CHAPTER – II

REVIEW OF LITERATURE

2.1 Introduction

This Literature review is an assessment report on information found in the literature related to this research work namely cervical cancer. These reviews help us to describe, summarize, evaluate, substantiate and clarify the literature. This gives a theoretical base for the research and helps the authors to determine the nature of the research and it acts as a road map. The Literature survey of this research proposal inculcates current status of cervical cancer research along with different methodologies used, its findings from various authors with substantial outputs and results.

The Literature review provides the context for the research and justify our research. Our literature review on Cervical Cancer ensures that this research hasn’t been done before or that it is not a replication study. It clearly enabled us to learn from previous or existing theory on the subject of Cervical Cancer. It illustrated how the subject has been studied previously and highlight the flaws and outline gaps in previous research.

Chapter II reviews the literature about Supervised Models like Classification And Regression Tree (CART), Random Forest Tree (RFT), Naïve Bayesian, C4.5, ID3 Algorithms and also on Unsupervised Models like K-Means algorithms, that are the Data mining techniques used for Cervical Cancer prediction. It also reviews literature on various other methodologies used in prediction of cervical cancer cells. Again we reviewed on Data mining tools, exclusively used for Cervical Cancer prediction, its features, advantages and disadvantages. A thorough survey has been made on different cancer related research papers and bio-medical journals, which discusses about the implementation techniques and its experimental results with its limitations.
To predict the occurrence of Cervical cancer in women, various factors are identified, the opinions of medical practitioners, Scientists and researchers who developed various architectures, models, simulated systems, prototypes for reasonably accurate prediction have been carefully studied. Few of such literatures are summarized and highlighted in the oncoming paragraphs.

(i) Cancer Research
(ii) Techniques for Cervical Cancer Research
(iii) Data Mining Tools used on Cervical Cancer prediction
(iv) Data Mining Algorithms for Cervical Cancer prediction

2.2 Review on Cancer Research in India/Abroad

Prediction of Cervical Cancer is a major challenge for the scientific and medical community. This modeling involves a combination of computer models, observation and knowledge of Cancer predictions and data patterns. Using the above stated methods reasonably near accurate prediction can be done. A Cervical Cancer prediction model involves observation from different medical sources and knowledge of past and current cancer trends and patterns. Using these methods and with expertise of medical practitioner’s sharp knowledge, reasonably accurate prediction can be done.

A Literature review on prediction identifies what is already existing. It’s a research journey with several stages:

1. Framing a Research Question – What is required?
2. Searching Relevant bodies of literature – Justifies our Query
3. Synthesizing the research literature – validating our proposed work
4. Writing an assessment of the literature – a clear vision of our work
The following are the various literatures reviews on cancer contain in eight research papers:

**Mahmoudi et al. (2013)** [21] compared four classifiers on six datasets which used the selected genes by GA and PSO methods in wrapped model. The biggest challenge was there were more adequate noisy data. He used SVM, CART and ANN algorithms for Cancer classification.

**Dr. DSVGK Kaladhar et al. (2013)** [22] proposed a model to analyse the imbalance problem in predicting cancer cells from breast, embryo, leukemia etc. The thickness, marginal adhesion along with other factors determined the gene association with cancer cells.

**Shon, Ho Sun et al. (2007)** [23] analyzed and classified leukemia patients using open data. The gene used in the experiment were 7,129 in total 72 samples, 38 training data and 34 test data to analyse minority genes.

**Vikas Chaurasia1et al. (2014)** [24] presented an enhanced diagnostic system for detecting breast cancer based on Rep Tree, RBF Network and simple logistic respectively. Yugoslavia database was used to evaluate the proposed system accuracy and performance level and was found to be 74.5%.

**Frank Emmert-Streib et al. (2006)** [25] predicted an innovative method to select genes from DNA microarray data. Sample Genes were selected and tissues of a certain tumor stage are represented. The tissues were compared and found correlated with cancerous tissues. The top rated tissues of the cells were identified to be cancer cells with tumor.

**Tanzeem Khan, et al. (2014)** [26] Feature Selection is done by implementing the techniques like genetic algorithm and SVM classification for Cancer detection.
R.K.Kavitha, et al. (2014) [27] analyses the performance of naïve Bayesian and C4.5 Algorithm in predicting the survivable rate of breast cancer patients. At the end of analysis, C4.5 proves better performance and SEER Dataset is used for analysis.

