Journal of Applied Technology and Innovation (eISSN: 2600-7304)
vol. 1, no. 2, (2017), pp. 10-27

Data Mining Techniques in Diagnosis of Chronic Diseases

Keerthana Rajendran
Faculty of Computing, Engineering & Technology
Asia Pacific University of Technology & Innovation
57000 Kuala Lumpur, Malaysia
Email: keer.abhitham@gmail.com

Abstract - Chronic diseases and cancer are raising health concerns globally due to lower chances of survival when encountered with any of these diseases. The need to implement automated data mining techniques to enable cost-effective and early diagnosis of various diseases is fast becoming a trend in healthcare industry. The optimal techniques for prediction and diagnosis vary between different chronic diseases and the disease related-parameters under study. This review article provides a holistic view of the types of machine learning techniques that can be used in diagnosis and prediction of several chronic diseases such as diabetes, cardiovascular and brain diseases, chronic kidney disease and a few types of cancers, namely breast, lung and brain cancers. Overall, the computer-aided, automatic data mining techniques that are commonly employed in diagnosis and prognosis of chronic diseases include decision tree algorithms, Naïve Bayes, association rule, multilayer perceptron (MLP), Random Forest and support vector machines (SVM), among others. As the accuracy and overall performance of the classifiers differ for every disease, this article provides a mean to understand the ideal machine learning techniques for prediction of several well-known chronic diseases.

Index Terms - Data mining, Healthcare systems, Machine Learning, Big data analytics

1. Introduction

The evolution of healthcare industry from the traditional healthcare system to the utilization of Electronic Health Records (EHR) system has introduced the concept of big data in the healthcare sector. Big data is defined in terms of 4Vs, which represent the volume, variety, velocity and veracity. The large amount of data generated through omics data such as genomic, proteomic, transcriptomic, epigenomic and metabolomic, as well as EHR data from clinical records, administrative records, charts and laboratory test results, contribute to the copious volume of data in the healthcare industry. Recently, social media data are also being integrated into the EHR system to analyse the patient behaviour. The data are produced in variety of formats from unstructured, semi-structured to structured data with errors such as missing values. Different data have different velocity of generation, so their acquisition time and frequency are largely

varied. Moreover, these data are obtained from diverse sources whose reliability is not authenticated (Raghupathi & Raghupathi, 2014; Auffray et al., 2016). By employing data mining methods into big healthcare data (BHD), several patients can be assessed at the same time and better care can be given based on improved understanding of patient medical profile. Some of the benefits of applying data mining in healthcare are as follows (Durairaj & Ranjani, 2013):

- Optimized management of hospital resources
- Better understanding of patients to improve customer relation
- Detection of fraud and abuses found in insurance and medical claims
- Control the widespread of hospital infections and identify high-risk patients
- Enhanced patient care and treatments through healthcare decision support system

Data mining is defined as the process of identifying unknown patterns, relationships and potentially valuable information from huge datasets with the use of statistical and computational approaches. The primary tasks of data mining are to build predictive and descriptive models as illustrated in Figure 1 (Durairaj & Ranjani, 2013). Data mining techniques that are used commonly in healthcare include classification, decision tree, k-Nearest Neighbour (k-NN), support vector machine (SVM), neural network, Bayesian methods, regression, clustering, association rule mining and Apriori algorithm. These machine learning techniques are helpful in assessing the risk factors such as socioeconomic and environmental behaviour of individuals besides their medical profiles in diseases, especially chronic illnesses and cancer (Tomar & Agarwal, 2013; Dey & Rautaray, 2014; Ahmad, Qamar & Rizvi, 2015).

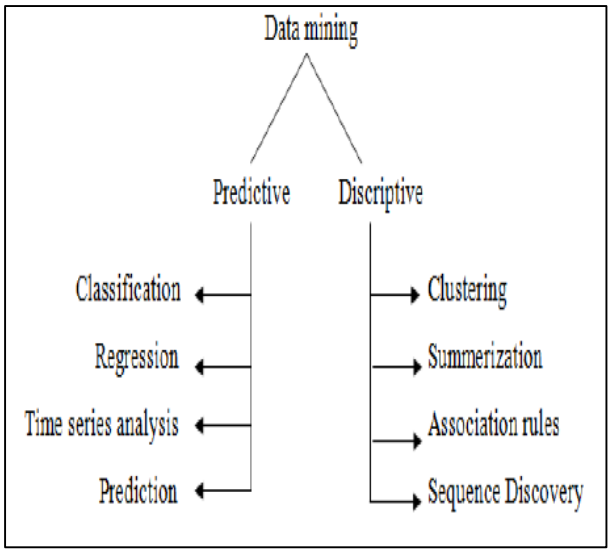

Figure1: Predictive and descriptive data mining techniques (Durairaj & Ranjani, 2013)

This review article focuses on the application of data mining techniques in chronic diseases, with a central focus on different cancer subtypes. The sections are segmented as follows: Section 2 highlights the importance of data analytics in healthcare in general, the current trends of data analytics and some of the data mining methods used in health

informatics. Section 3 elucidates on the existing data mining techniques used in identification of most known chronic diseases such as diabetes disorder, cardiovascular disease, brain disorder and chronic kidney disease (CKD). Each of these disease is explained as a sub-section of Section 3. Section 4 converges its attention to the employment of machine learning techniques in diagnosis, prognosis and prediction of various cancer types which are breast cancer, lung cancer and brain cancer, divided into sub-sections. Section 5 discusses the overall review of the involvement of data mining approaches in chronic diseases and cancer in terms of their limitations and benefits. Section 6 is the conclusion which provides an insight on the challenges of machine learning application in healthcare industry.

