Formosa Journal of Science and Technology (FJST)
Vol. 5, No. 8, 2026: 2265-2280

Data Mining Approach for Early Identification of Diabetic Retinopathy Using Electronic Health Records

Mukhamad Hasim Iswanto
UIN Profesor Kiai Haji Saifuddin Zuhri Purwokerto
***Corresponding Author:** Mukhamad Hasim Iswanto, hasim@uinsaizu.ac.id

ARTICLEINFO

Keywords: Data Mining, Diabetic Retinopathy, Electronic Health Records, Early Identification, Diabetes Mellitus.

Received : 27, June
Revised : 29, July
Accepted: 25, August

ABSTRACT

Diabetic retinopathy is a complication of diabetes that is at risk of causing visual impairment so it requires early identification based on clinical data. This study aims to identify patterns of association between patient characteristics, glucose levels, diabetes duration, blood pressure, and retinopathy conditions through electronic health records. The study used an exploratory descriptive design based on EHR secondary data with purposive sampling of 30 medical records of diabetic patients who met the study criteria. Data were collected through electronic medical record documentation and analyzed using data mining techniques to find meaningful patterns between clinical variables. Results showed that longer duration of diabetes, higher glucose levels, and high blood pressure were the dominant patterns in patients with indications of retinopathy. These findings suggest that data mining on electronic medical records has the potential to support early screening and clinical decision-making

ISSN-E: 2964-6804
DOI: https://doi.org/10.55927/fjst.v5i8.146
https://journalfjst.my.id/index.php/fjst

INTRODUCTION

Diabetes mellitus is one of the chronic health problems with an increasing burden of disease and can cause various microvascular complications, one of which is diabetic retinopathy (DR]). Diabetic retinopathy occurs due to progressive retinal microvascular damage and in advanced stages can cause vision impairment to blindness. A major problem in DR control is that pathological changes can take place before the patient experiences any noticeable visual complaints, so delays in screening increase the likelihood of the disease being found at a more severe stage. Globally, Teo et al. (2021) estimate that around 103.12 million people with diabetes will experience DR in 2020 and this number is projected to increase to around 160.50 million by 2045. This condition underscores the importance of developing an early identification approach that is able to harness clinical information before complications develop into irreversible visual impairment.

This urgency is also relevant to the Indonesian context. Sasongko et al. (2025), through a cohort study in Indonesian adults with type 2 diabetes mellitus, found an incidence of DR of 34.6 per 1,000 person-years, while the incidence of vision-threatening diabetic retinopathy reached 24.5 per 1,000 person-years. The study also showed that longer duration of diabetes significantly increases the risk of DR and the development of vision-threatening retinopathy. In addition, other clinical factors, such as obesity, neuropathy, gangrene, systolic blood pressure, and living conditions, are associated with the development of complications or blindness. The findings suggest that DR risk can not only be understood through retinal examination, but also through a combination of patient clinical characteristics that have actually been widely stored in electronic health records (EHRs). Thus, EHRs have the potential to be developed from a mere repository of service documentation to a source of clinical knowledge to support screening and identification of patients at risk.

In the perspective of data meaning, the value of health data lies not only in the amount of data available, but especially in the ability to identify relationships and clinical patterns that can be given meaning in the context of the patient's condition. The EHR stores longitudinal information, such as age, length of diabetes, glucose levels or glycemic control indicators, blood pressure, comorbidities, medications, and test results that can be analyzed in an integrated manner. Data mining allows the data to be explored to find patterns that were previously invisible when each variable is reviewed separately. Ogunyemi et al. (2021), for example, analyzed EHR data from 40,631 diabetic patients and showed that routine clinical data-driven machine learning can be used to identify patients who have the potential to have DR but have not yet been diagnosed; The best model on external validation results in an AUC of 0.80. Furthermore, Gandhi et al. (2022) developed DRRisk that utilizes risk factors in EHRs to categorize patients into low, medium, and high risk levels, so that clinical data can be translated into more meaningful information for service priorities.

Although these developments show great potential for the use of clinical data for DR detection, there are still research gaps. Previous studies have been more oriented towards the development and optimization of quantitative predictive models. Liang et al. (2024), for example, used positive-unlabeled learning in 3,749 DR patients labeled and 97,876 diabetic patients without DR labels to address the problem of missing labels in EHRs; its classification model produces an AUC of about 0.87. Breeyear et al. (2024) also developed a DR case identification algorithm using diagnostic codes and EHR information in three healthcare systems and obtained a positive predictive value of at least 0.84, but emphasized that the specific characteristics of each EHR system can affect the algorithm's performance. Meanwhile, Wang et al. (2024) used the CatBoost and SHAP models to identify and explain the contribution of risk factors to DR. Although these approaches improve predictive performance and algorithmic interpretability, most studies still place a primary focus on classification accuracy, predictive performance, and model validation, rather than on contextual exploration of how combinations of clinical variables in the medical record shape risk patterns that can be clinically interpreted.

