Clustering and Classification Techniques Based on Machine Learning Algorithms (Hybrid Convolutional Neural Network) for Diagnosis of Diabetics using Genomic Database

Table of Contents

1.0 Introduction

Nowadays, advances in biological and medical technologies have been providing us explosive volumes of biological and physiological data, like electroencephalography, genomic, medical images and protein sequences (Cao et al., 2018). Learning from these data facilitates the understanding of human health and disease. Particularly, Machine Learning (ML) helps to perform predictive analysis or pattern recognition on large data. Also, it offers a range of alerting and risk management decision support tools, targeted at improving patients' safety and healthcare quality. With the need to reduce healthcare costs and the movement towards personalized healthcare, the healthcare industry faces challenges in the core areas namely, electronic record management, data integration, and computer-aided diagnoses and disease predictions. Machine learning offers a wide range of tools, techniques, and frameworks to address these challenges (Nithya, 2016). For this purpose, artificial neural networks, deep learning-based algorithms show great promise in extracting features and learning patterns from complex data (Cao et al., 2018).

2.0 Background of the study

In a database system, mining of data is the important stage of information discovery which is the extraction of unique, implicit and potentially useful information from data (Han & Kamber, 2001; Huang et al., 2014). The difference amongst discovering knowledge and mining of data is that the utilization of different intelligent algorithms to excerpt patterns from the data whereas information discovery is the complete process that is involved in discovering knowledge from data. The primary objective is to abstract high-level information from low-level data (Dev et al., 2016). By applying machine learning and data mining methods in research is a key approach to utilizing large volumes of available diabetes-related data for extracting knowledge. The severe social impact of the specific disease renders data mining is one of the main priorities in medical science research, which unavoidably generates huge amounts of data. Undoubtedly combination of machine learning and data mining approaches is a great concern when it comes to diagnosis, management and other related clinical administration aspects.

In addition, the prevalence of diabetes mellitus is increased very fast in past few years, and it's now become a global health problem. International Diabetes Federation (IDF) estimated that worldwide 381.8 million people are affected by diabetes, and about 591.9

million people will be affected by this disease with a huge hike of 55% by 2030 (Beagley et al., 2014). Diabetes mellitus is simply excess of blood sugar level. Pancreas organ, a fish-shaped gland with its function of insulin secretion often associated with diabetes mellitus. Pancreas organ performs both the functions, exocrine and endocrine of glans. It produces and releases the digesting juices to the intestine as well as, controls the blood sugar level by producing glucagon and insulin (Friis-jensen, 2007; Samant & Agarwal, 2018). Hence, the framework of this study, efforts were made to review the current literature on machine learning and data mining approaches in Diabetes research (Cao et al., 2018; Kavakiotis et al., 2017).

3.0 Problem statement

Over the past decades have witnessed a massive growth in biomedical data, like protein structures, genomic sequences and medical images, due to the advances of high-throughput technologies. This deluge of biomedical big data necessitates effective and efficient computational tools to store, analyze, and interpret such data (Cios et al., 2005; Asgari & Mofrad, 2015). Consequently, the abundance of data has strengthened considerably data-oriented research in biology. In such a field, one of the most important research applications is prognosis and diagnosis related to human-threatening and/or life quality reducing diseases. One such disease is Diabetes Mellitus (DM) (Kavakiotis et al., 2017).

By applying machine learning and data mining techniques in DM research is a key approach to utilizing large volumes of available diabetes-related data for extracting knowledge (Cao et al., 2018; Kavakiotis et al., 2017). From the review of literature in healthcare applications that are used for diseases diagnosis or prediction does not support real-time use which enable the stockholders to access them anytime and anywhere (Abdelaziz et al., 2017; Sharma, 2016; Kumar et al., 2014; Prerana et al., 2015; Tintu & Paulin, 2013; Hamad, 2016; Arjun & Anto, 2015). However, the time delay represents a big challenge for the most of stakeholders in healthcare applications that run the medical requests on a cloud computing environment (Abdelaziz et al., 2018). For this purpose, researcher suggests, deep learning-based algorithmic frameworks shed light on these challenging problems (Cao et al., 2018). The goal of this research is to provide the novel framework based on Clustering and Classification (hybrid machine learning) for Diagnosis of Diabetics using Genomic Database in the healthcare field.