2. Data Analytics in Healthcare

Data analytics has a profound use in healthcare, especially in machine learning for the application of descriptive, prescriptive and predictive analytics. Medical data source ranges from EHR, genomic profiles to administrative and financial data, resulting in surplus of data which require extensive application of data mining algorithms to extract valuable output and make informed clinical decisions. In healthcare, the key function of data mining techniques involves the determination of rates of mortality due to disease risk factors and forecast of diseases at an early stage. In a book chapter by Hersh (2017), employment of data-driven measures for diagnosis and tailored treatment of diseases in precision medicine have also been highlighted. Genome-wide association studies (GWAS) incorporates genomic data into the EHR to identify disease disorders and obtain genome-level information. This information is processed and transformed into structured formats for computational analysis using statistical techniques and machine learning algorithms which results in insights gained from the disease prediction models. Inevitably analytics in healthcare data comes with several barriers such as misinterpretation of the transformed data due to coding leading to false positive results and ethical issues raised on the privacy and security of personal data as well as the rights of data access and sharing.

Another review article by Raghupathi & Raghupathi (2014) has carried a similar view point on the involvement of big data in healthcare along with its benefits and features, and conceptual frameworks and tools for data analytics used in healthcare. The vast amount of data generated in various structures and formats are growing exponentially and inability to detect accuracy of these data call in for the need of big data analytics in healthcare. Focusing more into the benefits of data mining in healthcare, the authors highlighted that data analytics can curtail the cost of treatment and diagnosis of patients, reduce trial and error practice in clinical trials by using data analytical tools and algorithms, estimate population health trends, recognize patients who are prone for re-admission, mitigate fraudulence and misuse, enable real-time update of patient conditions and perform genome-based analytics for precision medicine. The

architectural skeleton highlighted how multitude of data obtained from various sources in a raw form can be placed in a data warehouse to allow data transformation. These transformed data then go into a selection of tools/platforms for applied analytics. Hadoop is a prominent platform to analyse and process big data. Steps involved to apply big data analytics in healthcare are conception of project, manifestation of proposal, implementation of methods such as data compilation and processing, and lastly, deployment of the results. Obstacles such as privacy, data security, real-time data evaluation, quality and governance regarding big healthcare data should be brought to light.

The article by Dinov (2016) showcased the possible techniques to conquer these barriers so that convoluted data can be transformed into comprehensible format for analysis. To achieve an automatically processed decision, quantitative and structured format of BHD are vital. To derive statistically valuable data, some measures that can be used include visual mining, text analytics, information retrieval, data standardization and predictive modeling markup language (PMML) to interpret and analyse medical information. Social network analytics, which can be displayed in terms of nodes and edges, are used to identify relationships and patterns in the datasets. Multitude techniques such as k-NN, Gaussian mixture modeling (GMM) as well as un-supervised, semi-supervised supervised machine learning algorithms are utilized to segment, group and organize complex data. Incomplete records which might have missing values occurring at random or not random can be handled via logistic regression. Exploratory and explanatory analyses can be used to analyse and display incongruent data using cloud-based dashboards. Along this, predictive analytics is the crucial task of data mining in BHD. Programming languages such as SQL and NoSQL besides cloud computing are advances to refine BHD. Open-source platforms like Apache, Hadoop, MapReduce and Spark are freely accessible to analyse large data in the healthcare sector.

As healthcare information are known to exist in copious amount, it is necessary to construct a standardized series of steps to analyse these data and interpret them into knowledgeable output. A well-accepted process that is followed in healthcare data mining is Knowledge Discovery (KDD) which involves the interpretation of large volume of data and identifying a pattern in the data to attain an insight that improves decision-making, as shown in Figure 2 (Ahmad, Qamar & Rizvi, 2015). In KDD, selection of target data from the database is the first step, followed by data pre-processing where any unwanted data are filtered and the noisy data are eliminated. The raw and unstructured data are then transformed into a structured format for analysis. Data mining is the key process in KDD which involves descriptive and predictive analytics incorporating numerous algorithms and statistical measures to discover trends and build prediction models. Data mining is primarily assigned to carry out two approaches known as static end-point prediction and temporal data mining which consists of classification, regression, association rule learning, cluster analysis, hidden Markov model (HMM) and temporal association rule mining (TARM) on the transformed datasets. The interpreted

outcomes allow informed decision-making. Examples of data mining application in healthcare are artificial neural network of human brain, besides decision tree and nearest neighbour for classification and prediction. In precision medicine, genomic data incorporated into the EHR were analysed using clustering to identify cancer subtypes and provide tailored treatments (Taranu, 2015; Wu et al., 2017).

