Diabetes Case Reports
ISSN: 2572-5629
El-Sappagh and Elmogy, Diabetes Case Rep 2016,1:1
DOI: 10.4172/2572-5629.1000102
Research Article
Open Access

A Decision Support System for Diabetes Mellitus Management

Shaker El-Sappagh[1] and Mohammed Elmogy[2]*

[1]Faculty of Computers and Information, Minia University, Egypt
[2]Faculty of Computers and Information, Mansoura University, Egypt

Abstract

Diabetes mellitus is considered as a dangerous chronic disease. Diagnosis is the first step in its management. Clinical decision support system (CDSS) for diabetes diagnosis improves its detection and decreases the opportunity for its complications. However, its diagnosis is a theory-less problem. Case-based reasoning (CBR) is a problem-solving paradigm that uses past experiences to solve new problems. Integration of CBR and formal ontologies enhances the intelligence of this paradigm. Utilizing patients' electronic health records (EHRs) for building case-base knowledge solves the problem of knowledge acquisition bottleneck; however, preparation steps are required. Moreover, using standard medical ontologies, such as SNOMED-CT, enhances the interoperability and integration of CDSS with the healthcare system. If ontology-based CBR systems utilize vague or imprecise knowledge, the semantic effectiveness is further improved. This paper proposes an advanced and complete fuzzy-ontology-based CBR framework that manages and utilizes imprecise knowledge. We implement the most critical steps in CBR (i.e., case representation and retrieval). The implemented framework has been tested on the diabetes diagnosis problem using a case-base of 60 real cases from The EHR of the Mansoura University Hospitals, Mansoura, Egypt. The proposed system has an accuracy of 97.67%.

Keywords: Diabetes diagnosis; Clinical decision support system; Medical ontology; Case-based reasoning; Fuzzy ontology; Description logic

Introduction

Diabetes is one of the most serious chronic diseases. According to American Diabetes Association (ADA), it imposes a significant economic burden on the countries. Healthcare expenditures, due to diabetes, account for 11% ($465 billion) of the total healthcare expenses in the world in 2011 [1]. By 2030, this number is projected to exceed $595 billion. Worldwide, there are approximately 366 million people with diabetes and it is estimated that 552 million will be affected by 2030 [2]. WHO (*http://www.who.int/en/*) projects that diabetes is the seventh leading cause of mortality in 2030. How to decrease these threats is a critical issue. The early diagnosis is the first and most critical step in diabetes management process because it can prevent its long-term microvascular complications like retinopathy, nephropathy and neuropathy, and cardiovascular diseases. About 183 million people, or half of those who have diabetes, are unaware they have the disease [2]. The patient can be affected by diabetes for 9-12 years before being diagnosed [3]. As a result, at diagnosis time, complications often exist. According to ADA, if diabetes can be early diagnosed, the lifestyle, blood glucose control, and pharmacologic interventions are effective in controlling diabetes and reducing its related complications.

There are many clinical practice guidelines (*http://guidelines.diabetes.ca/fullguidelines/*) for standardizing the diagnosis process; however, these guidelines are long text documents, which are difficult to be used by a physician at the point of care. Many AI techniques have been utilized to enhance the diabetes diagnosis process such as rule-based reasoning and artificial neural network [4]. However, the results for early detection of diabetes based on these systems are not highly accurate. For example, rule-based systems are not suitable for the ill-formed, difficult to formulate, and experience-based problems. It can be difficult for an expert to transfer their knowledge into distinct rules, and many rules can be required to be valid for a system [5]. The management and maintenance of large rule-based are not an easy process. Moreover, Alves et al., [6] have asserted that neural networks are not the optimum choice for implementing diagnostic systems for medical problems. Using Clinical Decision Support System (CDSS) at the point of care integrated with the Electronic Health Record (EHR) system can improve the early detection of diabetes.

CBR is considered the most suitable AI technique for building experience-based CDSS systems [7]. It depends on collecting the previous experience of a medical expert in the form of cases and uses this knowledge for inference. CBR has many advantages for medical diagnosis problems. For example, the EHR stored data, such as symptoms, medical history, physical examinations, lab tests, diagnoses, treatments, and outcomes for each patient, can be used to define the case-base knowledge. This formulation solves the knowledge-acquisition bottleneck problem found in other AI techniques. Many researchers utilize CBR for diabetes diagnosis [8,9]. However, the diagnosis diabetes accuracy is still not encouraging. The purpose of this study is to provide a significantly advanced step in the CBR system developments.