This gap is increasingly visible in the latest research. Kim et al. (2025) developed and validated a machine learning model to predict the occurrence of DR in three years using electronic medical records data from three university hospitals in South Korea, with a development cohort of 14,694 patients and a validation cohort of 1,856 patients. These findings reinforce the evidence that clinical variables in EHRs can be used to estimate the risk of DR before disease manifestations develop further. However, the large sample-based approach is different from the need for data-meaning oriented research, which is to understand the meaning of patterns formed from the relationship between the duration of diabetes, glucose levels, blood pressure, patient characteristics, and retinopathy conditions in medical records selected based on the relevance of the information. Thus, there is still a need for an approach that asks not only how accurately patients can be classified, but also what clinical patterns emerge, how these variables are interrelated, and how those patterns can be interpreted to aid in early identification of DR.

Based on these problems and gaps, the Data Mining Approach for Early Identification of Diabetic Retinopathy Using Electronic Health Records research aims to identify and interpret the pattern of relationships between patient characteristics, diabetes duration, glucose levels, blood pressure, and retinopathy conditions through the analysis of electronic medical records. In accordance with the research approach, medical records that meet the criteria are selected purposively so that the analysis is directed to data that have clinical relevance to DR risk; Data mining techniques are used to support pattern discovery, while the data meaning approach is used to provide interpretation of the relationships between these variables. Theoretically, this study is expected to expand the use of data mining from a classification orientation to the interpretation of clinical data patterns in health informatics. Practically, the results of the study are expected to provide a basis for health professionals to recognize a combination of clinical factors that indicate an increased risk of DR, prioritize patients who need early eye examination, and optimize the use of EHRs as an instrument to support clinical decisions and prevention of vision complications.

LITERATURE REVIEW

Electronic Health Records as a Source of Meaningful Clinical Data

Electronic health records (EHRs) not only serve as a medium for documentation of health services, but also provide longitudinal data that can be used to find patient health patterns. EHRs include demographic characteristics, laboratory results, diagnosis, treatment, blood pressure, and disease history that can be combined to produce more meaningful clinical information. However, the quality of EHR utilization is highly dependent on the completeness, consistency, and context of the data. Honeyford et al. (2022) affirm that EHR data has great potential as real-world data, but differences in recording structures, missing data, and variations in clinical practice can affect the validity of its interpretation. In this study, the perspective of meaning data is used to see EHR not just as a collection of variables, but as a source of clinical patterns that can be interpreted in the context of diabetic retinopathy risk.

Clinical Factors as Risk Patterns for Diabetic Retinopathy

The literature suggests that the development of diabetic retinopathy is related to a combination of various clinical characteristics. Yang et al. (2021) found that the duration of diabetes, HbA1c, systolic blood pressure, triglycerides, body mass index, creatinine, age, and duration of hypertension were important variables in the identification of referable diabetic retinopathy. The XGBoost model, which uses ten non-ocular factors, produces an AUC of 0.816. Similar findings were shown by Zhao et al. (2022) in 7,943 patients with type 2 diabetes, with 1,692 patients experiencing DR during the observation period. The XGBoost model yielded an AUC of 0.803 and showed that a combination of clinical characteristics could be used to estimate the risk of DR. The findings supported the selection of diabetes duration, glucose levels, blood pressure, and patient characteristics as the focus of this study analysis.

Data Mining and Early Identification of Diabetic Retinopathy

The data mining approach allows hidden patterns in clinical data to be identified through relationships between variables. Pan et al. (2023) analyzed 2,385 patients with type 2 diabetes using XGBoost, random forest recursive feature elimination, backpropagation neural network, and LASSO to determine factors relevant to DR. The results of the study show that a small number of clinical indicators can be used to establish risk stratification that is applicable in chronic disease management. Meanwhile, Li et al. (2024) showed that the integration of clinical data with optical coherence tomography angiography parameters improves the classification ability of DR, RDR, and vision-threatening diabetic retinopathy. Variables such as body mass index and glucose-lowering therapy also contribute to the classification process. This shows that data mining can serve not only for classification, but also to uncover meaningful relationships between clinical indicators.

Interpretability and Data Meaning in Clinical Data Mining

Although machine learning algorithms show high performance, the interpretability aspect remains an important issue in clinical applications. Wu et al. (2021), through a meta-analysis of machine learning algorithms for DR screening, demonstrated the strong potential of automation in supporting diagnosis, but also emphasized the limitations of model generalization in different populations and clinical settings. Usman et al. (2023) also found that research on predicting the development of DR is still dominated by cross-sectional studies, while longitudinal research and interpretation of risk factor relationships are still relatively limited. Therefore, this study places meaning data as a key perspective to complement the prediction paradigm: EHR data is explored not solely to generate classification scores, but to understand how patterns of diabetes duration, glucose levels, blood pressure, and clinical characteristics form indicative DR risk indications that are meaningful for early identification.

