6.1 Introduction

Cervical Cancer as discussed earlier is the second deadliest diseases in the world among women. Various experiments and research activities were carried out to identify cervical cancer at the earliest stage as it is hidden in women till it becomes mature and incurable. Accurate prediction of occurrence of cervical cancer has been the most challenging and toughest task in medical data mining, because of the non-availability of proper dataset. Many researches had been done to develop different techniques that can solve problems and improve the prediction accuracy of cervical cancer through images. But in this research work, the prediction of cervical cancer is with NCBI (National Center for Biotechnology Information) data set. However, this is the only system designed exclusively for cervical cancer prediction with datasets.

This research has described the prediction of cervical cancer in two stages Normal Cervix or Cancer Cervix of women with data mining algorithms in congestion with machine learning to obtain the reasonable accuracy. This dissertation describes a finite, well defined GDS3233 database which is well suited for cancer prediction with data mining and machine learning technique.

Different data mining techniques like CART, RFT and machine learning K-Means with RFT are examined to find accurate results for cervical cancer. Several experiments were conducted using these algorithms with MATLAB (2013 version) coding. A refined algorithm is implemented to test the NCBI datasets which enable the user to test the data directly by implementing it in MATLAB tool. The major objective is to increase the accuracy of results
and improve the efficiency of prediction of cancer cells from normal cells. The achieved prediction performances are comparable to the existing techniques.

This research work is an attempt to find out the solution for cervical cancer and give awareness to the women regarding the health issues with regards to menopause and problems faced during this stage. A discussion with learned professors in Radiological Department of JIPMER Hospital is carried out at the beginning stage and analyzed about the testing of Cervical Cancer. As stated, in the previous chapters a sample of 350 records (250 for training and 100 for testing) are selected. The results obtained through biopsy test were put through statistical analysis and was given through MATLAB for algorithm testing. For the better understanding the results were divided and presented under following three heads are CART, RFT and K-Means with RFT

Even though, in the recent past, some progress has been achieved, yet there are still more remaining challenges for developing better prediction algorithms. This research work is presented in the form of prediction tree, to show the status of disease i.e. normal or cancer cervix using four phases.

The first phase contains the collection of database with 61 attributes and 22,283 records from National Centre for Biotechnology Information (NCBI). The Second phase employs preprocessing of the NCBI dataset using normalization. In third phase, this preprocessed data is applied to data mining algorithms CART, RFT and Hybridization of K-Means and RFT. The validation of the optimum algorithm with higher accuracy in prediction level is done in the fourth phase.

To enhance this accuracy level of the algorithms are compared and with their efficiency and various experimental prediction factors are used. They are analyzed and discussed with charts and tables.
6.2 Statistical Parameters of Prediction Factors

The predictions of cervical cancer need to be most accurate as it plays a vital role in the life of a human being and encourage her to take further tests for cervical cancer and treat them at an earlier stage. Hence various criteria were discussed in predicting accuracy of cancer cells in the cervical region of women. After careful analysis of various factors for accurate prediction, three major factors were finalized. They are

- Sensitivity
- Specificity
- F-Score

There are various reasons to select these factors for deciding the prediction accuracy of algorithms to detect cervical cancer.

6.2.1 Sensitivity

Sensitivity is a statistical measure of the performance of binary classification test, also known as classification function in statistics. Sensitivity is used to measure the proportion of positives that are correctly identified as the same. Hence it is also known as true positive rate. It has a big application in Medical field to identify and predict the formation of diseases in an accurate manner. Few of the important applications of sensitivity in medical field is to identify the percentage of sick people identified as having any disease following a condition or symptom and percentage of cancer patients identified through change in cell structure respectively.

