CHAPTER II
LITERATURE REVIEW

2. Introduction

This chapter discusses about various existing research works in the disease diagnosis using data mining approaches. This study reveals the importance of detection of disease in earlier stage is a critical task in the field of medical science. The detailed review of literature is described in the upcoming sections. Real-life data are frequently imperfect and erroneous. This may lead to incomplete and vagueness. Missing attribute value is considered as one of the forms of data incompleteness Manjula KR et al., 2015. Eradicating erroneous data is an important step when the heterogeneous data sources are combined.

In the data warehouse, the dirty data files are seen prevalently. Incomplete information and naming attribute conventions or the two common problems in the data warehouse. In supervised classification, the missing value treatment of LDA classifier is considered only as two classes in multivariate mode Tresp V et al., 1995. K-NN imputation techniques are introduced for the missing value using the supervised classification techniques. Data mining Edgar Acunal E et al., 2004 and Microarrays Scheel I et al., 2005 are the two statical applications dealing with missing value. These applications include supervised classification as well as unsupervised classification (clustering) Sharma S et al., 2001.

2.1 Missing values

Missing values are replaced by zero in the Micro arrays Troyanskay O et al., 2001. Data analysis techniques are mainly based on the assumption using intelligent techniques Maksood FZ et al., 2016. Trials and tribulations could not derive conclusion on incomplete knowledge. In the medical data analysis, the most common intelligent techniques used are the Bayesian classifier Kharya S, Soni S et al., 2016 , Genetic Algorithms Li D, Gu H, Zhang L et al., 2010, Decision Trees Madadipouyal K et al., 2015, Fuzzy theory Liao Z, Lu X, Yang T et al., 2009 and Krishnapuram R Keller JM., et al., 1999 and Neural Network Rajdev Tiwari, Manu Pratap Sing et al., 2010, Lim CP et al., 2005 and Ho Tim TN et al., 2006. This paper handles the problem of uncertainty while imputing the missing values.
Christina Catley et al. 2009 have protracted their work later with admiration to multi-dimensional medical data. They apply temporal data mining techniques to emerging data streams using the same “Neonatal Intensive Care” dataset. Their approach helped in clinical investigations on multidimensional time-series data. Hoda Meamarzadeh, Khayyambashi and Saraee., 2009 apply temporal data mining techniques, that help in determining hidden relationships in medical data sets. These authors believe that medical data has much temporal evidence that needs to be exploited in order to get cleverness on temporally frequent medical events. Mining temporal relational rules is their main focus. They present the rules that have been mined in the form of a graph for further processing.

The temporal intervals are used to know the temporally frequent events that exist in medical data. Their work helps in finding the temporal frequency of early detection of high risk patients, births, deaths, and pre-mature newborns. Archana, Ren Shuxia and Zheng et al., 2010 have planned fuzziness approach for mining indeterminacy temporal data. Using fluffiness concept, the authors discovered the degree of indeterminacy in patient records of medial data set. The model is helpful to demonstrate the proof of concept.

2.2 Classification Algorithms:

In 2013, D.Keranahanirex and Kalimuthu, have concentrated on thyroid dataset with multi objective problem of diagnosing and classifying thyroidism. An approach for finding thyroid attack is done by multi classification approach which achieves higher accuracy to enhance the classification and detection of thyroid disease. In 2012, M.R.Nazaril Kousarrizi, F Seiti and M.Teshnehlab, propose various methods for features subset selection in thyroid disease diagnosis and also handle a critical problem of classification of thyroid type.

