# A Review on Data Mining in Healthcare

**Ogundele I.O, Popoola O.L, Oyesola O.O, Orija K.T**

*Abstract*— **Data Mining is an advancing area in healthcare. Health data requires analytical methodology in identifying vital information that are used for decision making. Data mining is important for the healthcare sector in identification and detection of diseases, help researchers to make effective healthcare policies, develop recommendation systems and health profiles for patients. There are difficulties in evaluating the large data generated in the healthcare sector that are used to discover knowledge and find patterns for decision making. Healthcare data needs to be analyzed accurately in diagnosis, management and treatment of diseases. Data mining applications in health could have tremendous usefulness and potentials in healthcare industry. In this paper, we reviewed data mining techniques, its processes, tools, related works in healthcare system. The paper conclude with health analytics stages, data (source and transformation), and specific areas of application in data mining. The knowledge will help to reduce unnecessary spending and make accurate decision from the volume and complexity of the heath care data available.**

*Index Terms*— **applications, treatment, healthcare, techniques.**

## I. Introduction

The healthcare sector requires data mining in discovery of knowledge and finding patterns for decision making. Data mining is the most advancing field of study which requires finding useful and meaningful details from a large data. Health data requires analytical methodology in identifying vital information that are used for decision making. Detection, prevention and management of diseases including fraud in the health insurance, reduce spending in the solution of medical care are some of the importance of data mining. It also help researchers to make effective healthcare policies, develop recommendation systems and health profiles for patients. Volumes of data are generated in the healthcare industry that needs a database system to be stored for proper diagnosis and treatment of patients [1]. The volumes of data are complicated and complex to analyze so as to make meaningful decision about the patient health status. Data can consist of the cost for treatment, hospital, medical claims, patients, doctor, history etc. due to the complexity of the data, data mining tools is necessary for analyzing and discovering knowledge from the data to enhance the processes of the patient and management. The result of using data mining in healthcare sector is for classifying diseases of the patients and to assist in treatment and management of diseases. It helps to predict the duration of admission of patients in hospital, diagnosis of patients and accurate management information system. Currently, technologies and data mining techniques helps to reduce spending and evaluate the features that are responsible for diseases [2].

Application of data mining have relevant use in healthcare. It is important to collect, store, prepared and mined data, to make healthcare data clean and correct. Clinical practices and standardization of distributing data across organizations to help in healthcare data mining technologies [3].

Data mining promises great benefits in healthcare sector. The slow advance of technology and complexity of the volume of data make implementation of data mining strategies difficult. Till date, data mining in most part remains an academic practice. Data mining techniques had been used by academicians such as Neural Networks, decision trees, Support Vector machine, Naïve Bayes and genetic algorithm to write and publish research papers.

Data mining processes can be fully or partially automated to analyze the volume of data that are uncertain such as cluster of data, anomaly detection or outliers and data dependencies. Input data of patients are collected into the database based on the dataset features which are further used for analysis in diagnosis in other to obtain more accurate prediction outcome for decision making. That data mining process are data collection preparation, data collection, data preprocessing, and data transformation but do not include knowledge extraction and evaluation steps.

Healthcare data are stored in electronic format all over the world in health organization. The format of the data contains patient's details which are of vast data. Due to the increase in in data, there exist complexity and complications. It can be worrisome when using traditional methods in analysis this set of data to generate meaning knowledge from it. The field of mathematics, computer and statistics makes it easy to discover meaningful information from volumes of complex data which makes data mining to be of great benefits to the healthcare sector.

Data mining extracts meaningful information from complexity of data which were in a raw form. Numerous benefits are provided with the use of data mining in healthcare such as detection of fraud, detection of abuse of drugs, proper diagnosing of patients, treatments, early detection of diseases, survivability of patients etc.

Data mining techniques have been applied by various researchers. Such techniques are classification, association, clustering etc. the techniques play a vital role in the healthcare industry to support decision making, proper diagnosis, selection of treatments and prediction [3].

In this paper, we reviewed techniques of data mining in healthcare, tools used for data mining. The paper conclude with health analytics stages and specific areas data mining application in healthcare. The knowledge will help to reduce unnecessary spending and make accurate decision from the volume and complexity of the heath care data available.

## II. Data Mining Processes

The availability of the volume of data generated in healthcare industry need to be transform into a meaningful information for decision to occur. Data mining provides a great promise in analyzing complexity of data to generate information. The process of data mining helps to discover knowledge which are done in seven steps starting from selection stage to knowledge discovery.