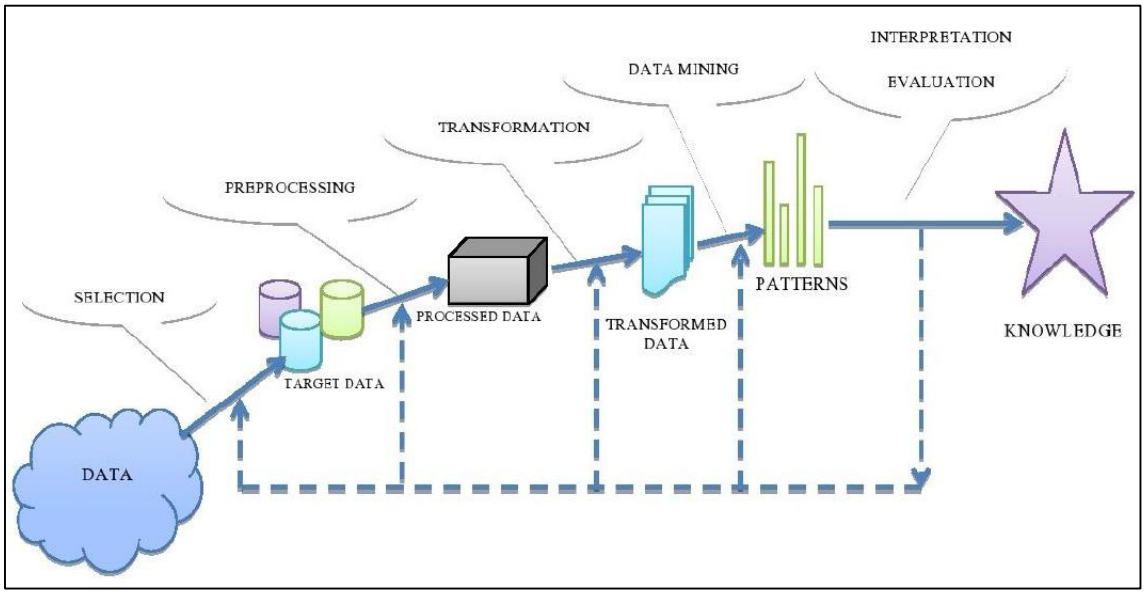

Figure 2: KDD Process (Ahmad, Qamar & Rizvi, 2015)

Application of computational knowledge in healthcare is called health informatics which is transpiring into a demanding field requiring data experts due to the evolution of big data in healthcare. Within this field, there are several subdomains such as bioinformatics, medical image informatics, clinical informatics and public health informatics which employs various levels of data generation from molecular, tissue, patient and population level data. These data are utilized to address the research questions posed in order to find answers for clinical, human biological and epidemic queries. Herland, Khoshgoftaar & Wald (2014) have elucidated on the various data mining techniques applied at each data levels in health informatics. At molecular level, patient cancer gene expression profiles were analysed to group leukaemia sub-types and to forecast the recurrence of colorectal cancer (CRC) among the subjects at the initial phase using classification and support vector machines (SVM) techniques. Tissue level involve methods like feature abstraction and selection applied on highly dimensional human brain images and MRI of brain tissue samples to develop an extensive neural network. Besides, Fuzzy Decision Tree (FDT) and classification were used to predict the occurrence of Alzheimer's disease at distinct stages. The third level uses the patient records tested using various scoring systems under classification technique and logistic regression to build prediction models for patient readmission rate, fatality estimate and life span of patient. Moreover, Alternating Decision Tree (ADT) and Principal Component Analysis (PCA) were also used for prediction based on patients' physiological

parameters, derived from EHR. At population level, text analytics were employed on social media data such as Twitter, internet search engines and messaging applications to provide patients with facts on illnesses. Real-time epidemics tracking and prognosis of a population health were determined using data mining techniques such as decision trees, SVM, Naïve Bayesian and logistic regression analysis.

3. Data Mining in Chronic Diseases Identification

The current trends in data mining field that are available to diagnose different chronic diseases that are the common causes of death worldwide such as diabetes, cardiovascular disease, brain disease and chronic kidney disease are illuminated in this section.

3.1 Diabetes Disorder

Diabetes mellitus is a condition where the body is unable to produce sufficient insulin, resulting in increased blood sugar level. Various risk factors contribute to diabetes which can be used as measures to predict diabetes. The article by Renuka Devi & Maria Shyla (2016) highlighted on diabetes mellitus discussed various data mining algorithms used on the Pima Indian Diabetes Dataset to determine the occurrence of diabetes. The data was cleaned to eliminate noise and replace missing instances. A total of six physiological variables were used to diagnose type 1 and type 2 diabetes, besides gestational diabetes. Some of the techniques used to classify the diabetes-related attributes include Naïve Bayes, Random Forest, Modified J48 Classifier, SVM, k-NN, genetic algorithm, etc. Software tools such as Weka, MATLAB, Tanagara, RapidMiner, etc. were used to perform data analytics operations containing all the machine learning techniques and statistical algorithms. Upon comparison across all the techniques, it was found that the highest prediction accuracy of 99.87% was achieved using Modified J48 Classifier with the aid of Weka and MATLAB tool.

Another study on type 2 diabetes mellitus by Hu et al. (2016) explained that impaired glucose tolerance (IGT) is a causative factor of this disease as well as atherosclerosis which can lead to cardiovascular disease. Hence, data mining techniques were employed to identify the patterns among IGT patients with an elevated risk of atherosclerosis. ACT NOW, a clinical trial dataset, was used as the training dataset where it was processed using imputation and categorical attributes were created for the disparate variables. First, feature selection using Fisher score was assigned to choose the attributes with most significance. Probabilistic Bayesian classifiers were used to generate the prediction model which has been proven to work well on small-scale, multimodal training and test datasets in other reported studies. Two more classification techniques known as multilayer perceptron (MLP) and random forest (RF) were adopted to compare the accuracy of the prediction, which was determined using Brier score and receiver

operating characteristic curve (AUC). The best predictor was found to be the Naïve Bayes with feature selection holding a better performance with 89.23% accuracy in comparison to the other two approaches with about 88% accuracy level each. Naïve Bayes technique proved to enable the determination and prognosis of IGT subjects who encounter rapid progression of atherosclerosis.

