (RNN)

Sangapu Venkata Appaji¹, R Shiva Shankar², K.V.S. Murthy³, Chinta Someswara Rao⁴

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

[chinta.someswararao@srkrec.ac.in, shiva@srkrec.ac.in, kvssmurthy75@srkrec.ac.in, venkataappaji.sangapu@gmail.com]

Abstract -- Cancer is a consortium of diseases which comprises abnormal increase in cells growth by having potential to occupy and attack the entire body. According to study breast cancer is the most likely occurs in the women and which became the second biggest cause of women death. Due to its wide spread and importance some of the researchers work on this, but still there is a need to improvement. During this work in order to partially fulfill this proposed technique of deep learning along with RNN in predicting breast cancer disease which will help the doctor while diagnosis the patient. To assess the efficiency of the proposed method we used breast cancer data belong to UC Irvine repository. Precision, recall, accuracy and f1 score of proposed method shows good scores and proposed technique performs well Consortium.

Key words: Cancer; Breast Cancer, Deep learning; RNN;

1. INTRODUCTION

Breast cancer disease is a malignant growth which generated from Breast fleshy tissue [1]. Breast malignancy can be identified by an abnormality in the Breast, an adjustment fit as a fiddle, dimples on the skin, and areola producing liquid, changes happened in the areola in recent past, or a red or covered fix of the skin [2]. Dangerous causes for breast cancer progress in female comprises obesity, absence of physical exercise, etc., [2, 3].

Results for breast cancer growth fluctuate contingent upon the malignancy type, degree of sickness, and individual's age.[4] Existence after Breast cancer in the created realm are tall,[5] having somewhere in the percentage range of 80 and 90 of those in Britain and the America alive for something nearer to 5 years.[6,7].

Breast cancer growth had reached top disease among females in India having age balanced level as high as 25.8 per 100,000 ladies and death 12.7 per 100,000 ladies [8].

Information gathered from different most recent national cancer related to breast growth libraries are looked at for rate, death rates. Age balanced frequency rate of cancer because of breast carcinoma was reached at higher rate as 41 for every 100,000 ladies in Delhi, followed via Chennai (37.9), Bangalore (34.4) and Thiruvananthapuram District (33.7).

Its projection for India amid time spans 2020 recommends the numerical value will reach high. Better wellbeing mindfulness, accessibility for its growth broadcast projects, cure offices would reason a great, positive clinical depiction in India.

Due to these statistics and previous work it very necessity to work in this area and detect the disease in early stage that may key to preventing damage.

2. LITERATURE SURVEY

Some of the work is done in this area is discussed here.

Lavanya and Rani [9] described ensembles of decision tree created by capturing accomplishes improved than the single decision tree model, the performance of ensembles of classifier in Breast cancer forecast has not frequently been discovered.

Kashish Goyal et al., [10] described method for identifying status of the cancer in patient, likely in kind state or virulent state. Data used was taken from WISCONSIN dataset from UC Irvine machine learning warehouse. WISCONSIN dataset comprises patients, whose life at risk of cancer development or have already revived cancer.

Cirkovic BR et al., [11] described the data mining practical methods on Breast cancer patients in order to estimate survival rate and disease deterioration. Authors performed comparison on popular machine learning models and they concluded that classifiers will help doctors in breast cancer survivability and recurrence.

Kate RJ and Nadig R. [12] described prediction of survivability of the patients with Breast cancer by applying machine learning techniques. Authors also evaluated models individually and also with combination.
In this work, authors [13] collected data from 100 peoples' which is combination of both cancer and non-cancer. Later applied the K means method to split the data into relevant and non-relevant. Authors used WEKA to measure the risk factors and raking. Finally, predicted the Breast Cancer risk level using Lotus Notes.

Ahmad LG et al., [14] studied data mining techniques applications later they developed the breast cancer predictive models for recurrence. Authors performed their study on a set of patients by follow up them for two years.


Chaurasia et al., [16] presented a model identifying the cancer in breasts with RepTree, RBF Network and Simple Logistic. During this work authors used Simple Logistic for measurement of feature reduction, also used RepTree and RBF for breast cancer prediction.

3. METHODOLOGY

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

Pre-processing: It is significant stage in the process of data mining and is particularly applicable to data mining and machine learning methods. Data collection procedures are often freely controlled that may lead to create out-of-range values, wrong combinations of data, null values, missing values etc., But we need the standardization of data for machine learning models. Hence we need to preprocessing the data.

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

Model Training: The procedure for training of Machine Learning model includes learning information from training data. The training data must contain the correct answer, which is known as a target attribute.

Model testing: The testing dataset is given to simulation model for testing purpose.

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

3.1. Recurrent neural network (RNN)

In this paper, we used recurrent neural network (RNN)[17,18,19, 20, 21], which is special category of artificial neural network in which links among the nodes generates a directed graph along with a temporal arrangement.

RNNs were utilized in deep learning and in the improvement of techniques which pretend the movement of neurons in the brain of human being. RNN nodes are more dominant than other models for predicting the outcomes because these models used back propagation. The architecture of our method is shown in Fig.2. This method having one input layer consists of 30 input nodes, three hidden layers consists of 64, 128 and 256 nodes respectively and one output layer consists of one node 0 or 1. relu activation function is used in hidden layer and dropout about 0.25 is used.
4. DATASET

During this research, the dataset used for Breast cancer was downloaded from the UC Irvine repository. This dataset has 569 instances and 31 attributes, out of which the first attribute, diagnosis, is considered as the target variable and the rest of the attributes are considered as input variables.


Fig 3 has shown the histogram of 2 to 31 features, from Fig 3, it is observed that every feature is important. Fig 4 shows the Bar plot of attribute 1, and from Fig 3, it is observed that class 0 having 357 instances and class 1 having 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 discussed the outcomes of 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)”.

\[
\text{Accuracy} = \frac{TP + TN}{P + N}
\]

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

\[
\text{Precision} = \frac{TP}{TP + FP}
\]

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

\[
\text{Recall} = \frac{TP}{P}
\]

**Confusion matrix**: It shows the number of correct and incorrect predictions with count values and broken down by each class. It provides us understanding not only into the mistakes being performed by a classifier however more importantly the categories of mistakes that are being performed.

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

<table>
<thead>
<tr>
<th></th>
<th>precision</th>
<th>Recall</th>
<th>f1-score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.98</td>
<td>0.97</td>
<td>0.98</td>
<td>66</td>
</tr>
<tr>
<td>1</td>
<td>0.96</td>
<td>0.98</td>
<td>0.97</td>
<td>48</td>
</tr>
<tr>
<td>avg/total</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
<td>114</td>
</tr>
</tbody>
</table>

![Confusion Matrix - Test Data](image_url)

Fig 6 confusion matrix
6. CONCLUSIONS

Recent statistics of cancer told that Breast cancer growth had reached top disease among females in India having age balanced level as high as 25.8 per 100,000 ladies and death 12.7 per 100,000 ladies. Due to its importance, in this paper we are taken breast cancer dataset from UCI repository having 569 instances out of which 357 instances having 0 class and 212 instances having class 1. This paper discusses the diagnosis breast cancer disease by making use of deep learning methods called RNN which will helps doctor while giving the medicine. Experimental results have shown that, the RNN model exhibits the 97% of f1 score.

REFERENCES