**Selection** Certain parameters are used to pick the data which is the first stage of the data mining.

**Preprocessing** This stage nullify some of the parameter that are not needed. It helps to have a clean and correct data.

**Transformation** It changes data that are necessary for specific solution. Data particular to that problem are transformed.

**Data mining** this stage helps to discover knowledge from complexity of data. It is called a knowledge discovery stage.

**Interpretation and evaluation** the information generated from data mining stage are evaluated. The evaluation of the data yield discovery knowledge from the complexity that will be useful for decision making.

.

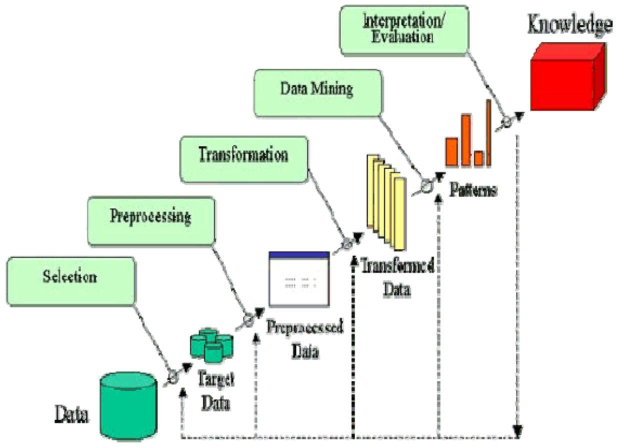

Fig 1: Application of Data Mining–A Survey Paper [4].

## III. DATA MINING TECHNIQUES IN HEALTH CARE

Supervised and unsupervised learning are the two classes of data mining techniques. Supervised learning involves a teacher that helps to learn. The learning predict an outcome based on certain criteria. Examples of such learning are classification, regression.

Similarly, unsupervised learning is a techniques that does not involves a teacher. It outline class of data without his assignment. Common example is the clustering. Table 1 shows the summary of the supervised and unsupervised learning.

Table 1: Characteristics and techniques of the unsupervised and supervised learning [5].

| | **Characteristics** | **Techniques** |
|---|---|---|
| Supervised Learning | • involves teacher<br>• identify class<br>• Mostly used | • Classification<br>• Statistical Regression |
| Unsupervised Learning | • No teacher involve<br>• Classes are defined<br>• Not frequently used | • Clustering<br>• Association Rule |

The data mining techniques are applied in the large complexity data to discover knowledge in data. Classification, regression, association and clustering have been used in this regard.

The different classification algorithms mentioned below in fig 2 are used analysis of various diseases.

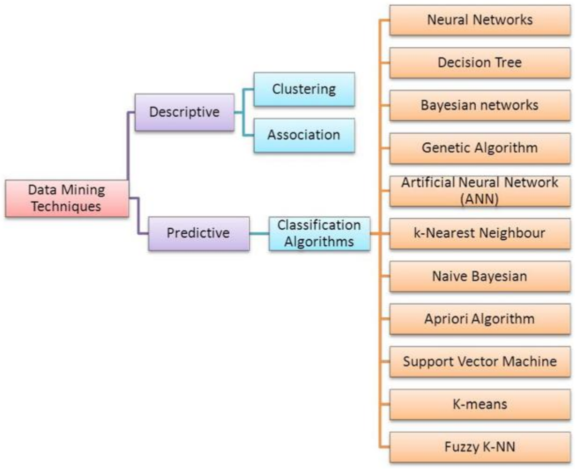

Fig 2: Different Data mining techniques in healthcare [6].

**Clustering:** it identify a categories of data in a finite set which describe the task [7]-[8]. It can be used to predict an outcome. K-means and x-means are some of the algorithms that have been used in clinical process and diagnosing results.

a. **Partitioned Clustering:** classify related dataset features into different group and analyst to known the clusters that are generated. Algorithm of the partitioned clustering helps to predict and diagnosed symptoms for a particular diseases. It is divided into two types been k-medoids and k-means.

b. **Hierarchical Clustering:** it cluster dataset features in form of hierarchy that are used for prediction in healthcare. It can breakdown its process either from top to down or from down to up methods. Strategies used in the clustering are classified into two. (1) Agglomerative and (2) Divisive. The outcome of the clustering are often presented in a dendrogram.

c. **Density Based Clustering:** the clustering method compliment the disadvantages of the partitioned and hierarchical clustering. It cluster together similar points that are collected in a datasets, points that are closer to each other. This detect anomaly in points that are gapped within themselves (whose close neighbors are too far away). It is the most used clustering algorithm because of the efficiency.