Building the case-base knowledge and implementing an accurate case retrieval algorithm are the main tasks of diagnostic CBR systems because the retrieved case can be used directly without adaptation of the CDSS suggested decisions. However, building a complete and consistent case-base knowledge, which covers all patient medical cases, is a challenge. In addition, implementing a semantic retrieval algorithm, which measures the clinical distance between two cases, is another challenge. We propose in this paper a set of three preparation steps to convert EHR data into CDSS knowledge including data preprocessing, data encoding, and data fuzzification. The data preprocessing phase solves the following issues: handling of missing data, feature selection,

***Corresponding author:** Mohammed Elmogy, Faculty of Computers and Information Technology Department, Mansoura University, Egypt, Tel: 00201098889791; E-mail: melmogy@mans.edu

Received: January 15, 2016; **Accepted:** February 11, 2016; **Published**: February 15, 2016

Citation: El-Sappagh S, Elmogy M (2016) A Decision Support System for Diabetes Mellitus Management. Diabetes Case Rep 1:102. doi: 10.4172/2572-5629.1000102

feature weighting, data normalization, data aggregation, data discretization, coding of unstructured data, and outlier detection and prevention. This step is handled by a set of machine learning algorithms.

To enable (1) the integration between EHR and CBR systems, (2) facilitate the collection of case from different EHR sites, (3) support the creation of Knowledge-Intensive CBR (KI-CBR) systems, (4) support the creation of distributed CBR systems, and (5) support the implementation of semantic retrieval algorithms, the case-base knowledge needs the encoding process by using a domain unified medical terminologies. If these terminologies are modeled in a standard medical ontology such as SNOMED CT (SCT), LOINC, UMLS, DO, GO, or ICD, the resulting knowledge base (i.e., case-base) and its retrieval function will be improved significantly [10]. However, such ontologies are very huge, which affects the performance of the CBR retrieval algorithm. Creating a domain-specific reference set is required [11]. The semantic interoperability between CBR and EHR systems requires the storage of encoded case-base knowledge in a standard and portable data model. The most popular data model in the medical environment is the HL7 RIM v3(http://www.hl7.org/implement/standards/product_brief.cfm?product_id=77). As a result, a mechanism must exist to convert the case base structure into a standard form [12].

The third preparation step of EHR data is the fuzzification step. As Zadeh [13] argued much of the knowledge that humans acquire through experience be perception-based and thus subject to imprecision and inaccuracy. Such knowledge, when not treated in some suitable way that can consider and convey its inherent imprecision, usually leads to reduced effectiveness of the knowledge-based systems that use it. As some of the case-base knowledge need encoding, some case-base features, i.e., numerical features, need fuzzification steps. The fuzzified case-base knowledge will better represent patients and will enhance the similarity measures implementation.

The most recent advances in CBR systems implementation are based on ontology, and it creates KI-CBR systems. They can play many roles in CBR such as background domain ontology, case-base ontology, semantic similarity measurement, and for sharing and reusing of CBR knowledge. With respect to the diabetes diagnosis, researchers have made an effort towards diabetes ontology development [8]. Nevertheless, the literature of ontology-based CBR for diabetes is not rich with studies. Jaya and Uma [14] have examined the roles of ontologies in diabetes diagnosis CBR systems.

Crisp ontologies have proved its applicability in CBR environment; however, these ontologies cannot represent and reason about vague knowledge. Vagueness can be handled using fuzzy set theory to create a fuzzy ontology. The lack of representation of this knowledge in ontological form restricts the effectiveness of these systems because they did not take advantage of the reasoning capabilities that ontologies provide. The fuzzy ontology focuses on assigning a meaning to the fuzziness of the ontology's components. For diabetes, the existing fuzzy CBR systems have not used fuzzy ontology or even crisp ontology as background domain knowledge or case-base ontologies [8]. On the other hand, ontologies and fuzzy logic have been utilized in diabetes in other domains such as rule-based expert systems [15-17]. In other words, fuzzy ontologies have not been used in any medical CBR systems especially diabetes diagnosis, and there are no studies in the literature that proposed a medical fuzzy case-base ontology especially for diabetes diagnosis. This paper proposes a novel fuzzy KI-CBR framework that handles and exploits imprecise and encoded medical knowledge through the effective integration of fuzzy logic in the ontology-based CBR paradigm. Fuzzy case-base ontology, standard domain ontology, and a fuzzy semantic retrieval algorithm are proposed and integrated to build an intelligent CBR for diabetes diagnosis. At this end, the remainder of the paper is organized as follows: Section 2 discusses related work. Section 3 defines CBR. Section 4 is the dataset description. Section 5 discusses case structure. Section 6 is the proposed system. Evaluation is discussed in Section 7. Finally, Section 8 concludes the paper and highlights future work directions.

Related Work

CBR is one of the most suitable AI technique for ill-formed problems such as diabetes diagnosis [6,7]. Building effective CBR systems for these problems faces many challenges where all of them must be handled to build an intelligent CDSS system. For example, the CDSS system must be interoperable with the EHR system; it must support the collection of cases from the distributed healthcare environments in a meaningful form; it must support the sharing and reuse of knowledge. To achieve these goals, a standard case-base structure and contents must be prepared. Moreover, ontologies, standard medical terminologies, and fuzzy ontologies play critical roles in all phases of CBR system.