Shikha Agrawal et.al (2015) [28] conducted a Survey on Neural Networks used in prediction of cancer in human body to develop a cost effective, user-friendly software. Two algorithms are used: FLANN which showed an accuracy of 63.4% whereas PSO-FLANN showed an accuracy of 92.36% respectively.

2.3 Review on Techniques for Identification of Cervical Cancer

National Cancer Institute, India in 2009 gives the definition for Cancer as “to a class of diseases in which a cell or a group of cells divide and replicate uncontrollably, intrude into adjacent cells and tissues (invasion) and ultimately spread to other parts of the body than the location at which they arose (metastasis)”. The Cervical Cancer is defined as “a disease in which the cells of the cervix become abnormal and start to grow uncontrollably, forming tumors”.

When this survey was done, we found that in cancer prediction, 80% of authors have taken only PAPSMEAR Tests with Image Processing as the prime attribute to predict cervical cancer. The remaining authors used HPV tests, Fuzzy K-Means etc., to conduct their research.

The various research works floated by various authors on techniques for cervical cancer is furnished here under.

Abhishek Das et al. (2011) [29] revealed the right identification of cervical tissues in the cervical region that contains cancer cells. The novelty of the research gives automatic identification of tissues within the cervix. It also analysed the significant image regions of cervical cancer that gave accurate results with a dataset of 210 normal cervigrams and 42 acetowhite cervigrams.
Das et al. (2011) [30] proposed an image pre-processing method to remove specular reflection and detect ROI for further detection. The major task was digital colposcopy combined with recent advances. The dataset contains 210 normal cervigrams and 42 acetowhite cervagrams.

Dr. L. B. Mahanta et.al (2012) [31] in their work proposed an approach to analyse and find abnormal cervical cells based on nucleus/cytoplasmic (N/C) ratio that is considered as one of the most important morphological features to distinguish normal and abnormal cells in the cervical region. The research used MATLAB to analyse the PAP smear images. The research analysed 100 cervical cells where 50 of them are normal and 50 of them are abnormal based on the papsmear test conducted among them. The conclusion of abnormal ratio was made based on the N/C ratio indicating normal cells for less than 0.05 and abnormal cells for greater than 0.05 respectively.

Kavishvar Kathy et.al (2012) [32] proposed A Clinical Decision support with automated text processing for Cervical Cancer screening. In this paper cervical cancer screening is used to interpret free text papanicolar reports are prepared for 73 patients the method showed two cases identified with diseases. The challenge was into propose the model for more than a single institution at the same time. Future implementation is recommended through natural language processing technique.

Anas Quteishat et al. (2013) [33] presented a Neural Network based system for classifying cervical cells as normal, low-grade squamous intra-epithelial lesion and high-grade squamous intra-epithelial lesion. The cervical cells are segmented using Adaptive fuzzy moving k-means clustering algorithm in first stage, feature extraction process in second stage and classification using fuzzy min-max in the third stage respectively.
Jyotismita Talukdar et.al (2013) [34] presented in his research paper that FCM Clustering algorithm is most advantageous than hard clustering methods to predict cervical cancer cells using Papsmear images. The FCM technique is applied on Papsmear images with three validity measures as partition coefficient, partition entropy and compactness and separation function respectively. In addition, a Shape analysis test is carried out to correlate the clustering result to clinical paradigms of Papsmear images. The conclusion of the research revealed that drawbacks are negligible compared to the various advantages of FCM Clustering algorithm in detecting cervical cancer.

G.SankaraSubramaniam et al. (2014) [35] proposed an automated method with genetic algorithm to test the diabetic dataset with biochemical parameters of cervical cancer. The dataset was found to be more reliable and also difficult at the same time as it produced a miss classification error rate of 26.20% and there are lots of missing values in it.

R.Vidya and Dr. G. M. Nasira (2014) [36] insisted the need for computerization in the early detection of cancer cells in the cervical region of women and also on the importance of data mining algorithms to achieve accuracy of results.

P.Priya et al. (2014) [37] suggested an image segmentation method to observe the lesion in cervical cancer. Three algorithms Watershed segmentation algorithm, k-means clustering segmentation algorithm, expectation maximization segmentation algorithm are used to find the results. As a conclusion, expectation maximization segmentation algorithm has higher error rate of 11.7787 to detect the abnormal region in cervix after diagnosis.

