Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 13013 - 13020
Received 05 March 2021; Accepted 01 April 2021.

An Extensive Survey on Heart Disease Prediction

S. Jeya Selvakumari[1], Sowmya Fernandez[2], J.Arockia Jeyanthi[3], Pa.Andal[4]

[1]Research Scholar, Manonmaniam Sundaranar University, Tirunelveli, India&Assistant Professor, Department of Computer Science, Kamaraj College, Thoothukudi, India (Affiliated to M.S University, Tirunelveli)
[2] Assistant Professor, Department of Computer Science, A.P.C.Mahalaxmi Colllege for Women, Thoothukudi
[3]Assistant Professor, Department of Computer Science, A.P.C.Mahalaxmi Colllege for Women, Thoothukudi
[4] Assistant Professor, Department of Computer Science, Kamaraj College, Thoothukudi.
Email: [1]jeyrobin@gmail.com, [2]sfsoris@gmail.com, [3]arockiajeyanthi@gmail.com, [4]andal77@gmail.com

ABSTRACT

Prediction of occurrences of heart diseases an important work in medical field. A growing number of studies use various methods and datasets for heart disease prediction. Still, a substantial part of my research work lacks categorization and systematization. Hence, there is a necessity to take stock of current knowledge in this field. In this sense, this paper carries out a review of the literature on heart disease prediction. Analysis of a total of 150 papers addressing heart disease prediction using Data Mining (DM), Machine Learning (ML) and Deep Learning (DL) methods published between 2010 and 2020 was carried out. The purpose and the contribution of this paper is to provide a clearer picture of the subfields in heart disease prediction by concentrating on two aspects. First, a recent research to categorize the main areas of specialization in heart disease prediction is reviewed and the challenge in each category has been addressed. Second, the thematic analysis is carried out to identify a specific method under each category and a suitable method is also recommended. Despite the large number of publications, this present study identifies deep learning methods with unstructured dataset by providing better results. This study helps researchers for understanding the research gap for their future study in heart disease prediction.

Keywords
Data mining, Machine learning, Deep Learning, Heart Disease

INTRODUCTION

Globally, Cardiovascular Diseases (CVDs) has become the primary cause of death [43]. The people who die of CVDs are comparatively higher when compared to other diseases. As the corona virus disease 2019 (Covid-19) pandemic has spread around the globe, there has been growing CVDs recognition[46]. As per the report of the European Society of Cardiology (ESC), 26 million adults globally are diagnosed with Heart Failure, whereas every year 3.6 million are diagnosed newly. 17–45% of the patients suffering from HF die within the first year and the remaining die within 5 years. When a proper technique is chosen at a correct time, the morbidity rate can be decreased.

The result produced by ML techniques will be more accurate as compared to DM since ML is an automated process. DM method uses a database or a data warehouse server with a data mining engine. Useful information were extracted using the pattern evaluation techniques. Whereas ML uses neural networks, predictive model and automated algorithms to make the decisions.

ML techniques produce more accurate result when compared to DM techniques [47]. It is because DM method can't work without the involvement of human. ML algorithms always require structured data, while DL networks rely on layers.

The intention of this survey paper is make the researcher to do research on heart diseases prediction using DM or ML or DL approach. This paper presents an extensive survey on approaches for heart disease prediction between 2010 and 2020.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 13013 - 13020
Received 05 March 2021; Accepted 01 April 2021.

This paper is organized in the following manner: Section II contains brief introduction about various heart disease prediction methods analysed. Section III provides the literature review of heart disease prediction. Section IV shows the comparative analysis of various techniques for prediction of heart disease. The future scope and conclusion is given in Section V.

HEART DISEASE PREDICTION METHODS

2.1. Data Mining

DM is the method of natural growth of information technology and used in variety of fields of technology, business and science. Data sets of prediction methods are normally increasing from day to day. With these large amounts of data, the useful information hidden in these large amounts of data has to be extracted. Since, the knowledge is becoming increasingly important in today's competitive world [28].

