A SURVEY ON DATA MINING TECHNIQUES IN HEALTHCARE

V. Krishnakumar[1], Dr. V. Sangeetha[2]

1Research scholar, Department of Computer Science, Periyar University, Salem, India

2Department of Computer Science, Periyar University Constituent College of Arts and Science, Pappireddipatti, Dharmapuri, India

kichuveera@gmail.com, sangee759@gmail.com

Abstract - There are some models are available in the field of healthcare to diagnose or predict the dengue fever, early stages.Process of these model to diagnose or predict the dengue fever is difficult and met major drawbacks such as less accuracy,inadequate performance, time, and cost factor is too high also absence of advanced technique and tools. Predictive model havebeen created to solve drawbacks of previous models. This predictive model come up with data mining, techniques, algorithm,tools and big data. There are some predictive model introduced in the field of healthcare using classification technique butinadequate performance in both diagnosis and predict of dengue fever. This survey paper is organized based on data miningtechniques, algorithms and tools built on the list of attribute. The attributes are technique, algorithm, type, approach, interface,IDE, data set collected from, data set collected year, format of the data set, data set size, method for removing missing values,predictive model technique, time complexity, tool, pros, and future works. Based on the existing works, results are dissimilar, accuracy level and outcome of the diagnosis or prediction level is compete each other to prove its better performance. If theappropriate techniques and algorithms are used will improved the performance as well as accuracy levels. This paper gives thesummary of the data mining concepts, techniques, algorithms, tools, challenges and discusses future work of data miningtechniques.

Keywords –Data Mining, Techniques, Algorithm, Predictive model, Tools

I.INTRODUCTION

Data and information have play vital role as a resources formost of the organizations [1] [2]. In early days data producedmechanism are very less, with the minimum amount of dataprocessing speed is very high. Particular data after enter intoprocess level gives output as information and store thatparticular information in any one of the storage devices usingdata base or data warehouse technology. Storage space wasadequate, so each and every industry

start to use technology forgetting better result. After few years data produced mechanismare increased such as internet, social media, and mobile dataand others. Using these mechanism data have been growingultra-fast manner. In this scenario storage space was inadequateand get better knowledge from data was difficult. In recentyear's industries like educational organizations, bank/financebased detail, healthcare sectors, working group of crimedetections and all others are collecting more data continuouslythan human expected limit. Especially in the field of healthcare,health related data which means, the patient information(including patient id, patient name and complete detail of individual patient), symptoms, prescription details, function ofthe human part like, heart, kidney and all other parts as mediafile and issues in healthcare are growing faster thantechnologies. From the huge collection of data or informationcould not found better result using existing technologies. In theview of previous technology have lots of drawbacks such as,storage space, size of the data, types of the data and format ofdata, accuracy, time and cost being present. Data mining comeup with solution with its techniques and algorithm for previoustechnology drawbacks. Big data come to play to solve storagerelated issues.This paper confers the idea of different types of technique andalgorithms and also describes as a report of survey data miningtechniques and algorithms.This paper is structured as follows. The related work discussedin section 2 and section 3 gives the process of data miningconcepts and section 4 describes data mining in healthcare fordengue and section 5 discussed classification technique used indengue prediction. The analysis briefed in section 6 and section7 provides the conclusion and future work.

II. RELATED WORK

Data mining with analysis and analytics in the field ofhealthcare is very fast growing. The existing research worksand various technology backgrounds done in the datamining with heal the care are abridged in this section. Thissection gives an overview of data mining techniques andalgorithm used by the following authors.D. Usha rani surveyed the data mining tools and techniques inmedical field in 2017 and deliberated about data mining toolsand techniques to be improved for reduce cost and time fromhuman resources and capability. Also, use the combination of adata mining techniques than particular technique applied for aspecific diseases diagnosing or predicting diseases inhealthcare sector could produce more advance level of positiveresults [3].A. Shameem Fathima *et al* reviewed knowledge discovery, datamining, and the purpose of the classification methods todiagnosis and prognosis used for

