CHAPTER II

LITERATURE SURVEY

Extensive studies have been made on the early diagnosis and treatment of cervical cancer with reference to demographic factors, clinical data, genomics, imaging data and so on. Classification, clustering, association mining, artificial intelligence, fuzzy logic and so forth have been applied on cervical cancer data for better diagnosis, treatment and predictions. All these studies have been made in view of the increase in epidemic of the cancer.

Data Mining has been applied in various domains for different kinds of analysis and for pattern recognition. Data mining has been applied on the e-governance data as presented in [4] and for various other kinds of analysis. There is a need for identifying efficient data mining algorithms with very high accuracy in the health care domain as the cost involved is very high. This has been discussed broadly in [90], and has come with a comparative study that talks of the necessity of designing efficient data mining algorithms in order to decrease the cost of health care services.

A new feature selection algorithm Fast Correlation Based Feature Selection was developed to be applied on high-dimensional data. Fast filter method is used to identify relevant and redundant features. Comparative study made with other methods using high dimensional data and found to be effective and efficient as presented in [58].

Patterns identified from the database of cervical cancer patients using clustering techniques. The purpose of the study was to discover patient groups with high risk of the disease, identify the most influential factors in the diagnosis of the disease and identifying treatment in that region. K-means clustering algorithm has been applied on the data to discover relevant patterns. Different clusters were generated according to the risk of the disease which helps the medical professional to identify the right treatment proposed by Thangavel in [57].

The paper in [54] talks about integrative computational biology that has been used to comprehensively analyze and interpret biological data and thus helps in diminishing the challenges to high throughput cancer biology. This approach has been applied on the integration of data, analyzing the network, applied on the databases. Broadly this approach had a tremendous impact in identifying new genes for cancer biology.

Computational tools developed to detect and classify the cells of the transformation cervix zone. The zone is characterized using color and texture descriptors in MPEG-7 standard, independent of the particular shape thereby achieving higher precision rates as indicated in [64].

Patterns for cervical cancer are identified based on demographic, environmental and genetic risk factors that are associated with the disease according to *Seung Hee Ho* in [105]. Comparative analysis made on logistic regression and Chi Squared Automatic Interaction detection (CHAID) decision tree algorithms and finally proposed using induction technique for analysis of risk factors for the proper management of cervical cancer.

A hybrid system was designed for detecting the different stages of cervical cancer with the help of Genetic algorithms based on rough set theory and the ID3 algorithm. Sub networks were generated initially using the concepts of rough set theory and the ID3 algorithm. These sub networks are then integrated to form the final network using the genetic algorithm. The performance has been enhanced which is measured in terms of the classification accuracy, the size of the network and the time taken to train the network as compared to MLP has been proposed by *Pabitra Mitra* in [84].

The processes of cervical cancer screening from sample collection to classification have been identified. The report gives a very broad overview of what tools to be used for each step of the screening process and usage of computational tools for centralized analysis of the samples given by *Patrik Malm* in [86].

Artificial neural networks have been applied to generate the network of nodes and decision trees have been used along with a statistical method, logistic regression, on breast cancer dataset (comprising of 200,000 cases) to develop prediction models for the survivability of breast cancer. 10-fold cross validation has been used as a measure to estimate the prediction models for performance analysis. *Dursun Delen* in his paper in

[20] indicates that C5 decision tree has been found to be the best predictor followed by artificial neural networks.

A filter that performs faster for removing the removing the redundancy on modified Kolmogorov-Smirnov Class Correlation based Filter has been designed utilizing the class label information while comparing the feature pairs. The purpose of this algorithm is in reducing significantly the initial space created in wrapper-based feature selection problems for high-dimensional problems say *Marcin Blachnik* in [65].

MLP with back propagation and Radial Basis Function (RBF) Neural Network have been applied for classifying date fruits, define a set of external quality features from the shape and color for different types of date fruits for testing the effectiveness of neural network models for image classification. RBF performed better than MLP neural network as given by *Khalid Alrajeh* in [48]. Bayesian decision theory applied for automated detection of atypical cells in a cervical pap smear. Two most suitable decision rules have been generated for classification of the atypical cells say *Oliver* in [82].

