Chapter 2
LITERATURE REVIEW

2.1 Application of Data Mining in Healthcare Domain

2.2 Data Mining and Diabetes : Literature review

2.3 Review and Analysis

2.4 Conclusion
The chapter contains the detailed overview of the application of data mining in healthcare domain with regard to its available literature. A thorough review of literature has been provided regarding use of data mining techniques in prediction of diabetes.

2.1 APPLICATION OF DATA MINING IN HEALTHCARE DOMAIN:

Devoid of any issue, the healthcare services system are in a great need of data mining today as it was just about a decade back. Researchers had put numerous arguments in support of data mining in healthcare sector. According to (Kob \textit{et al}, 2005), “the application of data mining in healthcare system is becoming more and more popular, if not essential”. They stated the example of healthcare insurers using data mining for reduction of medical insurance fraud and abuse. As applicable in healthcare fraud detection, the data mining can be applied in the commercial world; take example of fraudulent credit card transactions. As indicated by recent reports data mining has been successfully applied for healthcare fraud and abuse detection.

Stating another reason that necessitates the need of applying data mining in healthcare sector is magnitude of data that gets generated throughout the healthcare transactions. The data generated is extremely complicated and capacious that conventional methods can’t analyze and process it.

As per (Biafore, 1999) the employing of data mining for analysis of such an enormous quantity of data is more and more becoming crucial for the reason that financial anxiety has heightened the necessity in favor of taking decisions which are based on analysis of financial and clinical data.

Numerous other factors have been given that have boosted the attractiveness of data mining in health care domain, one such factor amongst them is non-invasive diagnosis and decision support. There are several diagnostic and clinical procedures that are invasive, expensive and painful to patients for instance of conducting a biopsy in women to identify cervical malignancy.

(Thangavel, \textit{et al}, 2006) Used k-means clustering intended for investigation of cervical cancer patients. It was accomplished that clustering had superior
prognostic results than existing medical opinion. A set of fascinating attributes were recognized which doctors can use as additional support for making decision regarding recommending a patient suspected of having cervical cancer for biopsy or not.

Healthcare system is a data concentrated progression, as simultaneously many processes are running and fresh data gets generated after every second. With largest consumers of public fund it has an intensive research field. As computers and new algorithms have evolved to a big extent, there has been a boost in the application of computer and its emerging tools in the healthcare industry. These emerging technologies has led to amalgamation of healthcare and computing ensuing in formation of health informatics.

Defining data mining in healthcare system (Peyman M, et al, 2013) explained it as a significant data collection of different disease. Healthcare and medicines are one of the most essential sections in industrial society.

The explanation of healthcare informatics has been given as the evolution of scientific discipline dealing in collection, storage, repossession, communication and finest use of healthcare allied data, information and knowledge. The study of this field is applied to medical care, nursing, public health and biomedical research, these all are devoted for the enhancement of patient care and public health. Some applications of data mining are:

- To predict healthcare cost.
- To find out the treatment of disease.
- To identify and forecast the disease of any kind.

In healthcare sector data mining is one of the fastest expanding areas moreover it covers an enormous range of applications and research. By applying available smart algorithms and machine learning intelligence, we can make available quality healthcare services using a variety of problem solving and decision making systems. In healthcare domain we can characterize decision support system as a system which is based on knowledge and supports information based system and assists in decision making activities.
According to (Pragnyaban M, et al, 2012) the availability of the clean healthcare data is the pivot for the success of healthcare system. Some areas where data mining is applied in healthcare are:

- For uncovering of deception and misuse by healthcare insurers.
- Customer relation management (CRM) decisions are taken by healthcare organizations.
- It is used by physicians for identification of efficient treatments and most excellent clinical practices.
- Better, more efficient and more affordable healthcare services are provided to patients.

Below are some of the challenges that data mining has in the medical domain:

- To dig out knowledge and provisions of scientific decision making ability for diagnosing and treatment of disease.
- The detection of patterns of successful medicinal therapies for different ailments.
- At the present time too many disease markers are available.
- A mammoth amount of data is currently being collected in the form of texts, graphs, and images.
- Behavior of noisy, inconsistent and incompleteness of medical data to be processed.

The existing applications of data mining in healthcare domain are:

- Prognosis: prediction of future outcome based on preceding data and present condition.
- Therapy: selection of treatment methods presently on hand keeping in sights, its effectiveness and suitability to patients.
- Diagnose: the classification and recognition of patterns in multivariate patient attributes.

