AN EFFECTIVE RULE BASED APPROACH FOR IDENTIFICATION OF COMORBIDITY PATTERNS IN DIABETIC PATIENTS

Bramesh S M
Assistant Professor, Department of Information Science and Engineering, P.E.S. College of Engineering,
Mandya - 571 401, Karnataka, India
smb@pesce.ac.in

Anil Kumar K M
Associate Professor, Department of Computer Science and Engineering, Sri Jayachamarajendra College of
Engineering, Mysore - 570 006, Karnataka, India
anilkm@sjce.ac.in
https://sites.google.com/site/anilkmsjce/home

Abstract
Diabetes is a chronic (long-lasting) disease and its incidence is quickly growing, in both developing and
developed nations. Diabetes patients have more risk of developing multiple conditions than the ones
without diabetes. In clinical literature, this phenomenon in general is known as comorbidity. The current
majority works have made great progress in extracting comorbidities patterns, but these works still have
a few limitations: first, little is known about comorbidities patterns; second, identifying top interesting
comorbidities patterns. The purpose of this research is to identify top-k diabetes-specific disease
comorbidity patterns from large clinical datasets. We have used the International Classification of
Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) codes to recognize patients diagnosed with
diabetes between 2001 and 2019 in Medical Information Mart for Intensive Care (MIMIC) datasets.
We have extracted diabetes-specific disease comorbidity patterns using Association Rule Mining (ARM)
techniques, namely Apriori, Frequent Pattern Growth (FP-Growth) and Frequent Pattern Maximal (FP-
Max) to address the first limitation. We have proposed an effective or a novel approach to identify top-k
association patterns of diabetes-specific disease comorbidities to address the second limitation. We have
also continued our investigation to find differences in gender-specific comorbidities patterns and major
diabetes subtypes-specific comorbidities patterns. Chronic kidney disease was the top most common
comorbidity (Support: 13.20%; Confidence: 85.51%, Lift: 3.766), followed by Atrial fibrillation
(Support: 13.72%; Confidence: 59.87%, Lift: 1.778). Also, we found that the diabetes-specific disease
comorbidities patterns are gender-specific and major diabetes subtypes-specific. Finally, identified top-k
comorbidities patterns are validated with medical literature. Notably, our results uncover novel
interesting and clinically meaningful comorbidities patterns and thereby assist clinical practitioners to
prescribe an optimal care for diabetic patients.

Keywords: Diabetes; Comorbidity patterns; Apriori; FP-Growth; FP-Max; ICD-9-CM codes.

1. Introduction
The word Diabetes mellitus is derived from the Greek word diabetes, which means “to pass through” and the
Latin word mellitus, which means “sweet”. An examination of history reveals that the word diabetes was used
for the first time around 250 to 300 BC by Apollonius of Memphis. Ancient civilizations of India, Greece and
Egypt have found the sweet nature of urine in this condition, and thus the spread of the word Diabetes Mellitus
has come into existence. In 1889, Mering and Minkowski wrote about the duty of the pancreas in the diabetes
pathogenesis. Collip, Banting, and Best at Toronto University in 1922, extracted the hormone insulin from
cow’s pancreas and hence paving the way for a successful diabetic treatment in the same year. Over the years,
outstanding work has been done and many discoveries, along with management strategies, have been developed
to address this growing problem.
Unfortunately, even these days, diabetes is one of the largest prevalent chronic disorders in the country and around the world [1][2]. Globally, 415 million people suffer from diabetes (8.8 percent of the world’s total population) and by 2020, according to the International Diabetes Federation, the number of people suffering from diabetes will reach 642 million [3]. In addition, the World Health Organization (WHO) reported that in 2019, diabetes was the 9th leading cause of death, with about 1.5 million direct deaths from diabetes.

There are several categories of diabetes, including type 1 diabetes, type 2 diabetes, maturity-onset diabetes of the young, neonatal diabetes, gestational diabetes, and secondary causes due to steroid use, endocrinopathies, etc. The major subtypes of diabetes are Type 1 Diabetes (T1D) and Type 2 Diabetes (T2D), which typically result from faulty insulin discharge (T1D) and or action (T2D). T1D occurs in adolescents or children, whereas T2D is believed to affect middle-aged and older persons with long-term hyperglycemia due to poor lifestyle and nutritional choices. However, all types of diabetes can lead to problems in many parts of the body and surge the overall risk of premature death. Stroke, heart attack, leg amputation, nerve damage, vision loss, and kidney failure are all possible consequences. Furthermore, uncontrolled diabetes during pregnancy surges the risk of fetal mortality and other problems.

Diabetes individuals are also more likely to have various diseases than those who do not have diabetes [4]. In clinical literature, this phenomenon in general is known as comorbidity. The presence of comorbidities complicates diabetes-related healthcare outcomes, care needs, treatment options and also associated cost. Therefore, accurate and clear understanding of the epidemiology of diseases that co-occur with diabetes is significant for setting goals of treatment.

Prior studies have shown that the majority of persons with diabetes have at the minimum one comorbid condition and 40 percent have three or more comorbid conditions; yet the viewpoint of the healthcare providers and treatment approaches are more focused on supervision of diabetes alone [5-8]. For the best possible healthcare delivery and the development of self-management solutions for the ever-increasing diabetes patient’s population, we want to figure out what the pattern is; the number and types of comorbidities, and how they affect diabetes care. These comorbidities may cause doctors to redirect their attention away from the diabetes known about comorbidity patterns. So, to progress the study of comorbidity, we need to find a method that extends beyond the examination of comorbid pairs.

Several studies conducted across India focused mainly on single comorbidities such as depression or hypertension [15-19]. Ramachandra et. al. [20], on the other hand, did a study on the occurrence of macro and micro vascular problems in T2D and they focused on this fact on wider aspect. Researchers have also studied the occurrence of hypertension and dyslipidemia in T2D. Similarly, various research focusing on metabolic syndrome have indicated the occurrence of dyslipidemia and hypertension among T2D [21-27]. Knowing the incidence of comorbid medical disorders in various demographic groups aids healthcare practitioners and policymakers in properly allocating health resources and tailoring diabetic care supervision to successfully use healthcare programs while reducing healthcare expenditure.

