CHAPTER – III

DEVELOPMENT OF A PREDICTION MODEL FOR CERVICAL CANCER USING CART ALGORITHM

3.1 Introduction

Chapter III focuses on the design and implementation of Cervical Cancer Model, overview of Data Mining and its Techniques like SVM, C4.5, ID3, and Naïve Bayesian with its features. Data Mining is a successful emerging technology that has made a revolutionary change in medical field. Data Mining is referred as [61] process of “extracting or mining knowledge from large amounts of data.” It is a methodology that emphasis on the searching of data from huge dataset by extracting knowledge from it. Hence Data Mining is chosen to predict the occurrence of Cervical Cancer.

Cervical Cancer, being the second largest cancer disease in the world, incorporates various research programs around the world. Hence computerized classification and regression methods are always used to classify normal cells from cancer cells. The classification and regression methods are available in data mining techniques under supervised and unsupervised models.

The architecture of Cervical Cancer prediction model is depicted in this chapter which clearly exhibits about the Cervical Gene attributes as input, the data mining techniques used and predicting accuracy as output. The Source, structure, attributes and the normalization of the Cervical Cancer data set used is clearly presented. The Data modeling techniques, its step-by-step implementation with appropriate figures, screenshots, table output and results are explained with neat illustrations. The MATLAB operations used in this data mining technique
are elaborately described and the desired results are obtained with prediction accuracy of the algorithm.

3.2 Cervical Cancer Dataset

3.2.1 Cervical Cancer Data Source

The database of high through gene expression information along with probe id information was available on the URL http://www.ncbi.nlm.nih.gov/geo. This dataset is a micro array type of data of the cervical cancer cell lines. The identifier of the database GDS3233 a complete soft file of this database is available in the file GDS3233.xls.

In this research, the data set are collected from National Center for Biotechnology Information (NCBI) [62]. It is an integral part of Unites States National Library of Medicine (NLM) which is a part of National Institutes of Health. The NCBI holds a collection of databases that contains biomedicine, biotechnology and also for bioinformatics tools and its services. It includes databases like Gene/bank DNA sequences and PubMed, a bibliographic database for the biomedical literature and also a database for BCBI Epigenomics respectively. All of such databases are viewed through online Search Engine.

The current data set is obtained from Gene Expression Omnibus [63] (GEO) database from NCBI where Cervical Cancer primary tumors [64] and cell lines are analyzed using Cervical Cancer Tumorigenesis. In Tumorigenesis, Chromosomal amplification is a common cellular mechanism of gene activation. To identify the Cervical Cancer progression, Chromosome 20 is used.

The Original Data on Cervical Cancer is obtained from the official website of National Centre for Biotechnology Information (NCBI).
The Data Source of the obtained Dataset from NCBI has the following information:

- DATABASE = Geo
- Database name = Gene Expression Omnibus (GEO)
- Database institute = NCBI NLM NIH
- Database email = geo@ncbi.nlm.nih.gov
- Database_ref = Nucleic Acids Res. 2005 Jan 1;33 Database Issue:D562-6
- DATASET = GDS3233
- dataset title = Cervical cancer tumorigenesis
- dataset description = Analysis of cervical cancer (CC) primary tumors and cell lines. Chromosomal amplification is a common cellular mechanism of gene activation in tumorigenesis; chromosome 20 is a commonly gained chromosome in CC. Results provide insight into the potential role of chromosome 20 gain in CC progression.
- dataset type = Expression profiling by array
- dataset_pubmed_id = 18506748
- dataset platform = GPL96
- dataset_platform_organism = Homo sapiens
- dataset_platform_technology_type = in situ oligonucleotide
- dataset_feature_count = 22283
- dataset_sample_organism = Homo sapiens
- dataset_sample_type = RNA
- dataset_channel_count = 1
- dataset_sample_count = 61
- dataset_value_type = count
- dataset_reference_series = GSE9750
- dataset order = none
- dataset_update_date = Jun 04 2008

Initially, the dataset obtained from NCBI is shown in the EXCEL format and organized in the form of table as shown in Figure 3.1. The Original Dataset from GDS3233 has obtained a total of 61 attributes with three states [65] of cervical cancer as shown in Figure 3.2 recognized during analysis. They are

1. Cervical Cancer Cell line, indicating the presence of Cervical Cancer that needs further testing and control,
2. Normal state showing absence of cervical cancer and
3. Cervical Cancer Primary Tumor, signifying the initial state of Cervical Cancer that can be cured with proper treatment.

Figure 3.1: Original Dataset from NCBI Incorporated as Excel Data

Figure 3.2 Three States of Cervical Cancer
The gene samples with attributes identified are shown in Table 3.1