2.2 DATA MINING AND DIABETES: A LITERATURE REVIEW.

Data mining is proving to be an incredibly advantageous in the field of medicinal data analysis, as it increases diagnostic accurateness, furthermore tends to
lessen the expenditure of patient treatment and saves human resources. Researchers have found a significant amount of curiosity in diabetic data mining and took it as apt disease to applying technology because of a variety of reasons.

- First reason being that the disease is very common and is found in every region and in every age group.
- Second there is a huge amount of data available.
- Third the disease afflicts high expenditure of treatment and the effect of this disease is on increase and till date there has been no permanent cure for the disease regardless of mammoth amount of money spend on its study and research.
- Fourth the diabetes can lead to tremendous amount of complications, so the medical practitioners may possibly be acquainted with how to raise and advance the outcomes of disease as much as possible. In such state of affairs data mining is the finest to go with. Prevention of diabetes at present is an area of attention in healthcare community.

The majority of research works done on diabetes have used Pima Indian data set for in their work. Almost every researcher has tried to build a model with an increased accuracy for prognoses of diabetes using various data mining tools.

(Polat, et al, 2007) combined PCA (principal component analysis) and adaptive Neuro-fuzzy inference system (ANFIS) for detection of diabetic disease, their aim was to increase the accuracy of prediction of diabetes by using their combination, this system worked in two phases. In phase I dimension reduction is done using PCA and in phase II diagnose of diabetes is done using ANFIS.

(Han J, et al, 2008), Applied ID3 algorithm to predict type-2 diabetes, the study concluded that plasma glucose attribute as key factor for diabetic condition.

In their research work (Polat, et al., 2008) used GDA (Generalized Discriminant Analysis) and (LS-SVM) Least Square Support Vector Machine on diabetic data set to diagnosis diabetes the system worked in two stages and had an accuracy of 82.05% using 10-fold cross validation.
In their research work (Kahramanli, et al., 2008) intended the designing of hybrid system which could be used for prediction of diabetes as well as cardiovascular disease. The applied methods were ANN and FNN. The system could predict the diabetes with an accuracy of 84.24% on Pima dataset.

(Temurtas et al. 2009) designed a model for diabetes diagnosis using Multi layer neural network and trained it with Lavenberg-Marquardt algorithm. The model was trained on Pima dataset and had a predictive accuracy of 79.62%.

(Jianchao et al. 2009) applied decision tree based model for prediction of diabetes, the main aim here was to find the relation of each attribute in diabetes, and the method has an accuracy of 80%.

Hybrid prediction model for prediction of type-2 diabetes was proposed by (Patil, Joshi, et al, 2010). They used K-means clustering and C4.5 algorithm and applied K-fold cross validation.

(Esin, et al., 2010) proposed a model that has two phases based on Linear Discriminant Analysis and Adaptive Network Based Fuzzy Interference System. In phase I feature variable is separated from positive and negative cases of diabetes. In phase II the healthy case and the feature variable obtained in phase I, are input conditions for ANFIS classifier. The dataset used was Pima dataset and the accuracy obtained was 84.6%.

(Stavros et al., 2010), applied eClass method – a fuzzy rule based systems new architecture to design a predictive model for two most common diseases viz, diabetes and dermatological disease, the model designed for diabetes had an accuracy of 79.37%.

(AA Aljarullah, 2011), applied decision tree for diagnosis of diabetes mellitus (Type II), the dataset used was Pima diabetes dataset. The proposed model had an accuracy of 78.17%.
For improving the accuracy of prediction of diabetes mellitus (Aliza Ahmad et al. 2011) made a comparative study between decision tree and neural networks. Both the algorithms were tested on Pima dataset however J48 based decision tree had a highest accuracy of 89.3%.

(K. Rajesh et al., 2012), applied C4.5 decision tree algorithm besides other algorithms like KNN, SVM, Naïve Bayes, and LDA to find the best algorithm for prediction of diabetes, the outcome of C4.5 decision tree algorithm was found to be highest with an accuracy of 91%.

(Pouya Afrand et al. 2012), in their research work designed an expert system for diabetic prediction using novel concepts of artificial intelligence. The model that was developed was an improved version of XCS (extended classifier system). This new improved method had a predictive accuracy of 91.32%.

(Fayssal Beloufa et al., 2013), proposed a novel model for diabetic classification. They modified artificial bee colony algorithm by adding a mutation operator to enhance its diversity, but the solution quality was kept uncompromised. This new modified algorithm as a tool was used for creation and auto optimization of member functions and rule bases from data directly. The method was proven to be highly effective in diabetes classification. The model had an accuracy of 84.21%.

