CHAPTER 9

Accomplishments of the Objec ves and Future Scope

9.1. Introduc on

Numbers of experiments, statistical analyses, and simulations on available data sets have been attempted, using different tools to achieve our research goal. Efforts have been made to address the purpose and focus of our study along with basic problems as set by our research goal in each chapter of the thesis. Research aims provide direction towards exploring, investigating, and analyzing various problems of interests and also the way to achieve it. This is one of the most important features of designing and implementing research work. Different steps to solve specific problems are clearly defined in chapters of the thesis.

9.2. Accomplishments of the Objec ves

During past few years of research work our intention was to come into the conclusion to achieve the target as mentioned in section 1.3 of chapter 1. Attempts have been made to explain the present states of the domain to realize the need of the proposed system. To fulfill the needs and to accomplish our objectives, we have developed intelligent automated models to diagnose diseases like breast cancer, cardiac arrhythmia, and hepatitis disease. We also tried to develop an intelligent system to diagnose kidney disease []. Various sets of features for the said diseases were proposed during analysis to diagnose the said diseases. The proposed systems may be helpful to doctors as it may assist them in making decisions. Even the system may be used for teaching as well as training purpose.

Statistics in relation with the present scenario of diseases such as breast cancer, cardiac arrhythmia, and hepatitis disease have been presented in global as well as local perspective. Efforts have been made to study all useful information in relation with these problems. We have also investigated mortality rate, reasons, and diagnosis procedures for the said diseases. We have discussed different issues as related to present healthcare services, which demands the need of an alternative healthcare management. These discussions ultimately inspired us to fix up our

objective for developing intelligent systems by using available information [Chapter 2 for information].

After determining the objectives we studied various soft computing techniques, which are used in the domain of disease diagnosis. In this context fuzzy expert system (FES), genetic algorithm (GA), and artificial neural network (ANN) are discussed in chapter 3. Previous approaches to implement these techniques in the said domain are specified in tabular form. Different forms of ANN architecture and various activation functions are also discussed. Training or learning algorithms such as incremental backpropagation learning network and Levenberg – Marquardt algorithm are also presented. From our study it was revealed that hybridization approaches such as combination of ANN with GA, ANN with FL (fuzzy logic), and FL with GA shows better performance in many studies as limitations of one may be removed by others.

Data mining does not mean the steps in the data analysis only. Data preprocessing is also regarded as one of the data mining steps, which is needed to make sense of data. Recently, the most challenging task is to extract right knowledge from huge raw data. Real world data or raw data are noisy, inconsistent, and incomplete. As the primary step to accomplish our objectives, different steps in the process of data preprocessing are explained in chapter 4. Good quality data preparation is one of the important issues for the development of valid and reliable intelligent models. Data cleaning, data integration, data transformation, and data reduction are the main objectives of data preprocessing. Data cleaning can be accomplished by removing noise, redundancy, and inconsistencies in the data set. Missing values in the data set should also be taken care of in the data cleaning phase. Data are noisy due to the presence of outliers, which can be taken care of by using different methods, such as: binning, clustering, and regression. Sometimes data are collected from different sources, which may include multiple databases or flat files. These files should be integrated before analysis. In many cases data are transformed to make it convenient for data analysis. Different methods to extract features from the data set for dimensionality reduction are also presented. Missing value handling techniques have been analyzed and implementation of the suitable technique has been made as illustrated in subsequent chapters. In chapter 4 initiatives have been taken to consider data mining approaches to obtain patterns in the data set, which have been integrated with the decision support system to extract useful information. All of these data preprocessing steps have been taken care of during the development phase of automated decision support system to make it valid and reliable and also to avoid ambiguous results.

