Smart Health Prediction System with Data Mining

Wilson Wibamanto, Debashish Das, Sivananthan A/L Chelliah

School of Computing & Technology, Asia Pacific University of Technology & Innovation, Kuala Lumpur, Malaysia.

ABSTRACT

Background: The digital technology era demands the world to provide an excellent health system, to ensure the citizen and community to be alive and healthy.

Purpose: This study proposes the application of data mining algorithm for health prediction that can eventually shape a suitable health prediction system for patients. Although health care is available to everyone in the world, there is still no healthcare system that is completely reliable and accurate to carefully diagnose a patient on their current health issues. Even though some hospitals are well equipped to provide the best healthcare services to its citizens, some of the hospitals are still lacking in certain qualities. Consequently, patients are doubtful and uncertain when it comes to picking which hospital suits them.

Problem: Numerous issues are faced by patients pertinent to hospitals such as being unable to provide medical services, insufficient number of qualified medical staffs, poor communication between doctors and patients, and unorganized health records and data. Eventually, these issues impede the opportunity for hospitals to handle both their management and their duties steadily to maintain the health of every citizen and community.

Conclusion: Patients need treatment and diagnosis that are accurate and precise for them to be able to recover back for their proper health and medical staffs are required to be well-equipped in their clinical knowledge and communication skills to carefully assess their patients to ensure good health. Therefore, application of data mining in health prediction is considered in this paper as the best practice to facilitate better healthcare system.

Key Words: Clinical prediction, Data mining, Smart health system, Medical service, Health prediction, Electronic health records

INTRODUCTION

Data mining can be described as a process of searching patterns or correlations from a large data sets to valuable information that can solve problems and predict outcomes. It involves analyzing certain amount of information to locate certain patterns of occurrence to predict future tendencies, using several processes of effective data collection, warehousing and computer processing. With this functionality, therefore, it serves a great purpose when it comes to predicting people’s health diseases especially on finding the correlation between the health information that has been given by both the medical staff and the patient. These finding may provide a beneficial advantage in the healthcare industry as it may be used to manage patients on their current health issues and for the doctor to alleviate them from their jobs.1

There are currently a lot of health institutions that have been developed such as hospitals and medical centres which are crucial to maintain and improve the health of the community around us. It is a prime establishment of giving proper health care especially for every one of us who have ever lived. For every illness and diseases that people may face today and sometime in the future, it is because of these medical institutions and all the doctors who worked at these places that have made our lives physically better and healthy. Although hospitals now are well-equipped with their staffs working, there are still known issues that persist that cause the staffs to make the poor clinical decision that affects a patient’s health such as the lack of qualified doctors, unorganized health information and poor communications between doctors and patients.1,3
With all of these issues mentioned, solutions are needed to be made and that is where a health prediction system should be implemented which could potentially eliminate these concerning issues. Hence, the current research proposes to apply data mining for smart health prediction.

**LITERATURE REVIEW**

Health care institutions are essential as it provides to every single person in the world proper health care. Its main purpose is to improve the current health of the community that we have shared and created. A health care institution such as hospitals or medical centres would essentially consist of numerous of doctors that were qualified and specialize on treating patients of their current illness that they endured and to restore them to proper health. Throughout this day and age, new technologies have been created and developed to improve people’s daily life and routine, especially for health care. Doctors and nurses were now guided by smart health prediction system on the purpose of storing medical information that may be used for research and diagnosis. A few years ago, doctors were expected to use their intuition and experience to handle every medical situation that different patients are facing every day. Although their current approach may have saved people’s lives back then, they are still prone to errors and wrongdoings that have endangered human life. It is without a doubt a heavy burden for everyone especially the medical staffs to understand that several decisions could heavily affect other people’s lives and health, it is also why such system itself proves to be vital on guiding medical staff to make a proper clinical decision to cure and restore the human health.

A smart health prediction system is defined as a healthcare system that is intended to assist health professionals in their decision-making process regarding medical situations. This system will provide the guidance and information needed for doctors to diagnose patient on their medical illness and it will eliminate the difficulties that the doctors needed to encounter, particularly in their clinical decision-making process. The system would require to gather a whole lot of medical information that are valuable to be used on predicting a patient’s health status, these patterns of information will be analyzed by using data mining techniques in order to find correlations and discover new pieces of information from unstructured data. By using data mining tools, it will not only be able to produce reliable results with less time consumption and complexity but also with smart decision-making and useful information.