Unfortunately, for good reason, researchers have only seen at comorbidity in incomplete ways: the combination of two or three disorders quickly becomes complex. For instance, consider the following scenario: you want to compare three diseases: heart disease, arthritis, and diabetes. We might inquire this basic question regarding relative association of the three diseases: Is it true that people with arthritis are more likely to develop heart disease or diabetes? There are 120 combinations of three diseases to analyze in a study with just ten chronic conditions and 45 pairs of comorbid. To conclude, identifying interesting patterns of associations from such a large number of combinations is a challenging task and checking whether the identified pattern of associations occurred more frequently than by chance alone is also a challenging task. As a consequence, little is known about comorbidities patterns. So, to progress the study of comorbidity, we need to find a method that extends beyond the examination of comorbid pairs.

Using association rule mining techniques is one solution to this problem. This kind of analysis is a descriptive technique that evaluates how elements are likely to occur together. Rule-based analysis could be effective for quickly and easily describing broad patterns of comorbidity. It is feasible to go beyond simple comorbid pairs using this method to acquire a wide overall picture of how diseases are related in a specific community and where a specific disease of attention appears in the pattern.

To our knowledge, no study has yet published the extraction of comorbidities patterns in MIMIC. The goal of this study is to look at the comorbidity patterns among a MIMIC representative sample of a diabetic patient. Specifically, our objectives are: (1) To report the conditions of comorbid, with diabetes as the index disorder, by using standard association rule mining techniques: Apriori, FP-Growth and FP-Max (2) To propose an effective approach to select top-k association patterns of comorbidities (3) To summarize the differences in gender-specific and also major diabetes subtypes-specific comorbidity patterns, and finally (4) To validate the extracted top-k comorbidities patterns with medical literature.

The paper is structured as follows. Section 2 presents the related work, Section 3 discusses the materials and methods used, Section 4 discusses the proposed methodology, Section 5 discusses the experimentation and results and lastly Section 6 concludes the paper.
2. **Related Work**

Diabetes patients have more risk of developing multiple conditions than the ones without diabetes. Comorbidity, defined as the presence of two or more disorders in the same patient [28][29], is a public health concern since it has significant implications for the healthcare system as well as patients [29][30]. According to numerous studies, the occurrence of comorbidity differs between ~20% and ~90% [31-34]. This dissimilarity is because of the population under study, together with additional characteristics of the study design, for instance the comorbidity definition. The occurrence of comorbidity rises with age, and it is not restricted only to the elderly population [34-37]. The accessibility of clinical data for data mining provides the chance to determine disease relations and patterns of comorbidity from the patient’s clinical history collected during repetitive medical care [38-40]. Hence many of the researchers have used clinical data to study comorbidities patterns. For example, Xueyan et. al. [41] (2022) have used association rule mining to extract comorbidities patterns in schizophrenia patients in Beijing, China. The study included data from 8,252 patients and they reported extrapyramidal syndrome (44.58%), constipation (31.63%), and tachycardia (19.13%) as the most frequently reported comorbidities in schizophrenia patients. Ahmad et. al. [42] (2022) implemented machine learning models (Logistic regression, Evmip functions, Classification and Regression algorithm) to predict the diabetic mellitus and cardiovascular diseases co-occurrence in people attending a screening program with high accuracy. Data from the Diabetes Complications Screening Research Initiative was used in the study, which contained over 200 variables from over 2000 participants. The predictive accuracy of the proposed machine learning model was found to be 94.09%. Meera et. al. [43] (2021) have used association rule mining to light on COVID-19 patients symptom patterns. In total, 1,560 COVID-19 patients were involved in this study and this study reveals that the most frequently reported symptom in COVID-19 patients was fever (67%), cough (37%), pneumonia (11%), and sore throat (8%). Chan et. al. [44] (2008) have used ARM to investigate the metabolic syndrome, which is related to other disorders and to comprehend the strength of association between hypertension, hyperlipidemia, and diabetes mellitus on patient’s records in Taiwan. From the study, it was observed that diabetes is connected to blear eyes and oral diseases and they also observed that metabolic syndrome patients have more connection with liver diseases than with diabetes patients. Harahap et. al. [45] (2018) have used K-means and association rule mining techniques to analyse patient prescriptions in order to detect the association between the medicines and disease that are used to treat the patient's illness. Here, to identify the association, patient prescription information from two hospitals were collected in 2015 and 2016. The K-means method is used to group the top ten diseases in the first step of their method. The Apriori method is then used to discover the relationship between drugs and diseases. Nahar and Phoebe [46] (2013) have used ARM to find factors that lead to heart disease in both men and women using the dataset, UCI Cleveland. Here, three methods namely, Predictive Apriori, Apriori, and Tertius, are examined to detect the factors. From the study, it was observed that females are at higher risk of having heart disease than males. Jia et. al. [47] (2020) have performed an association rules analysis on electronic medical records to explore patterns of comorbidity among older individuals with lung cancer in northeast China. In total, 1,510 patients were identified and studied here. They discovered that the most common comorbidities in older individuals with lung cancer are cerebral infarction, pneumonia, and hypertension, and that cardiovascular comorbidities are the most frequent comorbid combinations. Sea et. al. [48] (2020) proposed discriminative pattern-based features as an approach for improving readmission prediction. For the all-disease cohort, the proposed pattern-based model, enhanced performance of prediction roughly by 12% compared to baseline models, but it demonstrated minimal improvement for either the chronic kidney disease or diabetes patients. Jinghe et. al. [49] (2020), have proposed a new paradigm for learning sparse longitudinal representations of patients' medical records. The proposed paradigm is also compared with broadly used methods such as Bag-of-Pattern and Aggregated Frequency Vector in Sequences on actual electronic health record data, and the investigational findings revealed that the proposed approach achieves better predictive performance. Ma et. al. [50] (2017) proposed a microbe-based network of human disease based on the text mining approach. Here, the microbe-disease association dataset was used to build the network of disease. Here the intention was to find the relationships between microbes and disease genes, chemical fragments, symptoms, and drugs. In this study, the similarity between two diseases is identified using cosine similarity measure. Zhang et. al. [51] (2019) based on a weekly supervised learning approach, have proposed a disease-related gene mining method. Firstly, they screened the differentially expressed gene set. Secondly, they implemented support vector machine to forecast the disease-related genes in the differentially expressed gene set. They discovered that, when compared to other good techniques, the suggested model may effectively increase the disease-related gene prediction accuracy. Lan et. al. [52] (2018) proposed similarity computations to forecast the miRNA-disease correlations. Here, the miRNAs sequence and function information are used to calculate miRNA similarity and based on the disease semantic and function information, the similarity among diseases is computed. In this study, to deduce potential miRNA-disease connections, the data sources were combined using the kernelized Bayesian matrix factorization approach and also the experimental findings revealed that the proposed technique can successfully predict previously unknown miRNA-disease connections.
Also, most earlier studies, targets to find associations among diseases and have concentrated on revealing associations among a limited pre-defined specific condition [53–60]. For example, Farran et al. [59] (2013) targeted association between hypertension and diabetes, and one more work by Chen et. al. [60] (2015) discovered association between obesity and colorectal cancer.