arbovirus-dengue in 2011. Theproposed work was apply hybrid classification schemes andimplement data mining tools to analyze the data, evaluate thedata mining algorithm through these steps try to provide someperceptible health information mined by the data miningmethods [4].Subhash Chandra Pandey explored varies tasks of data mining,benefits and drawbacks of data mining techniques in the fieldof healthcare and compare the health related data, alsosuggested to get the better result, need to improve the datamining techniques in 2016 [5].Dr. A. R. PonPeriasamy and S.Mohan summarized thesignificance of data mining concepts applied in medicalhealthcare sectors, importance of preprocess task and analyzed the exclusivity of medical data mining in 2017 [6].Parvez Ahmad *et al* précised different types of data miningtechniques, applications and challenges in health careperspective in 2015. Above all discussed the feature selectionmethods which is one of classification techniques in datamining and importance of secure distributed healthcareenvironment [7].R. Naveen Kumar, M. Anand Kumar conceded the preprocess techniques produced a complete data set instead of using rawdataset which leads to delay in the overall process [8].S. Sharath *et al* surveyed clinical data mining concepts to minethe health related records. Innovation of data miningtechnology is distributed network which is used to share storagespace with legitimate channel [9].

III. DATA MINING

Data mining used to extract useful knowledge using hugevolume of data from past and present activities (collectivelycalled ‘large dataset’) to give better outcome. Data miningconsist different kinds of techniques and algorithm available todiagnose or predict the particular diseases. Classification is oneof the best technique in data mining to give accurate result inthe field of healthcare using some statistical languages.

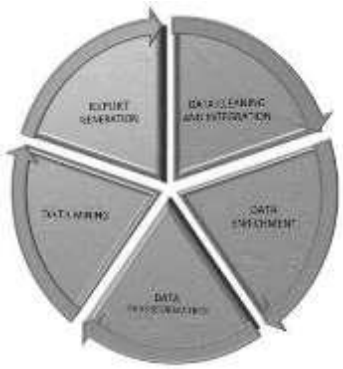

Fig 1: KDD Process Cycle

Description of KDD process as follows

1. Data cleaning and integration is the first in KDDprocess step to remove the no longer used datairrelevant or unwanted data from existing resources.Integration which means combine different formatdata as a same format.

2. Data enrichment is the second step in KDD process toincrease value of data with existing one.

3. Data transformation is the third step in KDD processto put the data in particular format that user want. Thisis also called format uniformities.

4. Data mining is the fourth step in KDD process toextract the knowledge from huge data repository.

5. Report generation is the fifth step in KDD process toget expected outcome.

IV. DATA MINING IN HEALTHCARE FOR DENGUE

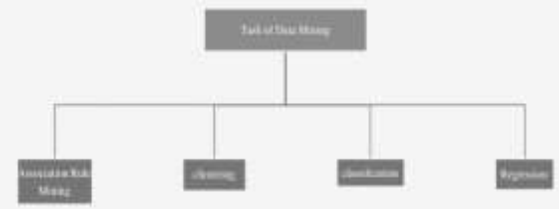

Fig 2: Data mining algorithms used for dengue prediction

1. Association rule mining work based on relationshipbetween the data points. For example Hospitals collectthe symptoms of the patient's data after collect thesymptoms can apply this ARM, the hospital can findthe number of patient affected by same kind of diseasefurthermore find the particular diseases spread inwhich place.

2. Clustering Analysis also called as cluster analysis.For example identify group of patients with sameproperties.

3. Classification is used to classify the patient bysymptoms. For example, patient age, masculinity,blood pressure, presence or absence of certainsymptoms, etc.

4. Regression also used in predict analysis based numeric values.

V. IMPORTANCE OF CLASSIFICATION ALGORITHMS.

One of the best algorithm among data mining algorithms toextract the exact well known category to expound the feature ofthe model from dataset. This classification algorithm has two phases such as i) Training and ii) Test phases. Evaluation hasfive constraint factors in classification is i) accuracy ration of the prediction from the dataset ii) time and cost iii) robustnessof all about dataset iv) model to determine of the data and v) understandability of particular model.

Different types of algorithm available with classification type based algorithm like, i) DT algorithm, ii) Rule-Based algorithmiii) statistical algorithm iv) Non-Linear algorithm and so on.

A. DT Algorithm is also called as decision tree algorithm. It consist two important algorithm such as i) iterativedichotomies 3 (ID3) and ii) c4.5(C4.5- Extension ofI D3 algorithm)
B. Rule based algorithm – CN2 and CL2
C. Statistical algorithm – cart, genetic algorithm
D. Nonlinear algorithm – neural network, nonlinearregression

VI.SURVEY ON DATA MINING TECHNIQUES

Different authors' to diagnose or predict the dengue fever usingdata mining techniques is analyzed and discussed in thissection. This survey is based on the following list ofparameters. The parameters are technique, algorithm, type, approach, interface, IDE, dataset, method for rmv, pmt, tc, tool, pros, future work. The above lists of parameters are shown inthe Table I.