### Table 3.1 Cervical Cancer Dataset Attributes with 61 Biopsy Gene Samples

<table>
<thead>
<tr>
<th>Sample</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSM246087</td>
<td>Cervical cancer cell line, C4-I</td>
</tr>
<tr>
<td>GSM246088</td>
<td>Cervical cancer cell line, CaSki</td>
</tr>
<tr>
<td>GSM246089</td>
<td>Cervical cancer cell line, C-33A</td>
</tr>
<tr>
<td>GSM246090</td>
<td>Cervical cancer cell line, HT-3</td>
</tr>
<tr>
<td>GSM246119</td>
<td>Cervical cancer cell line, SiHa</td>
</tr>
<tr>
<td>GSM246120</td>
<td>Cervical cancer cell line, SW756</td>
</tr>
<tr>
<td>GSM246121</td>
<td>Cervical cancer cell line, MS751</td>
</tr>
<tr>
<td>GSM246122</td>
<td>Cervical cancer cell line, ME-180</td>
</tr>
<tr>
<td>GSM246123</td>
<td>Cervical cancer cell line, HeLa</td>
</tr>
<tr>
<td>GSM246422</td>
<td>Normal cervix, commercial_Ambion</td>
</tr>
<tr>
<td>GSM246423</td>
<td>Normal cervix, commercial_Stratagene</td>
</tr>
<tr>
<td>GSM246484</td>
<td>Normal cervix, commercial_BioChain</td>
</tr>
<tr>
<td>GSM246485</td>
<td>Normal cervix epithelium_CaCX3</td>
</tr>
<tr>
<td>GSM246486</td>
<td>Normal cervix epithelium_CaCx4</td>
</tr>
<tr>
<td>GSM246487</td>
<td>Normal cervix epithelium_CaCx5</td>
</tr>
<tr>
<td>GSM246488</td>
<td>Normal cervix epithelium_03-3505</td>
</tr>
<tr>
<td>GSM246489</td>
<td>Normal cervix epithelium_03-4216</td>
</tr>
<tr>
<td>GSM246490</td>
<td>Normal cervix epithelium_03-4508</td>
</tr>
<tr>
<td>GSM246491</td>
<td>Normal cervix epithelium_03-4986</td>
</tr>
<tr>
<td>GSM247162</td>
<td>Normal cervix epithelium_03-5419</td>
</tr>
<tr>
<td>GSM247163</td>
<td>Normal cervix epithelium_03-5438</td>
</tr>
<tr>
<td>GSM247164</td>
<td>Normal cervix_03-5611</td>
</tr>
<tr>
<td>GSM247165</td>
<td>Normal cervix_03-5657</td>
</tr>
<tr>
<td>GSM247166</td>
<td>Normal cervix_05-15</td>
</tr>
<tr>
<td>GSM247168</td>
<td>Normal cervix_05-31</td>
</tr>
<tr>
<td>GSM247169</td>
<td>Normal cervix_05-446</td>
</tr>
<tr>
<td>GSM247171</td>
<td>Normal cervix_05-1308</td>
</tr>
<tr>
<td>GSM247173</td>
<td>Normal cervix_05-1352</td>
</tr>
<tr>
<td>GSM247174</td>
<td>Normal cervix_05-1981</td>
</tr>
<tr>
<td>GSM247175</td>
<td>Normal cervix_05-4602</td>
</tr>
<tr>
<td>GSM247188</td>
<td>Normal cervix_05-4615</td>
</tr>
<tr>
<td>GSM247189</td>
<td>Normal cervix_05-4959</td>
</tr>
</tbody>
</table>
### 3.2.2 Preprocessed Data Set

Initially the dataset is normalized by undergoing the process of pre-processing. The Figure 3.3 contains nearly 61 attributes (Biopsy Gene values) and 22,283 samples. Figure 3.4 contains

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSM247190</td>
<td>Normal cervix_05-5007</td>
</tr>
<tr>
<td>GSM247650</td>
<td>Cervical cancer_CC126</td>
</tr>
<tr>
<td>GSM247651</td>
<td>Cervical cancer_CC128</td>
</tr>
<tr>
<td>GSM247652</td>
<td>Cervical cancer_CC140</td>
</tr>
<tr>
<td>GSM247653</td>
<td>Cervical cancer_CC163</td>
</tr>
<tr>
<td>GSM247654</td>
<td>Cervical cancer_CC205</td>
</tr>
<tr>
<td>GSM247655</td>
<td>Cervical cancer_CC207</td>
</tr>
<tr>
<td>GSM247656</td>
<td>Cervical cancer_CC214</td>
</tr>
<tr>
<td>GSM247657</td>
<td>Cervical cancer_CC218</td>
</tr>
<tr>
<td>GSM247658</td>
<td>Cervical cancer_CC222</td>
</tr>
<tr>
<td>GSM247659</td>
<td>Cervical cancer_892T</td>
</tr>
<tr>
<td>GSM247660</td>
<td>Cervical cancer_1721T</td>
</tr>
<tr>
<td>GSM247661</td>
<td>Cervical cancer_1798T</td>
</tr>
<tr>
<td>GSM247662</td>
<td>Cervical cancer_1875T</td>
</tr>
<tr>
<td>GSM247663</td>
<td>Cervical Cancer_1981T</td>
</tr>
<tr>
<td>GSM247856</td>
<td>Cervical cancer_654T</td>
</tr>
<tr>
<td>GSM247857</td>
<td>Cervical cancer_841T</td>
</tr>
<tr>
<td>GSM247859</td>
<td>Cervical cancer_1434T</td>
</tr>
<tr>
<td>GSM247860</td>
<td>Cervical cancer_1509T</td>
</tr>
<tr>
<td>GSM247862</td>
<td>Cervical cancer_1898T</td>
</tr>
<tr>
<td>GSM247864</td>
<td>Cervical cancer_1900T</td>
</tr>
<tr>
<td>GSM247865</td>
<td>Cervical cancer_1907T</td>
</tr>
<tr>
<td>GSM247866</td>
<td>Cervical cancer_2035T</td>
</tr>
<tr>
<td>GSM247876</td>
<td>Cervical cancer_20-04T</td>
</tr>
<tr>
<td>GSM247877</td>
<td>Cervical cancer_56-04T</td>
</tr>
<tr>
<td>GSM247878</td>
<td>Cervical cancer_88-04T</td>
</tr>
<tr>
<td>GSM247879</td>
<td>Cervical cancer_98-04T</td>
</tr>
<tr>
<td>GSM247881</td>
<td>Cervical cancer_103-04T</td>
</tr>
<tr>
<td>GSM247883</td>
<td>Cervical cancer_CaCx54</td>
</tr>
</tbody>
</table>
i) GSM-(GEO-SAMPLE)-DNA value,

ii) Probe-ID,

iii) Gene Symbol (Identifier),

iv) TPM-VALUE (Transcripts Per Million)

v) Cell-Type

Among the data used, different variables are identified and their functionalities are classified based on the variables to be tested. The variables which are present as fields in the datasets are identified to predict the outcome of the result after data mining analysis. The dataset is tested with CART algorithm for execution. At this stage, the tool is used and the proposed algorithm is applied.