A few limitations observed from the existing literature are as follows:

- Clinical datasets or survey data used are usually small (generally limited to a single medical facility or network) and hence, have a limited population and time period coverage.
- Due to HIPAA regulations, clinical datasets are not available to all researchers, thus the research openings in this domain are limited.
- Targets to identify diseases associations have focused on a limited pre-defined specific condition.

From the above limitations, we can say little is known about comorbidities patterns. Hence, unlike the previous works, we investigate the diabetes-specific disease comorbidities patterns using large MIMIC datasets namely MIMIC-III & MIMIC-IV. Specifically, we apply ARM techniques, namely Apriori, FP-Growth and FP-Max in a non-traditional manner to identify an association across ICD-9-CM codes assigned and reported in the EHRs of diabetic patients in both datasets. Also, we propose an effective approach to select top-k association patterns of diabetes-specific disease comorbidities. We also investigate to find the existence of gender-specific comorbidities patterns and major diabetes subtypes-specific comorbidities patterns. Finally, we validated the diagnosis codes (conditions) grouped in comorbidities patterns with medical literature.

3. Materials and Methods

3.1. Datasets

We used the MIMIC-III & MIMIC-IV datasets for the study. Both MIMIC-III and MIMIC-IV, are large, de-identified and publicly accessible collection of medical records, which was approved by the Institutional Review Boards of Beth Israel Deaconess Medical Center (BIDMC, Boston, MA, USA) and the Massachusetts Institute of Technology (MIT, Cambridge, MA, USA) [61][62]. Each record in the MIMIC-III dataset includes ICD-9-CM codes and some records in the MIMIC-IV dataset includes ICD-10-CM codes also, which identify diagnoses and procedures performed. The MIMIC critical care dataset is unique and notable for the following reasons:

- It is the only critical care database of its sort that is freely available.
- The dataset spans more than a decade and includes detailed information about each patient's care.
- Once a data usage agreement is signed, analysis is unfettered, allowing for global education and clinical research.

The key characteristics of the datasets is shown in Table 1. An example of the MIMIC - III dataset is shown in Table 2.

<table>
<thead>
<tr>
<th>Key Characteristic</th>
<th>MIMIC-III Dataset [61]</th>
<th>MIMIC-IV Dataset [62]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of diagnostic records (only ICD-9-CM) of diabetic patients</td>
<td>1,99,964</td>
<td>9,47,244</td>
</tr>
<tr>
<td>Number of distinct hospital admissions of diabetic patients</td>
<td>14,222</td>
<td>68,118</td>
</tr>
<tr>
<td>Number of distinct people diagnosed with diabetes</td>
<td>10,318</td>
<td>26,115</td>
</tr>
<tr>
<td>Number of distinct male people diagnosed with diabetes</td>
<td>5,898</td>
<td>13,784</td>
</tr>
<tr>
<td>Number of distinct female people diagnosed with diabetes</td>
<td>4,420</td>
<td>12,331</td>
</tr>
<tr>
<td>Contains data from year</td>
<td>2001 - 2012</td>
<td>2008 - 2019</td>
</tr>
</tbody>
</table>

Table 1. Key characteristics of the datasets.

<table>
<thead>
<tr>
<th>ROW_ID</th>
<th>SUBJECT_ID</th>
<th>HADM_ID</th>
<th>SEQ_NUM</th>
<th>ICD9_CODE</th>
<th>GENDER</th>
</tr>
</thead>
<tbody>
<tr>
<td>1523</td>
<td>117</td>
<td>140784</td>
<td>1</td>
<td>5715</td>
<td>F</td>
</tr>
<tr>
<td>1524</td>
<td>117</td>
<td>140784</td>
<td>2</td>
<td>7895</td>
<td>F</td>
</tr>
<tr>
<td>1525</td>
<td>117</td>
<td>140784</td>
<td>3</td>
<td>7054</td>
<td>F</td>
</tr>
<tr>
<td>1526</td>
<td>117</td>
<td>140784</td>
<td>4</td>
<td>2875</td>
<td>F</td>
</tr>
<tr>
<td>1527</td>
<td>117</td>
<td>140784</td>
<td>5</td>
<td>4280</td>
<td>F</td>
</tr>
</tbody>
</table>

Table 2. An example of a MIMIC-III dataset.

Each record shown in Table 2 corresponds to a diagnosis information of a patient (identified by subject_id) and contains their hospital admission id (hadm_id), diagnosis code (ICD9_Code), diagnosis codes are ordered by a unique sequence number (seq_num) and unique record identifier (row_id).
3.2. Data pre-processing

The datasets were pre-processed using the following steps:

(1) Study Population:
Since, we aimed at finding the comorbidities patterns in diabetic patients, we selected the records of diabetic patients from the datasets. Here, we have selected the diabetic patient records, whose diagnosis details are represented in ICD-9-CM codes. Based on this criterion, 1,99,964 number of diagnostic records of diabetic patients were selected from MIMIC-III dataset and 9,47,244 number of diagnostic records of diabetic patients from MIMIC-IV dataset.

(2) Dealing with Missing Values:
From the exploratory data analysis, we observed that there were zero missing values in MIMIC-IV dataset and very few missing values in the diagnosis codes columns in MIMIC-III dataset, hence we removed those records for further analysis.

(3) Representing ICD-9-CM diagnosis codes defining diabetes:
ICD-9-CM diagnosis code from 250.0x to 250.9x and from 250.x0 to 250.x3 represent diabetes. In both datasets ICD-9-CM codes are represented as a sequence of characters without decimal digits and since we aimed at identifying the comorbidities patterns in diabetic patients we represented the ICD-9-CM diagnosis codes representing diabetes as “25000” for further analysis.

4. Proposed Methodology

4.1. System model

Our problem can formally be defined as follows:
Let us assume, a set of diagnosis codes \( C = \{C_1, \ldots, C_M\} \) and Patient dataset \( P_D = \{P\} \) are given, where each patient record \( R \) is a subset of diagnosis codes \( P \subseteq C \). Then, an association rule is defined as an implication:

\[ X \Rightarrow Y \tag{1} \]

Where \( X \subset C \), \( Y \subset C \), and \( X \cap Y = \emptyset \).
There are two elements of these rules:
Antecedent (IF): This is a disease / group of diseases that are typically found in the datasets.
Consequent (THEN): This comes along as a disease with an Antecedent / group of Antecedents.