4.0 Related work

The previous studies related to machine learning technique are discussed as follows:

A study by Tomar and Agarwal (2013), discovered the efficacy of an assortment of data mining techniques like association, classification, clustering and regression in health sphere. Authors briefly explained different data mining techniques with their intrinsic worth and demerits. Authors suggested that before applying the classification techniques, one should preprocess the data, i.e. data should be normalized by removing all types of redundancies as it may degrade the execution time. The authors recommended cross-validation method in the classification process. Clustering is beneficial if there is no or missing information. Finally, authors suggested using a fusion of classification, clustering and association to get better mining performance.

Peissig et al. (2014) introduced techniques to automate the learning process and improved the model performance. The techniques fall into three categories: first, selection of training set examples without expert (physician) involvement to provide supervision for the learning activities; second, left-censoring of background data to identify subgroups of patients that have similar features denoting the phenotype; and third infusing borderline positive examples to improve rule prediction. The suggested methods reduce the expert (physician) time and enhance attribute awareness in the EHR-driven phenotyping process.

A research by Ahmad et al. (2015) have analyzed the convenience of data mining techniques in healthcare. Authors mentioned that data mining techniques plays a noteworthy role in nurturing the momentous volume of data into useful information. Authors discussed the role of classification, clustering, SVM, NN and Bayesian methods in mining the data related to breast, lung cancer, heart diseases, smoking behavior, thyroid, dengue, Alzheimer's, diabetes etc. Authors also mentioned some of the challenges that researcher have faced while mining the data related to healthcare industry. Author suggested that medical industry has to focus on developing better and accurate medical information system by using different data mining techniques.

A study by Nilashi et al. (2017) proposed a new knowledge-based system for diseases prediction using clustering, noise removal, and prediction techniques. They utilized Classification and Regression Trees (CART) to generate the fuzzy rules to be used in the knowledge-based system. Further, test this method on several public medical datasets.

Results on Pima Indian Diabetes, Mesothelioma, WDBC, StatLog, Cleveland and Parkinson's telemonitoring datasets show that method remarkably improves the accuracy of diseases prediction. The results showed that the combination of fuzzy rule-based, CART with noise removal and clustering techniques effective in diseases prediction from real-world medical datasets.

A study conducted by Yassin et al. (2018) systematic review aims to help researchers in innovating and developing CAD systems to assist the medical society in detection/diagnosis and early treatment of breast cancer. From the review, they found the usage of artificial intelligence methods is increasing because of the effectiveness in classification and detection schemes assisting experts in the medical field. In the future work, it is recommended to have standardized public image databases that contain images from different image modalities for the same case to support the dependency of more than one image modality in classification task and combine information from multiple views. It will be wealthy if they contain DNA sequence of cases. This will enable CADs to provide results that depend on different perspectives concerning different modalities and even sequences. Also swarm intelligence is worth studying as it was rarely applied in the investigated publication in CADs systems. Developing MLT-CAD system that combines more than one image modality is a necessity.

4.1 Gaps identified

From the review of literature article, it is observed, the suggested approach may also be applicable to other diseases classification problems which include datasets with same nature used in this study. However, there is still plenty of work in conducting researches on clustering, noise removal and fuzzy rule-based techniques for disease diagnosis in order to exploit all their potential and usefulness. In the future, more attention required it should be paid to the datasets for disease classification and prediction using the incremental machine learning approaches. Hence, need to evaluate this method on additional datasets and in particular on large datasets to show the effectiveness of the method for computation time of large data. In addition, investigates that how the proposed method can be extended to be applicable to the other types of datasets in medical domain.

5.0 Research Aim

The goal of this research work is to develop and implementation of novel framework on the basis of Clustering and Classification which relies on hybrid machine learning technique for Diagnosis of Diabetics using Genomic Database in the healthcare field.

6.0 Motivation and objective of this research

The primary objective of this research is to show the contribution of data mining in assessing the lifestyle based diseases. The effort is made to review the literature article whose research work is concentrated for both experts as well as patients. The key points (disease, methodology, results, accuracy) of the different research works along with the use of tool or techniques will be highlighted. Finally, the focus is to determine the areas that require more attention to data mining techniques along with machine learning technique.