3.2 Cardiovascular Disease

Heart disease is one the most common chronic diseases that causes fatality in adults. Heart disease prediction and their risk factors analyses have been reported in several articles using machine learning algorithms such as SVM, decision tree, genetic algorithm, neural networks, Naïve Bayesian classifiers and Iterative Dichotomized 3 (ID3). But, association rule technique in cardiovascular disease prognosis has been barely explored. Hence, Khare & Gupta (2016) have incorporated association rule into their study to identify heart disease risk factors. The data was obtained from UCI open data repository. Training dataset was prepared by removing the missing observations and shortlisting the key attributes that are related to the analysis. The numerical attributes were nominalized to categorical attributes since Apriori algorithm of the association rule mining was used. This algorithm utilized candidate generation approach to discover the recurring (frequent) variable set. Rules were generated based on the presence of heart disease by focusing on the causative determinants for this disease, where classification association rules (CAR) were applied. Accuracy of the rule was validated in the training stage. The schematic representation of association rule method is shown in Figure 3. The results showed at least 85% confidence for all the rules generated. Attributes such as gender (male), older generation, increased serum cholesterol level, presence of asymptomatic chest pain and defective thalassemia are associated with the exposure to heart disease. While blood sugar level, an indirect measure of diabetes, was found to be negatively correlated to heart disease. This technique was found to be useful in focusing only on the primary risk factors of heart disease, which enables cost-effective diagnosis and time-efficient treatments.

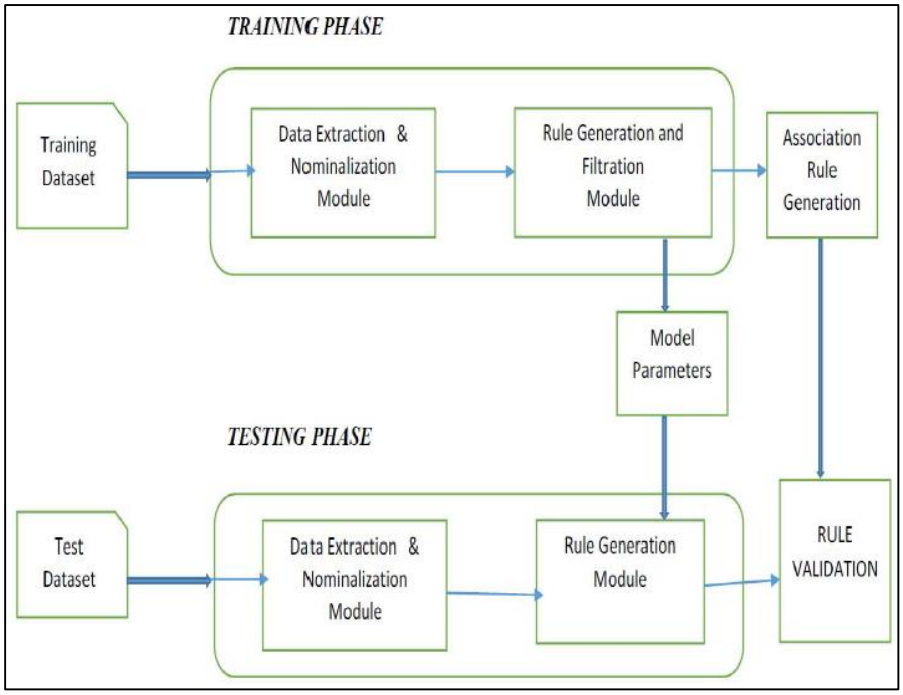

Figure 3: Association rule technique (Khare & Gupta, 2016)

3.3 Brain Disease

One of the well-known type of dementia among older people is Alzheimer's disease (AD). This neurodegenerative illness is one of the primary causes of fatality in the USA. Besides other external factors such as lifestyle and medical circumstances, genetic factors appear to have greater influence in the progression of AD. A study was conducted by Kumar & Singh (2016) to determine the participation of AD-related genes in the pathogenic pathway and diagnosis using decision tree method. The AD gene datasets were obtained from several online repositories and 2111 genes significant to the disease were selected. Feature selection of the variables were done using Chi-squared attribute evaluation and gain ratio (GR) methods. J48 algorithm found in Weka and C4.5 algorithm enabled in RapidMiner were used to achieve classification of the dataset and clustering of the genes were done through enrichment analysis. Mini Mental State Examination (MMSE) scores, one of the vital attributes, obtained through classification algorithm showed different values for distinct phases of AD. Upon classification of similar features, decision tree for the gene dataset was built using MMSE score as the root node. The roles of the genes were determined via enrichment analysis and 7 genes were found to be highly correlated with AD based on the association scores. The C4.5 algorithm proved to generate a better prediction accuracy with minimal error rate for prior AD diagnosis.