The current implementation makes use of the support, confidence and lift metrics to identify interesting association rules and the description of each metric is as follows:

**Support:** It informs us the fraction of patient records which contain disease ‘A’ and disease ‘B’. Basically, it tells us about the commonly diagnosed diseases or the combination of diseases diagnosed commonly.

\[
\text{Support} = \frac{\text{freq}(A, B)}{N} \tag{2}
\]

Where, ‘\( N \)’ is the total number of patient records.
So, with support, we can filter out the diseases that have a low frequency.

**Confidence:** Given the number of times disease ‘A’ occurs, it informs us how frequently disease ‘A’ and disease ‘B’ occur together.

\[
\text{Confidence} = \frac{\text{freq}(A, B)}{\text{freq}(A)} \tag{3}
\]

**Lift:** Lift shows the rule strength over the random occurrence of disease ‘A’ and disease ‘B’. It basically shows us how strong a rule is.

\[
\text{Lift} = \frac{\text{Support}}{\text{Supp}(A) \times \text{Supp}(B)} \tag{4}
\]

4.2. Architecture and Working

The architecture of the proposed work is shown in Fig. 1. As shown in Fig. 1, the MIMIC dataset's raw data is retrieved for data pre-processing such as cleaning, filtering, and transformation. The selected data set is next analyzed using the ARM algorithms, namely Apriori, FP-Growth and FP-Max to generate patterns that signify relationships among data. Form these patterns, interesting patterns will be selected using evaluation parameters like support, confidence and lift. Finally, filtered patterns are sorted using an effective approach to select top-k association patterns i.e., top-k association patterns of comorbidities, which can be used for designing the recommendation system and as a reference for the clinical practitioner and diabetic patient.
4.3. Finding diabetes-specific frequent diseases

In this research work, three well known frequent pattern mining techniques namely Apriori, FP-Growth and FP-Max algorithms have been used to find diabetes-specific frequent diseases patterns from both MIMIC-III and MIMIC-IV datasets.

4.3.1. Apriori

Given:
A set of diagnosis codes $C = \{C_1, \ldots, C_M\}$, Patient dataset $P_D = \{P\}$ and Minimum Support Threshold (MST).

Apriori says for generating diabetes-specific frequent diseases:
The probability that diagnosis code $C_i$ is not frequent is, if:
- $P(C_i) < \text{MST}$, then $C_i$ is not frequent.
- $P(C_i + C_j) < \text{MST}$, then $C_i + C_j$ is not frequent, where $C_j$ also belongs to $C$.
- If the diagnosis code set value is less than the minimal support, all of its supersets will also be less than the minimum support, and can thus be ignored.

The algorithm for generating diabetes-specific frequent diseases patterns is given below.
4.3.1. Apriori

Input: Patient dataset PD, minimum_support
Output: diabetes-specific frequent diseases

\[ L_1 = \{ \text{large 1-itemsets} \} \]

\[ \text{foreach (} n = 2; \ L_{n-1} \neq \emptyset; \ n++) \]

\[ C_n = \text{apriori}_\text{gen}(L_{n-1}) \]; // Novel candidates

\[ \text{foreach transaction } t \text{ in } D \]

\[ C_t = \text{subset}(C_n, t) \]; // Candidates in \( t \)

\[ \text{foreach candidate } c \text{ in } C_t \]

\[ c.\text{count} = 1 + c.\text{count}; \]

\[ \text{end} \]

\[ L_n = \{ c \in C_n | c.\text{count} \geq \text{minimum_support} \}; \]

\[ \text{end} \]

\[ \text{Answer} = \bigcup_n L_n; \]

4.3.2. FP-Growth

Given:
A set of diagnosis codes \( C = \{ C_1, \ldots, C_m \} \), Patient dataset \( PD = \{ P \} \) and minimum support threshold.

FP-Growth says: Grow long patterns from short ones using local frequent items only.
- “pqr” is a frequent pattern
- Obtain all transactions containing “pqr”, i.e., project DB on pqr: DB|pqr
- “s” is a local frequent item in DB|pqr -->pqrs is a frequent pattern

The algorithm for generating diabetes-specific frequent diseases patterns is given below.

4.3.2.1. Algorithm

Input: Patient dataset PD, minimum_support
Output: diabetes-specific frequent diseases

//Initial Call: \( R = \text{FP} \text{Tree}(PD), \ P = \emptyset, F = \emptyset \)

FP_Growth(\( R, P, F, \text{minimum_support} \))

Eliminate non-frequent diagnosis codes from \( R \)

if Is_Path(\( R \)) then /*Add \( R \) subsets into \( F \)*/

\[ X = Y \cup P \]

\[ \text{support}(X) = \text{minimum}_{\text{ef}} \{ \text{count}(X) \} \]

\[ F = F \cup \{ (X, \text{support}(X)) \} \]

else /* for each frequent diagnosis code \( i \), process projected FP-Trees */

\[ \text{foreach } i \in R \text{ in increasing order of support } (i) \]

\[ X = \{ i \} \cup P \]

\[ \text{support}(X) = \text{support}(i) \text{/} \text{sum of count}(i) \text{ for all nodes labelled } i\text{/} \]

\[ F = F \cup \{ (X, \text{support}(X)) \} \]

\[ R_X = \emptyset \text{ /*Projected FP_Tree for } X*/ \]

\[ \text{foreach path Path_From_Root}(i) \]

\[ \text{count}(i) = \text{count of } i \text{ in path} \]

Add path, without \( i \), into FP_Tree \( R_X \) with count, count\( (i)\)

end

if \( R_X \neq \emptyset \) then FP_Growth(\( R_X, X, F, \text{minimum_support} \))

end

4.3.3. FP-Max

FP-Max is a variant of FP-Growth and it is also used to generate maximal frequent patterns.

FP-Max property:
If a diagnosis code set \( X \) is frequent and there is no frequent super-pattern containing \( X \), it is said to be maximal.