The clustering techniques is summarized in table 1 into advantages and disadvantages.

Table 2: Summary of the clustering techniques, benefits and drawback

| **Methods** | **Benefits** | **Drawback** |
|---|---|---|
| K-means clustering | • Simple clustering approach<br>• Efficient<br>• Less complex method | • Requires number of cluster in advance<br>• Problem with handling categorical attributes.<br>• With non-convex shape might be difficult to discover. |

| | | |
|---|---|---|
| | | • The result of the outliers may differs. |
| Hierarchical clustering | • Implementation is not difficult<br>• Exceptional graphical abilities.<br>• Clustering figures need not to be specify | • Have cubic time complexity in many cases so it is slower<br>• It allows user to undo decision<br>• Gaps in between points might be difficult to discover |
| Density based clustering | • Clustering figures need not to be specify<br>• Unequal shapes can be taking care off.<br>• Complicated data are easily handles | • Can handle points of different gaps<br>• The more the data, the better the results. |

**Association:** finds relationship from datasets by classifying the data in other to predict and to give better outcome.it is used where better accuracy is required. The two types of the functionalities are classification and association rule mining. In association, no teacher is involve because it's an unsupervised learning where no attributes is needed in discovering the rule. Similarly, classification uses supervised learning algorithm to predict and classify dataset features from unknown data. Association classification encompasses two stages listed below:

i. Set of predetermined classes
ii. Analysis of error rate.

**Classification:** the data set are known. For example, framework for classification and prediction of patient's survivability from the previous knowledge for a period of years. Software used in classification can learn from the dataset to predict future occurrence. In classification the datasets features can be classified as low, moderate, high and very high based on the symptoms of the diseases diagnosed. The forms of learning used in classification is supervised that involves teachers having known data label. In healthcare industry, the classification is the most widely used methods for detection, prediction and optimization.

a. **K-Nearest Neighbour (K-NN):** It uses classification and regression techniques and is easy to use method. The output depends on whether k-NN is used for classification or regression.

In KNN, new data introduced into the database are analyzed by finding the subset of that set of data to get the optimal solution to be able to predict an accurate result. The research in [9] conducted on the software of nearest neighbor method that can be used as a yardstick to predict heart diseases. The research work produce an efficient result of the diagnosed diseases, the result generated have accuracy of 97.4% which is higher than other machine learning techniques used.

b. **Decision Tree (DT):** Decision tree is a tree-like model of data in the database. It is applied in operation research and machine learning to make decision that will to meaningful conclusion and also for classification in data mining to extract knowledge in clinical data [10]-[12]. It consist of three types of nodes which are represented as decision nodes, chance nodes and end nodes. Many researchers have used decision tree for prediction of future occurrences and this have helped to improve the accuracy of the result generated.

c. **Artificial Neural Networks (ANN):** ANN is a biological neural network that has a unique characteristic to learn and carry out parallel processing of information. The network connection and strength is determined by the activation function which can be either linear or non-linear. Each layer is a subset of the processing element in the neural network. It comprises of layers which are input layer, hidden layers and output layers. The hidden layer is between the input layer and the output layer [13]. Examples of some of the activation which were earlier mentioned are linear (binary, sigmoidal) and non-linear (tan hyperbolic sigmoidal functions). By adjusting the weight, the neural network adapts itself to learn and optimise to produce the desired output. ANN has been used in healthcare sector by applying the classification methods [14]. The main characteristic feature of the ANN such as ability to learn makes it possible to diagnose patients accurately [15]. The use of ANN can help predict the best practice in management of diseases based on the symptoms.

d. **Bayesian Classifier:** [16] and [17] predict diseases by using Bayesian classifier method which is important in the healthcare field. . In medical science, Bayesian Classifier is based on probability theorem and can be used as the logical process of performing medical diagnosis, particularly in automated medical diagnosis decision support systems [18]. It can handle an arbitrary number of independent variables whether continuous or categorical.

e. **Support Vector Machine (SVM):** its applied using classification and regression to evaluate data in a supervised learning form. It divides in hyperplane by classifying. It is easier to implement, it divides into two classes of hyperlane/line. SVM can optimize processes which makes it to be more efficient and effective. It has been significantly applied in healthcare for classification of images, identification of features for prediction [19]. Difficult quadratic programming can be solved with the use of SVM [20]. Table 3 shows the summary of the benefits and drawback of the data mining techniques.