2.2. Machine Learning

ML is the scientific discipline that focuses on how computers learn from data. It arises at the intersection of statistics, which seeks to learn relationships from data and computing algorithms. The unique computational challenges of building statistical models from massive data sets, includes billions or trillions of data points [26]. The types of learning used by computers are conveniently sub classified into categories, namely supervised learning and unsupervised learning.

2.3. Deep Learning

Deep Learning aids the computer to build complex concepts out of simpler concepts. DL is an approach to ML that has drawn heavily on our knowledge of the human brain, statistics and applied math over the past several decades. In recent years, DL has seen tremendous growth in its popularity and usefulness in training deeper networks.

LITERATURE SURVEY

For the past two decades, prediction of heart diseases using various techniques has been an on-going effort. Most of the works have implemented techniques, namely SVM, Neural Networks, Regression, Decision Trees, Naïve Bayesian classifiers, K-means, Association Rules under DM. Under ML classifiers, namely Boosting, PSO+ADT, Fuzzy Rule, Genetic Algorithms, ANFI, MLP, Random Forest (RF), ANN is used. Under DL techniques, namely CNN, LSTM, NLP, RNN, Gated Recurrent Unit (GRU), BiLSTM, RCNN and Conditional Random Fields (CRF) are applied. Table I provides a brief description of some works.

Table 3.1. Brief Description of heart diseases prediction works

S.No	Reference and Year	Dataset	Technology Used	Scenarios
1	Minas et.al., [28] 2010	A total of 528 cases were collected from the Paphos district in Cyprus.	C4.5 Decision Tree Algorithm	Significant practical difficulty for decision tree learning is over fitting. To avoid over fitting pruning is implemented. Bottom-up pruning algorithm is implemented using Laplace error estimation. Larger

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 13013 - 13020
Received 05 March 2021; Accepted 01 April 2021.

				dataset is needed for further investigation.
2	SreeHariRao et.al.,[1] 2013	STULONG dataset	PSO+ADT	Time complexity of imputation algorithm would be $k * (O(m1 * (n - 1) * m) + 3 * O(m1))$. The presence of the linear trend between the time taken and the varying database sizes ensures the numerical scalability of the performance of Predictive Risk Assessment of Atherosclerosis (PRAA) methodology.
3	Roohallah et.al.,[2] 2013	ZAlizadehSani with 303 patients and 54 features.	SMO, NB, Bagging and NN	SMO, NB, Bagging and NN classification algorithms were used to analyse the dataset. Using these classifiers, 94.08% accuracy is reported.
4	Jayshril et.al.,[26] 2014	Cleveland heart disease database	MLP	98% accuracy is reported. This technique provides good results compared to other methods.
5	Chaitanya et.al.,[22] 2015	The second track of the 2014 i2b2.	Rule-based combination of regular expressions and UMLS concepts	An error analysis is performed and found that most errors in the test set originated from the MEDICATION tag as number of annotations for the MEDICATION tag is substantially larger compared to other tags. 90.7% accuracy is reported.
6	Wang et.al.,[35] 2015	~400,000 primary care patients from Geisinger Clinic	NLP	Demonstrated that both unstructured and structured information in EHR can facilitate early detection of HF as early as two years prior to diagnosis.
7	Purushottam et.al.,[4] 2016	Dataset from Cleveland	Decision Rule	KEEL (Knowledge Extraction based on Evolutionary Learning) is being utilized for implementation. This method achieved an accuracy of 86.3% in testing phase.
8	Riccardo et.al.,[40] 2016	The Mount Sinai Health System	CNN, LSTM	Showed that the feasibility of using DL to predict patients' diseases from their EHRs.
9	Priyanga et.al.,[6] 2017	Cleveland Dataset	Naïve Bayes weighted Approach	When the attributes of the given input are more, Naïve Bayes classifier is a suitable technique for classification of heart disease. An accuracy reported in NBwa is 86%.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 13013 - 13020
Received 05 March 2021; Accepted 01 April 2021.