In 2013, Tian D.Gledson et al., A modification is done on conventional association rule mining by introducing the Bayesian network based association rule mining. In this work, the association rule follows two interesting measures namely, Bayesian confident and Bayesian lift. In 2014, K.S.Kumar and R.M Chezian., the authors have propose two important modes for predicting hypothyroid. Those modes are K-nearest neighbor and support vector machine. From the simulation result, it is absorbed that the performance of KNN towards the prediction of thyroid diseases is better than SVM.
2.3 Feature selection Algorithm

In 2012, the authors S.W.Mendre and R.D Raut introduced some algorithms that work on images. It holds the process of pre-processing segmentation feature extraction, feature selection and classification one after the other for thyroid diseases diagnosis. In recent years, various images processing techniques are introduced and they are implemented in an effective way for detecting presents of thyroid. This work adopts fuzzy cognitive map based decision support system. In 2009 Senol,et al., the authors proposed fuzzy Nero based diagnosis system. This work integrates fuzzy logic with neural network to extract or discover knowledge from voluminous data which overcomes inconsistency among the data and extract useful patterns for producing accurate prediction.

Nikita Sigh and Alka et al., (2012) have absorbed their experiment that SVM performs better comparing to KNN and Bayesian. To conclude that the SVM is better classifier as compared to KNN and Bayesian, the KNN will find the nearest neighborhood automatically. It is represented by the graph and each vertices has an object. Bayesian based on the probability classification which gives the sample data belongs to a class. Nasrul et.al (2011) introduced a novel approach to find the lobes of thyroid in the thyroid ultra sound image using MATLAB.

There are two lobes in thyroid they are left lobe and right lobe. The given image is under goes the enhancement in their contrast in order to minimize the presence of speckles in image. The local region based active contour is used for segmenting the portion of thyroid region, thus the thyroid region is segmented in to right and left with the aid of active contour. A transfer view of thyroid image is used so the regions with depth are measured in taken into account. The calculated area is considered as pixel unit.

2.4 Features Evaluation Metrics

Deepika Koundal ET al. 2012 have presented some of the information which are already exists to formulate the diagnosis part simpler and efficient with the help of automatic tools. They also discussed about different evaluation metrics and feature trends. Edgar Gabriel et al. (2010) have modeled techier based segmentation of images in a parallel way which is a first critical step in the field of automated solution. The version of MPI code is implemented to use the distributed memory.
M. Savelon et al. (2005) have designed and developed variable background active counter model for detecting abnormal in ultra sound image. It also handles edge smoothing less and overcoming chances in topology. It produces more accuracy and reduces the effect of back ground noises in images.

Preeti Aggarwal et al 2011 suggested segmentation method with the help of automatic tools by applying segmentation algorithm both on CT and US images. With the help of the tools, they produced better results in segmentation images. Eystrants et al., 2010 have designed a detector for thyroid disease using the prototype of computer aided diagnosis. It detects nodular tissue in thyroid images and they are examined by capturing video during examination.

Mary et al. 2005 in their paper, handle features associated with cancer. They point out which nodules should be considered the aspiration and which kind of nodules of thyroid are not to be considered for fine needle aspiration. Various types of image classifiers namely, Genetic Algorithms (GA) Abdella M, Artificial Neural Networks (ANN) Alfonso Bastias, Fuzzy measures, Support Vector Machines (SVM), Fuzzy support Vector Machines (FSVM).

To overcome the problem of misclassification, ANN is used for feature extraction based on texture of the image and SVM is best suited for classification of high feature set of data. Won-Jin Moon et al. 2008, perform the experiment in their work for diagnosing the presence of benign and malignant of thyroid nodules. This analysis is done using the tissues as standard reference. The detection of calcification, shape, edges and echogenicity are the important factors for finding whether the nodule under observation is malignant or benign.