Figure 3.3 consist of 61 samples with 22,283 records. Every row represents the gene dataset as cervical cancer cell line. The three states are normal cervix, cervical cancer cell line and cancer cervix. Figure 3.4 geo dataset along with id reference and gene names. GSM-GEO dataset contains gene name which has distribution mean as numerical value ranging from LSB to MSB.
3.3 Overview on Data Mining Algorithms

Data mining is defined as “finding hidden information in the database and it is called exploratory data analysis, data driven discovery and deductive learning”. Data mining access of a database differs from traditional access in three ways: Query, Data Access and Output.

3.3.1 Applications of Data Mining in Medical Field

Data mining has been one of the common technology that is applied in Computer Field especially in Medical informatics like healthcare organizations, health informatics, epidemiology, patient care and other monitoring systems where large volume of analysis are required. Different types of algorithms in data mining have significantly assisted in analyzing medical data more clearly, by distinguishing pathological data from normal data, for supporting decision making as well as in visualization and identification of hidden complex relationships between diagnostic features of different patient groups.
Data mining can easily classify medical data and correlate the relationship among variables in a huge data set. To make this possible, data mining has a rich set of algorithms. Such Data mining algorithms for the analysis of microarray gene expression offer promise for precise, accurate, and functionally robust analysis of genomics data in cancer classification. The efficiency and scalability of the presented technique also makes it well suited to the domains of medical image analysis for feature extraction and clustering of similar feature based rules. In this research, data mining methodologies are applied to medical informatics of cancer with special reference to Cervical Cancer, for the successful research benefit of researchers, professionals and teachers.

3.4 Existing Cervical Cancer Prediction Models Using Data Mining Algorithms

Various data mining algorithms like CART, C4.5, ID3, RFT, SVM, Naïve Bayesian and Apriori algorithms are applied, analyzed and identified the best model based on predictive performance. Predictive performance is a way to explain the correlation in sample dataset and producing stable results across sample data. This prediction, however looks simple but the actual process is enormous and elaborative in nature.

There are different varieties of “competitive evaluation of model” techniques in data mining that helps the researcher to achieve the goal. It’s important to carry out an intensive evaluation to find out the best algorithm among them. In the above dataset, it was analyzed that a huge amount of dataset is involved out of which a sample of 61 samples are considered to identify matching patterns that formulates the difference of cervical cancer cells from normal cells in women.

The process of applying different data mining models to the same dataset for developing a required model and compare their performance is called “competitive evaluation of models”. The technique is applied on the biological datasets of women and analyzed to obtain accurate
results of presence of cancer cells in women even at the earliest stage. Hence prediction should be most accurate to accept and identify the models by applying appropriate algorithm.

In order to identify the best algorithms, various techniques [66] were used in data mining which includes the following:

a) **Bagging** – identified by voting process and also by averaging

b) **Boosting** – used to reduce primary bias and variance

c) **Stacking** – enables to create generalizations of data

d) **Meta-Learning** – meta-data from datasets applied with automatic machine learning algorithms

Various algorithms from data mining need to be used along with data set of cervical cancer to give accurate predictions of the disease in women. Numerous algorithms are identified in data mining so far. But only few algorithms are considered as accurate and effective according to all-time statistics of research organizations. Regarding bio-medical applications, deeper analysis on data mining algorithms has to be made as biological data are more complex and hard to be predicted with future knowledge. Some of the well-known Data mining algorithms used in analysis of data sets are furnished below:

### 3.4.1 SVM Algorithm

Support Vector Machine (SVM) algorithm is a supervised learning algorithm that are used in classification and regression techniques for analysis of data and recognize patterns. When training set samples are given, SVM automatically builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier [67]. SVMs can be used to solve a lot of real world problems like the following:
SVM can be used in categorization of text and hypertext as their application can reduce the need for labeled training instances in both inductive and transductive settings in a significant manner.

SVM can also be used to perform classification of images and provides highest accuracy.

SVMs are also useful in medical science to classify proteins with up to 90% of the compounds classified correctly.

Initially SVM algorithm has chosen to predict cervical cancer. Though there are lot of advantages in the SVM, it has also major setback in over fitting to any kernel in the operating system and it may result in software errors and problems. Thus this algorithm is not squarely suitable for in prediction of Cervical Cancer.

3.4.2 Naïve Bayesian Algorithm

Naïve Bayesian Algorithm is a machine learning algorithm that is developed based on Bayes theorem in mathematics. Bayes theorem is an accurate result provider and hence it can be used in all levels of prediction. However, Bayes theorem [68] cannot handle huge data without misperception. Hence this algorithm cannot be applied to huge and sensitive data where the results are applied on human beings. Hence this algorithm is not much preferred to be used along with NCBI datasets.

3.4.3 ID3 Algorithm

ID3 is a simple decision tree algorithm that is aimed to create a decision tree of given set, by using top-down greedy search to check each attribute at every tree node of the sample data. For building ID3 algorithm, decision tree consists of nodes and arcs or sweeps which connect nodes [69]. To make a decision, one starts at the root node, and asks questions to determine which arc to follow, until one reaches a leaf node and the decision is made.
The main ideas behind the ID3 algorithm are:

**Step 1:** Each non-leaf node of a decision tree corresponds to an input attribute, and each arc to a possible value of that attribute. A leaf node corresponds to the expected value of the output attribute when the input attributes are described by the path from the root node to that leaf node.

**Step 2:** In a “good” decision tree, each non-leaf node should correspond to the input attribute which is the most informative about the output attribute amongst all the input attributes not yet considered in the path from the root node to that node. This is because we would like to predict the output attribute using the smallest possible number of questions on average.

**Step 3:** Entropy is used to determine how a particular input attribute is useful to get the desired output attribute for a subset of the training data. Entropy is a measure of uncertainty in communication systems introduced by Shannon. It is fundamental in modern information theory.

But ID3 algorithm may not fetch the desired results as it may stop in-between huge process and executes again. Hence such an unpredictable algorithm cannot be applied for a huge dataset like NCBI datasets.

