Correlation analysis of laboratory blood tests and complications in Diabetes Mellitus using data mining technique

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

[1]Faculty of Medical Technology, Huachiew Chalermprakiet University

[2]Division of Medical Laboratory and clinical pathology, Chomthong Hospital

[3]Faculty of Science & Technology, Huachiew Chalermprakiet University

*Email : r_panthip@hotmail.com

Abstract

The laboratory blood tests in this study were collected from patients diagnosed with diabetes mellitus (DM) who were admitted to a public hospital in Chiang Mai Province. The aim of this study is to examine the correlation between laboratory blood tests and the incidence of complications associated with DM using a data mining technique. The J48 classifier was applied to construct a decision tree model and verify the precision of the simulation model. The laboratory blood tests were collected from 1,736 patients diagnosed with DM who were admitted to a public hospital in Chiang Mai Province, Thailand in 2020. The results showed that among the total DM cases, 54.55% were diagnosed with chronic kidney disease stage 2 (CKD 2) or higher, consistent with the microalbuminuria category where early and advanced kidney damage was observed in 66.82% of total DM cases. Interestingly, clinical factors such as BMI, eGFR and microalbuminuria are related to DM complications, particularly with respect to kidney dysfunctions. The decision tree model, simulated with the J48 classifier, achieves a high predictive accuracy with a correct classified instance rate of 87.44%. Verification parameters were used to validate the quality of the model in each class of chronic kidney dysfunction which classifier exhibited the high %True positive rates of more than 70% in all targeted CKD classes (CKD 1-5) and with Precision of more than 80% indicating low %False positive rate. These findings highlight the advantages of clinical data analysis using data mining techniques.

Keywords : Diabetes mellitus, Hemoglobin A1c, Data mining, Decision tree, J48 classifier

1. Introduction

Information from the 5-year National NCDs Prevention and Control Strategic Plan (2017-2021) reveals that Thai people have died from Non-communicable diseases (NCDs) including cerebrovascular disease, diabetes mellitus (DM), ischemic heart disease, and diseases related to chronic obstructive pulmonary disease (COPD), accounting for 320,000 people (~ 75% of the total death population) (Department of Disease Control, Ministry of Public Health, 2021, p.148). The situation of diabetes in Thailand, according to the National Health Examination Survey (NHES) plan, indicates that the prevalence of diabetes among individuals aged 15 years and over has elevated from 6.9% in 2009 to 8.9% in 2014 (National Health Examination Survey Office, Health System Research Institute, 2016, p. 6). The WHO report in 2014 found that there were 4.8 million Thai adults with DM resulting in 76,000 deaths from diabetes-related causes, which is equivalent to over 200 people per day (World Health Organization report, 2014). DM is associated with various factors such as age, sex, and genetics, as well as dietary habits that are inconsistent with a healthy diet. The changing social conditions and hectic lifestyle often lead to rushed intake resulting in an imbalance in daily nutritional intake. Improper food consumption behavior and excessive consumption of high-sugar beverages are major contributors to obesity, ultimately leading to DM type 2 diabetes. DM is a serious public health problem in Thailand due to its association with unusually high blood sugar levels, which can lead to serious complications such as diabetic retinopathy, kidney complications, chronic kidney disease, and complications of large blood vessels (Chanlalit W, 2016, p. 38-39), (Sonthon P. *et al*., 2017, p. 9).

In order to establish guidelines for diagnosis and treatment, it is important to understand the relationship among various factors of the disease. It leads to more effective treatment manners. The currently used technique for data correlation is data mining. In this study, we analyzed the correlation between blood test results in patients with diabetes mellitus (DM) and the complications linked to DM using a supervised learning technique with a classifier approach to construct a decision tree model.

2. 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.

3. Materials and methods

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.

The ascertainment of diagnosis for diabetes for each patient is based on FBS level ≥ 126 mg/dL which associated with the HbA1c level of ≥ 6.5 mg%. HbA1C is a widely used marker of chronic glycemia, reflecting average blood glucose levels over a period of approximately 2 to 3 months (Bishop, 2018, p.764).

According to the American Heart Association, there are five range of BP: Normal (<120/80), Elevated (120-129 systolic and less than 80 mmHg diastolic), Hypertension stage 1 (130-139 systolic or 80-89 mmHg diastolic), Hypertension stage 2 (140/90 mmHg or higher), Hypertensive crisis (exceed 180/120 mmHg) (American Heart Association, 2023).

MAU is defined as persistent albuminuria in the range of 30-299 mg/24h or an albumin-creatinine ratio (ACR) of 30-300 □g/mg. Clinical proteinuria or macroalbuminuria is established with a microalbuminuria of ≥300 mg/24h or an ACR of ≥300 □g/mg. A healthy kidney does not let albumin pass into the urine or allows for less than 30 □g/mg (Bishop, 2010, p.326).