Table 3: Summary of the classification techniques, benefits and drawback

| **Techniques** | **Benefits** | **Drawback** |
|---|---|---|
| K-NN | • Simple to implement<br>• Efficient and effective in training data | • Data requires large space for storage in the database<br>• Data overfitting<br>• Delay in testing |
| Decision Tree | • Building decision tree does not requires prior knowledge<br>• Reduces anomaly and assign specific values to problem. | • Only requires one attributes.<br>• It generates categorical output<br>• Can be unstable because the data are dependent on the dataset features. |

| | | |
|---|---|---|
| | • The diversity of data can easily be processed<br>• Easy to understand<br>• Numeric and categorical data are only processed | • May suffer from overfitting<br>• Classifies by rectangular partitioning |
| Support Vector Machine | • Improve accuracy than other machine learning classifier<br>• It has a regularization parameter<br>• Uses kernel trick<br>• Defined by a convex optimization problem. | • Expensive to implement<br>• Problem need to be formulated as 2-class classification<br>• Consume time<br>• Only solve problem of binary numbers |
| Neural Network | • Result can be generated with incomplete information<br>• Information are easily store on the entire network<br>• Fault tolerance<br>• Can learn event to make decisions | •Requires higher processing power<br>•Data can have overfitting<br>• Difficult to discover the structure of the network. |
| Bayesian Network | • Easy to implement<br>• Can handle continuous and discrete data.<br>• Not sensitive to irrelevant features<br>• Lesser training data are required | • Dependency of the variable might result in inconsistency of the result<br>• Computational infeasible<br>• unautomated |

**Logistic Regression:** logistic regression is used in healthcare field for predicting diseases. It uses binary form to show relationship between a variable to analyze set of data given. It is mainly a statistical tool that is used in data mining. It only analyzed categorical data that are logistic and non-linear.

**Apriori Algorithm:** solve problem from the most complex to the least and group the data to be analyze. When all levels of the problem are solved, the algorithm will then terminates. It shows the connection between two inputs to separate the consistent and inconsistent of the inputs. Different types of apriori algorithm are used, like Hash table, transaction reduction, partitioning etc. [21]. [22] Considered apriori algorithm for creating associative rules that are used for medical billing. [23] Improved on the medical billing to effectively generate useful information. Apriori algorithm are applied in healthcare for predict diseases.

## IV. DATA MINING TOOLS USED IN HEALTHCARE

Data mining tools helps to analyzed volumes of complex data based on the dataset attributes that users specify in determining trends of occurrences. The software can be used for diagnoses, prediction, and management of diseases to extract knowledge and make decisions. Due to the availability of various software tools used, the choice of choosing appropriate software to solve a particular problem becomes difficult [24]. The most common data mining tools are explained below:

a. **WEKA (Waikato Environment for Knowledge Analysis):** WEKA is a program tools used in data mining processes. It is a software that is develop using java programming language that runs on different operating system. Weka compliment several data mining processes. The software can connect with the data directly or from the java code. It make use of the Graphical user interface (GUI) for access to his performance and functionality.

b. **KEEL(**Knowledge Extraction based on Evolutionary learning**)**: KEEL uses clustering, regression, classification to extract pattern from datasets. It is an open source software but source program may be hidden. Complete analysis can be performed using the KEEL data mining tools.

c. **R:** R is an open source program for computation and statistical analysis. R software is of great benefits to the research and development world and health industry. The software for development of R data mining tools are FORTRAN, C and R.

d. **KNIME:** Konstanz Information Miner, is an open source software that are used for analyzing and modelling data. Machine leaning and data mining features are supported with KNIME software. KNIME has been applied in clinical research, detection of diseases and classification. KNIME can generate a work processes that can documented in different format.

e. **RAPIDMINER:** RAPIDMINER, data mining, machine learning, text mining and business analytics are development by an organization that provides the same software. It is used in business, finance, banking, insurance, medical, and education in analyzing data which support data mining processes. It is an open source software used in various filed of human endeavor.