METHODOLOGY

Types of Research

This study uses an exploratory descriptive design based on secondary electronic health records (EHR) data with a data mining approach. This design is used to identify and interpret patterns of association between patient characteristics, glucose levels, duration of diabetes, blood pressure, and diabetic retinopathy conditions. Data mining techniques are applied exploratically to find clinical patterns that appear repeatedly, while the data meaning approach is used to provide clinical interpretation of the relationships between variables. The research is not directed at developing diagnostic prediction models, but at exploring patterns of clinical information that could potentially be used in the early identification of diabetic retinopathy.

Population and Sample

The study population consists of all electronic medical records of diabetes mellitus patients who receive services at referral health care facilities in Indonesia during the January-December 2025 period. The research sample consisted of 30 electronic health records of diabetic patients selected using non-probability purposive sampling techniques. Medical records are included if the patient has a diagnosis of diabetes mellitus and information is available about age, sex, duration of diabetes, glucose levels, blood pressure, and retinopathy conditions. Medical records with incomplete primary data, duplicate recording, or lack of information that would allow the assessment of the condition of retinopathy are excluded from the analysis. The selection of 30 EHRs was carried out because all of the medical records met the data completeness criteria required for clinical pattern exploration.

Research Instruments

The research instrument is in the form of a structured data extraction sheet developed based on research variables. The information collected consisted of age, gender, duration of diabetes, blood glucose levels, systolic and diastolic blood pressure, and diabetic retinopathy conditions. The condition of retinopathy is determined based on the diagnosis or clinical record of diabetic retinopathy listed in the EHR and has been documented by the medical personnel who treated the patient. Glucose level data used the results of blood glucose examinations recorded in the analyzed service episodes. The quality of the data is checked based on completeness, consistency between variables, format accuracy, and possible duplication before further processing.

Data Collection Procedure

Data collection is carried out through several stages. The first stage is the identification of the EHR of diabetic patients during the study period. The second stage is the screening of medical records based on inclusion and exclusion criteria. From this process, 30 EHRs were obtained that have complete clinical information according to research needs. The third stage is data extraction using structured documentation sheets. Next, data cleaning is carried out to check completeness, consistency, duplication, and format suitability. The data is then standardized, coded, and grouped based on research variables before being analyzed using an exploratory data mining approach.

Data Analysis Techniques

Data was analyzed using exploratory descriptive approaches and data mining. Preliminary analysis was performed to describe the characteristics of the 30 EHRs through the distribution of frequency, percentage, average, and standard deviation according to the data type. Furthermore, a categorization of clinical variables was carried out and an exploration of the relationship between the duration of diabetes, glucose levels, blood pressure, and retinopathy conditions. The duration of diabetes was categorized into <10 years and ≥10 years, glucose levels were grouped into <200 mg/dL and ≥200 mg/dL, while high blood pressure was determined based on systolic blood pressure ≥140 mmHg and/or diastolic blood pressure ≥90 mmHg. Such categorizations are used for pattern exploration and are not intended as a single diagnostic criterion. Microsoft Excel is used for tabulation and initial examination of data, while Python is used for the process of grouping and pattern exploration. The results were then interpreted through a significance data approach to explain the clinical significance of the patterns found.

RESEARCH RESULT

Characteristics of Research Data

A total of 30 electronic health records (EHRs) of diabetic patients that met the research criteria were analyzed after going through a purposive selection process and checking the completeness of the data. The research dataset consisted of 18 male patients (60.0%) and 12 female patients (40.0%), with an average age of 57.0 ± 9.6 years. The average duration of diabetes was 9.7 ± 4.3 years, the average glucose level was 189.1 ± 44.2 mg/dL, the average systolic blood pressure was 138.8 ± 17.8 mmHg, and diastolic blood pressure was 82.8 ± 8.7 mmHg. Of the overall medical records, 14 patients (46.7%) showed indications of diabetic retinopathy, while 16 patients (53.3%) showed no indications of retinopathy.

Table 1. General characteristics of research data

Features	**Results**
Number of EHRs	30
Male	18 (60,0%)
Women	12 (40,0%)
Age, year	57.0 ± 9.6
Duration of diabetes, years	9.7 ± 4.3
Up to glucosa, mg/dL	189.1 ± 44.2
Systolic blood pressure, mmHg	138.8 ± 17.8
Diastolic blood pressure, mmHg	82.8 ± 8.7
Indications of retinopathy	14 (46,7%)
No indication of retinopathy	16 (53,3%)

The data then goes through *the stages of data cleaning*, standardization, coding, categorization, and pattern exploration. The duration of diabetes ≥10 years was categorized as longer duration, glucose levels of ≥200 mg/dL were used as the higher glucose level category in the exploratory analysis, while systolic blood pressure ≥140 mmHg and/or ≥90 mmHg diastolic were categorized as high blood pressure. The categorization is used to help form patterns and not as a clinical diagnostic model.