The work Chang Pin Wei and Lion Ming Der et al., 2014 developed a model based on genetic algorithm for detecting breast cancer based on classification of patients with their expression and complexity by generating rules which yield better result compared to the existing approaches. WBC dataset is considered for simulation and its performance is compared with decision tree, ann, logistic regression and proposed genetic algorithm. The outcome of the simulation result shows that the prediction in positive attitude and the accuracy obtained in classification, motivate to evaluate and conclude that the genetic algorithm is a best accepted method in its accuracy of classification.
2.5 Fuzzy Rule

In 2010, Gandhi Rajiv K., Karnan Marcus and Kannan S., proposed particle swarm optimization technique to identify the presence of breast cancer. This paper integrates the concept of feature selection for generating the fuzzy rules optimal way and lessens the time complexity and resource management. The selected features are used by PCO for classifying the dataset with accurate results. The research work of Padmavathi J et al., 2011 use the RBF and MLP with logistic regression for finding the presence of breast cancer in an optimal way. It is noticed that RBF takes less time with good predictive functionalities.

The roughest based on hierarchical granulation is developed by Lee Heau Chul, Seohak Seon and Choi Chul., 2001 In this method, they accept to regulate the instructions with negligible attributes using lesser approximations of rough set theory the imitation is done on WBC dataset and the consequence examination exhibe that the proposed work twisted active result on cataloguing. In 2004 Hassanien Ella et al. propose a new approach by accomplishment timed dimensionality and normalizing of the WBC data with 360 instances. The preprocessed data is then observed by a rough set method for generating classification rules. To enhance the engendered rules the rough set decrease technique is useful to find all possibilities of the data which contains the minimal subset of attributes and determined optimized 30 rules out of 472.

In 2014 Sudhir D., Ghatol Ashok and Pande Amoel et al. apply SVM and ANN on the WBC data. The results of SVM have more accuracy while comparing ANN prediction models and this work mainly contributes the decision making for biopsy. In 2005, Jamarani S.M, Behnam .H and Rezairad G.A, accessed a progression for early breast cancer diagnosis by assimilating ANN and multi wave let based sub band image disintegration. The MIAS dataset is used for testing. The proposed approach is tested using the MIAS mammographic databases and the result shows that the multi wave performs better. The proposed approach could aid the radiologists in mammogram psychoanalysis and analytic verdict making.

S.Yasodha et al., 2012 have projected, to detect thyroid disease using CACC-SVM methods which is hybridization of class-Attribute Contingency Coefficient (CACC) and Support Vector Machine (SVM) for classification of thyroid. The proposed model achieves better accuracy compared to other models.
2.6 Disease Diagnosis

Alfonso Bastias et al., have modeled an artificial immune system based learning mechanism for diagnosing the medical dataset and it investigates thoroughly the working functionality of the proposed classifier. The proposed work improves the process of identification of type of thyroid disease very effectively. Gurmeetkaur et al., 2014 have intended to develop an artificial neural network of disease diagnosis. They present the model of three different architectures of neural networks to find the presence of thyroid. They hold difference in their architecture and features. Ali keles et al., 2009 focus on developing the expert system for disease diagnosis with the help of expert system. The expert system is normally a fuzzy system in this case neuro fuzzy system is used for generating the rules with more percentage of accuracy.

In 2005, Hoshi K, Kawakami J, Kumagai M et al., Analyzed that the neural network is adopt for regulating the Bayesian system with self-organizing map using multivariate analysis. The sickness due to any kind of diseases handle in Diagnosis Isa I S, Saad Z, Omar S et al., (2010) is using variant of network along with different dimensions, with the network related problems. Every network is related to choose the characteristics of the neural network which produces finest results. Polat K, Şahan S, Günes S., 2007 planned machine learning approach with simulated insusceptible greeting outline (AIRS) which is a preparation of context to examine the thyroid problem.

