10th National and International Academic Conference
Huachiew Chalermprakiet University

Correlation Analysis of Laboratory blood tests and complications in Diabetes Mellitus using Data Mining Technique

Panthip Rattanasinganchan[1], Kittipat Sopitthummakhun[2], Kritsanee Maneewong[3]

1 Faculty of Medical Technology, Huachiew Chalermprakiet University, Samut Prakan, Thailand
2 Faculty of Science and Technology, Huachiew Chalermprakiet University, Samut Prakan, Thailand
3 Division of Medical laboratory and clinical pathology, Chomthong Hospital, Chiang Mai, Thailand
*Corresponding Author's Email: r_panthip@hotmail.com Telephone: 084-234-2723

Rationale

Diabetes Mellitus (DM) is one of the major health challenges all over the world. The prevalence of diabetes among individuals aged 15 years and over increased from 6.9% in 2009 to 8.9% in 2014, with 4.8 million Thai adults having diabetes in 2014, resulting in 76,000 deaths from diabetes-related causes. Unhealthy dietary habits, rushed food consumption, and excessive high-sugar beverage intake contribute to obesity and type 2 diabetes, a serious public health problem associated with high blood sugar levels and various complications. In order to establish guidelines for diagnosis and treatment, it is important to understand the relationship among various factors of the disease. This will lead to more effective treatment manners. The currently used techniques for data correlation is a data mining. In this report, we analyzed the correlation between blood test results in DM patients and the complications that are linked to DM.

Research Objectives

To examine the correlation between laboratory blood test from DM patients and the incidence of complications associated through the high blood sugar using a supervised learning technique with a classification approach.

Methodology

Clinical consideration for laboratory blood tests in Patient with Diabetes Mellitus

The dataset acquired in this study was obtained from a public hospital in Chiang Mai Province, Thailand, in 2020. Each record contains several features, including age, sex, blood pressure (BP), body mass index (BMI), fasting blood sugar (FBS), hemoglobin A1c (HbA1c), total cholesterol (TC), triglycerides (TG), creatinine (Cr), estimated glomerular filtration rate (eGFR), and microalbuminuria (MAU). The study included a total of 1,736 patients who were referred based on their FBS and HbA1c levels. These patients were likely selected to assess the prevalence of diabetes or investigate diabetes-related factors within the study population.

Data workflow

The data mining workflow encompassed the following procedures: First, data collection involved gathering and compiling laboratory blood test data from patients with diabetes mellitus (DM) into a spreadsheet. Second, data preprocessing was carried out using Microsoft Excel and Notepad++. This step involved tasks such as data cleaning, removing duplicates, handling missing values, and transforming the dataset into a suitable format for analysis. Third, the converted dataset was utilized in the data mining software, called "Weka version 3.9.6". Finally, a supervised learning technique with a classification approach was employed to construct and evaluate the classifier model. The classifier model is involved the utilizing algorithms such as J48, RandomForest, TreeRandomTree, and LMT (Shama H and Kumar H, 2016, p. 2095) to develop and establish the validated model.

The construction of the classifier model was contingent upon a dataset consisting of a total of 1,736 instances. The dataset was subjected to a classifier algorithm using a 10-fold cross-validation manner, with each fold representing a "training set". Moreover, the entire dataset was utilized as a "testing set" for validation purposes. The evaluation of each classifier model from the training sets was based on the percentage of correctly and incorrectly classified instances. Further analysis of the best classifier model included the examination of metrics such as the %True Positive Rate (%TPR), %False Positive Rate (%FPR), and %Precision for each class. The workflow in data mining processing is depicted in Figure 1.