### 3.4.4 Apriori Algorithm

The Apriori algorithm learns association rules and is applied to a database containing a large number of transactions. The basic Apriori algorithm is a 3 step approach namely Join, Prune and repeat.

i) **Join.** Scan the whole database for how frequent 1-itemsets are.
ii) **Prune.** Those item sets that satisfy the support and confidence move onto the next round for 2-itemsets.

iii) **Repeat.** This is repeated for each item set level until we reach our previously defined size.

However, Apriori algorithm is an association algorithm and hence it can best suite business applications rather than bio-medical applications. Hence the algorithm is not considered for cervical cancer research to identify accurate predictions of pre-cancer cells in women.

### 3.4.5 C4.5 Algorithm

C4.5 constructs a classifier in the form of a decision tree. C4.5 is given a set of data representing things that are already classified [69]. This is a supervised learning technique, since the training dataset is labeled with classes. Various reasons make C4.5 different from other algorithms which includes the following:

- It uses information gain when generating the decision tree.
- It uses a single-pass pruning process to mitigate over-fitting.
- It can work with both continuous and discrete data.

However, C4.5 algorithms cannot make its impact on medical applications as it cannot make accurate results on complex biological data. Hence it is not selected for the proposed research on identifying accuracy of pre-cancer cells.

### 3.5 Implementation of Cervical Cancer Prediction Model Using CART Technique

Data mining models with complex and huge datasets requires coordinate efforts from various experts or knowledgeable persons throughout the entire organization. Hence biological dataset, which are the most complicated datasets in the world may require opinion of various
experts around the world. Various sources of information are available for cervical cancer in different parts of the world. However, National Centre for Biological Information (NCBI) provides more stable and refined datasets from experts and stakeholders around the world. Hence it is chosen to be tested with data mining models.

According to data mining literature, various frameworks have been proposed throughout the world to server as blueprints to gather datasets, analyze them, disseminating results and implementing them with experimental proof. Among several algorithms available in predicting knowledge for datasets, some algorithms always have a big impact on biological data. Various algorithms and their impact on biological data are analyzed in the literature section and it was found that CART, RFT and K-means algorithm can be best suited for biological analysis of data and predict results with utmost accuracy. According to the Phase II of the current research to identify cervical cancer cells in women, Classification and Regression Algorithm (CART) is recommended to test and analyze the NCBI dataset and predict the cancer cells in the cervical region of women.

3.5.1 CART Algorithm

In general, the value of a target or dependent variable is predicted based on the values of several input or independent variables in a data mining model using decision trees. CART decision tree has one such methodology to identify and classify to find the degree of relationship between the independent variables in a dataset.

The Classification And Regression Tree (CART) methodology was proposed by Leo Breiman, Jerome Friedman, Richard Olshen and Charles Stone in 1984 as an umbrella term to predict the knowledge based on following types of decision trees. CART builds Classification And Regression Trees (CART) for predicting continuous dependent variables and categorical or predictor variables and by predicting the most likely value of the independent variable. The
decision tree produced by CART is strictly binary contain two branches of each decision node. CART recursively partitions the record into subsets of record with similar values of target attributes. This algorithm grows the tree by conducting for each decision node an exhaustive search of all variables and all possible splitting values selecting the optimal split. It gives a prediction of normal cervix or cancer cervix in this research work.

CART is a rule based method that generates a binary tree, through a binary recursive partitioning process that splits a node based on “yes or no” answer of the predictors. CART builds the tree by recursively splitting the variable space based on the impurity of the variables to determine the split till the termination condition is met. The GINI impurity determines how often a randomly chosen element from the set would be incorrectly labeled if it were randomly labeled according to the distribution of labels in the subset.

The CART algorithm is applicable and found effective with huge data access and predicting accuracy [70] like bio-medical data applications. Hence it is selected for applying in NCBI data that can have very complex and unpredictable datasets. The results have to be checked for determining accuracy in predicting cancer cells in the cervical region of women.

There are two broad application operations performed on the datasets of the cervical cancer datasets used in the research. They are

- Classification Trees
- Regression Trees

3.5.1.1 Classification Tree

In this mode, the target variable is categorical and the tree is used to identify the “CLASS” within which the target variable of cancer cells would fall into.
In the above figure 3.5, the blended data are uniquely identified and classified. The classification is possible if there is no overlapping of data. Biological data are mostly independent as well as overlapped. Hence classification is possible for 60% of data whereas another method has to be associated to identify the data that are overlapped and cannot be classified. Classification analyses and identifies unique data and predict the nature of cells using gene Index values and TPM (Transcripts Per Million) Values. However, the classification method [71] is a direct method and cannot be used to isolate complex and overlapped data. Hence classification trees alone cannot be used to predict the cervical cancer cells and hence regression trees are also used to classify overlapped data components.

**3.5.1.2 Regression Tree**

In certain cases, the variables cannot be classified and grouped into classes. In some datasets, data will be combined together in such a way that it cannot be segregated. Hence the target variable being continuous, regression tree or prediction tree will be used to predict the occurrence of cancer cells as shown in the Figure 3.6.
The regression tree can identify the relationship between the two overlapped variables and thereby helps to identify normal data from affected data. This method is very convenient to be applied in NCBI data to identify cancer cells from normal cells in cervical region of women.

3.5.1.3 Gini Index Prediction

CART uses GINI Index to determine which attribute the branch should be generated. The strategy is to choose the attribute whose GINI index is minimum after splitting. CART uses GINI Index as measurement for data impurity for selecting attributes. The attribute with the largest reduction in impurity is used for splitting the node’s records. CART accept the record with numerical values and also handles missing attributes [72]. Gini Index is used for cost complexity pruning and also generate regression trees that is used for prediction. Assuming training set ‘T’ includes n samples, the target property has ‘m’ values, among them, and the $i^{th}$ value show in T with a probability $P_i$, so it’s GINI Index can be described as below.

$$GINI(T) = 1 - \sum_{i=1}^{m} P_i^2$$
Where T is the training set and P is the probability. Assuming that A be divided to \( q \) subsets \{T_1, T_2, T_q\} among them, Ti’s sample number is \( n_i \), so the GINI Index divided according to property A can be described below

\[
GINI(T) = 1 - \sum_{i=1}^{n} \frac{n_i}{n} GINI(T_i)
\]

CART divides the property which loads a minimum value after the division. The Gene Index thus created using CART algorithm is tested with the sample gene values from cervical cancer gene database of 61 samples. The database is usually identified as a manual curated catalog of experimentally validated genes that are expected to be involved in various stages of cervical carcinogenesis.