This context clearly indicates that sensitivity can be a big factor in predicting cervical cancer in women as it can indicate positive proportion as identified from the algorithm. After proper experiments using MATLAB 2013, the following results were identified shown in Table 6.1
Table 6.1 Sensitivity of Algorithms in Cervical Cancer Predictions

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>CART</td>
<td>1</td>
</tr>
<tr>
<td>RFT</td>
<td>1</td>
</tr>
<tr>
<td>RFT – with K Means learning</td>
<td>1</td>
</tr>
</tbody>
</table>

sensitivity = \( \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}} \)  \( (6.1) \)

The prediction on sensitivity of all three algorithms are calculated with equation (6.1) and its graphical representations as follows:

![Sensitivity graph](image)

**Figure 6.1 Cervical Cancer Prediction of Accuracy of Algorithms with Sensitivity factor**

Thus the sensitivity of all three algorithms are found to be equal in predicting cancer cells in the cervical region of women. Hence all accuracy for algorithms are considered to be 100% sensitive shown in Figures 6.1, 6.2.
6.2.2 Specificity

Specificity is also a statistical prediction of medical data similar to sensitivity. But Sensitivity measures the proportion of negative rates that are correctly identified as same. It is also known as true negative rate and hence it is the negation of sensitivity factor analysis. It avoids false positives as compared to sensitivity which avoids false negatives respectively. Hence this factor is also applied on cervical cancer prediction algorithms to identify the healthy persons who are not affected by cervical cancer. The specificity result of various algorithms are indicated in the table with formula as follows:

\[
\text{specificity} = \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}}
\]  

(6.2)

The following Table 6.2 indicates the prediction accuracy of specificity. The Figure 6.3 and Figure 6.4 indicate raise in accuracy of algorithms with specificity factor.
Table 6.2 Specificity Analysis of Cervical Cancer Algorithms

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>CART</td>
<td>0.73684</td>
</tr>
<tr>
<td>RFT</td>
<td>0.87500</td>
</tr>
<tr>
<td>RFT – with K means learning</td>
<td>0.93333</td>
</tr>
</tbody>
</table>

Figure 6.3 Chart to Indicate Raise in Accuracy of Algorithms with Specificity Factor

Figure 6.4 Chart Indicating the Formula and Result of Analysis for Specificity

\[
\text{Specificity} = \frac{T_N}{T_N + F_P}
\]
6.2.3 F-Score Prediction

The F-Score alternatively known as F-Measure is considered as the measure that is used in statistics oriented analysis and decision making to test accuracy of binary classification. The F-Score computes the final score by manipulating the precision (p) and recall (r) of the decision tree during execution where Precision $= \frac{TP}{TP+FP}$ and Recall $= \frac{TP}{TP+FN}$. The F-score is one of the best method to find the best result in any format by considering the weights of correct positive divided by positives to be returned. But the F1 Score gets completed only when the best values arrive at 1 and worst value arrive at 0 respectively as shown in Figure 6.5. It plays a vital role in analyzing the medical data as it has more accurate results compared to sensitivity and specificity. Hence this method of statistical measure is also used to test the best algorithm after the results are obtained in various different experiments.

![Figure 6.5 F-Score Prediction](image)

The results are observed and found that the RFT combined with machine learning algorithm K-Means provides the maximum accuracy as indicated in the Table 6.3.
Table 6.3 Comparison of Algorithms after Experiments Based on F-Score Statistical Measure

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>CART</td>
<td>0.84848</td>
</tr>
<tr>
<td>RFT</td>
<td>0.93333</td>
</tr>
<tr>
<td>RFT – with K Means learning</td>
<td>0.96552</td>
</tr>
</tbody>
</table>

Figure 6.6 Graph Showing the Performance of Algorithms

Figure 6.7 Pie Chart Indicating the Formula and Results of F-Score Analysis

Thus the three major factors were tested and analyzed to find out the best algorithm to detect cervical cancer with better accuracy rate represented in Figure 6.6 and Figure 6.7.
6.3 Comparison of Factors to Predict Accuracy

The detection rate is one of the best methods to indicate a better comparison of all medical algorithms for identifying best predictions. Hence detection rate is also used to predict the best of algorithms. It also indicates the proportion of individuals with a particular condition who is tested positive for that condition when measured by any relative algorithm or methodology. The detection rate also helps the researcher to predict the time span required for predicting the best algorithm possible for any huge data like NCBI datasets that can take much time to predict accuracy under normal circumstances. After analysis, the detection rates for various algorithms used in cervical cancer prediction is indicated below in Table 6.4. The Figure 6.8 and Figure 6.9 indicates the detection rate analysis of algorithms with sensitivity.

