International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 5.22 (SJIF-2017),e-ISSN:2455-2585
International Conference on
Recent Explorations in Science, Engineering And Technology (ICRESET'19)
Volume-5, Special Issue-March, 2019.

A Survey on Accuracy in Diabetics & Research and Predictive Re-surgery problems using Data mining techniques

Mr. A.P Christopher Arokiaraj, M.C.A, M.Phil.,(Ph.D).,

Department of Computer Science,KG College of Arts and Science

ABSTRACT: ***In our day to day life surgical procedures are associated with medicine, the same is the case for critical healthcare. The goal of this work is to review on the best works in Predictive Resurgery and to identify the most accurate method to predict Diabetes to assist health professionals in these areas in the field of biosciences. By applying various Datamining techniques it is possible to help the medicinal knowledge, to predict whether the particular patient should or should not be operated upon the same problem. In this study, some aspects such as history of the disease, hereditarial, and the age factor and some data classes were built to improve the models that has been already been formed. In addition, several models are also created that aims at predicting the re-surgery of patients. The metric used to get the sensitive datasets and the success rate of this approach is almost 90%.The modern advances in bioinformatics and health sciences have led to a considerable production of medicinal data, such as high throughput genetic data and clinical information, generated from large Health Related Electronic Records (HRERs) Diabetes mellitus is a metabolic disorder characterized by the presence of hyperglycemia due to defective insulin secretion, defective insulin action or both exerting significant pressure on human health across the world. The Diabetes research has led to the generation of massive volumes of data. A systematic review has been conducted in various applications of machine learning, techniques and tools in data mining in the field of diabetes research with respect to Prediction and Diagnosis, Complication due to Diabetes, Genetic Background and the surrounding environment, along with Health Care and Management. A wide range of machine learning algorithms were implemented in these approaches and in those findings indicate 85% of those used were characterized by supervised learning approaches and 15% by unsupervised ones mainly association rules. In addition, different data mining techniques used to uncover potential predictors of diabetes. Support vector machines is been suggested as the most accurate and popular algorithm. Clinical data sets are used considering the accuracy of data as input. This is achieved from the results by showing the performance of each classification algorithm through extraction of valuable knowledge.***

Keywords: Diabetes mellitus, Data mining prediction, DM, SVM, predictive re-surgery

I. INTRODUCTION

This study focuses on the use of classification approaches in order to predict the patients who are resurgeried together with the medical knowledge in view of assisting health professionals. The dataset used in this project was provided by standard hospital, however to improve the quality of the results, these have been modified. The strategies used were standardization of data to create the models, but without changing the accuracy of the results. The standardization of data is a set of rules that aims to reduce data redundancy and increase data integrity. This study was conducted by following the CRISP-DM methodology.

By applying Machine learning and data mining methods in DM 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 DM one of the main priorities in biological science research, which undoubtedly generates huge amounts of data. Therefore, machine learning and data mining approaches in DM are dealt with caution, when it comes to diagnosis, management and other related clinical administration aspects. This framework helps to review the recent literature on machine learning and data mining approaches related to diabetes research.

The review deals with background knowledge on machine learning and knowledge discovery in databases (KDD).Knowledge discovery in databases is precise process consisting of numerous distinct steps. Data mining is the nuclear step, which results in the discovery of hidden but useful knowledge from voluminous databases. Data mining is a non-trivial extraction of inherent formerly unknown and potentially valuable information about data.