The research objectives are,

- To categorize the content with respect to the behaviour informatics and cluster the data for analyzing their data pattern.
- To improve the diagnostic performance of current diagnostic methods for disease prediction by evaluating diagnostic information with supervised and unsupervised machine learning algorithms
- To validate the classification algorithm by two-layer nested cross-validation.
- To evaluate the performance of proposed approach by evaluating the parameters like precision, recall, F-measure, the accuracy with classification rate.
- To compare the performance of different classifiers and clustering algorithms on thesis diabetes datasets.

7.0 Proposed method

In this research focus on a new knowledge-based system for Diagnosis of Diabetics using Genomic Database via clustering, noise removal, prediction and classification techniques. The contribution of research is carried out in threefold: First, the Prediction is possible by use of existing variables in the database in order to predict unknown or future values of interest. Second, the description focuses on finding patterns describing the data the subsequent presentation for user interpretation. Third, use Classification and rule-based decision tree approach to generate the fuzzy rules to be used in the knowledge-based system. The knowledge-based system can assist medical practitioners in the healthcare practice as a clinical analytical method. The data classification process involves learning and classification (Based on deep belief network). In Learning the training data are analyzed via classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. The classifier-training algorithm uses these pre-classified to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier (combination of deep belief network with rule-based decision tree approach).

In our study, randomly divided the whole data into two sets with a similar number of subjects from each class in each set named as training and testing. Two-layer nested cross-validation is employed to validate the classification algorithm. The classifier that performs best in the two-layer nested cross-validation is considered to be the optimal classifier, and the corresponding parameters will be used in the classification model (Deep Belief Network with rule-based decision tree approach) to classify the new test sample.

7.1 Proposed methodology implementation flow

Stage I: Data pre-processing (Structured or unstructured)

Stage 2: Classification (Early detection of disease or disorder)

Stage 3: Evaluation Methods

7.1.1 Pre-processing: Data Reduction

- **Instance Reduction:** The study combines instant selection (combines both noise and redundancy removal) with drift detection module to influence the usability of prototypes.
- **Pre-processing:** To remove the noise and artifacts we implemented; low pass filtering, removing missing values (zeros or negative), sampling the data, checking and removing outliers from the data set, and calculating statistical/descriptive values such as; maximum, minimum, mean, median, mode, standard deviation and range, in order to have a normalized data set throughout the diagnosis.

7.1.2 Classification

The proposed study adds class information to the discretization process (CNN based classifier) to accommodate for local drifts, where properties of only some class change.

- Deep Belief Network is an algorithm among deep learning. It is an effective method of solving the problems from neural network with deep layers, such as low velocity and the overfitting phenomenon in learning.
- Rule-based decision tree approach help preserve the error that can be back-propagated through time and layers. By preserving a more constant error, they permit recurrent nets to continue to learn over many time steps (over 1000), thereby opening a channel to link causes and effects remotely.

7.1.3 Evaluation criteria

The performance of the classifier is validated based on three aspects; sensitivity, specificity, and accuracy. Sensitivity measures the predicted output with respect to the change in input. In other words, sensitivity shows the ratio of the true positives that are correctly identified. This is a contrast with specificity, which measures the ratio of true negatives that are correctly identified. The relationship amongst the predicted value and the actual value is called accuracy. Accuracy measures how close the predicted value to the actual value.

The three aspects are measured using the following formulae:

$$Sensitivity(\%) = \frac{TruePositives}{TruePositives + FalseNegatives} * 100$$

$$Specifity(\%) = \frac{TrueNegatives}{TrueNegatives + FalsePositives} * 100$$

$$Accuracy(\%) = \frac{TruePositives + TrueNegatives}{TotalNumberofSamples} * 100$$

where the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) are explained in Table 2.