EHR contains the patient's current and history medical data. These data can be used as a complete source for building the CBR's case-base knowledge [18,19]. However, these transactional data must be carefully prepared because they are always incomplete, noisy, and unstructured [20,21]. As a result, data pre-processing steps are the first and the foremost to improve the accuracy of CBR systems [19]. The preparation steps must include EHR data contents and structure. *Contents preparation* includes data pre-processing, data encoding, and data fuzzification steps [21]. *Structure preparation* includes converting EHR database structure into a standard and unified case-base structure [10-12]. Abidi and Manickam [18] assumed that the structure of case-base is defined in advance, and they mapped the structure then contents of EHR to case-base. However, this assumption is not realistic. There are many studies to standardize the case base structure such as *HealthInfoCDA* [17]. However, the most famous standard for medical data storage and exchange is HL7 RIM v3. No studies have utilized this model in case base preparation. Moreover, Diabetes Data Strategy (*Diabe-DS*) project (http://wiki.hl7.org/index.php?title=EHR_Diabetes_ Data_Strategy) stated that standardization of EHR structure and content is not enough for diabetes. It has proposed a standard set of data elements for diabetes diagnosis. This standard set of representative elements has not been used yet in the implementation of any of the existing CBR systems.

Using ontology to represent the background domain knowledge supports the seamless integration between CBR system and EHR system, the implementation of semantically intelligent case retrieval algorithms, and the creation of encoded and standard case-base [11]. There are many studies for encoding the EHR contents in a standard and unified language [16]. In a medical environment, there are many standard medical ontologies in the literature, and SNOMED CT (http://ihtsdo.org/snomed-ct/) is the most comprehensive one. It contains more than 388,000 active medical concepts organized in 19 hierarchies, 1.14 million descriptions, and 1.38 million relationships. Silva et al., concluded that SCT is the most suitable ontology for coding of problem lists and diagnosis. The encoding process of medical data, using the standard ontology, requires an encoding methodology. There are some existing methodologies as the oneproposed by Lee et al., [20]. However, Lee's approach concentrated on the data cleaning,

normalization, and matching steps, and has not mentioned the physical storage structure of the data such as EAV; moreover, it has not defined whether the EHR data model is standardized using RIM or not. As a result, a new methodology needs to be created. El-Sappagh et al., [10] proposed an encoding methodology, which covers the existing methodologies limitations. Moreover, because SCT is a very massive ontology, a small fragment of a particular domain has to be established to enhance the retrieval algorithm performance [11]. There is no existing SCT reference set for diagnosis diabetes concepts. As a result, a method has to be defined for extracting the diabetes concepts from SCT and converting the resulting set into OWL 2 ontology. This ontology will be used by the retrieval algorithm to calculate the clinical distance between patients. El-Sappagh et al., [11] proposed a methodology for extracting diabetes concepts from SCT and converting it into an OWL 2 ontology.

Regarding the role of ontology in diabetes CBR systems, in the diabetes domain, ontology has been used in many CDSSs [8]. For example, Chen et al., [22] introduced an ontology for diabetes drugs and an ontology for patients' symptoms. These ontologies utilize Semantic Web Rule Language (SWRL) and Java Expert System Shell (JESS) to determine potential prescriptions for the patients. Rahimi et al., [23] developed a Type 2 Diabetes Mellitus (T2DM) Ontology (DMO) to diagnose and manage patients with diabetes. They proposed an algorithm to query the ePBRN data repository to diagnose T2DM. Sherimon et al., [24] proposed a dynamic adaptive questionnaire ontology for gathering the diabetic patient's medical history. Hayuhardhika et al., [25] developed an ontology of diabetes disease and used a weighted tree similarity algorithm for diagnosis. However, regarding diabetes diagnosis, none of these ontologies is designed for CBR, and few studies have used ontology in CBR [9]. In diabetes diagnosis systems, ontologies have not been utilized in neither case-base nor background knowledge or case retrieval. El-Sappagh et al., [26] proposed a diabetes diagnosis case-base OWL 2 ontology. This crisp ontology can be used to store and retrieve cases semantically. Nevertheless, an issue that the ontology-based CBR paradigm has not yet addressed is that of knowledge imprecision [27]. Medical data, such as diagnosis diabetes data, are mostly imprecise and experience-based; the success of CBR in this domain depends on how this issue is handled [13].

Fuzzy logic has a great role in diabetes CBR systems because medical data are imprecise in nature [21]. If it is not treated in some suitable way that can consider and convey its inherent imprecision, usually this leads to reduced effectiveness of the knowledge-based systems that use it. Fuzzy sets have been integrated with CBR to generate Fuzzy-CBR in many studies [28], and used for calculating the fuzzy similarity between cases [29]. Recently, Sohn et al., [30] integrated fuzzy-CBR reasoning with crisp ontology reasoning for personalized service in a smart home environment. However, this hybrid system has not benefited from fuzzy ontology reasoning capabilities in CBR system. There are no real studies in the literature for fuzzy-CBR systems for diabetes diagnosis. On the other hand, fuzzy logic has been utilized to build diabetes diagnosis CDSS using other AI techniques such as rule-based [15].