The key problem addressed in this review paper is that patients who might develop diabetes are not conscious of the associated high risks. Late or lack of diabetes diagnosis increases the chance of developing any disease due to cardio vascular complications. However, screening patients and detecting asymptotic disease such as diabetes might help in delaying its progression and preventing its complications, controlling the treatment, and reducing the costs of this preventable disease in the health care system. Furthermore, it is also beneficial for both public health and clinical practice. Demographic characteristics such as age, sex and race are non-modifiable risk factors of diabetes. Machine learning methods have shown their capabilities to effectively deal with large number of variables while producing powerful predictive models. They also embed variable selection mechanisms which can detect complex relationships in data. Supervised classification techniques are popular machine learning methods that aim to explain the dependent variable in terms of the independent variables. The review study takes clinical research dataset collected by The Henry Ford Exercise Testing (FIT) project and using it to investigate and find out the relative performance of various machine learning classification methods such as Decision Tree (DT), Naïve Bayes (NB), Logistic Regression (LR), Logistic model tree (LMT) and Random Forests (RF) for predicting incident diabetes using medical records of cardio respiratory fitness.

II. LITERATURE SURVEY

Ricardo Peixoto et al[1] predicting resurgery diagnosis uses CRISP-DM methodology which is partitioned into six tasks: Business understanding, Data understanding, Data preparation; Modeling; Evaluation and Implementation. This methodology facilitates the users with a structured approach to project planning and can serve as a common reference point in the use of Data mining. In the developing phase, the tool used for data exploration, preparation and creation of scenarios and collation of data was Oracle SQL Developer. In this approach a layer was developed with the aim of being able to translate business goals by classification techniques. The classification task aims to recognize, together with the data, the observations that have the same characteristics. The goal is to predict the class of an item from the database. If a data record contains the Region field, then some of the typical values of the field, such as North, South, East, and West can define the class. In the classification task, the most common techniques used are decision trees, support vector machines, neural networks, classifiers Bayes and genetic algorithms. In order to achieve the expected results were created 17 scenarios that allow doctors of the ICU's to understand the characteristics of the resurgeried patients. The first scenario was created using some selection criteria, such as physically or mentally handicapped patients or patients with heart, hepatic, chronic or respiratory insufficiency. In total, 136 models were generated. These models can be represented by:

DMM1={17 scenarios, 4 techniques, 2 sampling methods, 1 Representation Methods, 1 Target}

In another study, by I.Kavakiotis et al[2], employing machine learning and datamining techniques on diabetes research were identified. Two databases were searched: the one extensively used in biomedical sciences, PubMed and DBLP Computer Science Bibliography, containing more than 3.4 million journal articles papers and other publications on computer science. The main reason behind the utilization of DBLP was that there are certain high impact international scientific journals in computer science not indexed by PubMed, although in some cases, the proposed published methods are applied on biomedical datasets. A large number of factors are known to be important in the development and progression of DM. Obesity stands as a major risk factor, especially in T2D, given the strong causal relationship between that and the onset of DM. DM diagnosis is carried out through several tests [α glycate haemoglobin (A1C) test, random blood sugar test, fasting sugar test or oral glucose tolerance test]. There is evidence that in both T1D and T2D, early diagnosis and prediction of the onset of the disease are vital to the a) retardation of the progression of the disease, b) targeted selection of medication, c) prolonging life expectancy, symptom alleviation, and d) onset of related complications. Biomarkers are measurable indicators of a certain condition representing health and disease states typically measured in body fluids, encountered and thus determined independent of their etiopathogenic mechanistic pathway, and used to examine clinical and sub-clinical disease burden and the response to treatments. Current technologies such as metabolomics, proteomics, and genomics contribute to the development of ptheora of new biomarkers.

In case of DM, biomarkers may reflect the presence and severity of hyperglycemia or presence and severity of the related complications in diabetes. In case of disease prediction and diagnosis, numerous algorithms and different approaches have been applied, such as traditional machine learning algorithms, ensemble learning approaches and association rule learning in order to achieve the best classification accuracy. Most noted among the aforementioned ones are the following: Calisir and Dogantekin proposed LDA-MWSVM, a system for diabetes diagnosis. The system performs feature extraction and reduction using Linear Discriminant Analysis (LDA) method followed by classification using the Mortel Wavelet Support Vector Machine (MWSVM) classifier. Ensemble approaches use multiple research algorithms have proven to be effective way of improving classification accuracy. Rotation forest (RF), newly proposed ensemble algorithm, to combine 30 machine learning algorithms.