Different performance prediction parameters, such as – precision, correct classification accuracy, sensitivity, specificity, area under ROC curve, Kappa statistic, lift, cumulative gain, and probability threshold, are suggested in literatures to judge the performance of a decision support system before using it. Performance prediction parameters are metrics that evaluates model's capacity to minimize errors between predicted and observed values based on test data set. All of these methods along with their usage and limitations are studied in chapter 5. From our study it has been revealed that the selection of performance prediction parameters depend on data type as well as on type of model. Generally, the whole confusion matrix that is precision, accuracy, sensitivity and specificity of an intelligent model should be judged to evaluate the performance of the model without the knowledge of distribution of data. The area under ROC curve is also an important predictive measure to depict the performance of a model.

For the last few decades different attempts have been made in designing automated systems to solve classification problems in the domains as give below:

- classification of plant diseases
- diagnosis of crop management problems
- financial applications such as evaluation of credit and authorization
- in the field of computer science such as diagnosis of problems in hardware and software
- to detect problems in electrical, mechanical, and electronic machinery
- medical diagnosis

In the subsequent chapters we focused on one application domain of intelligent system namely, medical diagnosis. Attempts have been made to develop models that accept a set of observations or symptoms as input and respond as a human expert. Aims of our work are exploration and evaluation of various innovative approaches in developing such systems that can be implemented successfully.

Now-a-days breast cancer has become a global burden. Moreover, most of the general hospitals do not have infrastructures which are required to diagnose the disease. Delayed diagnosis may increase the risk spreading the cancer cells. A number of steps have been explained thoroughly in chapter 6 to develop automated intelligent system to diagnose breast cancer, which helps in reducing time required to diagnose the disease as well as the mortality rate due to this cause. A survey on various approaches to diagnose breast cancer using different machine learning

algorithms or methods is also summarized in this chapter. Data preprocessing steps on UCI breast cancer data set were described precisely in order to produce good quality data. For dimensionality reduction of the data set some feature selection techniques such as correlation-based feature selection (CFS), rough set (RS), and principal component analysis (PCA) were implemented. Corresponding artificial neural network architecture was chosen based on some evaluations. Finally, incremental back propagation learning network (IBPLN) and Levenberg – Marquardt (LM) algorithms have been used for classification. Different combinations such as CFS + LM, RS + LM, PCA + IBPLN, PCA + LM were attempted. Experimental results from 120 simulations for the combinations CFS + LM and RS + LM and 100 simulations for the combinations PCA + IBPLN, and PCA + LM were tabulated. At the end results were compared with the previous similar studies in terms of various performance prediction parameters such as correct classification accuracy (CCR), sensitivity, specificity, and area under ROC curve (AUC). From this work it was revealed that the combination PCA + LM shows better performance in terms of accuracy, sensitivity, specificity, and AUC. The highest performance was 100%. The lowest and average performances were 96.08% and 98.55% respectively for the combination PCA + LM. The results were comparable with previous similar studies. This work proposed to consider PCA derived extracted features before taking final decision by doctors. Most of the previous works did not mention about the results they obtained, whether it is derived from the best simulation or it is derived by averaging results of several simulations. For this reason the highest, lowest and average behaviors of decision support system were also proposed to be judged before using it.

Neuro-genetic fusion was attempted for extracting important sign, symptoms, and features from the UCI cardiac arrhythmia data set as well as to improve the level of accuracy of the intelligent system. Classification performance may be affected for huge data set. Dimensionality reduction is needed by selecting appropriate attributes to enhance the predictive performance of classifiers. The main advantage of genetic algorithm (GA) is that it is useful for optimization when the data set is huge. Our study presented four combinations namely, GA + IBPLN, GA + LM, CFS + IBPLN, and CFS + LM i.e. features are selected using genetic algorithm and correlation-based feature subset selection (CFS) method and incremental backpropagation learning network (IBPLN), and Levenberg – Marquardt (LM) algorithms were used as classifier. Different data preprocessing steps and fixing up artificial neural network architecture were made to improve the performance of the classifiers. Experimental results, represented in tabular form, showed that the predictive performance is better for the combination CFS + LM. The highest, lowest,