(Salim Amour et al. 2014), applied two of the best known data mining techniques for disease prediction viz, Naïve Bayes and Decision Tree on diabetic data inorder to predict chances of diabetic condition and comparison of best among two. At the end of the experimentation it was concluded that both the methods predict the diabetes with acceptable accuracy however Naïve Bayes proved better than Decision Tree. The accuracy was 76.30% and 73.83% for Naïve Bayes and decision tree respectively.

(Veena et al., 2015), combined preprocessing techniques Principal Component Analysis and Discretization method with three of the most popularly used mining techniques (Decision tree, Naïve Bayes, SVM) with intension of increasing predictive accuracy for diabetic prognosis. Models were built using all
the three classifiers among which Decision tree and Naïve Bayes emerged as the best classifier when combined with PCA and both had a high accuracy rate of 79.01%.

(Huzaifa et al. 2015), in their research work compared various variants of Decision trees, LSSVM, Genetic algorithm techniques for designing an expert decision support system for diabetic prediction, however at end of research it was concluded that LSSVM based Gaussian radial function proved to be the best among all for prognostic purpose. The model named as GA-LSSVM had an accuracy of 81.33%.

(DI Geroge et al. 2015), presented a survey about the techniques that could be used for the designing a data mining tool for prediction of diabetes, various techniques were tested for accuracy, however Partial Least Square-Linear Discriminate Analysis achieved a highest accuracy of 76.78%.

(Durairaj et al., 2015) presented a survey for selecting the best soft computing tool for prediction of diabetes. They used various algorithm inorder to achieve the required conclusion. After experimentation it was ANN with an accuracy of 89% being the best classification tool.

(Vaishali et al. 2015) Applied fuzzy logic on diabetic dataset with an aim to increase prediction rate for diagnosis of this disease using data mining techniques. The model was proven to be having an accuracy of 87.2% which was an improvement over other models.

(K. Saravananathan et al. 2016): in their research work analyzed various classification algorithms of data mining to analyze the diabetic data; the intension was choosing the best classifier that could predict diabetes. The dataset used here was collected form medical diabetic center. At the outcome of research decision tree algorithm with an accuracy of 67.15% was the top model for prognosis purposes.

(K. saravanapriya et al. (2017), in their research work applied various algorithms for classification of diabetes inorder to establish the best technique for
predicting diabetes; here RBF network was proven to be the best classifier for prognosis having accuracy of 80%.

(G. Krishnaveni et al. 2017), applied various data mining techniques for designing an novel technique for prediction of diabetes, the techniques applied were Discriminant analysis, Naïve Bayes, KNN, SVM. The aim here was to select the best classifier for designing the predictive model. The Discriminant analysis techniques had a slight edge on accuracy with 76.35 over others in this case.

In their research work (Jawad kamal et al. (2017) tried to analyze the symptoms of diabetic disease to predict its risk using data mining classification algorithms. The model had two phases in first phase frequent patterns related to diabetes were extracted from database using FP-growth algorithm and in phase two Decision tree was used for risk prediction of diabetes. The model had an accuracy of 78%.

In their research work (Musavir et al. 2017) tried to establish the best prediction model for diabetes prediction among logistic regression and ANN. The ANN was evaluated using its various variants. The research was concluded with ANN model with an accuracy of 80% being higher than logistic regression.

(Meraj Nabi et al 2017), in their research work analyzed various data mining algorithms to deduce the best classifier for predicting diabetes. Experiments carried out on Pima diabetic dataset revealed Logistic regression with an accuracy of 80.43% as the best classifier.

(Monisha et al., 2018) applied Naïve Bayes, Logistic regression and Extreme gradient boosting classification techniques on Pima diabetic dataset with an aim to design decision support system for diabetic disease prediction. The research concluded with extreme gradient boosting having achieved an accuracy of 81%.
Table 2.1: Literature review of various models