After the success of crisp ontologies in CBR environment, fuzzy ontologies can extend its benefits by integrating ontology reasoning with fuzzy reasoning capabilities [31]. For example, the physician can more easily define experience cases using natural-like language, cases can be indexed more efficiently, and finally fuzzy-semantic retrieval algorithms can be implemented. There are 17 formal definitions for fuzzy ontology [31]. One definition is an ontology that uses fuzzy logic to provide a natural representation of imprecise and vague knowledge and eases reasoning over it. Building a case-base fuzzy ontology is a challenge. Fuzzy Ontologies have been used in many non-medical CBR systems. For example, Alexopoulos et al., [27] proposed a fuzzy case-base ontology by utilizing fuzzy algebra. With respect to diabetes, it has utilized fuzzy ontologies in many domains such as [15]. Lee and Wang [15] proposed a five-layer fuzzy ontology and utilized it in a fuzzy expert system for diabetes management. As stated before, CBR is the most suitable mechanism for managing ill-formed problems as diabetes diagnosis. However, to the very well of our knowledge, there are no efforts in this direction. There is no fuzzy ontology-based CBR for diabetes, there are no studies for formal case-base fuzzy ontology construction, and there are no similarity measures, which utilize fuzzy description logic. Crisp ontologies are not suitable to address imprecise and vague knowledge, which is inherent in the real world domains [13]. The integration of fuzzy logic, CBR, and ontology generates Fuzzy-KI-CBR, which is a yet unstudied topic in the medical domains.

This study tries to build a diabetes diagnosis CDSS system based on CBR technique. The proposed framework will handle all the challenges previously identified. This framework is implemented using JCOLIBRI APIs (http://gaia.fdi.ucm.es/research/colibri/jcolibri) and other semantic APIs related to semantic programming (e.g., OWL API (http://owlapi.sourceforge.net/, http://protege.stanford.edu/)). Our ontologies are built using protégé 4.3. Finally, our proposal is evaluated and compared with other studies.

Case based reasoning

Generally, CBR is an AI technique for solving a problem by remembering similar past experiences [32]. For example, physicians look for groups of known symptoms and engineers take many of their ideas from previously successful solutions. The main concept of CBR is *"similar problems have similar solutions"* [32]. CBR knowledge is formed in a case-base of previous experiences (either success or failure). It does not depend on the explicit model of the problem as in rule base reasoning for the inference process, but it simply utilizes the experience captured in the same way the expert usually inputs and processes it. The newly solved problems can be added as a new experience in the CBR system's experience-base (case-base), which supports the auto-learning process. The CBR can be defined as a cyclic process named "the four Rs" [32], (figure 1): (i) *Retrieve* the most similar cases, (ii) *Reuse* the cases that might solve the problem, (iii) *Revise* the proposed solution if necessary, and (iv) *Retain* the new solution as part of a new case. The most important aspects of CBR system are the case-base knowledge representation and the case retrieval algorithm, and these are our contributions in the current paper.

A case-base CB is a finite set of cases $\{C_1, C_2, \ldots, C_m\}$, where m is the number of cases in the *CB*. *A case* is a case is a contextualized piece of knowledge representing an experience. The i^{th} experience case

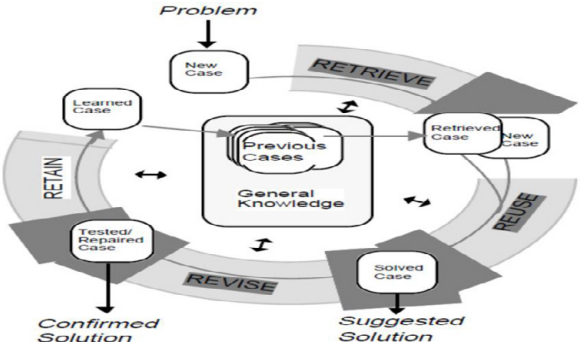

Figure 1: The CBR Cycle.

$C_i \in CB$ is formally defined As $C_i = \prec P_i, S_i \succ$, where P_i and S_i respectively represent the case problem description and the case solution features. *A case retrieval algorithm* is an algorithm that takes as input (query case or the new patient description C_q, case base or the domain expert knowledge *CB*, and features weighting vector $\vec{W}$); it calculates the level of similarity between C_q and every case in *CB*; and finally it returns the solution of the most similar cases. The k-nearest neighbour (k-NN) algorithm is the most applicable retrieval technique. Case base creation and semantic retrieval algorithm design are the most critical steps for CBR systems success.