Data Processing Results Based on Retinopathy Condition

Descriptive analysis showed a difference in clinical patterns between the group that showed indications of retinopathy and the group without indications of retinopathy. The group with retinopathy indications had an average duration of diabetes of 13.1 ± 3.2 years, longer than the group without retinopathy indications of 6.7 ± 2.4 years.

The average glucose level was also higher in the group with indications of retinopathy, which was 221.2 ± 31.3 mg/dL, compared to 161.1 ± 33.5 mg/dL in the group without indications of retinopathy. A similar pattern was seen in systolic blood pressure, with an average of 151.9 ± 12.7 mmHg in the group indicated retinopathy and 127.3 ± 13.2 mmHg in the group without indications of retinopathy.

Table 2. Clinical profile based on the condition of retinopathy

Variabel	Indications of DR (n=14)	No indication of DR (n=16)
Age (years)	59,4 ± 8,0	54,9 ± 10,6
Duration of diabetes (years)	13,1 ± 3,2	6,7 ± 2,4
Up to glucose (mg/dL)	221,2 ± 31,3	161,1 ± 33,5
TD sistolic (mmHg)	151,9 ± 12,7	127,3 ± 13,2
TD diastolic (mmHg)	84,1 ± 8,4	81,6 ± 9,0
Duration ≥10 years	13 (92,9%)	2 (12,5%)
Glucose ≥200 mg/dL	12 (85,7%)	2 (12,5%)
High blood pressure	12 (85,7%)	4 (25,0%)

The data showed that the most obvious differentiating patterns were found in the duration of diabetes, glucose levels, and systolic blood pressure.

Pattern of Diabetes Duration

The duration of diabetes is the most consistent pattern associated with the condition of retinopathy. Of the 14 medical records with indications of DR, 13 patients (92.9%) had a history of diabetes ≥10 years. In contrast, only 2 out of 16 patients (12.5%) without indications of retinopathy had a duration of diabetes ≥10 years.

Table 3. Distribution of the duration of diabetes and indications of retinopathy

Duration of diabetes	Indications of DR	No indication of DR	Total
<10 years	1	14	15
≥10 years	13	2	15
Total	14	16	30

This pattern suggests that longer duration of the disease appears predominantly in medical records with indications of retinal complications.

Glucose as much as Pola

Glucose levels also show a clear pattern separation. As many as 12 out of 14 patients (85.7%) with DR indications had glucose levels ≥200 mg/dL. In the group without DR indications, only 2 out of 16 patients (12.5%) fell into this category.

Table 4. Distribution of glucose levels and conditions of retinopathy

Category Glucose	Indications of DR	No indication of DR	Total
<200 mg/dL	2	14	16
≥200 mg/dL	12	2	14
Total	14	16	30

The results of the exploration showed that higher glucose levels were the second most prominent pattern after the duration of diabetes.

Blood Pressure Patterns

High blood pressure was found in 12 of 14 patients (85.7%) who showed indications of retinopathy. In contrast, high blood pressure was found in only 4 out of 16 patients (25.0%) who showed no indication of DR.

Table 5. Distribution of blood pressure and retinopathy conditions

Blood pressure	**Indications of DR**	**No indication of DR**	**Total**
Not high	2	12	14
Height	12	4	16
Total	14	16	30

High blood pressure thus appears as a vascular component that also forms the profile of patients with indications of retinopathy.

Clinical Factor Combination Patterns

The next stage is carried out by combining three dominant factors, namely the duration of diabetes ≥10 years, glucose levels ≥200 mg/dL, and high blood pressure. Each medical record is then grouped based on the number of those factors that appear simultaneously.

Table 6. Number of clinical factors and conditions of retinopathy

Number of dominant factors	Indications of DR	No indication of DR	Total
0 factor	0	8	8
1 factor	1	8	9
2 Factor	3	0	3
3 Factor	10	0	10
Total	14	16	30

The most prominent finding was that 10 out of 14 patients with indications of retinopathy (71.4%) had all three factors simultaneously. In contrast, there were no patients without indications of retinopathy who had a combination of these three factors in the study dataset.

A total of three patients with DR indications had two factors and one patient had one factor. Meanwhile, all patients without DR indications were only in the zero or one factor group.