An early diagnosis of chronic diseases can reduce fatality greatly and this is a demanding area of data mining application in healthcare. But, errors in diagnosis often lead to delayed treatments or administration of wrong medications. Thus, a study was done by Chase et al. (2017) on patients with multiple sclerosis (MS) to recognize the signs

and symptoms of the disease in an early stage using natural language processing (NLP) of the unstructured medical notes in the EHR. Medical records of patients with MS and a randomly selected population in a clinic at Columbia University Medical Center (CUMC) were used as datasets for this study. The MS-positive population data were classified to build a prediction model for pre-recognition of MS, while the random patient group data were used to determine unidentified MS. A classifier was trained to recognize the terms that are linked and not linked to MS. Employing Naïve Bayes algorithm under classification technique from Weka tool, differences between patients with MS and those without MS were determined. Terms with similar features or category were classed into a bucket. The model generated a sensitivity of 75% and specificity of 91% in MS-identified patients, while in the random group 81% sensitivity and 87% specificity of classification were achieved. This shows that the method is feasible for pre-diagnosis of MS prior to the detection of ICD9 code, which is a gene marker for MS. The classification of the random population proved to have assisted in identifying patients who may have MS, based on the signs and symptoms but are missing the neurological characteristics of MS. The limitations of the study include restricted number of patients, the classifier was developed based on majority of Hispanic female patients' population, and only one NLP system was used to identify the MS terms.

3.4 Chronic Kidney Disease

Chronic kidney disease (CKD) has recently become an increasing health concern worldwide, but there are yet to be many research papers on computerized diagnosis for this disease. UCI Machine Learning warehouse dataset was used in a study by Subasi, Alickovic & Kevric (2017) and well-studied techniques such as artificial neural network (ANN), SVM, k-NN, C4.5 decision tree and random forest (RF) were used to build prediction models for CKD analysis. Under ANN, MLP approach was used to identify the correlation between the 24 variables and 2 potential output results (with CKD or without CKD) via the network of neurons and nodes. Supervised SVM was used to assess classification and regression analyses by segregating the training group using hyperplanes. An improvisation of ID3 known as C4.5 decision tree was used on this CKD dataset as it can analyse numeric variables, while Classification and Regression Tree (CART) under random forest was implemented. The training set comprised of 90% of the dataset while 10% was used as test dataset. The outputs were displayed using a Confusion Matrix for various data mining techniques. The classification models were generated using Weka tool and the performances of the machine learning algorithms were statistically validated using total precision, F-measure and overall classification accuracy. All the techniques used yielded exceptional performance accuracies but RF classifier proved to have the highest classification performance rate, even compared to other techniques used in previous literature studies. Thus, RF was proposed as an ideal data mining technique to evaluate for an early diagnosis of CKD at a faster time.

4. Machine Learning in Cancer Diagnosis and Prediction

The data mining techniques and algorithms used to predict the occurrence of several types of the most common cancers such as breast, lung and brain cancers are explored in detail in this section.

4.1 Breast Cancer

Oftentimes, the gene expression arrays, also known as microarray profiling, were utilized predominantly in the diagnosis of cancer disease as genetic compositions have greater influence in the cause of cancer. But, by incorporating patient clinical records such as ultrasound images and laboratory results into the microarray data, this will enhance the decision-making process during cancer diagnosis. In the study by Gevaert et al. (2006), Bayesian networks are employed where it treats the clinical data and microarray analysis on the same level of importance in the prediction of breast cancer where the output is evaluated as either a poor or good prognosis. Patient dataset consists of the training set and testing set where each set has poor and good forecasts of breast cancer. Physical and biological parameters of patients were obtained from the clinical notes and were combined with microarray analysis dataset. The Bayesian network software used the pre-processed data as the input source. Three methods of combining clinical and microarray data were administered which were full, decision and partial integration and their performances were validated using Area Under the ROC (Receiver Operator Characteristics) Curve (AUC). Decision and partial integration showed a significantly varied ROC AUC compared to the other methods, thus they were utilized to build the models for the training dataset. Best Partial Integration Model (BPIM) was found to have better performance than Best Decision Integration Model (BDIM) as the former requires lesser number of genes when incorporating clinical data for forecasting the disease prognosis. The results also portrayed that the clinical and microarray variables in the Markov blanket enhances the model performance. Thus, BPIM under the Bayesian networks approach provides a mean for a cost-effective diagnosis of breast cancer while retaining the molecular level data.

Another research by Thomas et al. (2014) on breast cancer was done by incorporating the clinical and microarray data using a suggested data mining technique called weighted Least Square SVM (LS-SVM) classifier to result in an improved prognostic application in breast cancer therapy. The five datasets were obtained from the Integrated Tumour Transcriptome Array and Clinical data Analysis (ITTACA) repository, where 2/3rd of the population was used as training set and the rest for testing. Transformation of each dataset into a kernel matrix was done and an integration framework was developed. The paper elucidated on the Generalized Eigenvalue Decomposition (GEVD) and LS-SVM formulations, where they recommended a new machine learning technique, called weighted LS-SVM classifier, for data integration and classifications. The

performance rate of each dataset was compared for GEVD, kernel GEVD and weighted LS-SVM classifier using test AUC and Leave-One-Out Cross Validation (LOO-CV). The proposed classifier approach introduced an optimized single framework to resolve the issues of excessive cost and classification using heterogenous datasets and improved the prediction and treatment efficiencies for each patient.