Diabetes Dataset Description

The contents of a case-base must be defined in the first beginning of a CBR system. These contents determine all of the subsequent steps such as case-base ontology, case base fuzzy ontology, and case retrieval. After checking with the domain experts, CPGs, and handbooks of case histories in diagnosis diabetes domain, our case will contain the features described in Table 1. The paper uses a dataset from the diagnostic biochemical lab, AutoLab of Mansoura institution, Mansoura University, Mansoura, Egypt. This data was collected in the period from January 2010 through August 2013. The control subjects were healthy and recruited from the diagnostic biochemical lab and were matched by age, sex and ethnicity to the case subjects. The eligibility criteria for controls were the same as those for patients, except for having a cancer diagnosis. A short structured questionnaire was used to screen for potential controls based on the eligibility criteria. Analysis

Feature type	Feature name		Data type	Normal range	UoM	Min-mean-max	#
Demographics	Residence		P, C	{Urban, Rural}	-	-	1
	Occupation		P, C	{Farmer, Police…}	-	-	2
	Gender		P, C	{Male, Female}	-	-	3
	Age		P, N, F	-	year	29-48.117-74	4
	BMI		P, N, F	18.5 - 25	kg/m^2	20-33.117-45	5
Diabetes lab tests	HbA1C		P, N, F	<=5	mmol/L	5-6.373-7.4	6
	2h PG		P, N, F	< 139	mg/dl	165-202.733-235	7
	FPG		P, N, F	< 99	mg/dl	96-129.633-156	8
Haematological profile	Prothrombin INR		P, N, F	0 - 1	%	1-1.16-1.4	9
	Red cell count		P, N, F	4.2 - 5.4	10^6/cmm	3.8-5.194-5.88	10
	Hbg		P, N, F	12 - 16	g/dL	9.8-12.332-13.4	11
	Haematocrit (PCV)		P, N, F	37 - 47	vol%	31.1-35.215-36.8	12
	MCV		P, N, F	80 - 90	fl	26.8-71.908-76.4	13
	MCH		P, N, F	27 - 32	pg	3.3-25.47-29.4	14
	MCHC		P, N, F	30 - 37	%	1.8-35.465-41.7	15
	Platelet count		P, N, F	150 - 400	10^3/cmm	135-316.183-2000	16
	White cell count		P, N, F	4 - 11	10^3/cmm	6-8.055-9.2	17
	Basophils		P, N, F	0 - 1	%	0-1.013-5	18
	Lymphocytes		P, N, F	20 - 45	%	21.2-25.768-29	19
	Monocytes		P, N, F	2 - 10	%	1.7-2.942-4	20
	Eosinophils		P, N, F	1 - 4	%	1-1.897-3.4	21
Symptoms	Urination frequency		O	-	-	-	22
	Vision		O	-	-	-	23
	Thirst		O	-	-	-	24
	Hunger		O	-	-	-	25
	Fatigue		O	-	-	-	26
Kidney Function Lab tests	Serum potassium		P, N, F	3.5 - 5.3	mEq/L	2.4-3.767-4.3	27
	Serum urea		P, N, F	5 - 50	mg/dL	17-31.56-67	28
	Serum Uric acid		P, N, F	3.0 - 7.0	mg/dL	3-4.237-7.9	29
	Serum creatinine		P, N, F	0.7 - 1.4	mg/dL	0.9-1.35-3.6	30
	Serum sodium		P, N, F	135 - 150	mEq/L	134-137.833-158	31
Lipid profile	LDL cholesterol		P, N, F	0 - 130	mg/dL	50-94.917-170	32
	Total cholesterol		P, N, F	0 - 200	mg/dL	158-209.367-275	33
	Triglycerides		P, N, F	60 - 160	mg/dL	78-144.767-189	34
	HDL cholesterol		P, N, F	45 - 65	mg/dL	30-55.533-65	35
Tumor markers	FERRITIN		P, C	28 - 397	ng/mL	-	36
	AFP Serum		P, C	0.5 - 5.5	IU/ml	-	37
	CA-125		P, C	1.9 - 16.3	U/mL	-	38
Urine analysis	Chemical examination	Protein	O	-	-	-	39
		Blood	O	-	-	-	40
		Bilirubin	O	-	-	-	41
		Glucose	O	-	-	-	42
		Ketones	O	-	-	-	43
		Urobilinogen	O	-	-	-	44
	Microscopic examination	Pus	O	-	-	-	45
		RBcs	O	-	-	-	46
		Crystals	O	-	-	-	47

Liver function tests	S. albumin	P, N, F	3.5 – 5.0	g/dL	1.9-4.082-5.4	48
	Total bilirubin	P, N, F	0.0 – 1.0	mg/dL	0.8-1.317-3	49
	Direct bilirubin	P, N, F	0.0 – 0.3	mg/dL	0.3-0.533-1.6	50
	SGOT (AST)	P, N, F	0 – 40	U/L	35-54.567-165	51
	SGPT (ALT)	P, N, F	0 – 45	U/L	35-57.317-183	52
	Alk. phosphatase	P, N, F	64 - 306	U/L	170-214.2-360	53
	γ GT	P, N, F	7 – 32	U/L	18-35.833-98	54
	Total protein	P, N, F	6.0 – 8.7	g/dL	3.1-4.858-8.7	55
Females history	Amenorrhea	I	-	-	-	56
	Birth	I	-	-	-	57
	Dysmenorrhea	I	-	-	-	58
Diagnosis	Diabetes type	P, C	-	-	-	59
Nephropathy	Nephropathy check	I	-	-	-	60
Lipid disease	Hypercholesteremia's check	I	-	-	-	61
Cancer type	Tumor markers	I	-	-	-	62
Liver disease	Liver problem	I	-	-	-	63
Radiological examination	Radiological examination	I	-	-	-	64

Data type= {P=primitive, I= instance of SCT concept, N=numerical, C=categorical, F=fuzzy, O=ordinal}

Table 1: The patient attributes used to describe cases.