Table 1: Abbreviation for performance parameters

Abbreviation	Explanation
True positive (TP)	The number of people who actually diagnosed with disease or disorders ➢ Target output = 1 ➢ Network output = 1
False positive (FP)	The number of people who actually healthy but diagnosed with disease or disorders ➢ Target output = 0 ➢ Network output = 1
True negative (TN)	The number of people who actually healthy but diagnosed as healthy ➢ Target output = 0 ➢ Network output = 0
False Negative (FN)	The number of people who actually have the disease but diagnosed as healthy ➢ Target output = 1 ➢ Network output = 0

8.0 Chapterization

The organization the research will be,

Chapter 1 is the Introduction which presents the research background, problem statement, motivation, aim, objective and significance of the research.

Chapter 2 is a literature review that describes the concepts and definition of the previous studies along with the previous studies related diagnosis of diabetics using different database.

Chapter 3 is a research methodology which covers the detailed description of the proposed research design, approach, data collection, analysis and sample size of the data.

Chapter 4 is experimental results and analysis which discuss the analysis and validation of the collected data via the researcher.

Chapter 5 is a discussion, findings and conclusion which will cover the research findings, obtained results discussion and conclusion of the research. At last, we will discuss the recommendation of future studies.

References

Abdelaziz, A., Elhoseny, M., Salama, A.S. & Riad, A.M. (2018). A machine learning model for improving healthcare services on cloud computing environment. *Measurement*. [Online]. 119. pp . 117–128. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0263224118300228.

Abdelaziz, A., Elhoseny, M., Salama, A.S., Riad, A.M. & Hassanien, A. (2017). Intelligent Algorithms for Optimal Selection of Virtual Machine in Cloud Environment,Towards Enhance Healthcare Services. In: *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics*, [Online]. 2017, Germany: Springer, pp. 23–37. Available from: http://sci-hub.tw/10.1016/j.measurement.2018.01.022.

Ahmad, P., Qamar, S. & Rizvi, S.Q.A. (2015). Techniques of Data Mining In Healthcare: A Review. *International Journal of Computer Applications*. [Online]. 120 (15). pp . 38–50. Available from: http://research.ijcaonline.org/volume120/number15/pxc3904126.pdf.

Arjun, C. & Anto, S. (2015). Diagnosis of Diabetes Using Support Vector Machine and Ensemble Learning Approach. *International Journal of Engineering and Applied Sciences*. [Online]. 2 (11). pp . 68–72. Available from: http://sci-hub.tw/10.1016/j.measurement.2018.01.022.

Asgari, E. & Mofrad, M.R.K. (2015). Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics F. H. Kobeissy (ed.). *PLOS ONE*. [Online]. 10 (11). pp . e0141287. Available from: http://dx.plos.org/10.1371/journal.pone.0141287.

Beagley, J., Guariguata, L., Weil, C. & Motala, A.A. (2014). Global estimates of undiagnosed diabetes in adults. *Diabetes Research and Clinical Practice*. [Online]. 103 (2). pp . 150–160. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0168822713003847.

Cao, C., Liu, F., Tan, H., Song, D., Shu, W., Li, W., Zhou, Y., Bo, X. & Xie, Z. (2018). Deep Learning and Its Applications in Biomedicine. *Genomics, Proteomics & Bioinformatics*. [Online]. 16 (1). pp . 17–32. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1672022918300020.

Cios, K.J., Mamitsuka, H., Nagashima, T. & Tadeusiewicz, R. (2005). Computational intelligence in solving bioinformatics problems. *Artificial intelligence in medicine*. [Online]. 35 (1–2). pp . 1–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16095889.

Dev, S.K., Krishnapriya, S. & Kalita, D. (2016). Prediction of Heart Disease using Data Mining Techniques. *Indian Journal of Science and Technology*. [Online]. 9 (39). Available from: http://www.indjst.org/index.php/indjst/article/view/102078.

Friis-jensen, E. (2007). *Modelling and Simulation of Glucose-Insulin Metabolism*. [Online]. 2007. IMM. Available from: http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/5312/pdf/imm5312.pdf. [Accessed: 3 April 2018].

Hamad, A.M. (2016). Lung Cancer Diagnosis by Using Fuzzy Logic. *International Journal of Computer Science and Mobile Computing*. [Online]. 5 (3). pp . 32–41. Available from: http://sci-hub.tw/10.1016/j.measurement.2018.01.022.

Han, J. & Kamber, M. (2001). *Data Mining: Concepts and Techiques*. San Diego, USA: Morgan Kaufmann.