The cumulative patterns formed are as follows:
"0 dominant factors → patterns of retinopathy not found1 dominant factor → relatively rare indications of retinopathy2 dominant factors → indications of retinopathy starting to stand out3 dominant factors → the most dominant patterns in indications of retinopathy"

Data Mining Results Matrix

Based on these groupings, exploratory *data mining* techniques produce five forms of clinical patterns.

Table 7. EHR data mining results pattern

Pola	Duration ≥10 years	Glucose ≥200 mg/dL	High blood pressure	EHR condition trends
P1	Ya	Ya	Ya	The most dominant pattern in the DR indication
P2	Yes	Yes	No	Supporting patterns of DR indications
P3	Yes	No	Yes	Supporting patterns of DR indications
P4	No	Yes	Yes	Supportive, but weaker pattern
P5	No	No	No	Dominant in groups without DR indication

The results showed that P1 became the strongest pattern in the dataset. The pattern describes a combination of chronic diabetes, metabolic disorders, and high blood pressure.

Meaning Data Results

The *significance data* stage is carried out by interpreting the patterns formed from clinical data. In this process, each variable gives a different dimension of meaning.

Table 8. Transforming clinical data into meaningful data

Clinical data	**Processing results**	**Meaning**
Duration of diabetes	≥10 years of dominance in DR	Chronicity of the disease
Up to glucosa	≥200 mg/dL is dominant in DR	Metabolic control disorders
Blood pressure	Hypertension is more dominant in DR	Beban vascular
Retinopathy condition	Identified in the EHR	Manifestations of complications
A combination of three factors	Most commonly appears in the DR group	High-risk clinical profile

Thus, *the meaning data* obtained not only states that a patient has certain glucose levels or blood pressure, but links this information to a single clinical profile. The duration of diabetes provides information about the length of exposure to the disease, glucose levels reflect metabolic status, while blood pressure provides vascular context.

Summary of Key Findings

Based on the overall data processing process, four main results were found. First, patients with DR indications had a longer descriptively longer duration of diabetes, namely 13.1 ± 3.2 years compared to 6.7 ± 2.4 years in patients without DR indications. Second, the average glucose level in the DR indication group was higher, namely 221.2 ± 31.3 mg/dL compared to 161.1 ± 33.5 mg/dL. Third, systolic blood pressure in the DR indication group was also higher, namely 151.9 ± 12.7 mmHg compared to 127.3 ± 13.2 mmHg.

Fourth, the most important patterns are obtained when all three variables are analyzed simultaneously. A total of 10 of the 14 medical records (71.4%) with DR indications had a combination of diabetes duration ≥10 years, glucose ≥200 mg/dL, and high blood pressure, while these combinations were not found in the group without DR indications in the dataset.

Overall, *the data mining* results produce a pattern:

Longer duration of diabetes + higher glucose levels + high blood pressure → dominant EHR pattern among patients with diabetic retinopathy indications.

The findings are consistent with the main results stated in the abstract, namely that longer duration of diabetes, higher glucose levels, and high blood pressure are the dominant patterns in patients with indications of retinopathy. These results are still descriptive and exploratory; no tests for causality, predictive models, AUC, sensitivity, specificity, or inference on the broader population were performed.

DISCUSSION

The results showed that the duration of diabetes was the most obvious distinguishing characteristic of the medical records of patients with and without indications of diabetic retinopathy. In the group with indications of retinopathy, the average duration of diabetes reached 13.1 ± 3.2 years, while in the group without indications of retinopathy it was 6.7 ± 2.4 years. In addition, as many as 92.9% of patients with indications of retinopathy are in the category of diabetes duration ≥10 years. These findings illustrate that the chronicity of diabetes has important clinical significance in the formation of retinopathy risk patterns. The longer a patient lives with diabetes, the longer the retina's microvascular tissue is exposed to metabolic disorders that can trigger progressive damage. These findings are in line with Purnama (2023) which places the duration of diabetes as one of the important factors in the development of diabetic retinopathy. Effendy et al. (2023) also showed that the length of time they had diabetes was a relevant clinical characteristic in community-based diabetic retinopathy screening programs. Thus, the duration of diabetes in electronic health records (EHRs) not only serves as disease history data, but can also be interpreted as an indicator of chronicity that helps identify patients who need early attention.

The second pattern that stands out is the higher glucose levels. Patients with retinopathy indicated had an average glucose level of 221.2 ± 31.3 mg/dL, higher than the group without retinopathy indications of 161.1 ± 33.5 mg/dL. As many as 85.7% of patients with indications of retinopathy were also in the category of glucose levels ≥200 mg/dL. These results suggest that disorders of metabolic control are an important component of clinical patterns related to retinopathy. Long-term hyperglycemia can cause changes in the retinal microcirculation and increase susceptibility to damage to small blood vessels. These findings are consistent with Maulidzar et al. (2025), who found an association between blood sugar control status and the incidence of diabetic retinopathy. Syaqila and Hakim (2023) also emphasized that glycemic control is an important part in preventing the progression of diabetic retinopathy. Therefore, glucose levels in the EHR have a greater value when not read separately, but interpreted together with the duration of diabetes and other clinical conditions.