10	Yuan Luo[37] 2017	i2b2/VA relation	RNN	The word embedding on LSTM improved the performance of classification. An accurate relation classification is achieved by using Attentive pooling networks and Neural attention models.
11	MohdUsama et.al.,[39] 2018	Datasets contain real-life data of patients from China hospital.	RCNN	Here, fine-grain features are extracted, namely pulmonary infection, coronary atherosclerotic, and heart disease. An accuracy reported is 96.02%.
12	Beulah et.al.,[12] 2019	Cleveland	Ensemble Algorithms	When bagging, boosting, weak classifier with majority voting and stacking were used, the maximum accuracy improved were 6.92%, 5.94%, 7.26% and 6.93% respectively. The performance was improved by using best feature selection techniques and large datasets.
13	Binhua Wang et.al.,[30] 2019	Dataset from Chinese PLA General hospital	MT-DWNN	The AUC of the MT-DWNN is 0.9393 which is 0.01 improved than other techniques.
14	Shi et.al.,[46] 2020	Renmin Hospital of WuhanUniversity, from January 20, 2020, toFebruary 10, 2020.	Statistical Analysis	Showed that cohort study of 416 consecutive patients with confirmed COVID-19, cardiac injury occurred in 19.7% of patients during hospitalization, and it was one independent risk factor for in-hospital mortality.

Many DM and ML techniques are used for Cleveland data set, mainly for heart disease prediction [7-9, 24, 36]. The rules produced by the electrical and com recommended system are ranked as original rules, pruned rules, rules without duplicates, classified rules, sorted rules and polish are helped in heart disease prediction [4][5].
The deep features of heart sounds were extracted by Denoising Auto Encoder (DAE) algorithm as the input feature of 1D CNN45. The system in [38], [39], [41] demonstrates the feasibility of using DL to predict patients diseases from their EHRs.

OBSERVATIONS AND RECOMMENDATIONS

The magnitude of works analyzed on DL, ML and DM for heart disease prediction shows that there is a rampant and growing research works for heart disease prediction. The most constructive methods for heart disease prediction are observed and recommended.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 13013 - 13020
Received 05 March 2021; Accepted 01 April 2021.

Table I lists the representative of DM, ML and DL method on heart disease prediction published from 2010 to 2020 were analysed, along with the types of dataset and method used. Most of the ML techniques have utilized Cleveland data set, which is a structured data set. i2b2 data set is utilized by DL techniques, which is an unstructured dataset [2][3][8]. But, the highest accuracy is reported for Cleveland dataset using hybrid methods [6][12-13].

CONCLUSION

This paper describes the literature review of DM, ML and DL methods applied for heart disease prediction. Here, various DM, ML and DL techniques and research works are analysed, which is used in the heart disease prediction. Hybrid ML techniques with Cleveland data set have yielded better performance compared to DM techniques. Machine Learning techniques have reported highest accuracy results using structured dataset. The growing researchers can apply various DL techniques in the future to classify and predict heart diseases. DL techniques also improve high speed and accuracy**..**

REFERENCES

[1] V. SreeHariRao and M. Naresh Kumar, "Novel Approaches for Predicting Risk Factors of Atherosclerosis," IEEE Journal Of Biomedical And Health Informatics, 2013, Vol. 17, No. 1.

[2] Roohallah Alizadehsani, Jafar Habibi, Mohammad Javad Hosseini, Hoda Mashayekhi, Reihane Boghrati, Asma Ghandehariouna, Behdad Bahadorian and Zahra Alizadeh Sani, "A data mining approach for diagnosis of coronary artery disease," Elsevier, Computer Methods and Programs in Biomedicine, Vol. 111, No.1, pp. 52-61, 2013.

[3] Moloud Abdar, Sharareh R. NiakanKalhori, Tole Sutikno, Imam Much Ibnu Subroto, and Goli Arji, "Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases," International Journal of Electrical and Computer Engineering (IJECE), Vol. 5, No. 6, pp. 1569-1576, 2015.