Breast cancer has evolved as one of the most frequent cancer types globally among female and has contributed to major death rates. To subdue this concern, data mining tools can be implemented into the clinical system of cancer diagnosis at an early stage so that ideal treatments can be administered. Alickovic & Subasi (2015) explained that the traditional method of clinical diagnosis of breast cancer such as mammography can be supported with automatic diagnostic tools to assist in distinction between benign and malignant breast tumours. Two distinct datasets were acquired from UCI Machine Learning repository consisting of benign and malignant tumours data. The machine learning approaches used in this study are RF, MLP, SVM, C4.5 Decision Tree, Logistic Regression, Bayesian Network, Radial Basis Function Networks (RBFN) and Rotation Forest. Rotation Forest creates classifier groups by segmenting features into subsets and applying Principal Component Analysis (PCA) on each subset. Distinct rotations are formed from different feature set splits, resulting in variant classifiers with diversity and precision. Genetic Algorithms (GA) are used to select the key features from the breast cancer datasets as inputs to classifiers to enhance the classification accuracy of the multitude data mining techniques. Weka tool was used to execute these algorithms. AUC ROC was used to represent the classifiers' performances. The results portrayed that Rotation Forest, which is a Multiple Classifier System (MCS), with GA-based feature selection produced the greatest classification accuracy of 99.48% in the breast cancer dataset compared to other classification algorithms. This study recommended the employment of this novel technique by clinicians to correctly assess breast cancer diagnosis and enhance decision-making.

4.2 Lung Cancer

Lung cancer is another cancer type that is listed as the most occurring chronic disease worldwide. It is vital to infer an early diagnosis, identify the right type of lung cancer and provide accurate treatment to reduce the fatality rate among the carriers of this disease. An optimal solution to this is by employing various machine learning techniques as done by Podolsky et al. (2016) in their study to confirm an early diagnosis. This research used four open datasets that contain gene expression levels for various lung cancer types. The data mining methods used to analyse the data were k-NN with increasing degrees of freedom, Naïve Bayes classifier, SVM and C4.5 decision tree. Two of the datasets used 10-fold cross validation to be used as inputs into the algorithms while the other two datasets had pre-prepared training and testing sets. ROC AUC and Matthews Correlation Coefficient (MCC) were calculated to validate the performance of the algorithms. The

authors discussed that the SVM algorithm showed high performance values for two of the gene datasets which can be used to categorize lung cancer based on their histological variants at a great accuracy. The third analysed dataset was precisely distinguished between healthy lung and adenocarcinoma via all the data mining techniques except C4.5 decision tree. But, this technique was found to be compatible on the fourth dataset. This study concluded that SVM is the most ideal machine learning algorithm for an effective diagnosis of lung cancer and can predict tumour development and its metastasis.

4.3 Brain Cancer

Cancer of the brain is a very critical and life-threatening chronic disease as brain is part of the central nervous system of the body and any damages to brain activities will also affect the other parts of the body. Brain cancer exists as primary tumours (originated from brain cells) and secondary or malignant tumours (cancer cells that originate from other parts of the body and metastasize in the brain). A study was done by George, Jehlol & Oleiwi (2015) to classify Magnetic Resonance Images (MRI) to identify the brain cancer types and its subtypes using supervised machine learning approaches. C4.5 decision tree and MLP algorithms were used to perform automated classification of a type of benign cancer tumour and five variants of malignant cancer. Figure 4 illustrates the framework of the classification method used in this study. The MRI data were acquired from hospitals and from internet due to the lack of brain cancer data in open databases. Image pre-processing was done to overcome the problem of high dimensionality of data through sigma filtering to eradicate noise from MRI. The images were segmented using adaptive threshold method which partition the data based on grey or coloured images to generate binary image representations. Region detection was utilized to identify and separate the tumour region from the other objects in the MRI. The features were then extracted based on six shape properties related to brain images using MATLAB program. The data was split into 55% for training set and the remaining was used for test validation. The C4.5 algorithm displayed 91% accuracy while MLP yielded 95% accuracy for the overall brain tumour types. MLP was found to take more time to build the model compared to C4.5 decision tree algorithm. The study proposed to use a larger dataset with more features to improve the accuracy of diagnosis and performance of the classifiers.

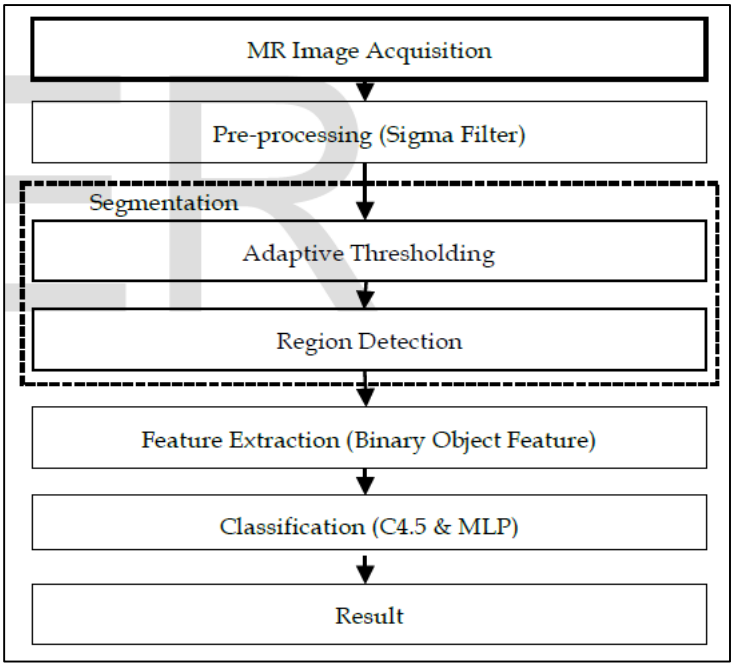

Figure 4: Brain cancer classification architecture (George, Jehlol & Oleiwi, 2015)

5. Discussion

The papers that have been discussed in the previous sections addressed few limitations while employing the data mining techniques. One of the issues encountered is that the size of the dataset is small, hence it was difficult to validate the performance and accuracy of prediction and diagnosis. The algorithm that gave the best result may work on a smaller dataset but the same technique might show different performance in a larger dataset for the same disease conditions. Thus, a larger sample size can yield better prediction accuracies with less errors (George, Jehlol & Oleiwi, 2015; Khare & Gupta, 2016; Hu et al., 2016).