and average classification accuracies for the combination CFS + LM is 95.12%, 80.49%, and 87.71% respectively. The sensitivity for the combination CFS + LM have reached 100% in a significant number of cases. Results were compared to the results of some recent works and it is evident that the proposed system provides better result in terms of correct classification accuracy. The work proposed to consider reduced feature subset, obtained by using CFS, before making decision by doctors. Since lowest performance of a disease diagnosis system should be a judging factor for evaluating the predictive performance, we proposed to judge the highest, the lowest and the average behavior of the automated disease diagnosis system before implementation [discussed in chapter 7].

Different soft computing methods to develop intelligent decision support system for diagnosing hepatitis disease are precisely explained in chapter 8. UCI hepatitis disease data set contains only 155 instances with a number of missing values. So, missing values in the data set were taken care of to avoid distorted analysis as well as to avoid biased result. Multiple imputations using EMB approach was implemented on the data set, which produces five completed data sets. Multiple imputations method was chosen as it is efficient for multivariate data set. Important features are selected from the data set by using correlation-based feature subset selection (CFS), and rough set (RS) techniques, implemented each of the 5 imputed data sets. For each of the said methods (CFS and RS) the common features, extracted from all imputed data sets, are considered as reduced feature subset for the corresponding method. Two different approaches have been attempted.

Firstly, all of the imputed files are analyzed using classifiers namely, incremental backpropagation learning network (IBPLN), and Levenberg – Marquardt (LM) algorithms. Different combinations, attempted in this approach, were represented as:

- CFS + IBPLN i.e. CFS is used for feature selection and IBPLN algorithm as classifier
- CFS + LM i.e. CFS is used for feature selection and LM algorithm as classifier.
- RS + IBPLN i.e. RS is used for feature selection and IBPLN algorithm as classifier
- RS + LM i.e. RS is used to select feature and LM algorithm as classifier.

For each of the said combinations the results were combined to obtain the final result. Experimental results have been presented in tabular form. From the result it

is evident that the combination RS + LM showed better performance in terms of accuracy, sensitivity, specificity and area under ROC curve.

Secondly, an integrated file has been generated from the 5 imputed data sets by using some assumptions as described in section 8.5. of chapter 8. IBPLN and LM algorithms were used for classification. In this approach we chose only the reduced features subset as obtained by implementing RS method. Two combinations, which have been attempted, were represented as:

- RS + IBPLN i.e. RS is used to select features and IBPLN algorithm is used for classification.
- RS + LM i.e. RS is used for feature selection and LM algorithm as classifier.

Experimental results for 100 simulations have been presented in tabular form, which revealed that the combination RS + LM for the later approach, where classification has been done on integrated file, performed better in terms of accuracy, sensitivity, specificity and area under ROC curve. A survey on similar studies is also summarized in this chapter. The highest, the lowest, and the average classification accuracy are: 100%, 86.96%, and 94.61% respectively, which outperforms all other methods. A set of features have been proposed to consider which would be important to doctors in decision making. It was also suggested to consider the lowest performance of intelligent model for disease diagnosis. Most of the previous literatures did not clearly specify about the results they obtained, whether it was the best simulation product or average of several simulations. For this reason, the highest, lowest and average behaviors of the proposed model have been presented in our work. At the same time it was also proposed to consider the highest, lowest, and average performance of a decision support system before using it.

Data sets, used in our study for classification, are free from outliers. Data preprocessing steps, one of the most important steps for the development of intelligent models, are precisely mentioned in each chapter. Uncertainty is a major issue in disease diagnosis. For this reason rough set theory (RST) was used to manage uncertainties effectively. Different methods have been deployed to find out important features from the data set as well as to reduce the dimension of the data set. The reduced data set improves the efficiency of the classifiers as it reduces the search space as well. Artificial neural network architectures for different cases have been fixed up after considering different factors such as number of inputs, number of outputs, and previous studies. Improved data mining approaches such as IBPLN and LM algorithms have been used as they are able to perform for nominal, ordinal, and real-valued attributes, which matches with the characteristics of medical

domain. These improved data mining approaches helped in achieving objectives of our work.