**MATERIALS AND METHODS**

This section explains the data mining processes and algorithms with its application in health prediction. It also analyses the prospects related to the application of data mining techniques in health prediction.

**Data Mining**

The health prediction system will rely on its implementation of data mining, which is referred to as mining knowledge and information from a large number of data sets. The medical industry is just one of many fields in society that collects a vast amount of information that can be utilized helpfully by data mining. Data mining can improve the medical industry by eliminating current health disparities by easily providing answers to complex medical cases to solve and eliminate any time consumptions created from making a clinical decision. Data mining is described as the process of searching certain patterns that comes in a database and to utilize that information to build predictive models, Figure 1. Its processes involve examining and selecting certain data from a vast data storage to uncover new and unknown patterns. Both M. Durairaj et al. mentioned that the results of data mining are of influence of statistics, databases, information retrieval, machine learning and algorithms.

Data mining is also a known process involve in Knowledge Discovery in the database which is the extraction of a large amount of data from databases. It used to operate and examine hidden patterns and relationship that can be found from a large amount of data for decision-making purposes. Knowledge Discovery in Database is done in 7 sequential steps as shown in Figure 2.

**Data Cleaning**

Data Cleaning is represented as the first step of KDD which requires to eliminate any data collected that is random, irrelevant or missing in values.

**Data Integration**

Data Integration is the second step after Data Cleaning which takes the data the has been filtered from the previous step to be combined into a meaningful and useful data.

**Data Selection**

Data Selection is defined as the process of which data that is relevant for the analysis is selected and retrieved from the collection of data.

**Data Transformation**

Data Transformation involves data being converted into forms that are required for performing different mining operations such as smoothing, normalization or aggregation.

**Data Mining**

Data mining consist of examining the data for any patterns or rules that are useful to be extracted and obtained.
Pattern Evaluation
Pattern Evaluation is defined as identifying patterns that represent knowledge based on a given measure.

Knowledge representation
Knowledge representation is the last step of KDD where it uses visualization tools and techniques to help users understand the knowledge from the result of data mining. There are various data mining techniques to choose from to turn raw data into something useful. These are 7 techniques that are most commonly used involved in data mining to choose.Δ

Classification
Classification represents a data mining technique that requires to collect various information and data for their attributes to be analyzed. Once the attributes have been identified, the data can be further categorized and managed, Figure 3.

Clustering
Clustering is a data mining technique that requires identifying data that relates to another according to its differences and similarities. It relies on a visual approach that shows the distribution of data in relation for people to understand.

Regression
Regression techniques involve identifying and analyzing the relationship between variables in a dataset. It is a technique that is used in aspects of data modelling. The relationship between variables may vary depending on its instances.

Outlier detection
Outlier detection or anomaly detection consists of observing data items in a data set for any anomalies which do not match a certain behaviour. For any anomalies that have been identified, it will become easier to understand the causes of these anomalies to prevent them.

Sequential Pattern
The sequential pattern is a technique that focuses on discovering similar patterns in a data transaction during certain periods. This technique is useful to uncover deviation in the data that is happening at regular intervals over time.

Prediction
Prediction simply involves analyzing events that are in the past to predict future events. So historical data that has been kept is used for examination to gain some insight that might be useful to predict what will happen in the future.

Association Rules
Association rules is a data mining technique that relates to statistics. It searches and indicates certain data for the association that may be linked together between two data set for discovering a hidden pattern.

For a healthcare prediction system, there are plenty of data mining algorithms to consider. Every data mining algorithm will eventually produce different result between one another and with these results, it is used to determine the effectiveness and accuracy of the system. The study proposed by V.Kirubha and S.Manju Priya describes an analysis of the application of data mining in the different medical domain and the algorithms that are used to predict different diseases. A table in Figure 3 shows the comparison of algorithms that were used for different disease prediction. With the result shown in Figure 3, it is sufficient to determine that data mining provides results that are good and useful in providing diagnosis for diseases when the correct data mining tools and techniques were applied.

Data Mining based Health Prediction System
For the past few years, the health industry has been growing significantly that leads to unsurmountable piles of data to be calculated. Various researchers have had different ideas and approaches that were distinguished by their choice of data mining modeling techniques and their focus on a particular disease. A decision support system was proposed to predict patients who are affected with swine flu using Naïve Bayes as the data mining modeling technique.Δ A testing dataset was used consisting of data with 100 patients with swine flu from various hospitals that are to be tested with this system. Figure 4 shows the following attributes that they used for the system and a list of swine flu symptoms to predict patients who have swine flu.