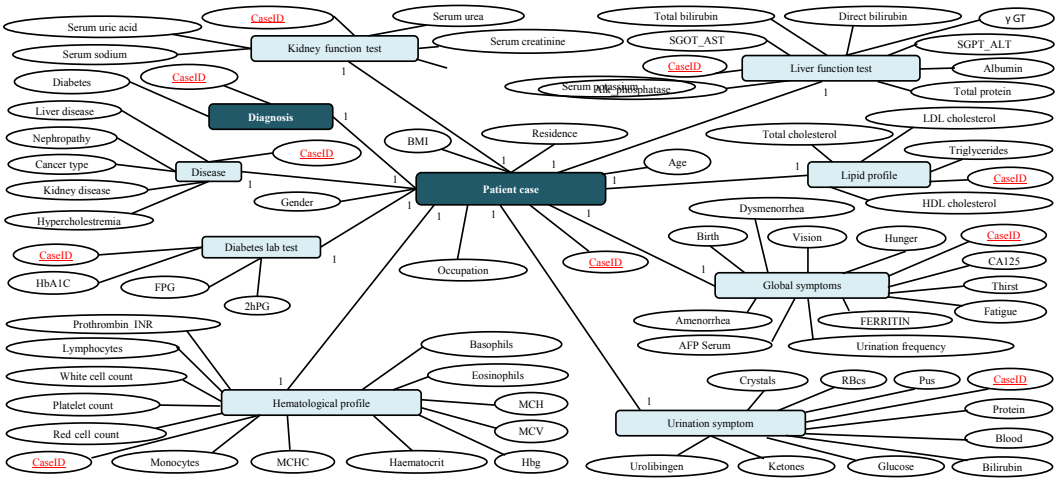

Figure 2: Diabetes diagnosis and other related complaints case base data model.

of the answers received on the short questionnaire indicated that 80% of those questioned agreed to participate in clinical research. A total of 67 eligible subjects were ascertained in the current study. However, seven control subjects were excluded due to limited blood samples for testing AFP. Blood samples (5 mL) were taken, centrifuged, and the serum separated and stored at 220 uC until analyzed. Serum samples were assayed for AFP by enzyme-linked immunosorbent assay with commercial kits (Abbott, North Chicago, IL), transferase (ALT) and aspartate aminotransferase (AST), with an auto-analyzer (Hitachi Model 736, Japan) and commercial kits. Our data set contains 70 features for describing diabetic patients and for linking diabetes with other disorders such as cancer, kidney diseases, and liver diseases. The data set is distributed as 33.3% pre-diabetic patients, 53% diabetic patients, and 13.7% normal patients.

The Structure of a Diagnosis Diabetes Case

Figure 2 shows an Extended Entity-Relationship (EER) model for all entities and attributes used in our data set. This data model is compatible with HL7 RIM. This compatibility facilitates the integration with EHR and supports the auto collection of cases. Moreover, this data model has been fuzzified with our proposed fuzzification methodology into a fuzzy EER model, and then converted to a fuzzy case-base database, which was the source of instances for our proposed fuzzy case-base ontology. These entities and attributes were enriched by entities and attributes in diabetes diagnosis CPGs as in the National Guidelines Clearing House (http://www.guideline.gov/). Entities and features related to diabetes treatment, medications, and drugs are out of scope.

Diabetes diagnosis cases are defined according to our data model. A case $C = \prec P, S \succ$ is defined as follows: P = *{LFT, LP, GS, A, B, R, G, O, KFT, LT, US, HP, DI}*

where *LFT*= liver function tests, *LP*= lipid profile, *GS*= global symptoms, *A*=age, *B*=BMI, *R*= residence, *G*=gender, *O*=occupation, *KFT*= kidney function tests, *LT*= lab tests, *US*= urination symptoms, *HP*= haematological profile, and DI= { *L* + *N* + *C* + *H* } where *L*= probable liver problem, *N*= probable nephropathy problem, *C*= probable cancer type, and *H*= probable hypercholesterolemia problem. *S (P)* is the solution part describes the diagnosis of diabetes including diabetic, prediabetic, gestational-diabetic, and prediabetic-gestational.