In Keleş A, Keleş A., 2008 expert system based thyroid disease is identified. The simulation result shows that the accuracy of detecting thyroid disease is above 95%. In Ozyilmaz L, Yildirim T., 2002 examine various neural network methods for diagnosis of thyroid disease. Five different variants of neural network is evaluated namely Fast Back-Propagation (FBP), Radial Basis Function (RBF), adaptive Conic Section Function Neural Network (CSFNN), Multilayer Perception (MLP) with Back–Propagation (BP) have been used for this purpose. In 2009 Temurtas et al. associate the thyroid ailment examination by using multilayer (NN), Probabilistic (NN) and learning vector quantization (LVQ-NN) neural systems. In 2009 Paliwal M., Kumar UA. Review on restorative diagnosis of disease with the help of feed forward simulated neural systems. In 2009 Sharpe PK., Solberg HE., Rootwelt and Yearworth M, approach two different types of ANN. ANN are examined and they are evaluated based on the
prediction capability of the disease. The multilayer based perceptron is used for evaluating the procedure with the learning vector quantization. In (2009) Kodaz H, Ozsen S, Arslan A, et al. devised a model for thyroid dataset analysis with the sensible classifiers and its performance is compared with other machine learning algorithms.

In 2014 Rastigi, Bhalia A et al., designed a neural network based on feed forward process which performs the cross validation technique very effectively. The determination of presence of disease is done very accurately using this technology. In (2012) Kulkarani, Sathish N et al., Propose the fuzzy hyper line section grouping neural system (MFHLSCNN) is designed for finding the thyroid alignment. The characterization and bunching are clearly defined in this. The proposed method’s performance is compared with MLP and it is determined that the time taken and accuracy is high while using the proposed method.

In 2012 Kousarrizi, Nazari MR., these authors highlighted a choice based technique for calculating the hereditary based issues in the problem of disease diagnosis. Bolster vector machine is used in this paper for practical outcome of the result. The proposed method is compared with traditional support vector and it is resulted that the bolster vector machine produces more promising result in diagnosing the diseases.

The linguistic and fuzzy sets are used in 2007 by R.Radha and S.P Rajagopalan. In order to determine the diabetic disease the minimum and maximum relationship in presence of uncertainty are used in the dataset. There are forty patients records are used to predict the possible outcome of the sign of the disease. P.Jeatrakul and K.W.Wong (2009) developed a classification based on binary neural network method which is used for overcoming the problem of detecting the diabetes. The valuation is performed using three different benchmark datasets collected from the uci machine learning repository.

In the work (2007) Q.Q.Zhou, M Purvish and N.Kasabov, developed fuzzy neuro system for finding the diabetes in its earlier stage with the concept of fuzzy membership values. This proposed work utilizes two different experimental results on medical datasets. In (1997) T.H Lin and V.W.Soo et al. anticipate a pruning technique which relies on the principal of minimal description. The main idea is replacing the theory of complexity with the prediction accuracy effectively. The proposed greedy method prunes the unwanted rules generated by the fuzzy association rule mining algorithm and the result shows more promising outcome while comparing the existing approaches used in this paper.
The authors F. Ensan, M. H. Yaghmaee and E. Bagheri (2006) developed a clustering process based on fuzzy which defines appropriate clusters using prototype sprit. Several experiments for algorithm evaluation are executed which shows an improved presentation evaluates to the distinctive widely used K-means clustering algorithm and data is taken from the UCI Machine Learning Repository.

The authors L. Carnimeo and A. Giaquinto (2006) proposed an automatic retinal diabetic detection using a multilevel perceptron neural network. To evaluate the optimal global threshold to minimize the pixel classification errors the network is trained by this algorithm. The performance of the proposed work improves the detection by adequate index based on neuro fuzzy subsystem.

### 2.7 Disease Detection algorithms

In 2012 Chen H. L., Yang and Wang et al., proposed an expert system with three different stages using support vector machine for disease detection and classification. The proposed method performs better than the other models. In 2014 Valanarasi Antony Santiagu Vaz, the authors developed a artificial neural system for identifying the presence of thyroid using MLPNN. The model is performed under tenfold cross validation with more than 90% accuracy compared to the three fold model.