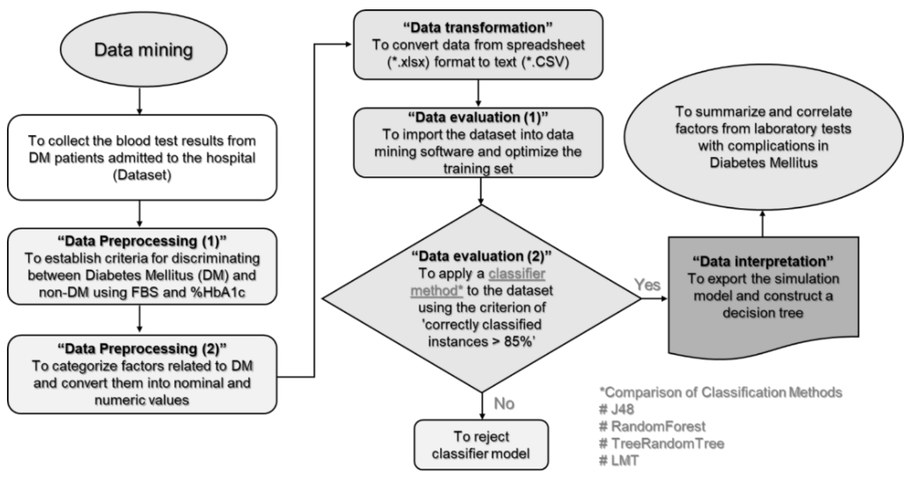

Figure 1. The workflow for data mining processing of laboratory blood tests from DM patients by the data mining technique.

Results

Data evaluation of DM condition and associated complication factors

The analyzed dataset focuses on the clinical factors of DM patients, comprising a total of 1,736 cases. In the case of TC and TG, the percentage of DM patients with normal levels is calculated to be 50.35% and 69.30%, respectively. Based on these findings, it can be suggested that there should be no correlation between the levels of TC and TG and the presence of DM. Investigation of the clinical factors that may be associated with DM conditions implied the particular interest in BMI, eGFR, and MAU. BMI, a commonly used factor for screening DM patients, is categorized into overweight and obesity, which together account for 66.53% of the total DM patient population. In contrast, BMI values within the normal and low range are calculated only to be 33.47% (Figure 3A), suggesting a significant correlation between elevated BMI and DM. eGFR and MAU are well-known factors associated with kidney failure. The analysis results showed that 54.55% of DM patients were diagnosed with CKD stage 2 or higher (Figure 3B). These results are consistent with MAU category, where early and advanced kidney damage was monitored in 66.82% of total DM cases (Figure 3C). The diagnosis of kidney failure correlates with Cr, MAU, and eGFR. Additionally, the classifier models further investigated DM complications and CKD factors to interpret the predictive model's accuracy for DM complications.

Investigation of the correlation between clinical factors and DM complications with the classifier model

The classifier model was applied to the "training set" using the method of 10-fold cross-validation. Each classifier model was evaluated based on the percentage of correctly and incorrectly classified instances. The J48, RandomForest, and LMT models achieved an accuracy of approximately 80% in correctly classified instances, with the percentage of incorrectly classified instances being less than 25%. The J48, RandomForest, and LMT classifier models were compared by evaluating them to the "testing set' using the entire dataset. The results revealed that the RandomForest model achieved perfect accuracy of 100.00% in correctly classified instances. However, it should be noted that this model did not report the factors related to CKD, and no simulation of the decision tree was provided. The LMT model was evaluated using the same methodology as RandomForest. The results indicated that the percentage of correctly classified instances was lower compared to the J48 model, while the percentage of incorrectly classified instances was higher than J48 due to these comparative parameters. Consequently, J48 was selected for further analysis of the correlation between DM complications and CKD.

Results (cont.)

Investigation of the correlation between clinical factors and DM complications with the classifier model (cont.)

The "testing set" derived from the J48 classifier model was employed to construct a decision tree, with the objective of integrating the correlation between clinical factors and complications pertaining to patients diagnosed with DM. In this decision tree model, priority rankings were assigned to Cr, eGFR, and MAU as factors associated with DM complications, specifically within the J48 pruned classifier tree. The simulation of the decision tree generated from J48 classifier model incorporates the clinical factors, including age, BMI, BP, Cr, MAU, and eGFR. The decision tree output revealed that the first root node was defined by Cr. The parameters obtained from the J48 classifier model for this dataset indicate a robust predictive model. The computed parameters derived from the J48 classifier, including %TPR (True Positive Rate), %FPR (False Positive Rate), and %Precision, substantiate the efficacy of the decision tree model within each class. Notably, the J48 pruned classifier model demonstrated a high %TPR rate exceeding 70% across all targeted classes (CKD stage 1-5), accompanied by a %Precision surpassing 80%, thereby aligning with a low %FPR.