S.B. Park et al. (2003) [38] suggested a solution to classify the false negatives from false positives of HPV during diagnosis. The results after experimentation showed high performance with regards to improvement in the HPV sequence database.
Derek Christopher Johnson et.al (2014) [39] studied the awareness of HPV and cervical cancer among women in Nepal region. It was found that the knowledge and awareness remained very low among them. It was also identified that women showed interest in acceptance of freely available HPV vaccine for children among those communities provided they get adequate information sharing about cervical cancer and HPV knowledge is provided to them.

Chandra J et.al (2015) [40] researched to handle the drawbacks of cancer research practice in India. The statement was proved that machine learning algorithms can be a better solution for identifying cancer cells anywhere in the body of a human being. After identification of Cancer cells, screening and testing will be done.

Ashwini B. Barbadekar et.al [41] proposed a research to compare the accuracy of SVM and ANN to identify the cervical cancer affected cells using Papsmear tests and other methodologies. After analyzing the classification results, the accuracy of SVM is found to be 92% whereas the accuracy of ANN is found to be 84% respectively.

Devi H.K.A (2006) [42] formulated several methods used for cervical cancer and discussed different classification techniques, support vector machines, fuzzy based techniques and texture classification to differentiate the normal and abnormal cancerous cells with image processing.

2.4 Review on Data Mining Tools Used for Detection of Cervical Cancer

The Cervical Cancer prediction is carried out with number of data mining tools from various authors as studied in the review of literature. However, Image Processing was the major area of study in those predictions and staging. Hence, in our research, we focused primarily on data sets obtained from biopsy features. To conduct this research successfully, the following data mining tools can be extensively used as follows:
The Data Mining Tools and predictions from various research papers are studied and submitted below:

**Márquez, Edna, et al. (2009)** [43] proposed a multi agent system for gene expression analysis with agents specialized in user interaction, data preprocessing, gene identification, tumor classification and database management. They used 116 genes includes in the similar ranking, the first 15 genes up regulated and the 5 down. The main proposal was the separation procedure for the platform agents to allow them to modify, add or delete or do statistical methods without affecting functionality.

**Santhanam.T, et al. (2010)** [44] used classification algorithm to study blood donor behavior using data modeling techniques implemented with WEKA. This research is conducted with original blood transfusion dataset and a new decision tree algorithm is predicted with CART. The experimental results predicted regular voluntary donors behavior with much accuracy.

**Agarwal, Subhash M., et al. (2010)** [45] manipulated a research proposal that validated various genes from the gene database. The research also created a catalog of such genes that is expected to be present in different stages of cervical carcinogenesis. They have complied 537 genes in their CCDB. The biggest challenge was bringing up the microRNA altered in cervical cancer.

**Horng, Jorng-Tzong, et al. (2012)** [46] idea that epistasis or gene – gene interaction is important in human biology. The dataset was provided by Tri-service general hospital by
applying four decision tree and Bayesian network. The dataset includes 720 records with 18 variables and hybrid dataset contain 238 records. The challenge was the dataset which remains insufficient for further analysis.

**Sinha, Vrushali Y et al. (2013)** [47] highlighted the prediction of unknown primary tumors through dataset. Multiclass classifier with random forest was used for classification. They used 18 attributes with 339 instances. SMOTE method is used for imbalanced dataset using WEKA tool. The major factor of research was enhancing automation of primary tumor prediction.

**Mohamad Saraee et al. (2004)** [48] conducted a survey and found that patient information system holding patient’s referral and medical history, cervical examinations can be used in associative rules mining, clustering and classification respectively. The classification and statistical analysis was applied and the result based on frequencies of attributes appeared.

**Denil, Misha, et al. (2014)** [49] set a goal to narrow the gap between theory and practice of RFT. They derived a new random regression forest algorithm. For this experiment, 2000 poses were taken for training and 500 poses for testing with Bagging and Boosting Techniques.

**Latha, D. Sowjanya, et al. (2014)** [50] made a comparative study with multiple classifiers to identify the stages of cervical cancer with dataset of 221 records which provides us a knowledge on various data mining models that are considered suitable for accurate prediction. The best accuracy was achieved by J48. The biggest challenge was predicting the best in sensitivity and specificity analysis.