**f. ORANGE:**Orange is an open-source software used in machine learning. It is characterized by two features of front end and back end. The front end uses visual programming while back end uses python libraries. It was developed using C++ and Python programming. In science, new machine leaning algorithms and techniques in genetics, medical can be tested using ORANGE. It was used in education to teach student biology, science and other related courses to medical.

## V. HEALTH ANALYTICS

Healthcare organization can improve their efficiency and effectiveness in the delivery of their duties as it relate to data extraction from patterns in other to making accurate decision by the use of information technology, computer based tools, mathematical computation, statistical tools etc. in health analytics decision making and problem solving is very vital for proper care of the patients. It enables the practitioners and health workers to make policy, improve working condition with d use of the technological base support tools.

Health analytics stages start with collection, preprocessing, transformation of health data. There are four types of health analytics which are discussed below.

a. **Descriptive analytics** is the most easy to use, simplest health analytic by every individual [25]. Data are easily calculated, interpreted, develop and implement in healthcare. Descriptive analytics uses graphical representation for better understand by practitioners in the field. Descriptive analytics gives overall details of number of patients treated, revenue generated, what are the symptoms of the patients, diseases diagnosed, how are they treated and managed to improve their condition. It gives a summary of the historical data for generate meaningful information.

b. **Predictive analytics** this gives a focus to the use of information. Predictive analytics which is used to identify future probabilities and trends to predict future occurrences. It ask a question of what could happen in the future. It gather information from historical background, learn patterns from the dataset to be able to predict the future by extract knowledge. Volumes of complicated data available in healthcare allow predictive analytics of the techniques used in data mining. Health professional could ask, what drugs to use for the treatment? Who could be affecting by this diseases next? To predict result of a patient and allocate resources appropriately.

c. **Prescriptive analytics** are applied when there are several options in the health problems or choice to deliver the best prescriptive analytics. Prescriptive analytics ask, how do we respond to those potential future events? This has been used in healthcare for treatment and drug prescription. Several drugs might be prescribed by weighing the pros and cons. It can be used to determine most accurate solution for a given problem in a dataset features.

d. **Discovery analytics** make use of the discovered knowledge to come up with new invention and innovation in the field of data mining healthcare. For examples, from previously known drugs to discover new drugs. This can help to discover new treatment, new diseases, and new medication from the known discovery. Data discovery is used in some healthcare organization to optimize processes that in a given dataset features. Many application of data mining involves all the four health analytics.

## VI. HEALTH DATA SOURCE AND TRANSFORMATION

**Health data sources:** Health data can be collected from both hospitals and healthcare practitioners. Hospital data include patient’s data, diagnosis, drug prescription and treatment data while healthcare practitioners’ data are data collected from government healthcare agencies such was World Health Organization (WHO) and other healthcare organizations. The most reliable healthcare data comes from governmental sources or healthcare professional organizations. The data comes from several sources of varying quality, inconsistent and incorrect that are of immense volume. The volume of the data need to be organized, structured and processed to get meaningful result. There are five most useful sources of data which are clinical data, claims data, patient-generated data and pharmaceutical data.

**Health data transformation:** data transformation is the process of changing data to information, usually from one format to another. The format from the source data are changed to the format of the desired information. Data needs to be transform before it can make more sense. Data in a raw state is not meaningful and useful until it’s transformed. Different approach can be employ to transform data such as (i) service-oriented architectural (ii) data warehousing. Data are made available in different form which are executed incrementally. Commonly used transformational languages are Perl, AWK, TEXT, XSLT and template languages and processors.

## VII. RELATED WORKS

Data mining play a vital role in healthcare industry. It is predominantly use for detection and prediction of diseases. Various researchers acknowledged the fact that there is a demonstrated need for the use of data mining in healthcare. This will help healthcare practitioner to effectively and efficiently deliver better services, most especially in prevention, prediction and management of diseases.

[26] The researchers’ works on innovative methods for detection of breast cancer using classification data mining techniques. The classification techniques and WEKA were used as data mining tools. The research examine the accuracy of various classification data mining techniques. A total of six hundred and eighty three data of ten datasets features were used to determine the accuracy of the data mining techniques used. Three data mining techniques were compared using WEKA and the result shows the Sequential Minimal optimization has better accuracy than other techniques.