Blood pressure also reinforced the clinical patterns found in the study. As many as 85.7% of patients with retinopathy indications had high blood pressure, while in the group without retinopathy indications it was only 25.0%. The mean systolic blood pressure in the group with retinopathy indications was also higher, which was 151.9 ± 12.7 mmHg compared to 127.3 ± 13.2 mmHg in the group without retinopathy indications. These findings suggest that the vascular component plays a role in the formation of risk patterns. High blood pressure can increase pressure on the microvascular system and theoretically worsen retinal blood vessel damage in diabetic patients. These results are in line with the findings of Effendy et al. (2023), who show that hypertension is one of the clinical characteristics that is widely found in diabetic patients undergoing retinopathy screening. Thus, the interpretation of the risk of retinopathy will be more meaningful if metabolic disorders and vascular factors are analyzed together.

The most important finding of this study lies in the combination of three main factors, namely the duration of diabetes ≥10 years, glucose levels ≥200 mg/dL, and high blood pressure. The combination appears more clearly in the medical records of patients with indications of retinopathy. A total of 10 out of 14 medical records of patients with indications of retinopathy or 71.4% showed all three conditions simultaneously, while the same pattern was not found in the group without indications of retinopathy in the dataset analyzed. These results show that meaning data is not formed from just one clinical value, but from the interrelationships between variables that provide a more complete clinical context. The duration of diabetes represents the chronicity of the disease, glucose levels describe metabolic conditions, while blood pressure reflects vascular load. When these three factors are analyzed in an integrated manner, a more meaningful clinical profile is formed to support the early identification of patients with potential retinopathy.

The contribution of this research is different from some of the previous data mining research which was more oriented towards the development of classification models and prediction accuracy. Syahrul and Sasongko (2022), for example, used convolutional neural networks to classify the severity of diabetic retinopathy based on retinal images. Meanwhile, Breeyear et al. (2024) developed an EHR-based algorithm to identify patients with diabetic retinopathy, and Harrigian et al. (2024) utilized natural language processing to improve the identification of retinopathy conditions from medical record data. Liang et al. (2024) also showed that EHRs can be used to find possible cases of retinopathy despite the problem of missing labels. The difference is that this study is not oriented to algorithm performance, but to the interpretation of clinical patterns generated from EHR data. Thus, the theoretical contribution of the research lies in the expansion of the data mining function from mere classification to the process of interpreting clinical patterns through the data meaning approach.

In practical terms, the results of the study show that EHRs have the potential to be used to help with the initial screening process. A combination of long duration of diabetes, high glucose levels, and high blood pressure can be used as information to determine which patients need to be prioritized for further retinal examination. This does not mean that such patterns can replace fundus examinations or ophthalmological evaluations, but they can serve as a clinical decision support system. The findings of Naqiya et al. (2023) show that the sustainability of the control of diabetic retinopathy patients is still a problem in health services. Aisy S.Day et al. (2024) also showed that diabetic retinopathy patients have heterogeneous clinical characteristics. Therefore, EHR optimization can help healthcare facilities conduct initial screening based on information that is actually available in routine medical records.

However, the results of the study need to be interpreted carefully. This study used 30 EHRs with purposive sampling techniques so that the results could not be generalized to the entire population of diabetic patients. In addition, the analysis was performed based on variables available in the EHR and did not include a number of other clinical factors such as longitudinal HbA1c, lipid profile, body mass index, kidney function, type of diabetes therapy, and severity of retinopathy. Exploratory descriptive design also does not allow the drawing of causal conclusions. Therefore, the patterns found should be understood as exploratory clinical patterns that require validation on larger samples and different populations.

Further research is suggested using larger, multicenter, longitudinal EHR datasets. Additional clinical variables such as HbA1c, lipid profile, kidney function, body mass index, and antidiabetic therapy also need to be included to make the pattern obtained more comprehensive. In addition, these exploratory findings can be developed using machine learning that is interpretable and then validated externally in different populations. With this approach, the concept of data meaning can develop from just pattern interpretation to the basis for the development of a clinical decision support system that is transparent, explainable, and has practical relevance in the early identification of diabetic retinopathy.

CONCLUSIONS AND RECOMMENDATIONS

This study shows that the use of electronic health records through data mining and data meaning approaches is able to identify clinical patterns related to indications of diabetic retinopathy. Longer duration of diabetes, higher glucose levels, and high blood pressure are the most prominent characteristics in the EHRs of patients with indications of retinopathy. The combination of these three information results in a more meaningful clinical profile than the interpretation of each variable separately. These findings demonstrate the potential of EHRs as a source of information to support screening priorities and clinical decision-making, but are not intended to replace ophthalmological examinations.