The system uses Java platform which access data from the database and SQL query language to build and access the models. It is proven that Naïve Bayes could identify all the significant medical predictors, though their research stated that it can be further improved with more data sets and attributes were provided for testing. Sellapan Palaniappan and Rafiah Awang attempted to develop a prediction system regarding to heart disease with motivation of facing the issue of healthcare organizations of providing quality services at an affordable cost. The system that they had use three data mining modeling techniques of which are Decision Trees, Naïve Bayes and Neural Network. The system they have proposed would provide answers on complex queries for diagnosing patient suffering from heart disease by assisting the practitioners to make intelligent clinical decision and providing effective and affordable treatment. They obtained a total of 454 testing dataset records from the Cleveland Heart Disease database to be tested with 3 different data mining modeling techniques, with 208 patients with heart disease and 246 without it, Figure 5.

As shown in the figure above, they used classification matrix to display the amount of correct and incorrect predictions,
comparing both the predicted and actual values in the data-set. The rows represent predicted values while the columns represent actual values (1 for patients with heart disease, ‘0’ for patients with no heart disease). The left-most columns show values predicted by the models. The diagonal values show correct predictions. They have summarized that Naïve Bayes appears to be the most effective and accurate modeling technique followed by Neural Network and Decision Trees. Another attempt was done to develop a system to predict lung cancer using Knowledge Discovery in Database (KDD) to extract implicit information from data in a database. The Knowledge Discovery proves that data mining is included as an important step in its process, Figure 6.

Four data mining models were used in this system which are IF-THEN Rule, Decision Tree, Bayesian classifiers and Neural Network. The system comes with a prediction approach based on statistical factors such as age, gender, symptoms and risk factors, tested in a historical lung cancer disease database. As more data mining models were applied, the researchers were able to compare each model with another, claiming that the Naïve Bayes to be the most effective in their system, followed by IF-THEN rule, Decision Trees and Neural Network.

From all these articles and research, the health prediction system that was proposed were only able to accommodate to particular medical diseases. The hospital and their staffs would require a system that could assist all patients for every illness that is common and universal to eliminate their troubles in decision-making and data collection. According to most of the article, such a system could be expanded upon by adding data mining techniques such as clustering or by integrating text mining into the system. In conclusion, a health prediction system has been proven to be resourceful and beneficial for all doctors and medical experts by eliminating time consumption and the troubles in their decision-making process while diagnosing a patient. Various data mining models have to be considered and compared in order to create an intelligent health prediction system. Based on the finding from the literature review, most health prediction system contains around more than one data mining algorithms to predict the diseases with all of it indicating Naïve Bayes algorithm to be the data mining algorithm that produces the best and most accurate result out of all the other data mining algorithm. Generally, the data mining algorithm will be considered and selected depending on the size of the dataset to be tested on its prediction accuracy.

**RESULTS AND DISCUSSION**

The result of the system created will consist of the diseases and its respective accuracy level the patient is currently suffering from. With data being analyzed following the steps of the process of Knowledge Discovery based on Figure 2.2, the accuracy level of a specific disease will be based on the various factors such as the patient’s medical history, age, gender and much more. The result from the following data mining will be used to assist the clinical doctors for them to be able to cure the patient depending on the diseases that have higher accuracy.

The implementation of a health prediction system will allow doctors and medical staffs to alleviate their efforts on their clinical decision-making process by simply inputting the user’s health data and symptoms that they are experiencing. The system will be implemented with data-mining algorithms that may cleverly deduce the disease that they bear by correlating the information given by the patient with the health information the doctors and medical professional provide and that is stored in a database, the entire process would efficiently reduce the time consumption and challenging efforts that doctors put themselves into for making a clinical decision. The system will also encourage patients and doctors to communicate by recommending patient to the doctors that are suitable to handle their diagnosis and are relevant in their specific medical fields. This project is intended to deliver a user-friendly system for patients and doctors to use of their demands for diagnosing illness and to provide suitable guidance on their current health issue that they are facing. The software itself will only be suited and installed on PCs. The proposed system needs to include the functionality:

**Patient Registration**

Patients would require registering themselves for the first time with their username and password to use the system.

**Patient Login:** Patients would require to login into their system with their username and password.

**Viewing Patient Details:** Doctors and patients may view details of one another to familiarize themselves.

**Disease Prediction:** Determine the illness/diseases that the user is trying to describe by going through several question and using data mining to pinpoint the most accurate symptoms.