[27] Focus on using data mining classification techniques to analyzed chronic diseases. The work predict chronic kidney diseases using data mining classification techniques such as ANN and Naives Bayes. Rapid miner tools was used to compare the two classification techniques and the results shows that Naives Bayes performs better than ANN.

[28] Research on the prediction of Dengue disease using WEKA as data mining tool. Some datasets features were used based on the symptoms of the diseases. They compared different data mining techniques with their own classifier algorithm. The result shows that Naïve Bayes have an accuracy of 100% and J48 has accuracy of 99.70%. Naives Bayes give a better prediction of Dengue diseases survival.

[29] The researcher’s uses patient’s datasets to predict heart diseases using classification data mining techniques. It determine the best framework that can give better accuracy in terms of performance for the diagnoses of the dataset feature input. Classification data mining techniques was used with WEKA to predict and interpret the result.

[30] The researcher uses A-priori and k-means algorithm to predict heart diseases and kidney failure. Predictive system is developed to evaluate data from the user which consists of 42 dataset features. Data are analyzed using data mining tools and the result were evaluated with operating characteristics (ROC) and calibration plots.

[31] The paper research on applying artificial neural Network (ANN) to an end stage kidney disease prediction. The system classify patients data based on the condition of the health using ANN for an end stage kidney diseases. The developed tools is useful in clinical practices to predict future outcome of the patients.

## VIII. SPECIFIC APPLICATION AREAS OF DATA MINING IN HEALTHCARE

The application of data mining in healthcare promises to advance clinical practice of diseases in diagnosis, treatment, prevention, prescription and optimization of the fast delivery to patient with these diseases. There are large data in health industry about patient status on diagnosis, treatment and cost which needs to be analyzed to extract meaningful information and knowledge from it. Data mining applications in healthcare are listed below.

a. Treatment effectiveness
b. Healthcare management
c. Fraud & abuse
d. Hospital Infection Control
e. Smarter Treatment Techniques

## IX. CONCLUSION

This paper reviewed previous works on data mining in healthcare. In this research work, we first discussed the background, definition, and processes of data mining, techniques used in healthcare, benefits and drawback of those techniques mentioned. Data mining tools used in healthcare to predict future outcome from information generated that assist health organizations to make decision. The descriptive, predictive, prescriptive and discovery analytics were introduced. Sources and transformation of health data were discussed. Related works of previous research were reviewed and specific application areas of data mining in healthcare were mentioned.

## REFERENCES

[1] D. P. Varghese and P. B. Tintu, “A Survey on Health Data using Data Mining Techniques‖,” *Int. Res. J. Eng. Technol.*, vol. 2, no. 07, pp. 56–2395, 2015.

[2] Z. S. Daliri, “Data Mining for Health Care Industry: A Practical Machine Learning Tool,” *Int. Res. J. Multidiscip. Stud.*, vol. 3, no. 4, 2017.

[3] M. Tomari, M. Razali, W. Zakaria, and W. Nurshazwani, “An empirical framework for automatic red blood cell morphology identification and counting,” 2015.

[4] A. Sharma, R. Sharma, V. K. Sharma, and V. Shrivatava, “Application of Data Mining–A Survey Paper,” *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 2, pp. 2023–2025, 2014.

[5] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[6] S. Patel and H. Patel, “Survey of data mining techniques used in healthcare domain,” *Int. J. Inf.*, vol. 6, no. 1/2, 2016.

[7] D. Vatsalan, Z. Sehili, P. Christen, and E. Rahm, “Privacy-preserving record linkage for big data: Current approaches and research challenges,” in *Handbook of Big Data Technologies*, Springer, 2017, pp. 851–895.

[8] R. Veloso *et al.*, “A clustering approach for predicting readmissions in intensive medicine,” *Procedia Technol.*, vol. 16, pp. 1307–1316, 2014.

[9] M. Shouman, T. Turner, and R. Stocker, “Applying k-nearest neighbour in diagnosing heart disease patients,” *Int. J. Inf. Educ. Technol.*, vol. 2, no. 3, pp. 220–223, 2012.

[10] N. Sharma and H. Om, “Data mining models for predicting oral cancer survivability,” *Netw. Model. Anal. Heal. Informatics Bioinforma.*, vol. 2, no. 4, pp. 285–295, 2013.

[11] K.-J. Wang, B. Makond, and K.-M. Wang, “An improved survivability prognosis of breast cancer by using sampling and feature selection technique to solve imbalanced patient classification data,” *BMC Med. Inform. Decis. Mak.*, vol. 13, no. 1, p. 124, 2013.