$S = \{DD\}$

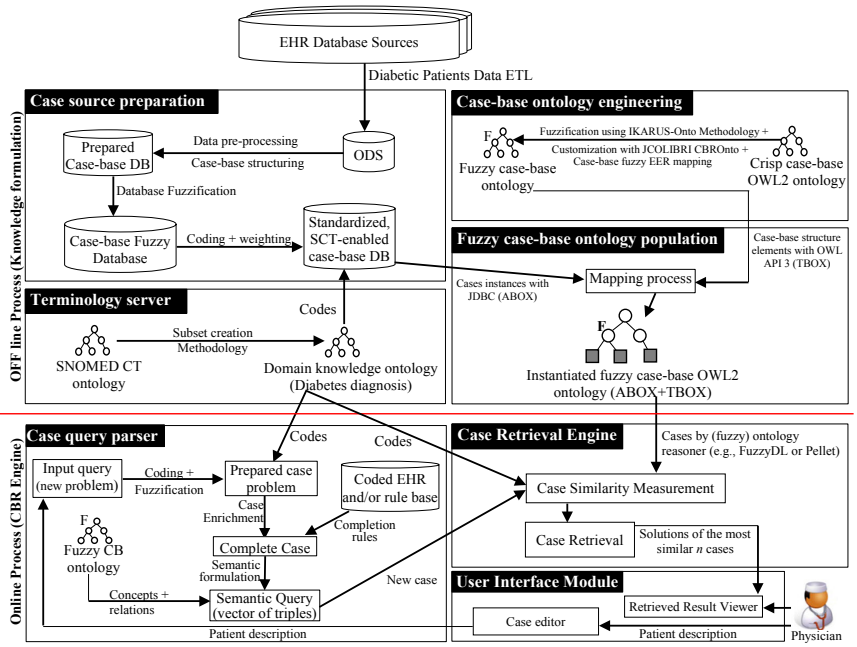

Figure 3: The proposed CBR framfvggework.

where DD= diabetes diagnosis. Our diagnostic features can be numerical features (e.g., age, lab tests, BMI, etc.), ordinal features (e.g., features in Global_symptoms table in Figure 2), and text features (e.g., sex, occupation, etc.). All these features have not been encoded in SCT concepts because their coding will not enhance the semantic retrieval algorithm of CBR. On the other hand, patient disorders are instance R semantic retrieval aspect. For example, if feature HbA1c=6.4 is encoded in SCT as |*43396009: Hemoglobin A1c measurement*|= 6.4, this code enhances semantic interoperability but does not efeatures, and we have mapped it to standard SCT concepts in another work [11]. We concentrated on the CBnhance semantic retrieval process in CBR. On the other hand, if the patient has a disorder such as a nephropathy, this concept has a long sub-tree of disorders (e.g., caliectasis, amyloid nephropathy, calyceal fistula, etc.), which can be described by different physicians. The semantic similarity of these concepts is critical in KI-CBR retrieval engine. Moreover, the case solution features are not encoded because these features do not participate in measuring the similarity between cases.

The Proposed Fuzzy KI-CBR Framework for Diabetes Diagnosis

This section provides a description of our proposed fuzzy-ontology based CBR system for diabetes diagnosis. The architecture of this system is shown in Figure 3. It has seven modules: Case source preparation, case base ontology engineering, terminology server, fuzzy case-base ontology population, case retrieval engine, case query parser, and user interface modules. The next sections describe the architecture of the proposed framework for details.

Case source preparation module

This module prepared the EHR raw data to a case-base structure and content. It collected the patient's features related to a diabetes diagnosis from distributed EHR systems and stored it in an Operational Data Store (ODS).

We have collected 60 cases, which describe diabetic patients, as shown in Table 1. These cases are descriptive of all types of cases. Next, these data were anonymized, cleaned, and normalized. Features' weights were calculated using machine learning algorithms including genetic algorithm, decision tree, and others. El-Sappagh et al. [21] proposed a case-base preparation process and applied it to the used case-base data. Moreover, the data were converted to a case base structure using our proposed standard data model [12]. In addition, the prepared case-base was coded according to SCT reference set that was created, which is specialized for diabetes diagnosis [11]. Finally, the encoded case-base was fuzzified in a fuzzy relational database as shown in Figure 4 according to the fuzzy features. Figure 5 shows a small sample of the fuzzified features. The resulting database is the source of instances (i.e., ABOX) for our proposed fuzzy case-base ontology.

Terminology server module

This module creates the domain background knowledge-ontology. This knowledge is critical in two places: (1) in semantic similarity measurement and (2) in semantic query formulation. The domain knowledge ontology can be built locally, or it can depend on a standard medical ontology such as SCT. Unfortunately, ontologies are typically created in an ad-hoc manner, which may influence the accuracy of the similarity calculations [14]. The second choice is better because existing clinical ontologies are mature, and they include all required medical concepts and relationships in a standard and globally agreed form. This standardization enhances the interoperability, reuse, sharing, and integration with the EHR environment. SCT was the terminology used in this study. Using the whole SCT in CBR affects the retrieval algorithm because it is a very large ontology. We have proposed a framework for collecting all diabetes diagnosis related concepts from SCT, and built its OWL 2 ontology (TBOX). Figure 6 shows a snapshot of the created ontology from protégé 4.3 [11]. This ontology only contains 550 concepts. Calculating semantic similarity using JCOLIBRI API depends on concept instances; however, SCT contains only concepts. We have solved this problem by creating an instance for each concept with