Secondly, heterogenous datasets with different attributes were compatible with different algorithms. Even for the same disease type, the performance of the algorithms was found to be varied for distinct datasets. This could be due to the criteria of the algorithms such as feature selection and filtering, size of the dataset, percentage of training and testing datasets, types of statistical measures used to evaluate the performance and high dimensionality such as medical visual images in the dataset. Hence, it is crucial to compare across the multitude machine learning techniques to attain an optimum result and reduce cost of diagnosis and treatment (Renuka Devi & Maria Shyla, 2016; Podolsky et al., 2016). Another limitation in a study highlighted that the population ethnicity and gender used in the dataset may have showed a biased result of the ideal data mining technique. The accuracy of the technique may be different if applied on different race or gender (Chase et al., 2017).

But majority of the articles have emphasized the primary benefits of using data mining techniques in assessing chronic diseases. The main advantage was that there was substantial reduction in cost incurred for diagnosis and treatment of the diseases by using machine learning methods. Time acquired for training and validating the algorithms were much lesser compared to clinical diagnosis. Further, the prediction models were generated in shorter time so the necessary treatments can be provided without any delay. Data mining also optimized the number of diagnosis required to medically assess the chronic diseases, thus the treatment costs would also be reduced significantly. Indirectly, this creates an affordable and high quality medical care for all and reduces the fatality rate globally (Gevaert et al., 2006; Khare & Gupta, 2016; Subasi, Alickovic & Kevric, 2017).

To overcome the limitations of the techniques, some articles have highlighted that two or more algorithms can be merged to produce new algorithms, which yielded better performance and precision results. More such approaches can be further explored for different chronic diseases, especially for cancer prediction and prognosis (Thomas et al., 2014; Alickovic & Subasi, 2015). Open-software tools that are free of charge such as Weka, MATLAB and RapidMiner have multiple in-built algorithms which were used in most of the papers to generate prediction models and to identify relationships and trends in chronic diseases. These tools can also be used by non-technical people who do not have IT background but have domain knowledge such as doctors or clinicians to make informed, evidence-based healthcare decisions (Kumar & Singh, 2016; Renuka Devi & Maria Shyla, 2016). Table 1 highlights the comparison of distinct data mining techniques in various chronic diseases as well as application of different algorithms for the same disease type.

Table 1: Comparison of data mining techniques and different algorithms in the literature

Chronic Disease	Author & Year	Contribution	Data Mining Technique
Type 1 & 2 diabetes mellitus & gestational diabetes	Renuka Devi & Maria Shyla (2016)	Applied data mining methods to classify diabetes-related attributes and predict occurrence of diabetes	Modified J48 classifier
Type 2 diabetes mellitus	Hu et al. (2016)	Enabled prognosis of IGT patients who encountered rapid progression of atherosclerosis	Naïve Bayes
Heart disease	Khare & Gupta (2016)	Identified primary heart disease risk	Association rule

		factors using machine learning techniques	
Alzheimer's disease	Kumar & Singh (2016)	Determined Alzheimer's correlated genes using decision tree for accurate prediction and prior diagnosis of disease	C4.5 algorithm
Multiple sclerosis (MS)	Chase et al. (2017)	Built a prediction model for pre-recognition of MS based on natural language processing of unstructured medical notes	Naïve Bayes
Chronic kidney disease (CKD)	Subasi, Alickovic & Kevric (2017)	Established a better classification and prediction technique for early diagnosis of CKD compared to previous works	Random Forest classifier
Breast cancer	Gevaert et al. (2006)	Identified a cost-effective diagnosis technique using minimum number of genes and clinical data	Bayesian networks
	Thomas et al. (2014)	Introduced an optimized single framework to improve disease prognosis in breast cancer therapy	Least Square Support Vector Machines (LS-SVM) classifier
	Alickovic & Subasi (2015)	Distinguished between benign and malignant cancer using classifier system with GA-based feature selection	Rotation Forest
Lung cancer	Podolsky et al. (2016)	Identified a technique for an effective diagnosis of lung cancer and can predict tumour	Support Vector Machines (SVM)

		development and its metastasis	
Brain cancer	George, Jehlol & Oleiwi (2015)	Built an automated system to classify medical that enabled identification of brain cancer types and its subtypes	Multilayer Perceptron (MLP) and C4.5

6. Conclusions

Big data analytics in the healthcare industry have given an insight and awareness towards the importance of data mining applications in diagnosis, prediction and prognosis of various illnesses and disorders. Past few years have seen an increase in the research for discovering ideal machine learning algorithms and tools for disease modelling. Chronic diseases such as diabetes, cardiovascular diseases, brain disease, CKD and various cancer types like breast, brain and lung cancers have caused many public health concerns globally and with the influence of machine learning approaches in the diagnosis of these diseases, the studies have proven to yield a reduction in the mortality rate. Yet, the full functionality of machine learning utilization remains a challenge due to the concerns in data security, privacy, integrity and exchange. Government legislations and encryption techniques are being employed actively on data analytics nowadays to mitigate these social and ethical issues.

Acknowledgements

I would like to express my deepest gratitude and appreciation to my lecturers, Prof. Dr. Logeswaran and Mr Manoj Jayabalan for their guidance and suggestions during this review writing. I would also like to acknowledge with much appreciation to all those who have helped me in completing this report successfully.