P.Manivannan, *et al* developed algorithm for dengue feverprediction using on k-medoid clustering based on unsupervisedalgorithm. In this research also used dwin's method for remove missing values. Time complexity calculated as O(n2) andpractically proved. Data set collected from Seremban DistrictHealth Office, Negeri Sembilan, Malaysia in the year of 2010-2013 size of this data was 171 attribute and 1910 records.Proposed work improve evolutionary patterns and applydengue fever prediction [10].

HUSAM, I.S *et al* created the model using statistical and expertverification tools. Predictive model techniques such as J48,DTNB and naïve bayes are applied in this model. Wrapperapproach applied for evaluate the future subsets and PSO, GA,RS algorithms also used. Compared with each other and foundresult. Dataset obtained from Public Health Department,Seremban, Negeri Sembilan, Malaysia, format of the was excel,year of the data 2003-2010, size the dataset is 20 attribute and6082 records [11].

Kamran Shaukat *et al* implemented predictive model usingclassification techniques and WEKA tool. Compared DT, NBand J48 with other. Algorithm such as, naive bay, J48, SMO,REP tree, and random tree also the purpose of interface pointof view in this research Explorer, knowledge flow andExperimenter interfaces are used. Dataset took from datasetwas collected from District Headquarter Hospital (DHQ)Jhelum. Format of the dataset was in comma separate value(.csv) it consists 18 attribute and 99 records [12].

Ashwini Rajendra Kulkarni *et al* created association rulegeneration for virus-related illness using association rulemining technique, algorithm like apriori, FP – growth,association rule generation and rapid miner text mining tool.This paper defined abstract level of association rule mining andaprioir algorithm and proved through implementation level.Dataset format was in excel. Future work of this paper isproduce the numerous patterns [13].

Ramandeep Kaur *et* al explained and evaluated dengue fever inhigh population area using cluster technique with R-Studio toolfor statistical report based on unsupervised learning. Usercreated dataset for this research with .csv file format it consistsof 10 attribute and 100 records [14].Nandini. V *et al* developed a system for detection andprediction of dengue fever using frequency analysis,classification and regression technique, interface designed aspatient and researcher GUI and supervised learning type.Varies predictive model technique and tools also used such astime series, exponential smoothing, moving average, simplelinear regression and tools like SAS and LibSVM. Format ofdataset is .csv in the year of 2010- 2015.This dataset consists of100 records. Future work is create desktop application andbrowser plugin file [15].

Shaufia*et al* discovered algorithm for identifying DHF and TFusing association rule mining technique with apriori, FPGrowthand Intersection Set Theory-Expand FP algorithm(ISTEFP).IST-EFP is also used to remove missing values from theparticular dataset. Suggested for future work is reduce thedimension of the dataset using Intersection

Set Theory-ExpandFP algorithm [16].M.V.Jagannatha Reddy *et al* developed expert system usingneural network technique, decision tree algorithm, MATLAB.

2013a tool. Dataset are obtained from Srinagarindra hospitaland ongklanagarind hospital, Thailand and remove the missingvalue manually. Future work taken from this paper is predictany types of fever [17].

M Krishna Satya Varma *et al* developed decision tree modelusing decision tree technique, ID3 algorithm based onunsupervised learning [18].

Dr. ArunKumar.P.M *et al* filtered related to dengue fever data using decision tree, support vector machine techniques as well as its algorithm, Netbeans acted as a user interface, tools forthis work was WEKA and predictive model technique calledfisher filtering and prediction. Important note in this work wasthis model worked in google application, cloud computing andDT. Future work will predict the environment factors [19].

VIII. CONCLUSION

Understanding of classification technique and developedpredictive model is depends on the dengue related data.Researcher has given overview of different predictive modelusing classification methods with their advantages anddisadvantages. Many approaches differ in the way of comparethe accuracy. Several approaches suggest additional techniquesand also advanced tools for diagnosis or prediction of denguefever. Preprocessing also suggested. Survey revealed thatconsidering classification methods choice is important,because, number of available predictive model provides samekind of functionality. In diagnosis or prediction of dengue feverstatistical language also involved. If the classification, preprocess, cloud computing storage purpose will help thediagnosis or prediction of dengue fever more efficient. The researcher will use data mining completely in predictivemodel classification that will provide tool to diagnose orprediction the dengue fever. Complexity of both time and spaceis not concentrate as much it is also considered for the futureresearch work.

REFERENCES

1.Klosgen W and Zytkow J M (eds.), “Handbook of data mining andknowledge discovery”, OUP, Oxford, 2002.