[4] Purushottam, Kanak Saxena and Richa Sharma,"Efficient Heart Disease Prediction System," Elsevier, Procedia Computer Science, Vol. 85, pp. 962-969, 2016.

[5] P. Priyanga and N. C. Naveen, "Web Analytics Support System for Prediction of Heart Disease Using Naive Bayes Weighted Approach (NBwa)," 2017 Asia Modelling Symposium (AMS), Kota Kinabalu, pp. 21-26, 2017.

[6] Amita Malav and KalyaniKadam, "A Hybrid Approach for Heart Disease Prediction using Artificial Neural and K-means," International Journal of Pure and Applied Mathematics, Vol. 118, No.8, pp. 103-110, 2018.

[7] Bo Jin, Chao Che, Zhen Liu, Shulong Zhang, Xiaomeng Yin and Xiaopeng Wei, "Predicting the Risk of Heart Failure with EHR Sequential Data modelling," IEEE Access, Vol. 6, pp. 9256-9261, 2018.

[8] Mohammad Shafenoor Amina, Yin Kia Chiam and Kasturi Dewi Varathan, "Identification of significant features and data mining techniques in predicting heart disease," Elsevier, Telematics and Informatics, Vol. 36, pp. 82-93, 2019.

[9] Senthilkumar Mohan, Chandrasegar Thirumalai and Gautam Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," IEEE Access, Vol. 7, pp. 81542-81554, 2019.

[10] Reeta Singh and E.Rajesh, "Prediction of heart disease by clustering and classification

techniques," International Journal of Computer Sciences and Engineering, Vol. 7, No. 5, 2019.

[11] C. Beulah Christalin Latha and S. Carolin Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," Elsevier, Informatics in Medicine Unlocked, Vol. 16, 2019.

[12] Yogita Solanki and Sanjiv Sharma, "Analysis and Prediction of Heart Health using Deep Learning Approach,"International Journal of Computer Sciences and Engineering Vol.7, No. 8, 2019.

[13] K. Subhadra and Vikas B, "Neural Network Based Intelligent System for Predicting Heart Disease," International Journal of Innovative Technology and Exploring Engineering (IJITEE), Vol. 8, No. 5, pp. 484-487, 2019.

[14] Amin Ul Haq and Jian Ping Li, Muhammad Hammad Memon, Shah Nazir and Ruinan Sun, "A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms,"Hindawi Mobile Information Systems, Vol. 8, pp. 1-21, 2018.

[15] Kaan Uyara and Ahmet İlhana, "Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks," Elsevier, Procedia Computer Science, Vol. 120, pp. 588–593, 2017.

[16] Oluwarotimi Williams Samuel, Grace Mojisola Asogbon, Arun Kumar Sangaiah, Peng Fang and Guanglin Li, "An Integrated Decision Support System Based on ANN and Fuzzy_AHP for Heart Failure Risk Prediction," Elsevier, Expert Systems With Applications, Vol. 68, pp. 163-172, 2017.

[17] Jesmin Nahar, Tasadduq Imam, Kevin S. Tickle and Yi-Ping Phoebe Chen, "Computational intelligence for heart disease diagnosis: A medical knowledge driven approach," Elsevier, Expert System with Applications, Vol. 40, No. 1, pp. 96-104, 2013.

[18] G. Thippa Reddy Gadekallu and Neelu Khare, "An Efficient System for Heart Disease Prediction using Hybrid OFBAT with Rule-Based Fuzzy Logic Model,"Journal of Circuits, Systems, and Computers Vol. 26, No. 4, 2017.

[19] Negar Ziasabounchi and Iman Askerzade, "ANFIS Based Classification Model for Heart Disease Prediction," International Journal of Engineering & Computer Science IJECS-IJENS, Vol. 14, No. 02, 2014.