In 2001, Rago T, Chiovato L and Grassco et al., proposed auto immune system for detection of thyroid disease using ultra sonography tool based on pattern matching and recognition. In 1991 Varde AS, Massey KL and Wood HC, Proposed a framework using expert system was developed for analyzing the unpredictable area of the thyroid disease. The result produces the encouraging output of the proposed system than the existing approaches. In (2012) Li L. N., Ouyang J. H and Chen H. I et al., A new principal component analysis is developed for finding the thyroid disease, the author used ELM based prediction and compared with the PCA as an extreme of learning machine algorithm.

In (2012) Azar At., Hassanien AE and Kim T. H, have produced linguistic fuzzy neural network based classifier with the help of selected features, the thyroid disease is detected. The issues and challenges in learning the patterns are mainly motivated and initiated in this work. Numerous diverse types of Neural Networks in 2013, Makas. H and Yumusak Are used for finding the thyroid disease with the usage of swarm optimization and the movement of the bird’s optimization approach. The neural networks with the uses of these techniques are achieved well
in predicting the output. In 2008, Papageorgious EI, Papandraianos NI and Apostolopoulos, proposed a new method for thyroid diagnosis based on cognitive maps using fuzzy was adapted for discovering the patterns. In 2011, Dogantekin E and Avci D planned an approach of finding the disease in universal level with the help of discriminant analysis and wavelet based SVM. In 2011, Aziz SB., establish ANN method for diagnosing disease using genetic algorithm. The simulation results are identified for bath testing and training set.

In 2012, Mendre W and Raut discovered a Fuzzy cognitive map which used decision support system and compared with other methods. The main goal of this research is to choose and mining appropriate attributes which heals highest classification accuracy. In 2012, Margret JJ, Lakshmipathi.B and Kumar SA, proposed an new technique for diagnosis of thyroid diseases using decision tree, by splitting their attributes into three. In this technique, 5 different principal for feature selection is used. They are indexing, gain ratio, calculation, chi-squared and finding likelihood proportion to build the tree.

In 2012, Pandey S., Gour DK and Sharma V, projected a classification model for diagnosing thyroid by adopting C4.5 classifier and random forest method. The proposed classifier produces high accuracy than the existing approaches. In 2013, Pandey S, Miri R and Tandan SR, established a classifier for purpose of hypothyroid disease with material mining procedures like variation in the classification techniques. In 2011, Sarasvathi V and Santhakumaran.A, developed a new simulated model of neural network with the concept of BP for getting the high specify and the accuracy. The error on the classification task is considerably reduced in this process. In 2012, Kaur Get al. developed back propagation with LVQ model for detecting illness.

The classification and prediction of disease is done very effectively. In 2013, Gharsehchopogh FS, Molany M and Mokri FD, planned MLPNN to categorize the thyroid disease using BP algorithm. In 2015, Prerana PS and Taneja K designed a BP based ANN model to predict thyroid. The model was trained with Marquette method and produces accurate results comparing to radiant descent algorithm. In this paper, using back propagation algorithm the thyroid disease is easy diagnosed and the approach is systematical calculated in early stage and it is accessible by neural network algorithm. In this field most of researchers are interested in using ANN algorithm for back propagation process because it used to predict the disease early. All the experimental dataset are reviewed and the result obtained initial. In this proposed model author
describes that the outcome of the ANN experimental data perform a good result, comparatively
the predication report of disease are worthy.

In 2010, Zhang K, Chai Y and Yang SX, established a classical model for more than one
disease diagnosis using SOM that is self-origination map. Pointing multi disease diagnosis a self-
organization map is established in this paper, tomato disease feature is developed and the feature
is created in-between mapping relationship and features. Her, in-curate algorithm for clustering
classification is formed and it is analyzed with som algorithm. In this review Eculidean
algorithm is optimized and taken as a main goal for discernment analysis report. adjacent
searching algorithm is also taken as an alternative report for discernment algorithm. Using all
these algorithms the mapping process for finding features become easy .The inputs are obtained
from feature and multi-diseases mapping model .The som has two different model of layer for
simulating multi disease. In 2008, Erol R, Ogulata SN and Sahin et al., explored the model for
multi layer perceptron and Radial Basis Function (RBF) to predict the presences of thyroid. This
model is designed to determine types of thyroidism in clinical studies.