Finally, Han et al. presented an ensemble learning approach, which transforms the "black box" of SVM decisions into comprehensible and transparent rules. Association rules are mainly employed to identify associations between risk factors in an interpretable form. In [10], authors applied association rules to detect combinations of variables or predictors frequently occurring together in diabetic patients. Simon et al. proposed Survival Association Rule (SAR) Mining, an extension to the traditional Association rule mining, capable of handling survival outcomes, make adjustments for cofounders and incorporate dosage effects. Finally, Batal et al. utilized temporal pattern mining for discovering predictive patterns in complex multi-variate time series data, to improve performance of existing classifiers.

III. MACHINE LEARNING AND KNOWLEDGE DISCOVERY

Machine learning is a term that deals with artificial intelligence along with the ways in which a machine learns from experience. The purpose of machine learning is the construction of computer systems that can acclimatize and learn based on the collection of experiences. A more detailed and formal definition of machine learning is given by Mitchel: A computer program is said to learn from a collection of experiences E given a set of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with familiarity of the given task.
Knowledge discovery in databases (KDD) is a field that comprises of theoretical aspects, methods and techniques, to make the data available meaningful and mine useful information from them. It is considered as a multi-step process that has various stages namely selection, pre-processing, transformation, data mining, interpretation and evaluation. The most imperative step in the KDD process is data mining, highlighting the application of machine learning algorithms in analysis of data. According to Fayyad et al. KDD is the nontrivial process identifying valid, novel, hidden useful information and ultimately understandable patterns in data.

A. Categories of Machine Learning Tasks

Machine learning tasks are characteristically classified into three wide-ranging categories. These are: a) supervised learning, in which the system refers to a function from labeled training dataset, b) unsupervised learning, in which the learning system tries to refer the configuration of unlabeled data, and c) reinforcement learning, in which the system communicates with a dynamic environment which has varying information.

1) ***Supervised Learning*** **:** Supervised learning deals with inductive learning using a function called target function, which is an expression of a model elaborating the information. The goal of this function is used to predict the value of a variable, called dependent variable or output variable, from a set of variables called independent variables or features or characteristics. The set of possible input values of the function are called its instances. Each case is described as a set of features. A subset of all cases for which the output variable rate is known as training data set. In order to arrive at the best target function, the learning system, given a training set, takes into consideration alternative functions called hypothesis and is denoted by h. In terms of supervised learning, there are two divisions of learning level tasks namely classification and regression. Models in classification focus on getting and predicting numerical values. The most common techniques are listed namely, Decision Trees (DT), Rule Learning and Instance based learning (IBL), such as (k-NN), Genetical Algorithm, Artificial Neural Networks (ANN), and Support Vector Machines (SVM).

2) Unsupervised Learning : When it comes to, unsupervised learning the system tries to find out the unknown composition of data or relations between variables. In that case, training dataset comprising of instances that lacking corresponding labels.

3) Association Rule Learning: Association Rule Mining came into existence much later than the concept of machine learning and is subjected to greater authority level from the research domain area of databases. It was proposed in 1990s as a market holder analysis, in which focuses to find correlations in the objects of a given database. It is based on the shopping cart ex-ample, the association rules can be of the form $\{X_1, \ldots, X_n\} \rightarrow Y$, which associates that if the user find all of $X_1, \ldots, X_n$ in a cart it is possible to find Y. The most well-known association rule discovery algorithm is Apriori algorithm.

Although, the association-rule mining came into existence as naïve market holder tool, it has become one of the most valuable tools for performing unsupervised exploratory data analysis considering a wide range of commercial and research areas that includes bioinformatics. Some of the well known applications in bio-informatics include biological sequence succession analysis, gene expression data analysis and other related analyses.

4) Clustering : Clusters are regarded as revealing patterns that occurs through clustering. It can be termed as the partition of a entire dataset into numerous groups of data, so that each examples from the dataset belonging to the same group are same as possible and the examples belonging to diverse groups differ as a great deal as possible.