[20] Edward Choi, Andy Sshuetz, Walter F Srewart and Jimeng Sun, "Using Recurrent Neural Network Models for Early Detection of Heart Failure Onset," Journal of the American Medical Informatics Association, Vol. 24, No. 2, pp. 361-370, 2017.

[21] Chaitanya Shivade, Pranav Malewadkar, Eric Fosler-Lussier and Albert M. La, "Comparison of UMLS terminologies to identify risk of heart disease using clinical notes," Elsevier, J Biomed Inform, Vol. 58, Suppl(Suppl):S103-S110, 2015.

[22] Thanat Chokwijitkul, Anthony Nguyen, Hamed Hassanzadeh and Siegfried Perez," Identifying Risk Factors For Heart Disease in Electronic Medical Records: A Deep Learning Approach,"Proceedings of the BioNLP, 2018.

[23] Kevin Buchan, Michele Filannino and Ozlem Uzuner, "Automatic prediction of coronary artery disease from clinical narratives," Elsevier, Journal of Biomedical Informatics, Vol. 72, pp. 23-32, 2017.

[24] Sungrim Moon, Sijia Liu, Christopher G. Scott, Sujith Samudrala, Mohamed M. Abidian, Jeffrey B. Geske, Peter A. Noseworthy, Jane L. Shellum, Rajeev Chaudhry, Steve R. Ommen, Rick A. Nishimura, Hongfang Liu and Adelaide M. Arruda-Olson, "Automated extraction of sudden cardiac death risk factors in hypertrophic cardiomyopathy patients by

natural language processing," Elsevier, Int J Med Inform, Vol. 128, pp. 32-38, 2019.

[25] Jayshril S. Sonawane and R. Patil, "Prediction of Heart Disease Using Multilayer Perceptron Neural Network," International Conference on Information Communication and Embedded Systems (ICICES2014), pp. 1-6, 2014.

[26] Mirpouya Mirmozaffari, Alireza Alinezhad and Azadeh Gilanpour, "Data Mining Classification Algorithms for Heart Disease Prediction", International Journal of Computing, Communications & Instrumentation Engg (IJCCIE), Vol. 4, No. 1, 2017.

[27] Minas A. Karaolis, Joseph A. Moutiris, Demetra Hadjipanayi and C. S. Pattichis, " Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining With Decision Trees," IEEE transactions on information technology in biomedicine, Vol. 14, No. 3, 2010.

[28] Syed Umar Amin, KavitaAgarwal and Dr.Rizwan Beg, "Genetic neural network based data mining in prediction of heart disease using risk actors," Proceedings of 2013 IEEE Conference on Information and Communication Technologies (ICT 2013).

[29] V. Krishnaiah, G. Narsimha and N. Subhash Chandra, 2015, "Heart Disease Prediction System Using Data Mining Technique by Fuzzy K-NN Approach," Springer, International Publishing Switzerland 2015, Emerging ICT for Bridging the Future , Vol. 1, pp. 371 – 384.

[30] Binhua Wang, Yongyi Bai, Zhenjie Yao, Jiangong Li, Wei Dong, Yanhui Tu, Wanguo Xue, Yaping Tian, Yifei Wang and Kunlun He, "A Multi-Task Neural Network Architecture for Renal Dysfunction Prediction in Heart Failure Patients With Electronic Health Records," IEEE Access, Vol. 7, pp. 178392-178400, 2019.

[31] José Antonio Sanz, Mikel Galar, Aranzazu Jurioa, Antonio Brugos, Miguel Pagola and Humberto Bustince, "Medical diagnosis of cardiovascular diseases using an interval-valued fuzzy rule-based classification system," Elsevier, Vol. 20, pp. 103 -111, 2013.

[32] Youness Khourdifi and Mohamed Bahaj, "Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization," International journal of Intelligent engineering & systems, Vol.12, No.1, pp 242-252, 2018.