The above property is the reason for using FP-Max to find diabetes-specific frequent diseases patterns because it has the potential to produce non-redundant frequent diseases with lengthy disease co-occurrence / sequence patterns. The algorithm for generating maximal diabetes-specific frequent diseases patterns is given below.
4.3.3.1. Algorithm

FP_Max (Input $T$: FP_Tree; Output $M$: MFI_Tree)

Variable
MFIT: MFI_Tree;
Front, Rear: Linked list of diagnosis codes;

if ($T$ comprises a single path $P$)
    Insert (Front $\cup P$) in MFIT
else
    foreach $j$ in header-table of $T$
        Append $j$ to Front;
        Construct the conditional pattern base $B[j]$ for $j$;
        Rear = {frequent items in $B[j]$};
        if Not (Front $\cup$ Rear in MFIT)
            Construct the FP_Tree $T$ [front];
            FP_Max ($T$ [front]);
            eliminate $j$ from Front;
    end
end

4.4. Finding association patterns of diabetes-specific disease comorbidities

After finding all the diabetes-specific frequent disease patterns, frequent disease patterns having length greater than or equal to two will be used to find patterns of diabetes-specific disease comorbidities. Specifically, the diabetes-specific disease co-occurrence association patterns. Disease co-occurrence association pattern in general, is in form rule $X \Rightarrow Y$, where $X$ and $Y$ are frequent diseases. The algorithm for finding diabetes-specific disease comorbidities association patterns is given below.

4.4.1. Algorithm

foreach (frequent diagnosis codes set $l_n, n \geq 2$)
call the procedure general_rules ($l_n, l_n$);
procedure general_rules ($l_n$: frequent diagnosis codes set, $a_m$: frequent $m$-diagnosis codes set)
    $X = \{(m-1)\)-diagnosis codes set $x_{m-1}, x_{m-1} \subset x\}$;
    foreach ($x_{m-1} \in X$)
        confidence = support ($l_n$) / support ($x_{m-1}$);
        if (minimum_confidence $\leq$ confidence)
            Output the rule $x_{m-1} \Rightarrow (l_n - x_{m-1})$ with support = support ($l_n$) and confidence = confidence;
            if (1 $< m-1$)
                call the procedure general_rules ($l_n, x_{m-1}$);
        end
    end

This algorithm's main goal is to locate all non-empty subsets of $l$ for each frequent diagnosis code set $l$. For such a subset $x$, the rule of the form $x \Rightarrow (l - x)$ is noted if the ratio of support ($l$) to support ($x$) is at least minimum confidence.

4.5. Finding top-k association patterns of diabetes-specific disease comorbidities

The diabetes-specific disease association patterns that passes the minimum confidence threshold and the minimum support threshold, will be filtered again using one more important threshold called minimum lift threshold, to avoid the patterns which occurs by chance. Finally, the filtered patterns will be sorted based on the procedure defined below in order to extract the top-k interesting association patterns of diabetes-specific disease co-occurrence.

4.5.1. Sorting Procedure

Given two association patterns, $r_i$ and $r_j$, $r_i$ has higher precedence than $r_j$ if the below conditions hold:

- $\text{lift}(r_i) > \text{lift}(r_j)$; or
- if $\text{lift}(r_i) = \text{lift}(r_j)$, but $\text{conf}(r_i) > \text{conf}(r_j)$; or
- if $\text{conf}(r_i) = \text{conf}(r_j)$, but $\text{sup}(r_i) > \text{sup}(r_j)$; or
- if $\text{sup}(r_i) = \text{sup}(r_j)$, but $\text{size}(r_i) > \text{size}(r_j)$, i.e., the length of antecedent of $r_i$ is larger than $r_j$. 

DOI : 10.21817/indjcse/2022/v13i4/221304054
Vol. 13 No. 4 Jul-Aug 2022
1074
Here, the lift value is ranked first because it indicates the strength of an association pattern over the random co-occurrence of disease(s) $X$ and disease(s) $Y$. The confidence value is ranked second since it indicates how closely diseases are linked to one another. High confidence indicates that disease(s) $Y$ is strongly associated to disease(s) $X$. The support value is ranked third since it indicates how many patients have developed disease co-occurrence association pattern. High support indicates diseases that occur together in several patients. Next, the association pattern length is ranked fourth because a longer pattern contains more information than a shorter one.

5. Experimentation and Results

5.1. Experimental Setup

To inspect the patterns of diabetes-specific disease comorbidities, three experiments were conducted. All experiments were implemented in Python language using “mlxtend” [63] as one of the main libraries. The details of the experiments are described as follows.

The first experiment compares the performance of well-known frequent pattern mining techniques like Apriori, FP-Growth and FP-Max, when generating frequent diseases on the MIMIC-III dataset with diverse minimum support thresholds. The thresholds of minimum support are set from 1% to 10%. All three frequent pattern mining techniques are assessed by using the number of frequent diseases and computation time. Based on the results of this experiment, we decided to set the minimum support threshold as 10% in order to filter out the uninteresting patterns extracted from MIMIC-III dataset.

The second experiment compares the performance of well-known frequent pattern mining techniques like Apriori, FP-Growth and FP-Max, when generating frequent diseases on the MIMIC-IV dataset with diverse minimum support thresholds. The thresholds of minimum support are set from 5% to 12%. All three frequent pattern mining techniques are assessed by using the number of frequent diseases and computation time. Based on the results of this experiment, we decided to set the minimum support threshold as 12% in order to filter out the uninteresting patterns extracted from MIMIC-IV dataset.

The third experiment is concerned with applying Apriori, FP-Growth and FP-Max algorithms to extract the top-k patterns of diabetes-specific disease comorbidities from the MIMIC-III dataset using the parameters setting as shown in Table 3. After then, using the sorting procedure described in section 4.5.1, the top-k patterns of diabetes-specific disease comorbidities are chosen. Here diabetes-specific disease association patterns are also studied based on gender and also based on major diabetes subtypes.

The fourth experiment is concerned with applying Apriori, FP-Growth and FP-Max algorithms to extract the top-k patterns of diabetes-specific disease comorbidities from the MIMIC-IV dataset using the parameters setting as shown in Table 3. After then, using the sorting procedure described in section 4.5.1, the top-k patterns of diabetes-specific disease comorbidities are chosen. Here diabetes-specific disease association patterns are also studied based on gender and also based on major diabetes subtypes.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Minimum Support in (%)</th>
<th>Minimum Confidence in (%)</th>
<th>Minimum Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIMIC-III</td>
<td>10</td>
<td>50</td>
<td>1.1</td>
</tr>
<tr>
<td>MIMIC-IV</td>
<td>12</td>
<td>50</td>
<td>1.1</td>
</tr>
</tbody>
</table>

Table 3. The default value for experimental parameters for each dataset.