[12] H. M. Zolbanin, D. Delen, and A. H. Zadeh, “Predicting overall survivability in comorbidity of cancers: A data mining approach,” *Decis. Support Syst.*, vol. 74, pp. 150–161, 2015.

[13] S. Gupta, D. Kumar, and A. Sharma, “Data mining classification techniques applied for breast cancer diagnosis and prognosis,” *Indian J. Comput. Sci. Eng.*, vol. 2, no. 2, pp. 188–195, 2011.

[14] A. E. Akinwonmi, “On the Diagnosis of Diabetes Mellitus Using Artificial Neural Network Model Artificial Neural Network Models,” 2011.

[15] B. Liu, M. Wang, H. Yu, L. Yu, and Z. Liu, “Study of feature classification methods in BCI based on neural networks,” in Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the, 2006, pp. 2932–2935.

[16] R. Armañanzas, C. Bielza, K. R. Chaudhuri, P. Martinez-Martin, and P. Larrañaga, “Unveiling relevant non-motor Parkinson’s disease severity symptoms using a machine learning approach,” *Artif. Intell. Med.*, vol. 58, no. 3, pp. 195–202, 2013.

[17] S. Bandyopadhyay *et al.*, “Data mining for censored time-to-event data: a Bayesian network model for predicting cardiovascular risk from electronic health record data,” *Data Min. Knowl. Discov.*, vol. 29, no. 4, pp. 1033–1069, 2015.

[18] B. Zheng, S. W. Yoon, and S. S. Lam, “Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms,” *Expert Syst. Appl.*, vol. 41, no. 4, pp. 1476–1482, 2014.

[19] P. J. García-Laencina, P. H. Abreu, M. H. Abreu, and N. Afonoso, “Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values,” *Comput. Biol. Med.*, vol. 59, pp. 125–133, 2015.

[20] B. R. Devi, K. N. Rao, S. P. Setty, and M. N. Rao, “Disaster prediction system using IBM SPSS data mining tool,” *Int. J. Eng. Trends Technol.*, vol. 4, pp. 3352–3357, 2013.

[21] Y. Ji, H. Ying, J. Tran, P. Dews, A. Mansour, and R. M. Massanari, “Mining Infrequent Causal Associations in Electronic Health Databases,” in *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, 2011, pp. 421–428.

[22] B. M. Patil, R. C. Joshi, and D. Toshniwal, “Association rule for classification of type-2 diabetic patients,” in *Machine Learning and Computing (ICMLC), 2010 Second International Conference on*, 2010, pp. 330–334.

[23] U. Abdullah, J. Ahmad, and A. Ahmed, “Analysis of effectiveness of apriori algorithm in medical billing data mining,” in *Emerging Technologies, 2008. ICET 2008. 4th International Conference on*, 2008, pp. 327–331.

[24] R. Mikut and M. Reischl, “Data mining tools,” *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 1, no. 5, pp. 431–443, 2011.

[25] A. Perer, “Healthcare analytics for clinical and non-clinical settings,” in *Proceedings of CHI Conference*, 2012.

[26] V. Chaurasia and S. Pal, “A novel approach for breast cancer detection using data mining techniques,” 2017.

[27] V. Kunwar, K. Chandel, A. S. Sabitha, and A. Bansal, “Chronic Kidney Disease analysis using data mining classification techniques,” in *Cloud System and Big Data Engineering (Confluence), 2016 6th International Conference*, 2016, pp. 300–305.a

[28] K. A. Shakil, S. Anis, and M. Alam, “Dengue disease prediction using weka data mining tool,” *arXiv Prepr. arXiv1502.05167*, 2015.

[29] H. D. Masethe and M. A. Masethe, “Prediction of heart disease using classification algorithms,” in *Proceedings of the world Congress on Engineering and computer Science*, 2014, vol. 2, pp. 22–24.

[30] A. Chaudhary and P. Garg, “Detecting and Diagnosing a Disease by Patient Monitoring System,” *Int. J. Mech. Eng. Inf. Technol.*, vol. 2, no. 6, pp. 493–499, 2014.

[31] T. Di Noia *et al.*, “An end stage kidney disease predictor based on an artificial neural networks ensemble,” *Expert Syst. Appl.*, vol. 40, no. 11, pp. 4438–4445, 2013.