[33] Svetla Boytcheva, Ivelina Nikolova, Galia Angelova and Zhivko Angelov, "Identification of Risk Factors in Clinical Texts through Association Rules," Proceedings of the Biomedical NLP Workshop associated with RANLP, pp.64-72, 2017.

[34] Qingcai Chen, Haodi Li, Buzhou Tang , Xiaolong Wang , Xin Liu , Zengjian Liu , Shu Liu ,Weida Wang, Qiwen Deng, Suisong Zhu , Yangxin Chen and Jingfeng Wang, "An automatic system to identify heart disease risk factors in clinical texts over time," Elsevier, Journal of Biomedical Informatics, Vol. 58, pp. S158-S163, 2015.

[35] Liang Yao, Chengsheng Mao and Yuan Luo, "Clinical text classification with rule-based features and knowledge-guided convolutional neural networks," BMC Medical Informatics and Decision Making, 2019.

[36] Yuan Luo, "Recurrent neural networks for classifying relations in clinical notes," Journal of Biomedical Informatics, Elsevier, Vol. 72, pp. 85–95, 2017.

[37] G. Maragatham and ShobanaDevi, "LSTM Model for Prediction of Heart Failure in Big Data," Springer, Journal of Medical System, Vol.43, No. 5, 2019.

[38] MohdUsama, Belal Ahmad, Jiafu Wan, M. ShamimHossain, Mohammed F Alhamid, and M. Anwar Hossain, "Deep Feature Learning for Disease Risk Assessment Based on Convolutional Neural Network With Intra-Layer Recurrent Connection by Using Hospital Big Data," IEEE Access, Vol. 6, pp. 67927-67939, 2018.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 13013 - 13020
Received 05 March 2021; Accepted 01 April 2021.

[39] Riccardo Miotto, Li Li and Joel T. Dudley, "Deep Learning to Predict Patient Future Diseases from the Electronic Health Records," Springer European Conference on Information Retrieval, pp. 768–774, 2016.
[40] Jingshu Liu, Zachariah Zhang and Narges Razavian, "Deep EHR: Chronic Disease Prediction Using Medical Notes," arXiv:1808.0[41]28v1 [cs.LG], 2018.
[41] Sebastian Gehrmann, Franck Dernoncourt, Yeran Li, Eric T. Carlson, Joy T. Wu, Jonathan Welt, John Foote Jr, Edward T. Moseley, David W. Grant, Patrick D. Tyler and Leo A. Celi, "Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives," PLoS One, Vol. 13, No. 2, 2018.
[42] MD Samiul Islam, Haider Muhamed Umran, Samir M and Mohammed Karim, " Intelligent Healthcare Platform: Cardiovascular Disease Risk Factors Prediction Using Attention Module Based LSTM," 2nd International Conference on Artificial Intelligence and Big Data, pp. 167-175, 2019.
[43] Manabu Torii, Jung-wei Fan, Wei-li Yang, Theodore Lee, Matthew T. Wiley, Daniel S. Zisook and Yang Huang, "Risk factor detection for heart disease by applying text analytics in electronic medical records," Elsevier, Journal of Biomedical Informatics, Vol. 58, pp. S164-S170, 2015.
[44] Jie Ji, Bairui Chen and Hongcheng Jiang, "Fully-connected LSTM–CRF on medical concept extraction," Springer, International Journal of Machine Learning and Cybernetics, Vol. 11, pp. 1971–1979, 2020.
[45] Shi S, Qin M and Shen B, "Association of cardiac injury with mortality in hospitalized patients with COVID-19 in Wuhan," JAMA Cardiol, Vol. 5, No. 7, pp. 802-810, 2020.
[46] Jie Ji, Bairui Chen and Hongcheng Jiang, 2020, "Fully-connected LSTM–CRF on medical concept extraction", Springer, International Journal of Machine Learning and Cybernetics, Vol. 11, pp. 1971–1979.
[47] Jiawei, MichelineKamber and Jian Pei, "Data Mining Concepts and Techniques", Third Edition, Elsevier.