**Search Doctors/Patients:** Doctors and patients may search for one another according to their specialty, diseases they had contracted and other references.

**Providing Feedbacks:** Doctors and patients may provide feedback that may serve as additional information to be viewed.

**Adding Diseases and Symptoms:** Administrators may add new diseases and symptoms into the system for doctors and patients to examine.

**Doctor Login:** Doctors are required to login with their username and password to use the system.
**Doctor Registration:** Administrators could add and register a new doctor into the system and assign their username and password.

**Admin Login:** Doctors are required to login with their username and password to use the system.

**View Diseases:** Administrators can view various disease details stored in the database of the system.

**Sharing Information:** Doctors could share information about a disease or patient to another doctor for verification.

**CONCLUSION**

The data mining can play a vital role in disease prediction to design a smart health prediction system. In medical diagnosis, data mining has been widely used for predicting diseases through diagnosis. However, no single data mining algorithm is best suited to resolve the prediction issues for healthcare data sets. In conclusion, the combination of several data mining or hybrid version of the data mining algorithm may be a better approach in designing health prediction system. The future research may be directed towards designing a better data mining based model that can address healthcare with real-time healthcare datasets. This study does not encompass the complete analysis of all existing data mining algorithms and real-time healthcare dataset. Besides, the proposed health prediction system is not built through the comparison of all the data mining algorithms available in the literature. However, future research may be directed towards the selection of the best suitable data mining algorithm through the analysis of all existing algorithms.

**ACKNOWLEDGEMENTS**

This paper is made possible through the help and support from everyone. Especially, we would like to dedicate our acknowledgement of gratitude towards the significant contributors. We would like to thank APU for most support and encouragement. Finally, we sincerely thank all the people, who provide advice and support.

**Conflict of Interest:** The authors involved in the current study does not declare any competing conflict of interest.

**Funding and Sponsorship:** No fund or sponsorship in any form was obtained from any organization for carrying out this research work.

**REFERENCES**

1. Scheffler RM, Arnold DR. Projecting shortages and surpluses of doctors and nurses in the OECD: what looms ahead. Health Econ Policy Law 2019 Apr;14(2):274-290.2.
2. Chowdhury SBS. A Review Paper on Health Data Integration with Secured Record Linkage in National Health Data Warehouse and Privacy & Compatible Solution for its Privacy and Security Problems in Bangladesh and other developing countries. 2017. International Islamic University Chittagong (IIUC). Available at: http://works.bepress.com/sayem-binsarwarchowdhury/2/
3. Tulsky JA, Beach MC, Butow PN, Hickman SE, Mack JW, Morrison RS, et al. A research agenda for communication between health care professionals and patients living with serious illness. JAMA Intern Med 2017 Sep;177(9):1361-1366.4.
4. Kuziemsky, C. Decision-making in healthcare as a complex adaptive system. Healthcare management forum. Healthc Manage Forum 2016 Jan;29(1):4-7.
5. 5. Lakshmi K, Ahmed DI, Kumar GS. A smart clinical decision support system to predict diabetes disease using classification techniques. ISRSET 2018; 4(1):1520-1522.
6. 6. Durairaj M, Ranjani V. Data mining applications in the healthcare sector: a study. Int J Sci Tech Res 2013; 2(10):29-35.
7. 7. Chaudhary BV, Radhika Y. A Survey on Applications of Data Mining Techniques. Int J Appl Engg Res 2018; 13(7): 5384-5392.
8. 8. Kirubha V, Priya SM. Survey on data mining algorithms in disease prediction. Int J Comp Trends Tech 2016; 38(3): 124-128.
9. 9. Thakkar, B.A., Hasan, M.I. and Desai, M.A. Health care decision support system for swine flu prediction using naive Bayes classifier. Int Conf Adv Recent Tech Comm Comp 2010. DOI:10.1109/ARTCOM.2010.98
10. 10. Palaniappan S, Awang R. Intelligent heart disease prediction system using data mining techniques. Conference: Computer Systems and Applications, 2008. AICCSA 2008. DOI: 10.1109/ AICCSA.2008.4493524.
11. 11. Krishnaiah, V., Narsimha, D.G. and D.N.S., Chandra. Diagnosis of lung cancer prediction system using data mining classification techniques. Int J Comp Sci Info Tech 201; 4(1): 39-45.
Figure 3: Comparison of data mining algorithms with different diseases.  

Figure 4: The flow of execution of the decision support system.

Figure 5: Result of classification matrix for all the three models.

Figure 6: Comparison using ODANB and NB.