Every data from dataset contains the following information:

i) Gene and Protein sequences
ii) Location
iii) Architecture
iv) Function
v) Chromosomal positions
vi) Accession number
vii) Gene
viii) CDS sizes, gene ontology and homology to other eukaryotic genomes.

Apart from above information, the original dataset from NCBI gene index is also associated along with the samples as source of external data.

### 3.5.2 Proposed Algorithm for Cervical Cancer Prediction Model

Predictive Models developed for a medical information that involves huge data for processing has to be designed carefully with right procedure to arrive at the optimal solution.
Hence CART algorithm is used as base for creating a new algorithm for predicting Cervical Cancer. The algorithm is modified according to the needs of the Cervical Cancer and coded based on the input of the NCBI Dataset of GDS3233. Based on the CART algorithm, we designed the following Prediction Algorithm of Cervical cancer.

**Input** : NCBI dataset – biopsy values  
**Output** : Prediction with Regression Tree output

BEGIN

Step 1: Start the process by loading the cervical dataset of GDS3233 into a database named CCPDB.

Step 2: From the database select the dataset TS that contains the values for training dataset.

Step 3: Load the training set TS and their respective classnames into TC for the database CCPDB.

Step 4: Generate a binary tree graph based classification tree with traindata TS and classnames TC.

Step 5: Select the testdata TES from database CCPDB and train the dataset TS.

Step 6: After the training load the TES and their respective classnames.

Step 7: Test the entire test data in an iterative manner varying from 1 to TES

Step 8: After the testing process, load tested data in a temporary variable DMV and then start the prediction.

Step 9: Store the predicted values in a new data variable DPV.

Step 10: If the predicted value DPV matches with temporary variable dataset DMV, summate the value of True Positive by 1 else Go to Step 7.

Step 11: The Gini Index is used to get the best split in binary tree.

Step 12: The Gini Index for True Positive (TP) is given as $Gini(TP) = \sum_{i=1}^{TES} TP$
Step 13: The Prediction Accuracy for the tested samples are calculated as follows

\[ \text{Prediction Accuracy} = \frac{\text{True Positives}}{\text{Total Size}} \times 100 \]

Step 14: Display the prediction accuracy of identification of records affected with Cervical Cancer

Step 15: Stopping Criterion, Exit

END

The Prediction algorithm creates a split by comparing the predicted value with the original value in the dataset. The classification And Regression Tree (CART) is applied with the specified algorithm to represent graphically the tree structure which might be used in pruning of data. In general, the splitting criterion helps the algorithm to identify the right testing samples for final execution that will produce utmost accuracy in prediction results.

3.5.3 Strength of CART Algorithm and Limitations

The CART algorithm is structured and comprises of a sequence of questions for which answers are predicted. The answer from one node again nodes to the next question and the process continues until all the data are being classified and represented in a tree like sequence. The terminal node is the node that has no more questions and hence the decision is arrived at that point. In cervical cancer data, the terminal part will be the identification of cancer or non-cancer cells in the cervical region of women.

It is a well-known fact that biological data always required best algorithm to find accurate results to predict the outcome of any potentially harmful and deadly disease like cervical cancer. Many researches proved that efficiency of an algorithm in prediction of a disease has to be based on the accuracy of results obtained. Hence the efficiency of CART algorithm needs to be accessed so that it can be further recommended for analysis or rejected in the beginning stage itself.
The Classification and Regression Technique (CART) algorithm in the proposed model had the following advantages

(a) **CART algorithm doesn’t need variables to be selected and defined in advance:**

To identify the cervical cancer in women, a huge database from NCBI was selected. However, no variable is identified earlier to initiate the results. The CART algorithm automatically selects the most significant variables and eliminates non-significant variables. This stage was tested in the research conducted in this chapter and the variables were selected easily and isolated from non-significant cervical variables.

(b) **CART algorithm has no assumption and operates on available variables very fast:**

In the current research to identify cervical cancer, the field values in the table are given as input directly without further refinement and result obtained at a very faster rate compared to other systems.

(c) **CART can easily handle outliers or noisy data with accurate separation of proper data from corrupted data:**

NCBI is a global database which is supposed to have relevant and non-relevant noisy data. Hence CART algorithm plays a pivotal role in segregating the corrupted or noisy data [73] from the NCBI database through regression analysis and classifies data in an effective manner.

(d) **The resultant of the CART Algorithm is invariant to monotone transformations of its independent variables:**

In NCBI database, various fields are subjected to changes during execution. But the accuracy of the resultant tree was not affected due to the changes in those fields with change as logarithmic or square root values respectively.

(e) **CART algorithm is always flexible and adjusts or modifies itself according to the changes over time:**

The algorithm focuses on the NCBI data to analyze the
presence of cervical cancer in women. It was identified that even if any changes are
tertained in the future, the algorithm can adjust to values spontaneously without
forther changes to original data. The classification tree was found to be good with
no changes in the organizational levels of the cervical cancer data.

After careful analysis of the best features of CART algorithm, a proposal to employ this
algorithm to identify presence of malignant or benign tumor in the cervical region of women
is always possible.

Various drawbacks were identified in the CART algorithm during the analysis of results
in NCBI data.

(a) Resultant CART tree for cervical cancer has unstable decision trees: Though
the accuracy level is over 80%, the resultant decision tree does not give accurate
decision trees. It is also identified that few fields are also modified which results in
loss of information. After testing, it is found that several observations have to be
eliminated as CART selects its own variables during classification.