A new method in classification of cervical cancer developed to decrease the processing time while maintaining the accuracy of detection. Before classification the image is optimized and labeled which will help in accelerating the segmentation and classification process proposed by *Bustanur Rosidi* in [10].

Novel data selection methods have been developed for selection of data to be trained with SVM. The first method is based on confidence measure, a statistical confidence that enables that data to be selected if it is part of the largest network/group indicating that it belongs to that class and none other. The second method is basically used for non-separable plane, select the data that has the smallest Hausdorff distance. These methods have been applied on breast cancer, liver cancer datasets and found to have good performance according to *Jiang Wang* in [36].

Pap smear images have been classified using global color and texture features using MPEG-7 descriptors. Once the cells are identified classification techniques like K-Nearest Neighbor, SVM have been applied. *Luz Helena Camargo Casallas* [64] analyzed images to find out the presence of HPV as the causative virus.

Feature selection is one major challenging problem in any domain. Identifying the most influential attributes is a very critical task. Kolmogorov-Smirnov Class Correlation-based filter has been developed to remove redundancy at faster pace. This enables the reduction of the initial space applied in wrapper based feature selection, which is suitable for high-dimensional problems say *Marcin Blachniki* [65].

High risk Human Papilloma Virus (HPVs) identified by computing the F-score measure using AdaCost technique. This helps in identifying the design to be followed for DNA-chips in order to diagnose the presence of HPV among cervical cancer patients. This approach reduces time and monetary cost according to *Seong-Bae Park* [106].

A text-based Clinical Decision Support System for processing cervical cancer has been designed. NLP has been used to develop the system. Rule-based and Guideline rule-based text processing techniques have been designed. Given the pap smear results and the guidelines for interpreting the results this system acts as a tool for the physician in making a decision and treating the patient accordingly *Wagholikar KB* [118] says.

Several learning algorithms have been compared and analyzed for their performance. No single algorithm has outperformed the others. In different scenarios algorithms performance was varying. One can conclude saying that the performance of the algorithms is dependent on the domain, volume, nature of the dataset, presence or absence of noise in the dataset. Accuracy of C4.5 was high on the unpruned tree; whereas neural networks scaled even in presence of noise. Similar studies have been performed to give a bird’s eye view on the type of algorithms available over different domains in the paper presented by *Shravya Reddy Konda* [110].

Several classification models have been applied on liver disorder dataset. Neural networks, Rough sets, Decision tree algorithms have been compared. One striking conclusion from this study is the accuracy of the performance of the algorithms varies along with the size of the dataset. Neural networks have outperformed the others and its performance is enhanced with the increase in the size of the dataset. Multi Layer Perceptron suits well for a large dataset according to *Peiman Mamani Burnaghi* [87]. Breast cancer data has been analyzed using different classification techniques on large scale data. Bayes network classifier has projected highest accuracy and took less time for computation. This study made by *Mohd Fauzi Bin Otham* [75] concludes that Bayes network is suitable for medical domain.

Comparing these studies one can conclude that no algorithm is suitable over all domains. But MLP has higher performance compared to others because of its scalability.

Picking the right features in order to classify the dataset is important to design a good classifier. C4.5 has embedded in it the feature selection process while constructing the decision tree. Contextual merit algorithm is another feature extraction algorithm. This identifies the distance between the instances in the Euclidean space and picks those features which are nearest. Correlated and irrelevant features reduce the performance of the algorithm. Pre-selection of the features enhances the performance of the algorithm proposed by *P Perner et al* [88].

Symmetrical uncertainty (a measure of information gain and entropy) is used as a goodness measure for selecting the features with high correlation or similarity. This is more applicable to high dimensional data says *Lei Yu et al* [59]. It is an inexpensive solution consuming less amount of time for its analysis.