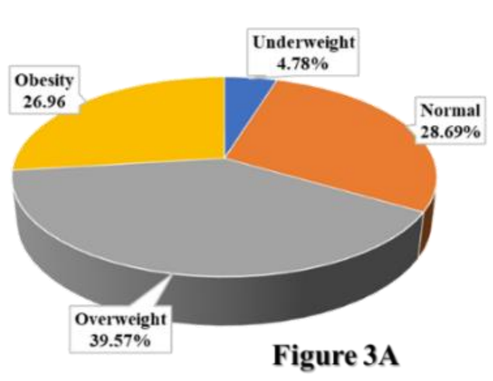

Figure 3A

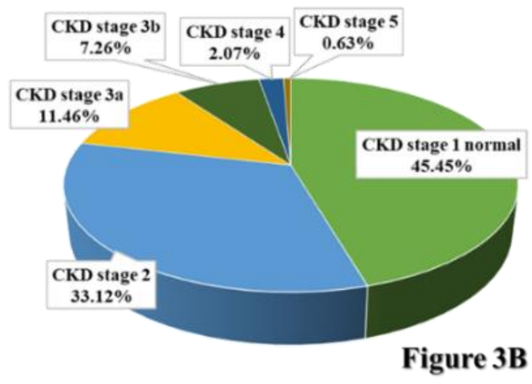

Figure 3B

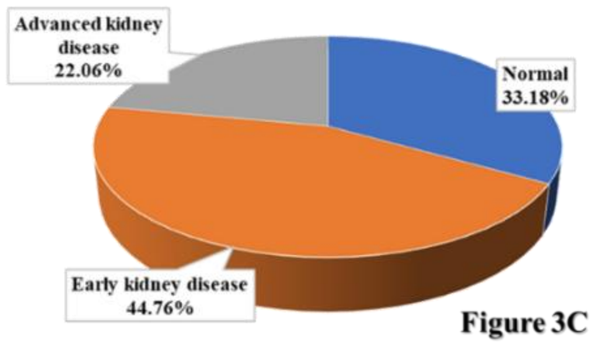

Figure 3C

Figure 3. Pie chart depicts the complications in DM patients such as Figure 3A (BMI), Figure 3B (eGFR), and Figure 3C (MAU) categories.

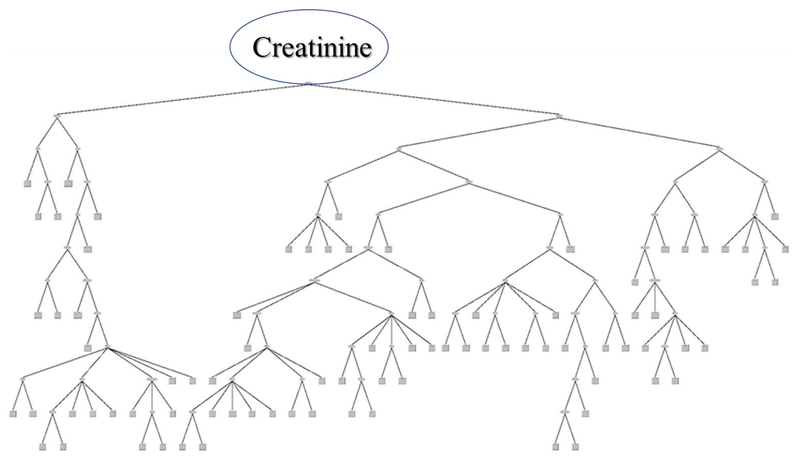

Figure 4 The simulation of decision tree from J48 classifier model composed of clinical factors such as Cr (defined as first root node), eGFR, and MAU in which associated with DM complications.