(b) CART algorithm used in cervical cancer splits the huge data based on a single
variable: Hence the resultant data is not expected to give the expected outcome if
the selected variable is not efficient to give accurate results. CART can easily
handle the splits but cannot produce reliable information to predict cervical cancer
exactly.

(c) CART algorithm couldn’t give a clean accurate result of 100% in predicting
cervical cancer in women: With reference to the first level test of NCBI dataset
using CART, the result is 83.87% accurate but it needs more accurate results as the
patients’ needs more accurate and exact predictions to have a trust in the
methodology.
3.5.4 Proposed Prediction Model for Cervical Cancer

A simple diagrammatic decision tree is proposed for cervical cancer prediction of cancerous and non-cancerous cells during the research with CART algorithm as shown in Figure 3.8. The main elements of CART are:

- Rules for splitting data at a node based on the value of one variable;
- Stopping rules for deciding when a branch is terminal and can be split no more; and
- Finally, a prediction for the target variable in each terminal node.

![Diagram of Proposed Decision Tree Using CART Algorithm to Predict Cancer Cells From Normal Cells](image)

Figure 3.7 A Proposed Decision Tree Using CART Algorithm to Predict Cancer Cells From Normal Cells

Based on the Decision tree sequences and predictions, the following model shown in Figure 3.8 is proposed for predicting cervical cancer in Women using CART Algorithm.
Figure 3.8 The Proposed Framework Model Using CART Algorithm to Predict Cancer Cells From Normal Cells
The Proposed Algorithm is scripted below that has to be tested with cervical cancer dataset GDS3233:

**Algorithm CCDTPREDICT** (Generate a Decision Tree from Training and Testing set of data to predict the Cervical Cancer)

**Input**
Global variable cervdata; // which contains the entire NCBI Dataset of GDS3233
Global variable classnames; // that holds the identifier for dataset
Global variable fieldnames; // which holds the column names of biopsy values.
Global variable ctree; // displays the graphical output of trained sets
Global variable traindata; // loads the training dataset from 61 biopsy values
Global variable trainclassnames; // reads classnames of training set values
Global variable testdata; // loads the testing dataset from 61 biopsy values
Global testclassnames; // reads classnames of testing data
Global tp; // True positives of results

**Output**
Global res; // stores the prediction tested values of cervical data
Global prate; // stores the computed prediction accuracy of cervical cancer

**Method:**
cervdata = load_dataset('GDS3233.xlsx'); // loads the cervical dataset GDS3233 into a variable
classnames = read_column(cervdata);
fieldnames = read_Identifier(cervdata);
traindata= cervdata ([1 to 9] && [34 to 44] && (10-19));
initialise i=1;
for (i=1; i<=9; i++)
    trainclassnames[i]=classnames[i];
for(i=34;i<=44;i++)
    trainclassnames[i]=classnames[i];
for(i=10;i<=19;i++)
    trainclassnames[i]=classnames[i];
load ctree; //loading the CART tree
view(ctree,’mode’,’graph’); // view the CART Tree
tree=apply classificationTree(traindata,trainclassnames);
testdata=[cervdata(20:33),cervdata(45:61)];
j=1;
for(j=20;j<=33;j++)
testclassnames[j]=classnames[j];
for(j=45;j<=61;j++)
testclassnames[j]=classnames[j];
load Testdataset;
tp=0;
for(i=1;i<=testdata;i++)
{
}
trec=1;
tdata=testdata[trec];
res=predict(ctree, tdata);
if(res==testdata[trec])
    tp=tp+1;
else
    print “Misclassification data”;
}
prate=(tp/testdata)*100;
print “Prediction Accuracy” + prate;
}

The NCBI dataset with Gene Index conditions and Biopsy GENEID samples are inserted into CART Algorithm. The proposed model of CART for cervical cancer accepts mRNA’s TPM Value given for implementation with PROB-ID features. The CART algorithm then splits the data into trees with root value. If dataset is blended with overlapping components, best split is achieved with Gene Index values. The decision tree is analyzed with disease status to find the GAS7 value. The GAS7 value is tested to check if it is Normal or Cervical cancer cells. The algorithm efficiency is tested and implemented using Data mining tools in a successful manner.

3.5.5 Implementation of Proposed Model using MATLAB Tool

In this research work, the methodology to predict cancer cells, MATLAB 2013 tool will be utilized. MATLAB is a fourth generation high-level Programming Language that is created by MathWorks is used as an interactive environment for numerical computation, visualization and programming.

There are few common characteristics of MATLAB as given below:

- It creates and gives an interactive environment for iterative exploration, design of products and solving complex problems.
- As it is designed for Mathematical Computations, it gives a rich set of mathematical functions for all complex mathematical and statistical problems.
• It supports user with built-in graphics for visualizing data and tools that can create user customized plots.
• It also have a good programming interface of development tools that can enhance the quality of code and maximizing performance [74].
• It provides tools for building applications with custom graphical interfaces.
• It has provisions to integrate MATLAB based algorithms with external applications and other languages like C, Java, Dotnet and MS Excel.

The MATLAB Integrated Development Environment (IDE) can be launched from the bin directory of the installed directory. The IDE contains the following components as indicated in the Figure 3.9. MATLAB development IDE can be launched from the icon created on the desktop. The main working window in MATLAB is called the desktop. When MATLAB is started, the desktop appears in its default layout.

![Figure 3.9 Various Panels in MATLAB](image)

MATLAB allows writing two kinds of program files —
- **Scripts** – script files are program files with `.m` extension. In these files, we can write series of commands, which can be executed together.

- **Functions** – functions files are also program files with `.m` extension. Functions can accept inputs and return outputs. Internal variables are local to the function.

The M Files can be created using MATLAB or any other editor and then scripts are created and written based on the created algorithm. After creation and code, the M file of the implementation is Run using the command prompt by specifying the filename or right-clicked to select the Run from the pop-up menu respectively. List of M-files are,

1. DemoCancerDetect.m
2. DemoCancerDetectfull.m
3. DemoCancerDetectRF.m
4. DemoCancerDetectRFfull.m
5. DemoCancerDetectRFfullkmeanslearning.m

The list of m-files is used in proposed work. These MATLAB script files loads GDS3233.xls database using load cervical data. This excel file contains 61x22,283 cell and records.