BMI is a measure of body fat based on height and weight. It is calculated as weight in kilograms divided by height in meters (kg/m^2). BMI is classified into four groups according to the Asian-Pacific cutoff points: underweight (<18.5 kg/m^2), normal weight (18.5-22.9 kg/m^2), overweight (23-24.9 kg/m^2), and obese (≥30 kg/m^2) (Lim JU, 2017, p.2466).

Glomerular filtration rate (GFR) or eGFR serves as the most comprehensive indicator of kidney function. Chronic kidney disease (CKD) is categorized into five stages based on the extent of kidney function, including CKD stage 1, CKD stage 2, CKD stage 3a, CKD stage 3b, CKD stage 4, and CKD stage 5 as shown in Table 1 (NICE, 2014, p.6).

Table 1 Categories for Glomerular Filtration Rate (GFR) as per kidney disease

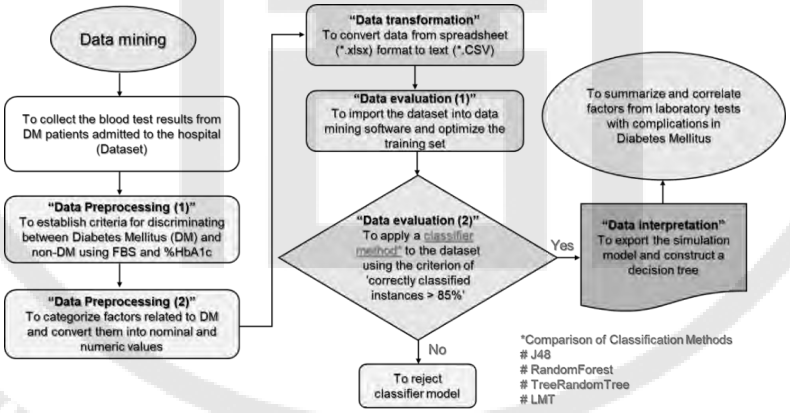
GFR category (CKD stage)	**GFR ($ml/min/1.73m^2$)**	**Terms**
CKD stage 1	90 or higher	Normal or high
CKD stage 2	60-89	Mildly decreased
CKD 3a	45-59	Mildly to moderately decreased
CKD 3b	30-44	Moderately to severely decreased
CKD 4	15-29	Severely decreased
CKD 5	Less than 15	Kidney failure

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.



4. Results

Overview of laboratory blood test in DM patients

The dataset was finally preprocessed with a total amount of 1,736 cases from DM patients. The age of DM patients ranged from 24 to 93 years old (Figure 2 depicts the distribution of age among DM patients). Patients between 24 and 59 years old accounted for 40.55% of the dataset, while patients aged 60 years or older accounted for 59.45%. The selected clinical data used to examine the correlation associated with DM complications consisted of BMI, BP, FBS, Cr, TC, TG, HbA1c, eGFR, and MAU. The criteria for identifying DM patients were based on an HbA1c level greater than 6.5 mg%.

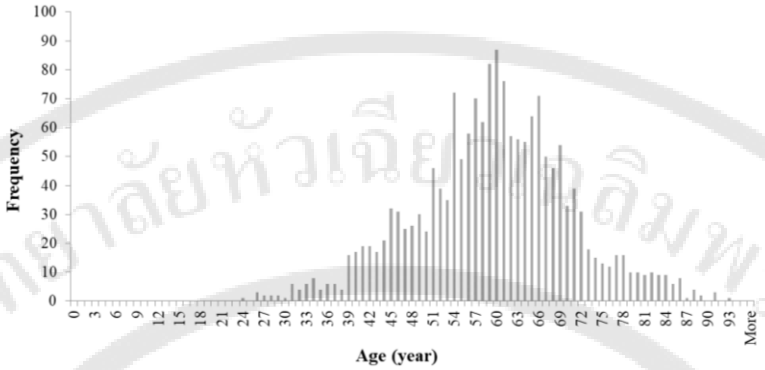


Figure 2 Histogram depicting the distribution of ages among diabetes patients that were admitted to a public hospital in Chiang Mai province.

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.

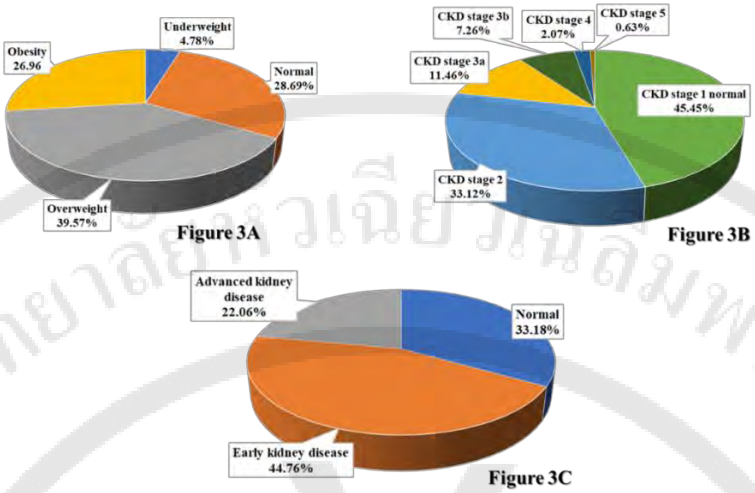


Figure 3 Pie chart depicts the complications in DM patients such as Figure 3A (BMI), Figure 3B (eGFR), and Figure 3C (MAU) categories.