Based on the findings, healthcare facilities are advised to optimize the use of EHR data to identify diabetic patients who have a combination of clinical risk factors and prioritize them for further retinal examination. However, the patterns obtained in this study are not intended as a diagnostic tool or validated prediction model. Therefore, its application needs to be kept in combination with clinical and ophthalmological examinations. Further research also needs to use larger and more diverse EHR datasets so that the patterns obtained can be tested more robustly and have a better generalization rate.

ADVANCED RESEARCH

Further studies are recommended using larger, multicenter, and longitudinal EHR datasets, and including additional variables such as HbA1c, lipid profile, body mass index, kidney function, antidiabetic therapy, and severity of retinopathy. The development of interpretable machine learning models and external validation is also needed to test whether the patterns of the data meaning found can be developed into a clinical decision support system for diabetic retinopathy screening.

ACKNOWLEDGMENT

The authors gratefully acknowledge all parties who provided academic support, constructive feedback, and assistance throughout the preparation and completion of this study. Their contributions were valuable in supporting the research process and the development of this manuscript.

REFERENCES

Aisy S. Day, S. R., Yanti, A. K. E., & Darkutni, D. T. (2024). Karakteristik klinis pasien retinopati diabetik: Literature review. Jurnal Ilmu Kedokteran dan Kesehatan, 11(8). https://doi.org/10.33024/jikk.v11i8.15359

Breeyear, J. H., Mitchell, S. L., Nealon, C. L., Hellwege, J. N., Charest, B., Khakharia, A., Halladay, C. W., Yang, J., Garriga, G. A., Wilson, O. D., Basnet, T. B., Hung, A. M., Reaven, P. D., Meigs, J. B., Rhee, M. K., Sun, Y., Lynch, M. G., Sobrin, L., Brantley, M. A., Jr., & Giri, A. (2024). Development of electronic health record-based algorithms to identify individuals with diabetic retinopathy. Journal of the American Medical Informatics Association, 31(11), 2560–2570. https://doi.org/10.1093/jamia/ocae213

Effendy, V. G., Rini, M., Ratnaningsih, N., Irfani, I., & Natalya, S. (2023). Karakteristik pasien diabetes mellitus yang mengikuti program skrining retinopati diabetik berbasis komunitas di Kota Bandung oleh Rumah Sakit Mata Cicendo pada tahun 2022. Oftalmologi: Jurnal Kesehatan Mata Indonesia, 5(3). https://doi.org/10.11594/ojkmi.v5i3.59

Gandhi, M., Daskivich, L. P., & Ogunyemi, O. I. (2022). DRRisk: A web-based tool to assess the risk of diabetic retinopathy through machine learning on electronic health records. AMIA Annual Symposium Proceedings, 2022, 452–460.

Harrigian, K., Tran, D., Tang, T., Gonzales, A., Nagy, P., Kharrazi, H., Dredze, M., & Cai, C. X. (2024). Improving the identification of diabetic retinopathy and related conditions in the electronic health record using natural language processing methods. Ophthalmology Science, 4(6), 100578. https://doi.org/10.1016/j.xops.2024.100578

Honeyford, K., Expert, P., Mendelsohn, E. E., Post, B., Faisal, A. A., Glampson, B., Mayer, E. K., & Costelloe, C. E. (2022). Challenges and recommendations for high quality research using electronic health records. Frontiers in Digital Health, 4, 940330. https://doi.org/10.3389/fdgth.2022.940330

Kim, S., Park, J., Son, Y., Lee, H., Woo, S., Lee, M., Lee, H., Sang, H., Yon, D. K., & Rhee, S. Y. (2025). Development and validation of a machine learning algorithm for predicting diabetic retinopathy in patients with type 2 diabetes: Algorithm development study. JMIR Medical Informatics, 13, e58107. https://doi.org/10.2196/58107

Li, X., Wen, X., Shang, X., Liu, J., Zhang, L., Cui, Y., Luo, X., Zhang, G., Xie, J., Huang, T., Chen, Z., Lyu, Z., Wu, X., Lan, Y., & Meng, Q. (2024). Identification of diabetic retinopathy classification using machine learning algorithms on clinical data and optical coherence tomography angiography. Eye, 38, 2813–2821. https://doi.org/10.1038/s41433-024-03173-3

Liang, Y., Wang, R., Wang, Y., & Liu, T. (2024). Estimating the prevalence of diabetic retinopathy in electronic health records with massive missing labels. Intelligence-Based Medicine, 10, 100154. https://doi.org/10.1016/j.ibmed.2024.100154