Information gain and fuzzy rough sets have been applied on demographic data related to cervical cancer. Using these approaches the risk factors that lead to the cancer have been identified. From the reduced set, rules have been extracted using fuzzy rough sets that demonstrate the risks of the onset of the disease. In this study *Kuzhali Vandar, et.*al [49] found out that a person having multiple sex partners and with HPV is risky of the cancer. One limitation in this study is that HPV is causative virus but a person is infected with the virus only after contracting the disease. So HPV cannot be the risk factor rather helps in confirming the presence of the cancer. In this thesis work has been done in this direction to identify the risks of the cancer as a means to evade the cancer.

Age, gender, education have been identified as the parameters for evaluating the risk of cancer among the South Indian women. J48 algorithm has been used to perform this analysis. This would be useful analysis to know the factors for the onset of the disease so that the government can take precautions to curb the spread of the disease.

This has been studied and researched by *P Ramachandran et.al* [96]. With this nature of study it would be very easy to educate the people on the factors related o cancer and hence eradicate the onset of disease by molding the habits of the people.

A review on the prevalence of cervical cancer among Indian women, impact of HPV vaccination and cost effectiveness, risk factors, age-wise incidence of the cancer has been made in the article by *Kaarthigeyan* [41]. The impact of administering HPV vaccination at a very early age, between 15-26 years, was very good. Different types of HPV vaccines are available both for male as well as female, but vaccines for male are not being administered except for Australia. Economically, HPV vaccination is not preferred in developing countries. Low cost solution being worked for an easy access to the vaccination.

Cancer prevalence and deaths among women are high among rural and illiterate as per the study made by *Rajesh Dikshit, et.al* [95]. An extensive study has been made on the cancer mortality in India. The analysis of the data concludes by showing the mortality rate among the rural illiterate women is high compared to the other factors. Age-specific and region-specific study reveals this fact. Hence this research is focused on cancer among women specifically cervical cancer as this is related to the economical, cultural, environmental conditions prevalent in an area.

Data mining techniques would be helpful in analyzing this data in order to prevent and diagnose the cancer at an early stage. Awareness programs are a means of creating awareness among the public in order to reduce the frequency of occurrence of the disease. Cervical cancer mortality rates are very low in developed countries because of the continuous awareness programs compare to the developing countries like India. The causes, symptoms, diagnosis, treatment modalities for cervical cancer have been studied, analyzed and compiled by *Jyotsna A Saonere* [40].

Labeled colour intensity, a method to pre process the images before classification for higher accuracy and low processing time proposed in paper by *Bustanur Rosidi, et.al* [10]. Pap smear is a screening test for diagnosing the presence of cervical cancer among women. A study in Africa found a correlation between the cases where cancer was suspected through visual inspection and the same being confirmed through Pap smear test in a paper presented by *Mbamara SU* [73].

HPV is one the causative virus for the spread of cervical cancer. HPV exists in different forms and it can be either high or low. Each form of HPV leads to a specific type of cervical cancer. Diagnosing the level of HPV form is another area of research. Kernel based Support Vector Machine (SVM) has been used for classifying forms of HPV as either high or low. High-HPV is risky for treatment, hence the necessity of detecting the type of HPV according to *Su Kim et al* [111]. HPV protein sequences were trained to identify the level using SVM.

MRI images analyzed to identify the invasion of parametria and the lymph node involvement in identifying the stage of the cancer. This analysis was performed on the pre-treated images. There was less involvement of lymph node and parametria in the pre-treated MRI images. Cost effectiveness of MRI to be evaluated if it has to replace the clinical staging system for cervical cancer say *Hyun Hoon Chung* [34]. Gene expression data of Lukemia patients has been preprocessed using mining techniques.

Discover and Mask, a knowledge discovery method applied on the microarray data to extract the genes that are considered to be informative. The genes with specified accuracy and matching the threshold have been picked up as the ones with high information. This method did yield a higher result in terms of the precision, recall and F-measure in the study made by *Fazal Famili* [22].