5.2. Experimental Results

Table 4 lists the number of frequent diseases that are extracted from the MIMIC-III dataset using Apriori, FP-Growth and FP-Max by changing minimum support threshold from 1% to 10%. It is clear from the Table 4, that for minimum support threshold values from 1% to 10%, the Apriori and FP-Growth has generated the same frequent diseases, whereas FP-Max has generated a smaller number of frequent diseases when compared with Apriori and FP-Growth.

On the other hand, from Fig. 2, we can observe that initially, i.e., when support is low Apriori algorithm takes more time for execution when compared with FP-Growth and FP-Max algorithms because the Apriori algorithm requires high computation to process very large sets, which is the result of choosing a minimum support value as low. As the support value increases the Apriori algorithm will take less time to process because as the support value increases, the size of the set that Apriori algorithm to process reduces. However, FP-Growth and FP-Max has taken almost the same time for execution because both algorithms are tree-based approaches and we can also observe that the execution time of both algorithms has decreased as the minimum support value is increased because the number of frequent diseases decreases as we increase the minimum support value. When compared to Apriori and FP-Growth algorithms, the FP-Max algorithm produced a smaller number of frequent diseases in a shorter amount of time, hence it was chosen for detecting frequent diseases in our work.
Table 4. Comparing the number of frequent diseases extracted from MIMIC-III dataset.

<table>
<thead>
<tr>
<th>Minimum Support in (%)</th>
<th>Apriori</th>
<th>FP-Growth</th>
<th>FP-Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3565</td>
<td>3565</td>
<td>1027</td>
</tr>
<tr>
<td>2</td>
<td>1041</td>
<td>1041</td>
<td>319</td>
</tr>
<tr>
<td>3</td>
<td>523</td>
<td>523</td>
<td>162</td>
</tr>
<tr>
<td>4</td>
<td>297</td>
<td>297</td>
<td>97</td>
</tr>
<tr>
<td>5</td>
<td>211</td>
<td>211</td>
<td>68</td>
</tr>
<tr>
<td>6</td>
<td>149</td>
<td>149</td>
<td>48</td>
</tr>
<tr>
<td>7</td>
<td>99</td>
<td>99</td>
<td>37</td>
</tr>
<tr>
<td>8</td>
<td>83</td>
<td>83</td>
<td>31</td>
</tr>
<tr>
<td>9</td>
<td>69</td>
<td>69</td>
<td>26</td>
</tr>
<tr>
<td>10</td>
<td>56</td>
<td>56</td>
<td>21</td>
</tr>
</tbody>
</table>

Table 5. Comparing the number of frequent diseases extracted from MIMIC-IV dataset.

<table>
<thead>
<tr>
<th>Minimum Support in (%)</th>
<th>Apriori</th>
<th>FP-Growth</th>
<th>FP-Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>381</td>
<td>381</td>
<td>124</td>
</tr>
<tr>
<td>6</td>
<td>261</td>
<td>261</td>
<td>84</td>
</tr>
<tr>
<td>7</td>
<td>197</td>
<td>197</td>
<td>71</td>
</tr>
<tr>
<td>8</td>
<td>147</td>
<td>147</td>
<td>51</td>
</tr>
<tr>
<td>9</td>
<td>109</td>
<td>109</td>
<td>39</td>
</tr>
<tr>
<td>10</td>
<td>95</td>
<td>95</td>
<td>34</td>
</tr>
<tr>
<td>11</td>
<td>79</td>
<td>79</td>
<td>28</td>
</tr>
<tr>
<td>12</td>
<td>63</td>
<td>63</td>
<td>21</td>
</tr>
</tbody>
</table>

Table 5 lists the number of frequent diseases that are extracted from the MIMIC-IV dataset using Apriori, FP-Growth and FP-Max by varying minimum support threshold from 5% to 12%. It is clear from the Table 5, that for minimum support threshold values from 5% to 12%, the Apriori and FP-Growth has generated the same frequent diseases, whereas FP-Max has generated a smaller number of frequent diseases when compared with Apriori and FP-Growth.

From Fig. 3, we can observe that initially, i.e., when support is low Apriori algorithm takes more time for execution when compared with FP-Growth and FP-Max because of the same reason as explained in the case of MIMIC-III dataset. However, in the case of MIMIC-IV dataset, FP-Max has taken less time when compared to FP-Growth algorithm because of the reasons (1) MIMIC-IV is a dense dataset and (2) FP-Max generates a small FP-Tree when compared with FP-Growth. When compared to Apriori and FP-Growth algorithms, the FP-Max algorithm produced a smaller number of frequent diseases in a shorter amount of time, hence it was chosen for detecting frequent diseases in our work.
Table 6 reports the top-3 most frequent co-occurrence patterns (comorbidities patterns) found from diabetic patients in MIMIC-III dataset by adopting our proposed methodology. For all top-3 patterns, a unique pattern number has been given and every pattern has been represented in two forms. For example, Pattern number 1 has been represented using ICD-9-CM code, i.e., (42731, 25000) => (4280) and the same pattern has been represented using corresponding ICD-9-CM title or description i.e., (Atrial fibrillation, Diabetes) => (Congestive heart failure, unspecified). This pattern reveals that patients who had developed Atrial fibrillation and diabetes were more likely to develop congestive heart failure.

<table>
<thead>
<tr>
<th>Pattern No.</th>
<th>Patterns</th>
<th>Support in (%)</th>
<th>Confidence in (%)</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(Atrial fibrillation, Diabetes) =&gt; (Congestive heart failure, unspecified) (42731, 25000) =&gt; (4280)</td>
<td>14.59</td>
<td>52.35</td>
<td>1.426</td>
</tr>
<tr>
<td>2</td>
<td>(Diabetes, Other and unspecified hyperlipidemia) =&gt; (Unspecified essential hypertension) (25000, 2724) =&gt; (4019)</td>
<td>13.57</td>
<td>55.87</td>
<td>1.191</td>
</tr>
<tr>
<td>3</td>
<td>(Diabetes, Coronary atherosclerosis of native coronary artery) =&gt; (Unspecified essential hypertension) (25000, 41401) =&gt; (4019)</td>
<td>18.13</td>
<td>53.42</td>
<td>1.139</td>
</tr>
</tbody>
</table>

Table 6. The most frequent co-occurrence patterns found from diabetic patients in MIMIC-III dataset.