B. Feature Selection

Feature selection is categorized as one of the most significant processes of the KDD's data transformation step. It is formally denoted as the procedure of selecting a subset of features from the provided predefined feature space, which is more applicable to and informative for the construction of a model. The advantages of feature selection are numerous and relate to different aspects in analysis of data, such as improved visualization and knowledge of data, reduction of computational time and duration of analysis, and better accuracy for data prediction.

There are two major diverse approaches in the process of feature selection. The first one is to construct an independent judgment, based on general data characteristics. The Methods that identify to this family come into this approach, hence called filter methods, because the feature set and is filtered away before the construction of the model. The second approach is to use a machine learning algorithm to assess different feature subsets and at last select the one with the best performance based on accuracy in classification. The latter algorithm can be proposed to be used in the end to build a better predictive model. Methods that come in this class are called wrapper methods, because the upcoming algorithm covers the whole feature in the selection process.

C. Diabetes Mellitus

Diabetes mellitus is a metabolic disorder characterized by the presence of hyperglycemia due to defective insulin secretion, defective insulin action or both exerting significant pressure on human health across the world. Insulin deficiency results in elevated blood glucose levels called hyperglycemia and impaired metabolism of carbohydrates, fat and proteins. DM is used the most complicated and the most frequent endocrine disorders, affecting more than 200 million people worldwide. Diabetes can be identified as diseases by estimating its dramatic rise in the upcoming years.

DM can be categorized into several distinctive types. Although, there are two most important clinical types, namely Type 1 diabetes (T1D) and Type 2 diabetes (T2D), according to the etiopathology of the disorder. T2D has been found out to be the most widespread form of diabetes, mainly characterized by resistance of insulin. The main causes of T2D include lifestyle of the individuals, physical activity, diet related habits and heredity, in contrast T1D is thought to be due to auto-immunological obliteration of the Islet of Langerhans possessing pancreatic-β cells. T1D affects approximately 10% of all the diabetic patients along the world, with nearly 10% of them developing idiopathic diabetes. The other applications of DM are divided on the basis of insulin secretion profile and /or that includes gestational diabetes, endocrenopathies, neo-natal, mitochondrial and diabetes that happens during pregnancy. The symptoms of DM are namely, polyurea, poly-dipsia, and significant weight loss among others. Diagnosis depends on blood glucose levels and the fasting plasma glucose level =7.0 mmol/L.

The progression of Diabetes Mellitus is strongly associated to several complications, mainly due to hyperglycemia. It is rather a familiar type of Diabetes that spans a wide range of heterogenous patho-psycological conditions. The most frequent complications are partitioned into micro- and macro-vascular conditions, including diabetic nephropathy, retinopathy, diabetic coma and related cardiovascular disease.

It is important to administer insulin as a main treatment of T1D, although insulin is also administered in certain cases of T2D patients, when the condition arises that hyperglycemia cannot be controlled through oral diet, weight loss, exercise and oral medication. Current medication is especially in need primarily in a) Saving one's life and curing the disease symptoms, and b) Anticipation of long term diabetic complications and/or elimination of numerous threat factors, thereby increasing longevity of the patient. The majority of the presence of anti-diabetic agents, however, exhibit abundant side-effects. In addition, insulin therapy is associated to sudden weight gain and hypoglycemic actions. Hence, anti-diabetic drug design and discovery is of great concern and concurrently a research challenge.

Although extensive researches are conducted in DM has provided extensive knowledge, over the past few decades, on the a) etiopathology genetic or environmental or external factors and cell related mechanisms), b) the corresponding treatment, and c) Screening and managing of the disease, there is still much to be the area of concern.