References

Ahmad, P., Qamar, S. & Rizvi, S. (2015). Techniques of data mining in healthcare: a review. International Journal of Computer Applications. 120(15). p. 38-50.

Alickovic, E. & Subasi, A. (2015). Breast cancer diagnosis using GA feature selection and rotation forest. Neural Computing and Applications. 28(4). p. 753-763.

Auffray, C., Balling, R., Barroso, I., Bencze, L., Benson, M., Bergeron, J., Bernal-Delgado, E., Blomberg, N., Bock, C., Conesa, A., Del Signore, S., Delogne, C., Devilee, P., Di Meglio, A., Eijkemans, M., Flicek, P., Graf, N., Grimm, V., Guchelaar, H., Guo, Y., Gut, I., Hanbury, A., Hanif, S., Hilgers, R., Honrado, Á., Hose, D., Houwing-Duistermaat, J., Hubbard, T.,

Janacek, S., Karanikas, H., Kievits, T., Kohler, M., Kremer, A., Lanfear, J., Lengauer, T., Maes, E., Meert, T., Müller, W., Nickel, D., Oledzki, P., Pedersen, B., Petkovic, M., Pliakos, K., Rattray, M., i Màs, J., Schneider, R., Sengstag, T., Serra-Picamal, X., Spek, W., Vaas, L., van Batenburg, O., Vandelaer, M., Varnai, P., Villoslada, P., Vizcaíno, J., Wubbe, J. & Zanetti, G. (2016). Making sense of big data in health research: towards an EU action plan. Genome Medicine. 8(1).

Chase, H., Mitrani, L., Lu, G. and Fulgieri, D. (2017). Early Recognition of Multiple Sclerosis Using Natural Language Processing of the Electronic Health Record. BMC Medical Informatics and Decision Making, 17(1).

Dey, M. & Rautaray, S. (2014). Study and analysis of data mining algorithms for healthcare decision support system. International Journal of Computer Science and Information Technologies. 5(1). p. 470-477.

Dinov, I. (2016). Methodological challenges and analytic opportunities for modeling and interpreting big healthcare data. GigaScience. 5(1).

Durairaj, M. & Ranjani, V. (2013). Data mining applications in healthcare sector: a study. International Journal of Scientific & Technology Research. 2(10). p. 29-35.

George, D., Jehlol, H. & Oleiwi, A. (2015). Brain tumor detection using shape features and machine learning algorithms. International Journal of Scientific and Engineering Research. 6(12). p. 454-459.

Gevaert, O., Smet, F., Timmerman, D., Moreau, Y. & Moor, B. (2006). Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. Bioinformatics. 22(14). p. 184-190.

Herland, M., Khoshgoftaar, T. & Wald, R. (2014). A review of data mining using big data in health informatics. Journal of Big Data. 1(1). p. 2.

Hersh, W. (2017). Healthcare data analytics. Health Informatics: Practical Guide for Healthcare and Information Technology Professionals, 6th Ed. Florida: Informatics Education. p. 62-75.

Hu, X., Reaven, P., Saremi, A., Liu, N., Abbasi, M., Liu, H. & Migrino, R. (2016). Machine learning to predict rapid progression of carotid atherosclerosis in patients with impaired glucose tolerance. EURASIP Journal on Bioinformatics and Systems Biology. 2016(1).

Khare, S. & Gupta, D. (2016). Association Rule Analysis in Cardiovascular Disease. In Proceedings of the 2016 Second International Conference on Cognitive Computing and Information Processing (CCIP). Mysore, India: IEEE. p. 1-6.

Kumar, A. & Singh, T. (2016). A new decision tree to solve the puzzle of Alzheimer's disease pathogenesis through standard diagnosis scoring system. Interdisciplinary Sciences: Computational Life Sciences. 9(1). p. 107-115.

Podolsky, M., Barchuk, A., Kuznetcov, V., Gusarova, N., Gaidukov, V. & Tarakanov, S. (2016). Evaluation of machine learning algorithm utilization for lung cancer

classification based on gene expression levels. Asian Pacific Journal of Cancer Prevention. 17(2). p. 835-838.

Raghupathi, W. & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. Health Information Science and Systems. 2(1).

Renuka Devi, M. & Maria Shyla, J. (2016). Analysis of various data mining techniques to predict diabetes mellitus. International Journal of Applied Engineering Research. 11(1). p. 727-730.

Subasi, A., Alickovic, E. & Kevric, J. (2017). Diagnosis of Chronic Kidney Disease by Using Random Forest. In Proceedings of the International Conference on Medical and Biological Engineering 2017 (IFMBE Proceedings, Vol 62). Singapore: Springer. p. 589-594.

Taranu, I. (2015). Data mining in healthcare: decision making and precision. Database Systems Journal. 4(4). p. 33-40.

Thomas, M., Brabanter, K., Suykens, J. & Moor, B. (2014). Predicting breast cancer using an expression values weighted clinical classifier. BMC Bioinformatics. 15(1).

Tomar, D. & Agarwal, S. (2013). A survey on data mining approaches for healthcare. International Journal of Bio-Science and Bio-Technology. 5(5). p. 241-266.

Wu, P., Cheng, C., Kaddi, C., Venugopalan, J., Hoffman, R. & Wang, M. (2017). -Omic and electronic health record big data analytics for precision medicine. IEEE Transactions on Biomedical Engineering. 64(2). p. 263-273.