Discussion

Diabetes Mellitus is one of the major health challenges all over the world. Prevention and prediction of diabetes mellitus is increasingly gaining interest in the healthcare community. There are several data mining techniques for diabetes prediction and course of progression. Among the various techniques for diabetes prediction and course of progression, decision tree is widely considered as one of the most powerful and effective methods for classification. There are many studies about prediction of parameters and diabetes. In 2016, Sajida P. conducted a study to classify patients with diabetes mellitus using risk factors such as age, sex, blood pressure, HDL, triglycerides, BMI and FBS. The data mining approach plays a crucial role in DM research, enabling the utilization of the vast amount of available data on DM and its complications. Dagliati A. and colleagues' methodology demonstrates the effectiveness of adopting data mining techniques in clinical medicine to develop models that leverage patient-specific information for predicting relevant outcomes (Dagliati a. *et al.*, 2018, p. 295-296). This report utilized the supervised learning technique with classifier approach, J48, and constructed the decision tree as a base learner, along with standalone data mining techniques. The results showed a significant difference in diabetes prevalence among different age groups, indicating that age is a significant influencing factor for diabetes. However, in this study, we predicted complication parameters for diabetes patients. We utilized data mining software for evaluation and interpretation using a predictive approach in Weka version 3.9.6. The decision tree was illustrated based on the J48 classifier algorithm. The result showed diabetes patients with > 6.5 mg% HbA1c associated with CKD, BP, and BMI. This study found that 55% of diabetic patients had kidney disease (Figure 3). Chronic kidney disease (CKD) commonly coexists with other conditions, including diabetes. Prolonged high blood sugar levels, caused by diabetes, can damage blood vessels and nephrons in the kidneys, leading to impaired function. Additionally, diabetes patients are prone to developing high blood pressure, which can also cause kidney damage. Obesity can lead to changes in the body's metabolism, causing fat tissue to release free fatty acids and glucose into the blood. Overweight (39%) and obesity (27%) are associated with diabetes. Diabetes reduces the body's ability to use nitric oxide, a molecule that helps blood vessels relax and promote blood flow. This can cause blood vessels to become less elastic and restrict blood and oxygen flow, increasing the risk of hypertension over time.

References

American Heart Association. Retrieved April 13, 2023, from Understanding Blood Pressure Readings | American Heart Association.
Bishop ML., Fody EP. Schoeff LE. 2010. Clinical Chemistry. Techniques, Principles, Correlations. 6th ed. Philadelphia, PA, USA:/Wolters Kluwer Lippincott Williams & Wikins.
Bishop ML., Fody EP. Schoeff LE. 2018. Clinical Chemistry. Techniques, Principles, Correlations. 8th ed. Philadelphia, PA, USA:/Wolters Kluwer Lippincott Williams & Wikins.
Chanlalit, W. (2016). Ocular complications from diabetes mellitus. Journal of Medicine and Health Sciences, 23(2), 36-45.
Dagliati A., Marini S., Sacchi L., Cogni G., Teliti M., Tibollo V., Cata PD., Chiovato L., Bellazzi R. (2018) Machine Learning Methods to Predict Diabetes Complications. 2018, Journal of Diabetes Science and Technology. 12(2), 295-302.
Department of Disease Control, Ministry of Public Health. (2021). Prevention and control of diseases and health threats plan in 5 years (2018-2022), Retrieved April 13, 2023, from https://ddc.moph.go.th/uploads/publish/1189320211018081803.pdf
Lim JU, Lee JH, Kim JS, Hwang YI, Kim TH.,Lim SY., Yoo KH., Jung KS., Kim YK., Rhee CK. Comparison of World Health Organization and Asia-Pacific body mass index classifications in COPD patients. International Journal of COPD 2017:12; 2465-2475.
National Institute for Health and Care Excellence. 2014. Chronic kidney disease in adults: assessment and management. www.nice.org.uk/guidance/cg182.
Shama, H., Kumar, H. (2016). A Survey on Decision Tree Algorithms of Classification in Data Mining. International Journal of Science and Research, 5(4), 2094-97.
Sonthon, P., Promthet, S., Changsirikulchai, S., Rangsin, R., Thinkhamrop B., Rattanamongkolgul, S., Hurst, CP. (2017). The impact of the quality of care and other factors on progression of chronic kidney disease in Thai patients with Type 2 Diabetes Mellitus: A nationwide cohort study. Plos One, 12(7), Published online 2017 Jul 28. doi: 10.1371/journal.pone.0180977
Thai National Health Examination Survey, NHES V. (2016). Thai National Health Examination Survey V Study Group Nonthaburi, Thailand, Retrieved April 13, 2023, from http://www.thaiheart.org/images/column_1387023976/NHES5_EGAT Meeting13Dec13.pdf
World Health Organization. (2014). Noncommunicable disease country profiles, Retriever April 13, 2023, from https://www.who.int/news-room/fact-sheets/detail/diabetes