D. A view in terms of Computational research into Diabetes

The study conducted envisions that machine learning and datamining can be used to increase the precision of the prediction of DM. The vast majority of the reported articles updated classification accuracy, above 80% in the prediction of DM. With regard to the prediction task, most of the classification algorithms has been implemented. Although, the most commonly used ones are SVM, ANN and DT. SVM has the upper hand and stands as a most successful algorithm in both biological and clinical datasets in DM. A vast majority of articles (85%) used supervised learning approaches namely in regression and classification tasks. The outstanding 15%, of them employed association rules primarily to study the relationship between biomarkers.

To be more specific, that deals with the task of evaluation, in the research conducted on the reports generated indicate subsets of biomarker features were evaluated through necessary procedures by splitting the dataset into training and test-set of data or through cross-validation.

It is worth highlighting, the fact that in many of the related studies, after feature/biomarker selection, researchers have been performed as a comparative analysis on different machine learning algorithms in order to evaluate the performance of prediction and finally chose the best ones. As a result, an algorithm with the superlative performance in one dataset could have lower prediction accuracy when compared to other algorithms in different datasets. SVM displayed the pinnacle with regard to the classification accuracy or the Area beneath the Curve (ABC).

The projection of the overall results on a wide variety of algorithms and techniques are used in DM research. Evidently, diverse machine learning tasks are used in diverse scientific questions, such as forecast on DM or association among biomarkers. We can identify at this end that classification and regression techniques are used for prediction tasks, to find out glucose levels and association rules were the condition occurs on dependencies between biomarkers. Fascinatingly, for every machine learning task, a wide variety of algorithms have been used in the current literature. The reason behind that the fact of the accuracy of the algorithm depends heavily on the kind of data used. Consequently, immense effort in research depends on the preprocessing of data, namely selection of the feature and then a range of algorithms applied to the processed data in order to recognize the most thriving one for the distinguished dataset. Furthermore, it is essential for machine learning studies that a dataset be adequately large enough for the algorithm to be trained properly. Although biomedical technology have arrived at the era of big data for numerous reasons, namely lesser cost of next generation sequencing or unified HRERs, datasets with great changeability in size are very frequent.

III. DISCUSSION

After the assessment of all the scenarios in the study, it is possible to see that some of the models strike the forecast of resurgeried patients in about 90% of the cases. These models having high probability have properties that can be viewed as significant note in predicting resurgery in patients. The knowledge that these models decipher into besides with the knowledge and experience of physicians can be decisive to the ICUs. However, the accuracy and specificity not have very significant results, which makes as to come to a conclusion that the data provided are only good at predicting the resurgeried patients. The created models cannot assure by 100 percent that a patient will be resurgeried, however, these models can be combined with the medical knowledge will indeed allow to make that prediction more accurate.

IV. CONCLUSIONS

In this area of study, a methodical effort was made to recognize and examine machine learning and data mining approaches applied on Diabetes Mellitus research. DM is swiftly emerging as one of the utmost global health challenges of the 21st century. Till date, there are significant works carried out in nearly all aspects of DM research and especially in biomarker detection and prediction-diagnosis. The arrival of biotechnology, with the enormous amount of data produced, along with the increasing amount of RHERs is expected to ascend to further comprehensively investigation toward diagnosis, etiopathophysiology and treatment of DM through employment of machine learning and data mining techniques in concentrated datasets that include clinical and biological information.

Finalized study, it is identified that the size information of each patient undergone resurgery is quite high. As such, this work permits that the conclusions through human observation it would be quite difficult to get the accurate data for exact prediction. The created models are not able to guarantee that the resurgery patients are the patients possessing the distinctiveness present in the built models. However, by accumulation this information to the valuable cognizance, it can be essential in the treatment of the patient helps the professional health in their decision making. By analyzing the work done, it is possible to visualize that the only practical work done to-date is the first part of this project, whose goal is to characterize the re-surgery patients in intensive care with at most accuracy. This work is important in terms of clinical knowledge, since it can help the physicians have updated knowledge and experiences. With this fact it is essential for the health of patients in ICUs. In terms of science, since resurgery problem was recently identified, this study allows us to realize that through a given dataset the patient's illness can be predicted and treated appropriately.