- First consider the data source processing contain 350 records used in database for a sample, select 250 training dataset and 100 testing set as shown in Table 3.2, 3.3.
- Apply CART algorithm on the NCBI dataset for the finding the root and leaf of the tree by using GINI index splitting criteria.
- Perform the same operations to each attribute up to last split and create decision tree.
- Each leaf stores a continuous-valued prediction and it is the average value of the predicted attribute for the training tuples that reach the leaf Regression and model
trees tend to be more accurate than linear regression when the data are not represented well by a simple linear model.

- Pruning method is used on the decision tree for better accuracy.
- In 350 records, 250 for training sets and 100 for testing set

Table 3.2 Sample Training Set Values

<table>
<thead>
<tr>
<th>S. No</th>
<th>Data Value</th>
<th>S. No</th>
<th>Data Value</th>
<th>S. No</th>
<th>Data Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>71</td>
<td>21.</td>
<td>308.2</td>
<td>41.</td>
<td>315.7</td>
</tr>
<tr>
<td>2.</td>
<td>0.5</td>
<td>22.</td>
<td>313.4</td>
<td>42.</td>
<td>422.3</td>
</tr>
<tr>
<td>3.</td>
<td>680.7</td>
<td>23.</td>
<td>211.4</td>
<td>43.</td>
<td>246.4</td>
</tr>
<tr>
<td>4.</td>
<td>865.7</td>
<td>24.</td>
<td>346.7</td>
<td>44.</td>
<td>141.9</td>
</tr>
<tr>
<td>5.</td>
<td>657.5</td>
<td>25.</td>
<td>307.6</td>
<td>45.</td>
<td>331.4</td>
</tr>
<tr>
<td>6.</td>
<td>426.8</td>
<td>26.</td>
<td>298.5</td>
<td>46.</td>
<td>577.6</td>
</tr>
<tr>
<td>7.</td>
<td>560.7</td>
<td>27.</td>
<td>343.4</td>
<td>47.</td>
<td>316.7</td>
</tr>
<tr>
<td>8.</td>
<td>310.3</td>
<td>28.</td>
<td>214.1</td>
<td>48.</td>
<td>187.2</td>
</tr>
<tr>
<td>9.</td>
<td>325.8</td>
<td>29.</td>
<td>216.1</td>
<td>49.</td>
<td>416.4</td>
</tr>
<tr>
<td>10.</td>
<td>391.2</td>
<td>30.</td>
<td>230.8</td>
<td>50.</td>
<td>272.7</td>
</tr>
<tr>
<td>11.</td>
<td>261.6</td>
<td>31.</td>
<td>322.2</td>
<td>51.</td>
<td>369.2</td>
</tr>
<tr>
<td>12.</td>
<td>344.9</td>
<td>32.</td>
<td>321.8</td>
<td>52.</td>
<td>327</td>
</tr>
<tr>
<td>13.</td>
<td>170.1</td>
<td>33.</td>
<td>262.5</td>
<td>53.</td>
<td>250.4</td>
</tr>
<tr>
<td>14.</td>
<td>204.9</td>
<td>34.</td>
<td>307</td>
<td>54.</td>
<td>311</td>
</tr>
<tr>
<td>15.</td>
<td>314.3</td>
<td>35.</td>
<td>307.9</td>
<td>55.</td>
<td>371.5</td>
</tr>
<tr>
<td>16.</td>
<td>189.6</td>
<td>36.</td>
<td>360.1</td>
<td>56.</td>
<td>426</td>
</tr>
<tr>
<td>17.</td>
<td>326.6</td>
<td>37.</td>
<td>544.3</td>
<td>57.</td>
<td>233.7</td>
</tr>
<tr>
<td>18.</td>
<td>249.6</td>
<td>38.</td>
<td>200.2</td>
<td>58.</td>
<td>326.1</td>
</tr>
<tr>
<td>19.</td>
<td>497.3</td>
<td>39.</td>
<td>338.3</td>
<td>59.</td>
<td>253.3</td>
</tr>
<tr>
<td>20.</td>
<td>408.3</td>
<td>40.</td>
<td>181.3</td>
<td>60.</td>
<td>277.1</td>
</tr>
</tbody>
</table>

Table 3.3 Sample Testing Set Values

<table>
<thead>
<tr>
<th>S. No</th>
<th>Data Value</th>
<th>S. No</th>
<th>Data Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>13.6</td>
<td>11.</td>
<td>17</td>
</tr>
<tr>
<td>2.</td>
<td>9.4</td>
<td>12.</td>
<td>10.6</td>
</tr>
<tr>
<td>3.</td>
<td>43.2</td>
<td>13.</td>
<td>15.3</td>
</tr>
<tr>
<td>4.</td>
<td>2.1</td>
<td>14.</td>
<td>9.1</td>
</tr>
<tr>
<td>5.</td>
<td>9</td>
<td>15.</td>
<td>31</td>
</tr>
<tr>
<td>6.</td>
<td>1.4</td>
<td>16.</td>
<td>21.2</td>
</tr>
<tr>
<td>7.</td>
<td>8.6</td>
<td>17.</td>
<td>15.4</td>
</tr>
<tr>
<td>8.</td>
<td>15.1</td>
<td>18.</td>
<td>4.2</td>
</tr>
<tr>
<td>9.</td>
<td>11.9</td>
<td>19.</td>
<td>31.4</td>
</tr>
<tr>
<td>10.</td>
<td>22.6</td>
<td>20.</td>
<td>9.4</td>
</tr>
<tr>
<td>S. No</td>
<td>Data Value</td>
<td>S. No</td>
<td>Data Value</td>
</tr>
<tr>
<td>-------</td>
<td>------------</td>
<td>-------</td>
<td>------------</td>
</tr>
<tr>
<td>21.</td>
<td>14.1</td>
<td>31.</td>
<td>58.7</td>
</tr>
<tr>
<td>22.</td>
<td>22.9</td>
<td>32.</td>
<td>21.4</td>
</tr>
<tr>
<td>23.</td>
<td>5.5</td>
<td>33.</td>
<td>28</td>
</tr>
<tr>
<td>24.</td>
<td>10.9</td>
<td>34.</td>
<td>29.7</td>
</tr>
<tr>
<td>25.</td>
<td>18.2</td>
<td>35.</td>
<td>69.1</td>
</tr>
<tr>
<td>26.</td>
<td>9.1</td>
<td>36.</td>
<td>23.2</td>
</tr>
<tr>
<td>27.</td>
<td>21.6</td>
<td>37.</td>
<td>32.8</td>
</tr>
<tr>
<td>28.</td>
<td>56.3</td>
<td>38.</td>
<td>37.6</td>
</tr>
<tr>
<td>29.</td>
<td>39.5</td>
<td>39.</td>
<td>48</td>
</tr>
<tr>
<td>30.</td>
<td>37.1</td>
<td>40.</td>
<td>15.7</td>
</tr>
</tbody>
</table>