2. Kantardzic M, “Data mining: concepts, models, methods, andalgorithms”, John Wiley, New Jersey, 2003.

3. D.Usha Rani "A Survey on Data Mining Tools and Techniques inMedical Field ", International Journal of Advanced Networking &Applications (IJANA) Volume: 08, Issue: 05 Pages: 51-54 (2017) SpecialIssue.

4. A. Shameem Fathima, D. Manimegalai and Nisar Hundewale"AReview of Data Mining Classification Techniques Applied for Diagnosisand Prognosis of the Arbovirus-Dengue", IJCSI International Journal ofComputer Science Issues, Vol. 8, Issue 6, No 3, November 2011 ISSN(Online): 1694-0814.

5. Subhash Chandra Pandey "Data Mining Techniques for Medical Data:A Review", International conference on Signal Processing,Communication, Power and Embedded System (SCOPES)-2016.

6. Dr. A. R. PonPeriasamy , S. Mohan "A Review on Health Data UsingData Mining Techniques", International Journal of Advanced Research inComputer Science and Software Engineering Volume 7, Issue 3, March2017.

7. Parvez Ahmad, Saqib Qamar, Syed QasimAfser Rizvi "Techniques ofData Mining In Healthcare: A Review", International Journal of ComputerApplications (0975 – 8887) Volume 120 – No.15, June 2015

8. R. Naveen Kumar, M. Anand Kumar "Medical Data Mining Techniquesfor Health Care Systems", International Journal of Engineering Scienceand Computing, April 2016.

9. S. Sharath, M. N. Rao, H. G. Chetan "A Survey on the principles ofmining Clinical Datasets by utilizing Data mining technique",International Journal of Innovative Research in Computer andCommunication Engineering (An ISO 3297: 2007 Certified Organization)Vol. 2, Issue 4, April 2014.

10. Kamran Shaukat, Nayyer Masood, SundasMehreen and UlyaAzmeen, " Dengue Fever Prediction: A Data Mining Problem ", Data Mining inGenomics & Proteomics (Dar et al., J Data Mining Genomics Proteomics)2015, 6:3.

11. M.V.Jagannatha Reddy and B.Kavitha, " Expert System to Predict theType of Fever Using Data Mining Techniques on Medical Databases" ,International Journal of Computer Sciences and Engineering Volume-03,Issue-09 2015.

12. M Krishna Satya Varma, N K KameswaraRao,"Dengue data analysisusing decision tree model", International Conference On Emerging Trendsin Science Technology Engineering and Management 09th & 10th,October 2015.

13. Nandini. V, and Sriranjitha. R and Yazhini. T. P," "Dengue Detectionand Prediction System using Data Mining with FrequencyAnalysis",Natarajan Meghanathan et al. (Eds): ACITY, VLSI, AIAA,CNDC pp. 53–67, 2016. © Computer Science & Information Technology(CS & IT)- CSCP 2016.

14. Shaufiah and BobySiswanto, "Association Rule Mining ForIdentifying Dengue Hemorrhagic Fever (DHF) and Typhoid Fever (TF)Disease with IST-EFP Algorithm", Fourth International Conference onInformation and Communication Technologies (ICoICT) 2016 IEEE.

15. P.Manivannan, Dr. P. Isakki @ Devi," Dengue Fever Prediction usingK-Medoid Clustering Algorithm", International Journal of InnovativeResearch in Computer and Communication Engineering, Vol. 5, SpecialIssue 1, March 2017.

16. Husam, I.S., abuhamad, azuralizaabubakar, suhailazainudin,Mazrurasahani&zainudinmohdali "Feature selection algorithms formalaysian dengue outbreak detection model (PemilihanCiriAlgoritmauntuk Model PengesananWabakDenggi)", SainsMalaysiana46(2)(2017): 255–265.

17. Ashwini Rajendra Kulkarni, Dr. Shivaji D. Mundhe," Data MiningTechnique: An Implementation of Association Rule Mining in Healthcare,International Advanced Research Journal in Science, Engineering andTechnology Vol. 4, Issue 7, July 2017.

18. Ramandeep Kaur, Gaurav Gupta, "Demographic Analysis of DengueFever Using Data Mining",International Journal of Advanced Research inComputer Science Volume 8, No. 7, July – August 2017.

19. Dr.ArunKumar.P.M, Chitra Devi.B, Karthick.P, Ganesan.M andMadhan.A.S , "Dengue Disease Prediction Using Decision Tree andSupport Vector Machine, "SSRG International Journal of ComputerScience and Engineering- (ICET'17) - Special Issue - March 2017.