Table 6.4 Detection Rate Analysis for Algorithms with Sensitivity Factor

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Detection Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>CART</td>
<td>0.838710</td>
</tr>
<tr>
<td>RFT</td>
<td>0.935484</td>
</tr>
<tr>
<td>RFT – with K means learning</td>
<td>0.967742</td>
</tr>
</tbody>
</table>

![Detection Rate Analysis](image)

Figure 6.8 Detection Rate Analysis of Algorithms with Sensitivity Factor
Figure 6.9 Pie Chart to Indicate the Detection Rate of Algorithms with Sensitivity factor

Based on the various factors and co-factors the accuracy of the algorithms is predicted using the following formula,

\[
Accuracy (A) = \frac{TN + TP}{TN + TP + FN + FP}
\]  

(6.3)

After Careful Examinations and Experiments, the sensitivity and specificity factor of the datasets are taken as major criteria for predicting the accuracy of algorithms. The three algorithms that are used in detecting cervical cancer in women using NCBI dataset is employed in the formula to calculate accuracy and after calculations the accuracy levels are determined. The various accuracy levels of algorithms are identified in the Table 6.5

Table 6.5 The Accuracy Levels of Various Algorithms Using Sensitivity and Specificity Factor

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CART</td>
<td>100</td>
<td>73.68</td>
<td>84.84</td>
</tr>
<tr>
<td>RFT</td>
<td>100</td>
<td>87.50</td>
<td>93.93</td>
</tr>
<tr>
<td>RFT – with K Mean</td>
<td>100</td>
<td>93.93</td>
<td>96.55</td>
</tr>
</tbody>
</table>
6.4 Summarized Interpretations

Cervical cancer is the second deadliest disease in women that affects many women just because they couldn’t identify it in the earlier stage. Hence this research is very effective to find the presence of cancer cells in the cervical region of women at the beginning stage itself. The result is obtained from three different algorithms during the research and the algorithm with utmost accuracy is opinioned as the best algorithm to predict the cancer cells. The interpretation and experimental verification of accuracy of algorithm is carried out under three different factors as sensitivity, specificity and F-Score respectively.

Finally, the sensitivity combined with detection rate factor and specificity is considered to find the accurate result for all the algorithms. After careful experimentation, the novelty of research, the hybrid algorithm is considered as the best compared to all other algorithms.

The Performance analysis detector is used to evaluate, allow comparisons where several ratios has been taken into account shown in the Table 6.6

<table>
<thead>
<tr>
<th></th>
<th>Cancer</th>
<th>Normal</th>
<th>Cancer Cervix</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP</td>
<td>Cancer</td>
<td>Cancer</td>
<td>Cancer Cervix</td>
</tr>
<tr>
<td>TN</td>
<td>Cancer</td>
<td>Normal</td>
<td>Normal Cervix</td>
</tr>
<tr>
<td>FP</td>
<td>Normal</td>
<td>Cancer</td>
<td>Cancer Cervix</td>
</tr>
<tr>
<td>FN</td>
<td>Normal</td>
<td>Normal</td>
<td>Cancer Cervix</td>
</tr>
</tbody>
</table>

