Chapter 7

CONCLUSION AND FUTURE SCOPE

7.1 Contribution to This Research Area
7.2 Conclusions
7.3 Future Scope
The thesis concludes with this chapter and contains the summary of the research findings and the inferences drawn from the research, the discussion about the answers provided to the research questions asked. Future scope for the potential research is also discussed in which work can be extended.

7.1 CONTRIBUTION TO THIS RESEARCH AREA

1. A thorough review of literature pertaining to our field of research was carried out in order to find out previously used best classifiers for model building. It was found that the model built by (Patil et al., 2010) had the best accuracy of 92.3% by using decision tree classifier. In this phase it was also deduced that research has been done on Pima diabetic dataset.

2. We developed a model using a primary dataset that was collected afresh from superb diagnostic center in Srinagar. The dataset had all the necessary parameters that are critical of diabetes and our dataset overcome flaws of other dataset like:
   - Pima dataset is a very old dataset collected in 1988 but appeared online in 1994.
   - The dataset didn’t contain vital parameters of HbA1c and waist thickness measures.
   - The dataset contains the only half measure of blood pressure viz. diastolic while systolic pressure is the integral part of the measure.
   - The dataset had the records of only female population, while as males too suffer from diabetes equally as appear on WHO bulletin from time to time.

3. We developed our model on dataset on all major machine learning classifiers and compared them for accuracy in order to establish the best five classifiers.

4. Using the identified best classifiers we improved the prediction accuracy of classifiers by addressing the problem of class imbalance using SMOTE.

5. We also developed a novel prediction model that is based upon Non-clinical parameters. The aim here is to design a predictive model which does not require a patient to undergo blood tests and diabetes is predicted using only non-clinical parameters. This experiment is done keeping in view that all
patients are not well off to perform the costly blood tests especially in developing and poor nations. It is the first attempt to predict diabetes using only non-clinical parameters. The results obtained are coordinal with an accuracy of 85.74%.

7.2 CONCLUSIONS

In the beginning of this research we had set some of objectives for research work, the main aim of this work was to design a prognostic model for diabetes prognosis which can aid physicians in decision making. The research began with the analysis of literature of both data mining and application of data mining in healthcare domain. After thoroughly analyzing the literature available, for the proposed solution we zeroed on few top data mining classifiers which have been thoroughly used in literature for the prognosis of diabetes the said techniques are recommended and have been proven successful for medical diagnosis in the available literature.

The models were built with experiments that were carried out on clinical dataset which was collected from authentic diagnostic centers under the expert medical supervision. Dataset contained 734 instances with eleven attributes. The models were built using Weka 3.8 machine learning software. During the course of study we carried out several experiments on the dataset using these data mining classification techniques. At first we conducted experiment on unbalanced dataset to select the top five classifiers based on their accuracy; apart from accuracy we also compared the performance of the developed models using other statistical values such as Precision, True Positive Rate, and False Positive Rate and ROC area.

After selecting the best model among the three basic classifiers, as part of study was to improve the performance accuracy of the models. During thorough study and analysis of literature it was found that the dataset are often imbalanced and there is always bias for majority classes in dataset, the classification results always tend to favor the result inclined towards majority class. So we removed the class imbalance from the dataset to make dataset bias free as class imbalance has been identified as a major lapse in diagnostic datasets. Another set of experiments
were conducted using the identified top classifiers. So as part of another experiment, we applied class rebalancing algorithm SMOTE on our dataset so as to end bias and again the best model *viz* decision tree was selected for the purpose, the prediction accuracy of the model got improved after class rebalancing, the obtained accuracy for the final model was 94.7% with an outstanding ROC area of 0.953. The developed model using decision tree and SMOTE can be differentiated from prior ones for the following reasons:

- A real life dataset was used for study.
- Imbalance was reduced by applying SMOTE rebalancing algorithm to enhance prediction accuracy of decision trees.
- The model we presented is a generalized model that can be used for both male and female population and has all primary factors as defined by International Diabetes Federation.
- We applied rebalancing algorithm for removing class imbalance as it was rarely done in previous studies which enhanced the Capability of the decision trees for diabetic prognosis.

The research work also developed a novel model for diabetic prognosis. The model is built with selected attributes, we removed all those attributes that required the patient to undergo painful blood tests and the model was built using completely non-clinical parameters. The developed model using bagging technique had an accuracy rate of 85.8% which is acceptable for a model, the model also performed well in ROC having value of 0.946. The developed model is special in an approach as it doesn’t necessitate lab parameter tests for prognostic purposes and predicts the disease with decent accuracy rate.

The performances of the models were evaluated using the standard. 10-Fold Cross Validation was adopted for randomly sampling the training and test data samples. All models performed well in predicting diabetic cases. The most effective model to predict patients with diabetic disease appears to be a Decision Tree-SMOTE classifier with a classification accuracy of 94.7013% after rebalancing the dataset; it was the enhancement in accuracy prediction from simple decision tree model.
This study established that data mining techniques can be used efficiently to model and predict diabetic cases. The outcome of this study can be used as an assisting tool by physician to help them to make more consistent diagnosis of diabetes. Furthermore, the resulting model has a high specificity rate which makes it a handy tool for junior doctors to screen out patients with high probability of having the disease and transfer those patients to senior doctor for further analysis. This study has established that data mining techniques are very much applicable in the prediction of diabetes disease and the resulting models of this study are worthy of clinical testing.

7.3 FUTURE SCOPE

The developed models could be considered for further enhancements as there is a possibility for expansion of the same if for some reason it s believed that some other classifier is better suited, we could compare it with our current models.

Researchers may further add more instances to our dataset which might be purposeful as strong results could be achieved with larger dataset. We can add some more attributes to dataset that also have some relevance to diabetes prediction take for example for our model with clinical parameters we can add lipid levels of blood which also has some relevance with diabetes, for model with non-clinical parameters we add some life style related parameters like intake of junk food or if a person takes daily exercises or not.

The model can also be enhanced by applying other classifiers that can predict the vulnerability of diabetic patient to other allied diabetes triggered disease like heart disease, renal disease etc. and patient can be forewarned about the complications and can be asked to change the lifestyle habits to lessen the complications of disease.

Our model can predict only type II diabetes. Researchers may further enhance this model for type I diabetes also We have used WEKA tool, researchers may further use some other data mining tools Scikit, Rapid miner, R, Orange, and Knime.