Huang, G., Song, S., Gupta, J.N. & Wu, C. (2014). Semi-Supervised and Unsupervised Extreme Learning Machines. *IEEE Transactions on Cybernetics*. [Online]. 44 (12). pp . 2405–2417. Available from: http://ieeexplore.ieee.org/document/6766243/.

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I. & Chouvarda, I. (2017). Machine Learning and Data Mining Methods in Diabetes Research. *Computational and Structural Biotechnology Journal*. [Online]. 15. pp . 104–116. Available from: http://linkinghub.elsevier.com/retrieve/pii/S2001037016300733.

Kumar, C.B., Kumar, M.V., Gayathri, T. & Kumar, S.R. (2014). Data Analysis and Prediction of Hepatitis Using Support Vector Machine (SVM). *International Journal of Computer Science and Information Technologies*. [Online]. 5 (2). pp . 2235–2237. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.662.340&rep=rep1&type=pdf .

Nilashi, M., Ibrahim, O., Ahmadi, H., Shahmoradi, L. & Farahmand, M. (2018). A hybrid

intelligent system for the prediction of Parkinson's Disease progression using machine learning techniques. *Biocybernetics and Biomedical Engineering*. [Online]. 38 (1). pp . 1–15. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0208521617301948.

Nilashi, M., Ibrahim, O. bin, Ahmadi, H. & Shahmoradi, L. (2017). An analytical method for diseases prediction using machine learning techniques. *Computers & Chemical Engineering*. [Online]. 106. pp . 212–223. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0098135417302570.

Nithya, B. (2016). An Analysis on Applications of Machine Learning Tools, Techniques and Practices in Health Care System. *International Journal of Advanced Research in Computer Science and Software Engineering*. 6 (6). pp . 1–8.

Peissig, P.L., Santos Costa, V., Caldwell, M.D., Rottscheit, C., Berg, R.L., Mendonca, E.A. & Page, D. (2014). Relational machine learning for electronic health record-driven phenotyping. *Journal of Biomedical Informatics*. [Online]. 52. pp . 260–270. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1532046414001543.

Prerana, T., Shivaprakash, N. & Swetha, N. (2015). Prediction of Heart Disease Using Machine Learning Algorithms- Naïve Bayes,Introduction to PAC Algorithm, Comparison of Algorithms and HDPS. *International Journal of Software Engineering*. [Online]. 3 (2). pp . 90–99. Available from: http://sci-hub.tw/10.1016/j.measurement.2018.01.022.

Samant, P. & Agarwal, R. (2018). Machine learning techniques for medical diagnosis of diabetes using iris images. *Computer Methods and Programs in Biomedicine*. [Online]. 157. pp . 121–128. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0169260717304649.

Sharma, M., Singh, G. & Singh, R. (2017). Stark Assessment of Lifestyle Based Human Disorders Using Data Mining Based Learning Techniques. *IRBM*. [Online]. 38 (6). pp . 305–324. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1959031817301112.

Sharma, S. (2016). Cervical Cancer stage prediction using Decision Tree approach of Machine Learning. *International Journal of Advanced Research in Computer and Communication Engineering*. [Online]. 5 (4). Available from:

https://www.ijarcce.com/upload/2016/april-16/IJARCCE 88.pdf.

Tintu, P. & Paulin, R. (2013). Detect Breast Cancer using Fuzzy C means Techniques in Wisconsin Prognostic Breast Cancer (WPBC) Data Sets. *International Journal of Computer Applications in Technology*. [Online]. 2 (5). pp . 614–617. Available from: http://sci-hub.tw/10.1016/j.measurement.2018.01.022.

Tomar, D. & Agarwal, S. (2013). A survey on Data Mining approaches for Healthcare. *International Journal of Bio-Science and Bio-Technology*. [Online]. 5 (5). pp . 241–266. Available from: http://www.sersc.org/journals/IJBSBT/vol5_no5/25.pdf.

Yassin, N.I.R., Omran, S., El Houby, E.M.F. & Allam, H. (2018). Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review. *Computer Methods and Programs in Biomedicine*. [Online]. 156. pp . 25–45. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0169260717306405.