In 2009, Zabidi A, Mansor W, Khuna LY, Yassin IM and Sahak R, proposed the
presentation of MLP in identifying between new born babies feeling with hypothyroidism
founding on their screams. The proposed technique specifies the MLP is higher to alternate strat-
egies. In 2010, Martins L.G.A and Monterio LHA examined intellectual control replicas for
bringing the beneficial pills to chief hypothyroidism patients, deprived of the thyroid gland.
MLPNN systems are used in predicting the hormones TSH and T4 in the blood.

In 2007, Kim YS and Yoon CN, established hybrid model proposed based system and
the decision making system of finding disease in the thyroid gland using the logistic regression.
Lymph node metastasis is mutual in all type of thyroid cancer. The administration of clinical
negative is provocative in research field. The student revels that the evaluated primary tumor size
is foreseen in LMN. Multivariate logistic regression analysis was used for DTC patients they
were treated with surgery between 2002 and 2012 in the Surveillance, Epidemiology, and End
Results database, to determine the association of tumor size at 10 mm increments with LNM is
used. In that study researcher found that the percentage of male sex, white race follicular
histology, gross extra thyroidal extension, lateral lymph node metastasis, and distant metastasis
gradually increased with size.
In 2013, Bhalla JS and Aggarwal A, projected two ANN’s based on marquardt bp and conjugate gradient based bp for predicting the patterns and categorizing the instances as positive or negative. Two different artificial neural networks are proposed for disease diagnosis, which uses Scaled Conjugate gradient backpropagation and Levenberg-Marquardt backpropagation algorithm for training the neural networks. The proposed model has been tested on a dataset about Thyroid disease collected from a local hospital.

These samples are first trained using Levenberg-Marquardt propagation and outcomes are measured, then the same samples are trained by means of Scaled Conjugate gradient backpropagation algorithm and results are noted. The algorithm used is capable of distinguishing amongst infected person or non-infected person. The results from the two models are compared and analyzed to show the efficiency of prediction by ANNs in medical diagnosis.

In 2015, Razia S, Narasingarao MR and Sridhar GR, proposed a two different neural system models for disease determination. MLP is trained and tested with Back Propagation algorithm while RBF is trained and tested with SPSS software. The consequence proves that MLP and RBF have likewise performed well in diagnosing the disease. In 2014, Qi Chena , Amand Whitbrook , Uwe Aickelin and Chris Roadknight proposed the Dempster-Shafer method is employed as the hypothetical basis for creating classification data systems. Testing is carried out using three popular levels of datasets that have two, three and four classes. In each case, a subset of the available data is used for training to establish thresholds, limits or likelihoods of class membership for each attribute, and hence create mass functions that establish probability of class membership for each attribute of the test data. Classification of each data item is achieved by combination of these probabilities via Dempster’s Rule of Combination.

Results for the first two datasets show extremely high classification accuracy that is competitive with other popular methods. The third dataset is non-numerical and difficult to classify, but good results can be achieved provided the system and mass functions are designed carefully and the right attributes are chosen for combination.

In all cases the Dempster-Shafer method provides comparable performance to other more popular algorithms, but the overhead of generating accurate mass functions increases the complexity with the addition of new attributes. Comparing all the level of dataset DSM proves better result .Overall; the results suggest that the D-S approach provides a suitable framework for
the design of classification systems and that automating the mass function design and calculation would increase the viability of the algorithm for complex classification problems.

2.8 Conclusion

This chapter discusses about the various existing work done in the field of medical science and disease diagnosis. The upcoming chapters overcome the still existing problem on detection of disease in earlier stages in an effectual way.