The proposed methodology is implemented in MATLAB 2013 and the design is constructed. The initial empty screen interface will be displayed on the MATLAB IDE as shown in the Figure 3.10, which is narrates the Initial screen shot before loading the data. We have the following buttons, (i) Load Dataset to load the GDS3233 Dataset (ii) Train Dataset to train the dataset (iii) Load Test Patient Data for testing the Dataset values (iv) Record number used to indicate the single record number (within sample limit) to test the patient status with test patients data button.

![Figure 3.10 Home Page of Proposed Model](image)

Figure 3.10 Home Page of Proposed Model
The Figure 3.11 describes, (i) Load Dataset- used to load the GDS3233 Dataset. It contain 61 biopsy values and 22,283 records. The Figure 3.12 explains the Training Process with random dataset sample. After loading the data values, Train Dataset button is clicked. After the training, regression and decision tree output is arrived indicating the completion of Training.

Figure 3.11 Loading the Training Dataset GDS3233

Figure 3.12 Training Phase of the Proposed System
The Figure 3.13 shows the status and vision after loading the Test Patient data for testing a single record. The Figure 3.14 discuss how the process of a single patient record is tested and the result is given as either cancer cervix or normal cervix. Specifically record number 10 is tested. The result shows that correctly decide the status for this particular record.
The Figure 3.15 discuss how the process of a single patient record is tested and the result is given as either cancer cervix or normal cervix. Specifically record number 6 is tested. The result shows that not correctly decide the status for this particular record.

![Figure 3.15 Processing of a Single Record](image)

**Figure 3.15 Processing of a Single Record**

The Figure 3.16 narrates the picture of testing the entire dataset as whole to test the dataset values of 61 biopsy features with 22,283 records. The Figure 3.17 shows the picture of after loading the data set to test the prediction for the entire 22,283 records with 61 attributes. The Figure 3.18 describes the Training over status after train all the records with binary tree output. The Figure 3.19 portrays the status of loading the Patient data for testing purpose. Then with the following Code execution of prediction is finalized and the prediction accuracy output will be displayed. The sample MATLAB coding for the above binary tree classifier CART algorithm is listed in appendix –I.
Figure 3.16 Home Page for the Process of all Records

Figure 3.17 Screen Shot of the Loading the Dataset GDS3233
Figure 3.18 Screen Shot of the Status of the Training Given to Records

Figure 3.19 Screen Shot of the Loading the Test Data
3.5.6 Solution of Cervical Cancer Prediction Model using CART

As discussed in the implementation model for cervical cancer, the prediction accuracy is calculated by using the formula:

\[
\text{Prediction Accuracy} = \frac{\text{True Positives}}{\text{Total Size}} \times 100.
\]

After careful testing methods applied to the datasets, the following prediction accuracy is obtained as result from CART algorithm. When we compare the actual values with the predicted values, we found that the prediction accuracy of CART is 83.87% accurate as shown in Figure 3.20 and Figure 3.21.

![Figure 3.20 Result of Database After Implementation](image)

![Figure 3.21 Status of the Prediction](image)
The final stage involved writing code in MATLAB 2013 version. After completion, a good prediction accuracy of 83.87% is achieved using CART Tree output with prediction of either normal cervix or cancer cervix shown in Graph 3.1.

Accurate prediction of occurrence of cervical cancer has been the most challenging and toughest task in medical data mining task in medical data mining because of the non-availability of proper dataset. Many researchers have been done to develop different techniques that can solve problems and improve the prediction accuracy of cervical cancer through images. But in our research work, the prediction of cervical cancer is with Numerical Data. NCBI (National Center for Biotechnology Information) data set has been used. However, no such system is designed exclusively till now for cervical cancer prediction with datasets.

Our research has described the prediction of cervical cancer in two stages i.e. Benign or Malignant of women with data mining algorithms, with reasonable accuracy. This dissertation describes a finite, well defined numerical dataset which is well suited for cancer
prediction. The CART algorithm is found to have its own limitations as directed in the steps followed for predicting cancer cells in the cervical region of women.

3.6 Summary

Thus Chapter III envisions design and implementation of CART Algorithm in prediction of Cervical Cancer and its implementation steps in MATLAB13. The Cervical cancer dataset used for this model is portrayed clearly along with its source of origin, structure and description. We performed preprocessing and data transformation to result in normalized dataset, as only valid dataset will yield accurate predictions.

This Chapter discussed about different perspectives of data mining and its techniques. The other data mining algorithms namely SVM, Naïve Bayesian, ID3 and Apriori are narrated briefly with its limitations.

The predicted results were compared with the actual values and its accuracy percentage is determined. The Models, thus constructed are validated based on the validation and performance operators and its output is obtained. From this, we found that CART algorithm prediction accuracy for Cervical Cancer in this research work is not that much accurate enough to predict the optimal solution. Hence another algorithmic technique using Random Forest Tree (RFT) will be discussed in Chapter IV.