**Harpreet Kaur et. al. (2014)** [51] CART with Genetic algorithm used to reduce the complexity from decision tree indexes with Pruning. Author explained how CART algorithm provide refined result based on cased fetched after applying genetic algorithms on it.
Vidya, R., and Dr. G. M. Nasira (2015) [52] presented a method using CART and RFT algorithm to predict the Cervical Cancer of Women with high accuracy. Cervical Cancer dataset is applied and implemented with MATLAB’13 for prediction.

2.5 Review on Data Mining Algorithms Used for Cervical Cancer Prediction

A Data mining algorithm can be identified as a set of heuristics and computations that can create a data mining model from the given data sets. Hence, creating a model employs the analysis of data that is provided through algorithm, patterns that can be used to implement the model according to research criteria. The Algorithm uses this analysis as result to find optimal parameters for the research problem. The parameters thus obtained can be applied to data sets and results extracted in the form of patterns that have detailed statistics.

Hence this data mining algorithmic technique is highly suitable for Cervical Cancer Prediction with higher accuracy which is evident from the oncoming review of literature from various authors.

Pabitra Mitra et al. (2000) [53] researched a combination of algorithm to detect and predict the different stages of cervical cancer. The result was proposed based on hybrid combination of genetic algorithms and data mining algorithms like ID3 algorithm. The novelty of the research employs a restricted mutation operator with network tuning of weights from genetic algorithms.

Kutiannan Thangavel et.al. (2006) [54] Investigated a problem in their study on how to support a physician’s decision of whether a biopsy is warranted or not by analyzing an existing method using clustering techniques (k-mean) using 50 attributes and the challenge was discovery of high risk groups.

Nester Jeyakumar et.al. (2012) [55] Improved Classifier performance through GA for cervical cancer prediction SVM and CART techniques are implemented. The research used
biochemical parameters of Cervical Cancer as dataset that is collected from Institute of obstetrics and gynecology, Egmore, Chennai. The samples from unidentified women genes from age 20 to 80 are used as test cases and the algorithm proved efficient with an improved accuracy of approximately 97%.

**Sinha, Vrushali Y et al. (2013)** [56] proposed a disjoint partitioning approach with random forest classifier using 6 datasets with various instances. It is trained using base decision tree. The greatest challenges were dataset imbalanced in nature and have binary classification.

**Delen, Dursun, et al.** [57] proposed a decision support system using image segmentation, feature extraction applying adaptive fuzzy K-mean. They used 500 cases. The main task was to classify the dataset.

**Lokanayaki, K, et al. (2013)** [58] modified the existing system of how data mining algorithms are utilized to overcome diagnosing disease. The major problem was with bio inspirational based techniques.

**Jaree Thongkam et al. (2013)** [59] studied various machine learning techniques like Support Vector Machine (SVM), C4.5, CART and K-Nearest Neighbor (KNN) to eliminate the outliers and factors affecting the ability to predict the survival of patients in 5-year cervical cancer survivability. The experiments with effectiveness of Adaboost, bagging and Bayesisan proved that data filtering with CART gives optimized results.

**Govind Maheswaran et. al. (2013)** [60] A comparative study of three algorithms K-means clustering algorithm, Enhanced K-means with Harmony Search K means to implement the Lloyd’s algorithm. For better time complexity Enhanced K means algorithm is good and for better accuracy Harmony search algorithm is suitable.
2.6 Summary

The effective study on literature survey helps the authors to redefine the problem or even change the research criteria as per the need. Thus one can find that a journey on research starts with a study on literature review or a survey and ends with the clear vision of the proposed work, its methodologies and its findings.

The Literature review process is iterative as the more we do, the more knowledge we can gain. Thus the literature review enlightens the researcher to frame the right question for the research to be conducted on the research area.

Thus in this Chapter II, a clear picture on the existing techniques of Cancer Prediction and how our proposed cervical cancer prediction model is supported by this literature survey has been pointed out clearly. Also, it gaudily describes about the literature survey on various supervised data mining algorithms like Classification And Regression Trees (CART), Random Forest Tree (RFT), Support Vector Machine (SVM), Naïve Bayesian, ID3, C4.5 and unsupervised K-Means Techniques for Cancer Prediction. A brief literature survey is discussed on other Cancer Prediction models. Based on the above literature survey in Chapter II, the research gap is analyzed and the proposed research work is finalized.