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% (Table 2). 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 a perfect accuracy of 100.00% in correctly classified instances (Table 3). 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.

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 (Figure 4). 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 (as shown in Table 4). Notably, the J48 pruned

classifier model has 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.

Table 2 The comparison of correctly and incorrectly classified instances from 10-fold cross-validation of "training set" in each classification approach: J48, RandomForest, TreeRandomTree, and LMT.

*Classifier model	%Correctly classified instances	%Incorrectly classified instances
J48	78.40	21.06
RandomForest	79.49	20.51
TreeRandomTree	74.42	25.58
LMT	79.09	20.91

* The number of correctly and incorrectly classified instances is calculated from a total dataset size of 1,726 instances

Table 3 The percentage of correctly and incorrectly classified instances using the "testing set" in each classification approach: J48, RandomForest, TreeRandomTree, and LMT.

*Classifier model	%Correctly classified instances	%Incorrectly classified instances
J48	87.60	12.40
RandomForest	100.00	0.00
LMT	81.99	18.01

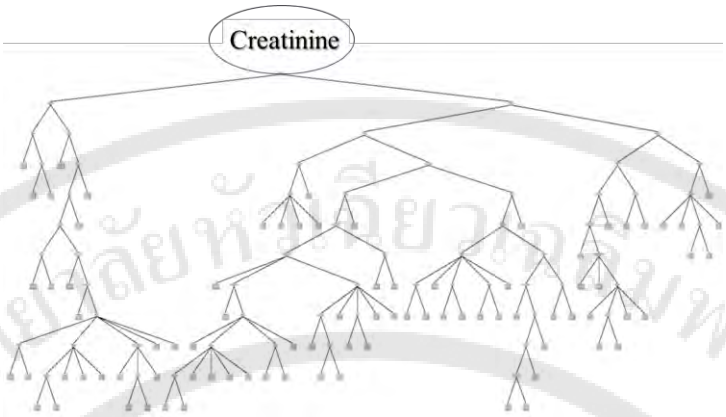


Figure 4 The simulation of decision tree from J48 classifier model composed of clinical factors such as Cr (defined as first root node), eGFR, and MAU in which associated with DM complications.

Table 4 Calculated parameters from J48 classifier model in each CKD class

Classifier output (Testing set)		**Total classified instance**	
Correctly classified instances		87.60%	
Incorrectly classified instances		12.40%	
Detailed accuracy by class			
Class	%TPR**	%FPR[#]	%Precision[$]
CKD stage 1 (Normal)	91.60	4.30	93.60
CKD stage 2	91.40	10.40	82.90
CKD stage 3a	70.80	2.40	81.70
CKD stage 3b	75.40	8.00	88.10
CKD stage 4	90.50	3.00	86.40
CKD stage 5	100.00	0.00	100.00

**%True positive rate (TPR) is calculated in percentage and defined as the probability that an actual positive will test positive in particularly targeted class.

[#] %False positive rate (FPR) is calculated in percentage and defined as the ratio between the total number of wrongly categorized as positive (called false positive) and total number of actual negative events in particularly targeted class.

[$]%Precision is the measurement of accuracy in positive prediction represented by the number of true positive predictions divided by the number of true positive predictions plus false positive predictions which is calculated in percentage.

5. 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.

6. Conclusion

Diabetes mellitus and its complications have become a public health problem, and current therapeutic policies need to be improved. Managing and treating diabetes mellitus is a challenge for researchers and healthcare personnel. The management and treatment of diabetes mellitus poses a challenge to researchers and healthcare personnel. This research strongly suggests the association of diabetes and CKD, BP, and BMI. Knowing about diabetes-related conditions such as kidney disease, high blood pressure, and high BMI (obesity) is important to educate diabetics about and increase awareness of related risks, such as weight control. Proper control of calorie

intake and exercise can help in managing these risks. It is also important to control sodium levels in the diet.

7. 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.

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/guldance/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

8. Author(s) Biodata (50 words)

Panthip Rattanasinganchan, Ph.D. (Biochemistry)

Academic rank: Assistant Professor

Faculty of Medical Technology

Huachiew Chalermprakiet University

Education: Ph.D. in Biochemistry from the Department of Biochemistry, Faculty of Science, Mahidol University, Thailand

Research Experience:

Antioxidants, Cancer Biology, Signaling Pathways, Data mining.

Research Assistant at Proteomics Core Facility, NIH, Maryland, USA