5.2.1. Finding Gender-specific comorbidities patterns in MIMIC-III dataset

For finding the Gender-specific comorbidities patterns in MIMIC-III dataset, we run the proposed model separately on 5,898 male diabetic patients and 4,420 female diabetic patients. Table 7 reports the top-4 most frequent co-occurrence patterns (comorbidities patterns) found from male diabetic patients in MIMIC-III dataset. Table 8 shows the top-3 most frequent co-occurrence patterns (comorbidities patterns) found from female diabetic patients in MIMIC-III. By observing the contents of Table 7 and Table 8, we infer that the male and female diabetic patients show many identical comorbidities patterns, for example, (atrial fibrillation, diabetes, congestive heart failure), (hyperlipidemia, diabetes, essential hypertension), and (hyperlipidemia, diabetes, coronary atherosclerosis of native coronary artery). One exception is that the male diabetic patients show the comorbidity pattern (hyperlipidemia, diabetes, coronary atherosclerosis of native coronary artery) i.e., the male diabetic patients who had developed hyperlipidemia and diabetes were more likely to develop Coronary atherosclerosis of native coronary artery, which is not found in female diabetic patients.
Table 7. The most frequent co-occurrence patterns found from male diabetic patients in MIMIC-III dataset.

<table>
<thead>
<tr>
<th>Pattern No.</th>
<th>Patterns</th>
<th>Support in (%)</th>
<th>Confidence in (%)</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(Atrial fibrillation, Diabetes) =&gt; (Congestive heart failure, unspecified)</td>
<td>14.90</td>
<td>51.26</td>
<td>1.422</td>
</tr>
<tr>
<td></td>
<td>(42731, 25000) =&gt; (4280)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>(Other and unspecified hyperlipidemia, Diabetes) =&gt; (Coronary atherosclerosis of native coronary artery)</td>
<td>12.38</td>
<td>50.17</td>
<td>1.308</td>
</tr>
<tr>
<td></td>
<td>(2724, 25000) =&gt; (41401)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>(Other and unspecified hyperlipidemia, Diabetes) =&gt; (Unspecified essential hypertension)</td>
<td>13.75</td>
<td>55.75</td>
<td>1.228</td>
</tr>
<tr>
<td></td>
<td>(2724, 25000) =&gt; (4019)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>(Coronary atherosclerosis of native coronary artery, Diabetes) =&gt; (Unspecified essential hypertension)</td>
<td>20.31</td>
<td>52.96</td>
<td>1.166</td>
</tr>
<tr>
<td></td>
<td>(41401, 25000) =&gt; (4019)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 8. The most frequent co-occurrence patterns found from female diabetic patients in MIMIC-III dataset.

<table>
<thead>
<tr>
<th>Pattern No.</th>
<th>Patterns</th>
<th>Support in (%)</th>
<th>Confidence in (%)</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(Diabetes, Atrial fibrillation) =&gt; (Congestive heart failure, unspecified)</td>
<td>14.19</td>
<td>53.90</td>
<td>1.434</td>
</tr>
<tr>
<td></td>
<td>(25000, 42731) =&gt; (4280)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>(Other and unspecified hyperlipidemia, Diabetes) =&gt; (Unspecified essential hypertension)</td>
<td>13.34</td>
<td>56.03</td>
<td>1.148</td>
</tr>
<tr>
<td></td>
<td>(2724, 25000) =&gt; (4019)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>(Diabetes, Coronary atherosclerosis of native coronary artery) =&gt; (Unspecified essential hypertension)</td>
<td>15.33</td>
<td>54.23</td>
<td>1.111</td>
</tr>
<tr>
<td></td>
<td>(25000, 41401) =&gt; (4019)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 9 shows the top-6 most frequent co-occurrence patterns (comorbidities patterns) found from diabetic patients in MIMIC-IV dataset by adopting our proposed methodology. For all top-6 patterns, a unique pattern number has been given and every pattern has been represented in two forms as explained earlier. For example, Pattern number 6 reveals that patients who had developed hypertension and diabetes were more likely to develop hyperlipidemia.

<table>
<thead>
<tr>
<th>Pattern No.</th>
<th>Patterns</th>
<th>Support in (%)</th>
<th>Confidence in (%)</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(Diabetes, Chronic kidney disease, unspecified) =&gt; (Hypertensive chronic kidney disease, unspecified, with chronic kidney disease stage I through stage IV, or unspecified)</td>
<td>13.20</td>
<td>85.51</td>
<td>3.766</td>
</tr>
<tr>
<td></td>
<td>(25000, 5859) =&gt; (40390)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>(Hypertensive chronic kidney disease, unspecified, with chronic kidney disease stage I through stage IV, or unspecified, Diabetes) =&gt; (Chronic kidney disease, unspecified)</td>
<td>13.20</td>
<td>58.13</td>
<td>3.766</td>
</tr>
<tr>
<td></td>
<td>(40390, 25000) =&gt; (5859)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>(Diabetes, Atrial fibrillation) =&gt; (Congestive heart failure, unspecified)</td>
<td>13.72</td>
<td>59.87</td>
<td>1.778</td>
</tr>
<tr>
<td></td>
<td>(25000, 42731) =&gt; (4280)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>(Coronary atherosclerosis of native coronary artery, Diabetes) =&gt; (Other and unspecified hyperlipidemia)</td>
<td>14.46</td>
<td>54.40</td>
<td>1.195</td>
</tr>
<tr>
<td></td>
<td>(41401, 25000) =&gt; (2724)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>(Diabetes, Esophageal reflux) =&gt; (Other and unspecified hyperlipidemia)</td>
<td>13.12</td>
<td>53.25</td>
<td>1.170</td>
</tr>
<tr>
<td></td>
<td>(25000, 53081) =&gt; (2724)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>(Diabetes, Unspecified essential hypertension) =&gt; (Other and unspecified hyperlipidemia)</td>
<td>22.42</td>
<td>50.55</td>
<td>1.110</td>
</tr>
<tr>
<td></td>
<td>(25000, 4019) =&gt; (2724)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 9. The most frequent co-occurrence patterns found from diabetic patients in MIMIC-IV dataset.

5.2.2. Finding Gender-specific comorbidities patterns in MIMIC-IV dataset

For finding the Gender-specific comorbidities patterns in MIMIC-IV dataset, we run the proposed model separately on 13,784 male diabetic patients and 12,331 female diabetic patients. Table 10 reports the top-4 most frequent co-occurrence patterns (comorbidities patterns) found from male diabetic patients in MIMIC-IV dataset. Table 11 shows the top-2 most frequent co-occurrence patterns (comorbidities patterns) found from female diabetic patients in MIMIC-IV. By observing Table 10 and Table 11, we infer that the comorbidities pattern like (diabetes, hypertension, hyperlipidemia) is common in both male and female diabetic patients. However, we found one interesting comorbidity pattern, especially in male diabetic patients, patients who had developed diabetes were more likely to develop chronic kidney disease.
We looked up whether these diagnosis codes (conditions) are known to coexist in the medical literature to see if our patterns were clinically valid. Table 12 lists for each conditions pair, the PubMed identifier of the research article that establishes the medical conditions between them. According to the medical literature, the majority of the condition pairs are familiar, confirming the claim that the proposed methodology sufficiently reveals real association among diagnosis codes.