REFERENCES

[1] Ricardo Peixoto, Lisete Ribeiro, Filipe Portela, Manuel Filipe Santos and Fernando Rua, Predicting Resurgery in Intensive Care- A data mining Approach, International Workshop on Healthcare Interoperability and Pervasive Intelligent Systems, Procedia Computer Science 113 (2017) 577-584.

[2] Ioannis Kavakiotis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas, Ioanna Chouvarda, Machine Learning and Data Mining Methods in Diabetes Research, Computational and Structural Biotechnology Journal 15 (2017) 104-116

[3] P.Suresh kumar and V.Umatejaswi, Diagnosing Diabetes using Data Mining Techniques, International Journal of Scientific and Research Publications, Volume 7, Issue 6, June 2017 ISSN 2250-3153.

[4] Manal Alghamdi, Mouaz Al-Mallah, Steven Keteyian, Clinton Brawer, Jonathan Ehrman, Sherif Sakr, Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford Exercise Testing (FIT) project, Pone journal, July2017,https://doi.org/10.1371/journal.pone.0179805.

[5] Tahani Daghistani, Riyad Alshammari, Diagnosis of Diabetes by Applying Data mining Classification Techniques-Comparision of Three Data Mining Algorithms, International Journal of Advanced Computer Science and Applications Vol.7, No.7, 2016

[6] G. Krishnaveni , Prof. T.Sudha, A Novel technique to predict Diabetic disease using data mining – classification techniques, International Conference on Innovative Applications in Engineering and Information technology (ICIAEIT-2017) Volume 3, Special Issue 1, March 2017

[7] Dr.M. Renukadevi, J Maria Shyla, Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus, International Journal of Applied Engineering Research, ISSN 0973-4562 Volume 11, Number 1 (2016) pp 727-730.

[8] Messan Komi , Jun Li ,Yongxin Zhai , Xianguo Zhang, Application of data mining methods in diabetes prediction, International Conference on Image, Vision and Computing(ICIVC),10.1109/ICIVC.2017.7984706, 2017 IEEE.

[9] Karkhanis Apurva Anant, Tushar Ghorpade,Vimla Jethani, Diabetic retinopathy detection through image mining for type 2 diabetes, International Conference on Computer Communication and Informatics (ICCCI), 10.1109/ICCCI.2017.8117738, 2017 IEEE.

[10] Gyoorgy Eigner, Katalin Koppány, Peter Pausits, Levente Kovács, Nonlinear identification of glucose absorption related to Diabetes Mellitus, International Conference on Intelligent Engineering Systems (INES), 10.1109/INES.2017.8118567, 2017 IEEE.

[11] V. Mareeswari, Saranya R, Mahalakshmi R, Preethi E, Prediction of Diabetes Using Data Mining Techniques, Research Journal of Pharmacy and Technology (RJPT), Vol 10, Issue 4, 2017

[12] Vrushali R. Balpande ; Rakhi D. Wajgi Prediction and severity estimation of diabetes using data mining technique, International Conference on Innovative Mechanisms for Industry Applications(ICIMIA),10.1109/ICIMIA.2017.7975526, 2017 IEEE.

[13] K R Pradeep, N C Naveen, Predictive analysis of diabetes using J48 algorithm of classification techniques, International Conference on Contemporary Computing and Informatics (IC3I), 10.1109/IC3I.2016.7917987, December 2016, IEEE.

[14] Saman Hina Anita Shaikh and Sohail Abul Sattar, Analyzing Diabetes Datasets using Data Mining, Journal of Basic & Applied Sciences, 2017, 13, 466-471

[15] P.Selvi, "An Analysis on removal of duplicate records using different types of Data Mining Techniques :A Survey", International Journal of Computer Science and Mobile Computing, November 2017.

[16] R.Sarala, "The process of Augmenting Data Warehouses with Big Data – A Survey" , International Journal of Multidisciplinary Educational Research, Volume: 6, Issue: 10, October 2017.