Our proposed models did not claim the highest performance achievers, but at the same time it was suggested to consider various observations that have been proposed, which would be worthwhile in making final decisions. The suggested intelligent systems may assist doctors during diagnosis of diseases. In addition it provides diagnosis results faster, which would be helpful for medical professionals. As a result early diagnosis of diseases and better treatment management can be done, which is necessary for augmenting survivability.

In order to fulfill our study we considered the best means of our objectives. An enormous effort has been provided to complete the objects successfully as well as to achieve the goals. We have consulted with physicians, studied various soft computing techniques used in medical diagnosis, analyzed different techniques used to handle missing values in the data set statistically, studied various performance prediction parameters, surveyed previous research works, analyzed other works, performed more than 2500 simulations (minimum 100 simulations for each combination for each data set) to obtain experimental results, made observations and comparative study.

9.3. Future Scope and Further Works

The objectives of our study is to model intelligent systems to diagnose diseases such as breast cancer, cardiac arrhythmia, and hepatitis disease using soft computing techniques as mentioned in earlier chapters. The target is to compare the outcomes with golden standard. Though we accomplish our objectives successfully, still we must say that the proposed systems will not completely replace the knowledge of domain experts rather it can be used as effective and reliable tools to assist health professionals in making decisions while diagnosing the diseases.

The successful implementations of soft computing techniques in related studies demand that impact of it will be growing rapidly in coming years. It plays an important role in designing decision support system, which can be considered as an important tool as well as the most demanded tool in various fields. At present, the use of such system is not compulsory rather it is used as a tool voluntarily. Therefore, further scope of this work is to improve the performance of the proposed systems to make it more acceptable. As a result it may replace domain experts as well as may be useful as a compulsory tool in the field of medical diagnosis.

In future we will be concentrating to design Web-based decision support system which has the ability to produce instantaneous diagnosis report for the diseases such as breast cancer, cardiac arrhythmia, and hepatitis disease based on certain desired symptoms or laboratory findings. The system may support both national and regional languages to interact while using the system.

Another future scope of our work in broader sense is to implement these intelligent systems into mobile applications. Acute scarcity of domain experts, modern equipments such as computers and associated peripherals, and proper maintenance of these equipments are major issues in the field of medical domain especially in underserved or rural areas. Most of people are now using mobile phones efficiently. Therefore people in general can easily access the facility of the system. As a result early diagnosis of diseases and better management for the treatment can be done immediately which are keys for survival of a patient.

Uncertainty in medical domain is a big issue, which may cause incorrect and distorted analysis while diagnosing diseases. It is observed that different diseases may have some common symptoms. Again, a disease may have more than one symptom. Sometimes prominence of secondary symptom may hide the original making the diagnosis process more complicated. Moreover, a patient may be suffering from more than one disease making the process of diagnosis to be difficult and confusing. Soft computing techniques possess the ability to handle uncertainty, imprecision, partial truth, and approximations efficiently. It is a science of reasoning, evaluations, and thinking, which can be used to build conceptual intelligences in machines. The purpose of soft computing techniques is to model cognitive behavior of human brain. For this reason the techniques used in our study may be implemented to diagnose other diseases.

From our work it is evident that Levenberg – Marquardt algorithm performs better as a classifier to diagnose diseases as mentioned. Future scope of our study is to identify the characteristics of the data set, based on which one can choose suitable classifier or classifiers to obtain better performance.

Another scope is to incorporate the decision support system into robot. In coming days domain experts can be replaced by robots. Fusion of such decision support systems and robot may be useful as it may compensate the scarcity of domain experts in related fields.