20. M.Bhavani, and S.Vinodkumar," A data mining Approach For PreciseDiagnosis Of Dengue Fever", International Journal of Latest Trends inEngineering and Technology Vol.(7)Issue(4), pp.352-359.

21. Chia-Bao Chu, Chao-Chun Yang, “Dengue-associated TelogenEffluvium: A report of 14 patients”, dermatologicasinica 35 124-126 Mar30, 2017.

22. Buchade Omkar, Dalsania Preet, Deshpande Swarada, DoddamaniPoonam,” Dengue Fever Classification using SMO OptimizationAlgorithm”, International Research Journal of Engineering andTechnology (IRJET) Volume: 04 Issue: 10 |Oct -2017.

23. JastiniMohd Jamil, IzwanNizalMohdShaharanee and Ve Chun Yung,“An Innovative Data Mining and Dashboard System for Monitoring ofMalaysian Dengue Trends”, Journal of Telecommunication, Electronicand Computer Engineering Vol. 8 No. 10.

S. No	Technique	Algorithm	Type/ Approach/ Interface/ IDE	Dataset					Method For RMV	PMT	TC	Tool	PROS	Future Work
				DSCF	Y	F	SDS							
							A	R						
1	K-Medoid [10]	Clustering	Unsupervised Algorithm	HCMC	2010 - 2013	-	171	1910	Dwin's	-	$O(n^2)$	-	Proof Based	Enhance evolutionary Patterns and apply hierarchical clustering Algorithms for dengue fever prediction.
2	Feature Selection [11]	PSO, GA, RS	Wrapper Approach	PHD, Seremban, neger seremban,	2003 – 2010	Excel	20	6082	-	J48, DTNB And Navie bayes	-	Statistical and expert verification	Comparison	-

				Malaysia										
3.	Classification [12]	Naïve Bays, J48 tree, SMO, REP tree & Random Tree algorithms	Explorer, knowledge flow and Experimenter interfaces	(DHQ) Jhelum	-	.csv	18	18	-	-	-	WEKA	Comparison of DT,NB and J48	-
4	Classification [13]	REP Tree, J48, SMO, ZeroR and Random Tree	-	Created	-	.csv	10		-	-	-	WEKA	i)High accuracy ii)Compared by plotting graphs and table	-
5	ARM[14]	Apriori,FPgrowth and association rule generation algorithm	-	-	-	Excel	-	-	-	-	-	Rapid Miner (Text Mining)	-	Apply text mining to produce the various patterns from healthcare data.
6	Clustering [15]	-	Unsupervised Learning	Created	-	csv	10	100	-	-	-	R - Studio	-	-
7	Classification [16]	Frequency Analysis, Classification	Supervised Learning/ Patient and		2010 - 2015	.CSV		100	-	Time series, Exponentia	-	SAS (use logistic	UMLS	Implement desktop app and

		Regression	Researcher GUI						l Smoothing , moving Average, SimpleLinear Regression		regression model) LibSVM		Browser plugin.
8	ARM[17]	Apriori and FP-growth		Hospital Medical Record	-	-	9	192	IST-EFP	-	-	-	Dimensional reduction. Predictive DHF and TF.	reduce dimension of the dataset using ISTEFP algorithm
9	Neural Network[18]	Decision Tree Algorithm	Decision Tree	Srinagarind ra Hospital and ongklanaga rind Hospital, Thailand	-	Categorical to Numerical data	>400	-	Manual missing value imputation	Multivarient Model	-	MATLAB 2013a	i)generate performance curves, ROC curves, Confusion curves for both training and test data. ii)an accuracy of 100.0% in children and adults using both	Extended to predict any type of diseases.

												clinical and laboratory features	
10	Decision Tree [19]	ID3	Unsupervised Learning	Health department, Hospital, Urban Local Body	-	-	6	142	-	-	-	-	supervised classifier model	-
11	Decision Tree,SVM and ANN[20]	SVM	Net Beans	-	-	-	18	108	-	Fisher Filtering and Prediction.	-	WEKA Feature Selection	Google Applications Cloud Computing DTS	Environmental factor prediction
12		Optimization [18]	Probabilistic Neural Network	GDS5093	-	-		56	-	-	-	Greedy forward selection algorithm	Implementation SMO	Non – Dominated sorting genetic algorithm – II

Table I - Analysis of various techniques and tools used in data mining for dengue fever