Soft computing methodologies applied on uterine cancer data in order to create a medical decision support system for better cancer management. Neural network, Genetic algorithm, rough set theory applied on the dataset to come up with a better model for decision support. With rough set theory and ID3 algorithm reduced model is obtained which is refined with the help of Genetic Algorithm. This integration aids in building an enhanced decision support system which would be a useful tool and an aid for identifying the stage of the cancer. The parameters that are closely associated with each stage of the cancer is picked with the help of this model. This is a very efficient process for rule extraction, but complicated system to develop and use say *Sushmita Mitra* [112].

Support Vector Machine (SVM) based feature screening method has been applied on the pap smear images to extract the features which would be closely related to the risk of the disease. Global relevance measure is computed for the boundary point values generated from SVM. These features are then ranked. Statistical significant test is applied on these ranked features to select the subset of features. This method developed by *Jiayong Zhang* [39] when applied on large dataset has generated good feature subset.

Cervical cancer can be detected using different techniques like pap smear, cytology, HPV detection, visual inspection, cervicography etc. Most of these techniques are not feasible to be implemented because of the cost involved. *Krishnakumar Duraiswamy* [53] talks about the different techniques and the need for developing a tool which can be easily used by the paramedical personnel to reduce the mortality rate. It gives a broad view for the necessity of developing computer based medical support systems that would be useful to the society in common.

Biochemical parameters like sugar, albumin, creatinine etc have been compared with cancerous and non-cancerous patients. This comparison has been made with SVM's and Genetic Algorithms (GA). The results projected those parameters which are closely associated with cervical cancer. Based on the presence and levels of the parameters a person can be diagnosed having the cancer. This would be a useful study to analyze the factors leading to the cancer according to *Nester Jeyakumar* [80].

Mark W Craven [70] has come up with an algorithm that deduces very comprehensible models for extracting rules from the trained neural network, which is very difficult to understand. This is coined as Trees Parroting Network (TREPAN) as it generates a tree based model for a better understanding of the network. The trees generated by this model are comparatively more accurate than the conventional tree based algorithms but the built is complex. They can be applied on high dimensional input spaces. This model emancipates the low understandability of a neural network and hence can be used extensively over any domain.

MLP and Kohonen's net have been used for rule generation and querying to infer decisions in *Sushmita Mitra's* thesis. This has been applied on speech, medical and synthetic data and found out to be effective models. The effectiveness of neuro-fuzzy model has been studied in depth to come up with comprehensible models with high accuracy and applicability. Max-min and product-probabilistic sum have been modeled on the conventional back propagation algorithm. Fuzzy implication operators like *and* and *or* have been employed on the network to incorporate mutual interaction for error propagation. This has also been applied on speech and medical data like cervical cancer to come up with precise rules for pattern recognition. The thesis concludes that fuzzy MLP is more suitable for complex feature spaces like medial domain.

Rule extraction algorithm designed from pruned neural network by *Rudy Setiono* [100]. This is applied on a three-layered feed forward neural network. The main goal of this work is to retain only small number of inputs connected to a hidden unit; thereby the accuracy can be enhanced for extracting the rules. With high predictive accuracy rates this algorithm is able to extract reasonably compact set of rules.

Fidelity-accuracy dilemma on rule extraction from neural networks has been addressed by *Zhia-Hua Zhou* [122] in order to address the issue of quality of the rule framework. Fidelity, accuracy, consistency, comprehensibility are the four parameters on which the rule extraction algorithms are evaluated. To address the issue of rule extraction using neural networks ACC (Accuracy, Consistency, Comprehensibility) framework would be suitable else the FCC (Fidelity, Consistency, Comprehensibility) framework is suitable for rule extraction for the neural network. Applying the right framework for the algorithm is necessary as this derives the quality of the algorithm.

Few citations have been made comprising of the current research that has been made. Only the striking works have been reviewed.