6.5 Conclusion

According to the results after experimentation, it is found that the novelty of the research in Cervical Cancer cells prediction, Hybridization of RFT with K-Means is the best algorithm with utmost accuracy in prediction. Thus the accuracy levels of various algorithms are designed and predicted using sensitivity, specificity and F-Score as prediction factors and presented with results to enable the researcher to conclude the best algorithm that can be used with NCBI database to give accurate prediction in the most successful manner.
CHAPTER – VII

CONCLUSION AND FUTURE ENHANCEMENTS

7.1 Conclusion

The main contribution of the thesis is summarized in this final chapter. Today, in the medical scenario there are enormous opportunities, specialities, advanced technologies and automated systems etc., available to diagnose or predict diseases in its early stage so as to prevent further damages. Though there are many supportive technologies or mechanisms available in the medical domain, still research is on progress for diagnosing and predicting diseases like cancer, tumor and heart problem etc., For giving better treatment like that of cancer, this research work focuses mainly on predicting the Cervical Cancer in its early stage for giving better treatment.

Cervical Cancer is one such prominent form of cancer that affects many women in India. Cervical cancer is a much complicated disease in women that shows no sign in its initial stage. Hence it is a challenging and complicated process for researchers that needs a very efficient and complex algorithm with combinations of expertise, observations and techniques extracted from various disciplines. The earlier research works made in cervical cancer detections are based on images taken during screening tests and manipulated using image processing techniques. From the literature review of various domains, it is found that data mining algorithms are efficient in predicting the diseases in medical field.

Our novel attempt to make use of MATLAB2013 tool efficiently and effectively for implementing the Data Mining Techniques, especially for prediction of Cervical Cancer, which is fully tailored and customized in simple step by step wizards, with all the mandatory and optional parameters. No wonder with MATLAB, being a server side development tool, supports decision making for the physicians and predication of cervical cancer for end users.
To experiment the proposed research work, a sample dataset was collected from National Centre for Biotechnology Information (NCBI) and its URL is, http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS3233

Initially the dataset is normalized by undergoing the process of pre-processing. The screenshot contains nearly 61 attributes (Biopsy Gene values) and 22,283 records. We have tried various techniques for Cervical Cancer prediction models and we have reviewed different data mining techniques, namely CART and RFT algorithms along with K-Means machine learning algorithm. Experiments were conducted using these two datamining algorithms and the predicted values were compared with the observed values. The results obtained from the proposed work is also validated and justified.

Our proposed Classification And Regression Tree (CART) algorithm for Cervical Cancer is capable of producing, a prediction accuracy of nearly about 83.87%. Our next proposed method with RFT produces an accuracy percentage of nearly 93.54% for this dataset. The hybridization of our proposed model (K-Means with RFT) producing a prediction accuracy of 96.77%. From the literature survey, it is found and justified that the accuracy obtained through our model is well accepted. Our comparative study and experimental results on the prediction of Cervical Cancer using CART, RFT, K-Means, confirms that hybridization (K-Means with RFT) is the best algorithm that yields better accuracy percentage and for NCBI dataset, we found that the result of Hybridization of algorithm is accurate to the extent of 96.77%. Further the work may be extended to improve the accuracy percentage of all other techniques too, and to integrate with soft computing techniques with respect to the performance and overhead issues and also to predict the period of invasion of Cervical Cancer. In future, the detection of a cervical cancer can be intensified in tracking the growth and the development of malignant tumour in the cervical region.
7.2 Scope for Future Research

The proposed model can be extended in predicting the presence of benign or malignant tumour in cervical region using Nano Technology. Earlier researches had proved that Nano technology can be applied in the prediction of various other diseases like brain tumour, breast cancer etc., Similarly, Nanotechnology can also be applied in Cervical Cancer Prediction and the accuracy can be increased using the hybridization of other algorithms can be implemented in future.

Thus, Cervical Cancer prediction module is a disease monitoring, mitigation and management service and it is a preventive and defensive system too. Thus, our work will render timely, lifesaving Cervical Cancer Treatment that will kindle millions of women to save themselves for the enlightenment of their family and the society as well.