Maulidzar, R., Meylani, N. R., Alex, Asroruddin, M., Elida, S. Y., & Fitrianingrum, I. (2025). Hubungan status kontrol gula darah dan kejadian retinopati diabetik. Oftalmologi: Jurnal Kesehatan Mata Indonesia, 7(2), 75–83. https://doi.org/10.11594/ojkmi.v7i2.89

Naqiya, W., Virgana, R., & Kartasasmita, A. S. (2023). Gambaran kepatuhan berobat pasien retinopati diabetik di PMN RS Mata Cicendo 2021–2022. Oftalmologi: Jurnal Kesehatan Mata Indonesia, 5(2), 1. https://doi.org/10.11594/ojkmi.v5i2.54

Ogunyemi, O. I., Gandhi, M., Lee, M., et al. (2021). Detecting diabetic retinopathy through machine learning on electronic health record data from an urban, safety net healthcare system. JAMIA Open, 4(3), ooab066. https://doi.org/10.1093/jamiaopen/ooab066

Pan, H., Sun, J., Luo, X., Ai, H., Zeng, J., Shi, R., & Zhang, A. (2023). A risk prediction model for type 2 diabetes mellitus complicated with retinopathy based on machine learning and its application in health management. Frontiers in Medicine, 10, 1136653. https://doi.org/10.3389/fmed.2023.1136653

Purnama, R. F. N. (2023). Retinopati diabetik: Manifestasi klinis, diagnosis, tatalaksana dan pencegahan. Lombok Medical Journal, 2(2). https://doi.org/10.29303/lmj.v2i1.2410

Sasongko, M. B., Febryanto, G. A., Haryanto, S., Wardhana, F. S., Lestari, Y. D., Puspita, A., & Widyaputri, F. (2025). Incidence and progression of diabetic retinopathy and blindness in Indonesian adults with type 2 diabetes. PLOS ONE, 20(8), e0322093. https://doi.org/10.1371/journal.pone.0322093

Syahrul, F. H., & Sasongko, P. S. (2022). Penerapan convolutional neural network untuk klasifikasi tingkat keparahan retinopati diabetik pada penderita diabetes melitus. Jurnal Masyarakat Informatika, 13(1), 1–14. https://doi.org/10.14710/jmasif.13.1.42354

Syaqila, N. H., & Hakim, A. W. (2023). Retinopati diabetik proliferatif: Faktor risiko dan penatalaksanaan. Jurnal Pandu Husada, 4(1).

Teo, Z. L., Tham, Y.-C., Yu, M., Chee, M. L., Rim, T. H., Cheung, N., Bikbov, M. M., Wang, Y. X., Tang, Y., Lu, Y., Wong, I. Y., Ting, D. S. W., Tan, G. S. W., Jonas, J. B., Sabanayagam, C., Wong, T. Y., & Cheng, C.-Y. (2021). Global prevalence of diabetic retinopathy and projection of burden through 2045: Systematic review and meta-analysis. Ophthalmology, 128(11), 1580–1591. https://doi.org/10.1016/j.ophtha.2021.04.027

Usman, T. M., Saheed, Y. K., Nsang, A., Ajibesin, A., & Rakshit, S. (2023). A systematic literature review of machine learning based risk prediction models for diabetic retinopathy progression. Artificial Intelligence in Medicine, 144, 102617. https://doi.org/10.1016/j.artmed.2023.102617

Wang, X., Wang, W., Ren, H., Li, X., et al. (2024). Prediction and analysis of risk factors for diabetic retinopathy based on machine learning and interpretable models. Heliyon, 10(9), e29497. https://doi.org/10.1016/j.heliyon.2024.e29497

Wu, J.-H., Liu, T. Y. A., Hsu, W.-T., Ho, J. H.-C., & Lee, C.-C. (2021). Performance and limitation of machine learning algorithms for diabetic retinopathy screening: Meta-analysis. Journal of Medical Internet Research, 23(7), e23863. https://doi.org/10.2196/23863

Yang, C., Liu, Q., Guo, H., Zhang, M., Zhang, L., Zhang, G., Zeng, J., Huang, Z., Meng, Q., & Cui, Y. (2021). Usefulness of machine learning for identification of referable diabetic retinopathy in a large-scale population-based study. Frontiers in Medicine, 8, 773881. https://doi.org/10.3389/fmed.2021.773881

Zhao, Y., Li, X., Li, S., Dong, M., Yu, H., Zhang, M., Chen, W., Li, P., Yu, Q., Liu, X., & Gao, Z. (2022). Using machine learning techniques to develop risk prediction models for the risk of incident diabetic retinopathy among patients with type 2 diabetes mellitus: A cohort study. Frontiers in Endocrinology, 13, 876559. https://doi.org/10.3389/fendo.2022.87655