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Author</th>
<th>Method</th>
<th>Dataset</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Kemal Polat et al. (2007)</td>
<td>PCA-ANFIS</td>
<td>Pima</td>
<td>89.47%</td>
</tr>
<tr>
<td>2</td>
<td>Han, et al. (2008)</td>
<td>Decision tree</td>
<td>-do-</td>
<td>80%</td>
</tr>
<tr>
<td>3</td>
<td>Kemal, et al (2008)</td>
<td>GDA-LSSVM</td>
<td>-do-</td>
<td>82.05%</td>
</tr>
<tr>
<td>6</td>
<td>Esin, et al. (2010)</td>
<td>LDA-ANFIS</td>
<td>-do-</td>
<td>84.61%</td>
</tr>
<tr>
<td>7</td>
<td>AA. Aljarullah (2011)</td>
<td>Decision Tree</td>
<td>-do-</td>
<td>78.17%</td>
</tr>
<tr>
<td>9</td>
<td>Fayssal Beloufa (2013)</td>
<td>Modified Artificial bee colony algorithm</td>
<td>-do-</td>
<td>84.21%</td>
</tr>
<tr>
<td>10</td>
<td>Dewani, et al. (2014)</td>
<td>Naïve Bayes and Decision tree</td>
<td>-do-</td>
<td>76.30% for Naïve Bayes</td>
</tr>
<tr>
<td>11</td>
<td>V. Anjali (2015)</td>
<td>PCA-Naïve Bayes, PCA-Decision Tree, PCA-SVM</td>
<td>-do-</td>
<td>79.01% for PCA-Decision Tree &amp; PCA-Naïve Bayes both.</td>
</tr>
<tr>
<td>12</td>
<td>Huzaifa et al. (2015)</td>
<td>GA-LSSVM</td>
<td>-do-</td>
<td>81.33%</td>
</tr>
<tr>
<td>13</td>
<td>Vaishali et al. (2015)</td>
<td>Fuzzy Logic</td>
<td>-do-</td>
<td>87.2%</td>
</tr>
<tr>
<td>14</td>
<td>Durairaj et al. (2015)</td>
<td>SVM, KNN, Decision Tree, ANN</td>
<td>-do-</td>
<td>89% on ANN</td>
</tr>
<tr>
<td>15</td>
<td>Saravananathan et al. (2016)</td>
<td>Decision tree, CART, SVM,KNN</td>
<td>Medical Diabetic Center dataset</td>
<td>67.15% on Decision Tree</td>
</tr>
<tr>
<td>16</td>
<td>Saravanapriya et al. (2017)</td>
<td>Naïve Bayes, Decision Tree, MLP, SVM, KNN, RBF Network.</td>
<td>Pima</td>
<td>80% on RBF neural network</td>
</tr>
<tr>
<td>17</td>
<td>Krishnaveni et al. (2017)</td>
<td>Discriminant analysis, KNN, Naïve Bayes, SVM</td>
<td>-do-</td>
<td>76.16 on Discriminant analysis</td>
</tr>
<tr>
<td>18</td>
<td>Jawad kamal et al. (2017)</td>
<td>FP growth , Decision Tree</td>
<td>Privately obtained dataset (Source not defined)</td>
<td>78% on decision tree</td>
</tr>
<tr>
<td>19</td>
<td>Musavir et al (2017)</td>
<td>Logistic regression, ANN</td>
<td>Pima</td>
<td>80% on ANN</td>
</tr>
<tr>
<td>20</td>
<td>Meraj et al. (2017)</td>
<td>Naïve Bayes, Logistic regression, random forest, decision tree</td>
<td>-do-</td>
<td>80.43% on Logistic regression</td>
</tr>
<tr>
<td>21</td>
<td>Monisha et al. (2018)</td>
<td>Naïve Bayes, Logistic Regression, Gradient Boosting</td>
<td>Pima</td>
<td>81% on Gradient Boosting</td>
</tr>
</tbody>
</table>
2.3 REVIEW ANALYSIS

Analyzing the above review following can be interpreted.

- All most all the researchers have used Pima diabetic dataset which has the following drawbacks or limitations:
  (1) The dataset contains the data of only female instances.
  (2) The dataset misses out on crucial HbA1c value.
  (3) The dataset contains only half measure of Blood pressure.
  (4) The dataset does not contain waist thickness measure.

- Some researchers have taken a different dataset without revealing its source.

- Many researchers have used same dataset with same classifiers but obtained different accuracy rates e.g. Han et al and Rajesh et al both have used decision tree classifier on Pima dataset but obtained varied accuracy of 80% and 90% respectively.

2.4 CONCLUSION

The chapter contains details about use of data mining techniques in healthcare and work done on diabetes. In the beginning of we reviewed about the work that has been done in healthcare sector with application of data mining by analyzing various techniques. In next section of the chapter the work that has been done on diabetic research using data mining techniques has been thoroughly reviewed and discussed ending with analysis of findings.