**Table 12. Medical significance of association between pairs of condition grouped in comorbidities patterns found from the MIMIC datasets.**

<table>
<thead>
<tr>
<th>Condition 1</th>
<th>Condition 2</th>
<th>Supporting Literature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atrial fibrillation</td>
<td>Congestive heart failure</td>
<td>20347787 [64]</td>
</tr>
<tr>
<td>Hyperlipidemia</td>
<td>Essential hypertension</td>
<td>1927888 [65]</td>
</tr>
<tr>
<td>Coronary atherosclerosis</td>
<td>Disorder of kidney</td>
<td>24527682 [66]</td>
</tr>
</tbody>
</table>

5.2.3. Finding T1D and T2D comorbidity patterns in MIMIC datasets

We also investigated the T1D and T2D comorbidity patterns in MIMIC datasets, by running the proposed model separately on T1D and T2D patients. Table 13 reports the top-1 most frequent co-occurrence patterns (comorbidities patterns) found from T1D patients in MIMIC-III dataset. Table 14 shows the top-4 most frequent co-occurrence patterns (comorbidities patterns) found from T2D patients in MIMIC-III dataset. Similarly, Table 15 reports the top-2 most frequent co-occurrence patterns (comorbidities patterns) found from T1D patients in MIMIC-IV dataset. Table 16 shows the top-2 most frequent co-occurrence patterns (comorbidities patterns) found from T2D patients in MIMIC-IV dataset.

From the investigation we found that T1D patients were at more risk of developing Retinopathy and Polyneuropathy, which is unlikely among T2D patients. T2D patients were at more risk of developing hyperlipidemia, hypertension, Atrial fibrillation, etc. Here we have encoded all T1D ICD-9-CM codes to ‘25001’ and all T2D ICD-9-CM codes to ‘25000’.

**Table 13. The most frequent co-occurrence patterns found from T1D patients in MIMIC-III dataset.**

<table>
<thead>
<tr>
<th>Pattern No.</th>
<th>Patterns</th>
<th>Support in (%)</th>
<th>Confidence in (%)</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(T1D, Background diabetic retinopathy) =&gt; (Polyneuropathy in diabetes)</td>
<td>12.75</td>
<td>55.07</td>
<td>2.262</td>
</tr>
</tbody>
</table>

**Table 14. The most frequent co-occurrence patterns found from T2D patients in MIMIC-III dataset.**

<table>
<thead>
<tr>
<th>Pattern No.</th>
<th>Patterns</th>
<th>Support in (%)</th>
<th>Confidence in (%)</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(Other and unspecified hyperlipidemia, T2D) =&gt; (Unspecified essential hypertension)</td>
<td>14.73</td>
<td>57.56</td>
<td>1.160</td>
</tr>
<tr>
<td>2</td>
<td>(Coronary atherosclerosis of native coronary artery, T2D) =&gt; (Unspecified essential hypertension)</td>
<td>19.41</td>
<td>55.01</td>
<td>1.109</td>
</tr>
<tr>
<td>3</td>
<td>(Atrial fibrillation, T2D) =&gt; (Congestive heart failure, unspecified)</td>
<td>16.05</td>
<td>52.61</td>
<td>1.377</td>
</tr>
<tr>
<td>4</td>
<td>(Acute kidney failure, unspecified, T2D) =&gt; (Congestive heart failure, unspecified)</td>
<td>12.55</td>
<td>51.25</td>
<td>1.342</td>
</tr>
</tbody>
</table>
References


6. Conclusion and future scope

Today people with a sedentary lifestyle are prone to many diseases. One of the common ones is diabetes, the studies suggest people diagnosed with diabetes will have at least one other condition that can influence the self-management of diabetes and its progression. In this research work, we have studied the patterns of comorbidity among diabetes patients. We have found that the chronic kidney disease was the top most common comorbidity (Support: 13.20%; Confidence: 85.51%, Lift: 3.766), followed by atrial fibrillation (Support: 13.72%; Confidence: 59.87%, Lift: 1.778). Also, we found that the diabetes-specific disease comorbidities patterns are gender-specific and also major diabetes subtypes-specific, for example, the results revealed some novel, interesting information like male diabetic patients who had developed diabetes were more likely to develop chronic kidney disease, which is not found in female diabetic patients and another insight i.e., type 1 diabetes patients were at more risk of developing Retinopathy and Polyneuropathy, which is unlikely among type 2 diabetes patients and so on. Finally, extracted top-k comorbidities patterns are validated with medical literature. Our proposed approach is useful for building a recommender system that suggests health conditions that are expected to coexist with a patient's existing diagnosis to a clinical practitioner. Finally, as conditions linked with other disease are coded in a similar fashion within electronic health records, our proposed approach can be implemented to similar research concerns for diseases other than diabetes.


Chan, C. L., Chen, C. W., & Liu, B. J. (2008, June). Discovery of association rules in metabolic syndrome related diseases. In 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence) (pp. 856-862). IEEE.


Authors Profile

Bramesh S.M. is currently doing Ph.D. under the supervision of Prof. Anil Kumar K.M. in part-time, in the research center SJCE, Mysuru under the Visvesvaraya Technological University, Belagavi. He is working as an Assistant Professor, in the Department of Information Science & Engineering, P. E. S. College of Engineering, Mandya, Karnataka, India. He has teaching experience of 14 years and research experience of 5 years. His research interest includes Text mining, and Data mining. He has published nearly 8 Research papers in National and International proceedings.

Dr. Anil Kumar K.M. is currently working as an Associate Professor, in the Department of Computer Science & Engineering, JSS Science and Technology University, Mysuru, Karnataka, India. He did his postdoc at Deakin University under Professor Jemal Abawajy and Ph.D. from the University of Mysore under the supervision of Prof. Suresha, Chairman, DOS in Computer Science. He has teaching experience of 23 years and research experience of 16 years. His research interest includes Text mining, Sentiment Analysis, Data mining, Opinion mining, Web Mining, Data Analytics, Computer Networks, and Cyber Security. He has received 5 grants from different Government and Private funding agencies for Research & Development. He has Published nearly 50 Research papers in National and International proceedings.