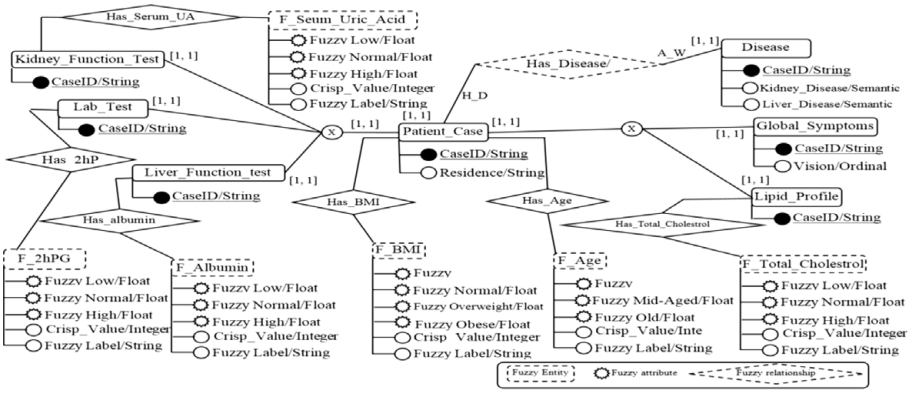

Figure 4: The extended Case base fuzzy EER model.

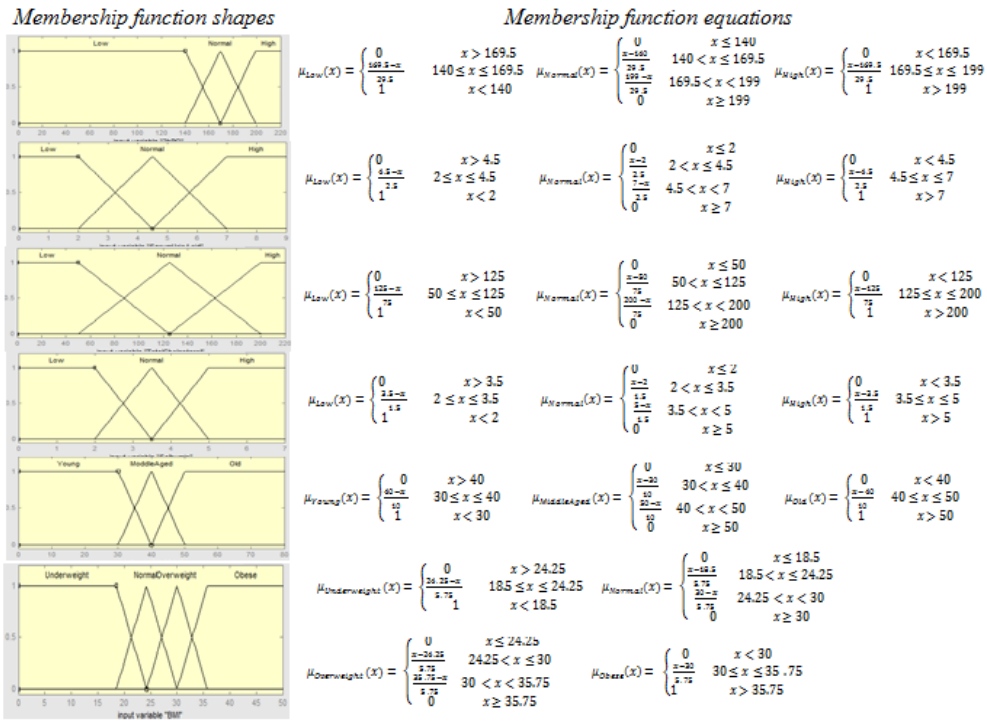

Figure 5: The numerical attributes fuzzification.

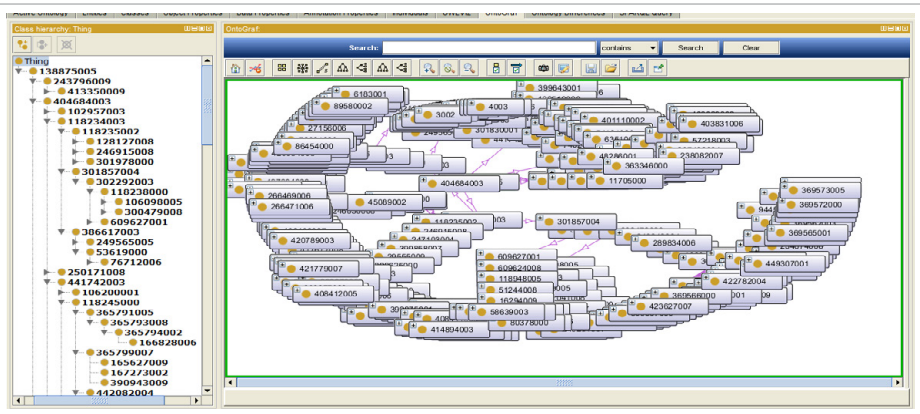

Figure 6: The proposed SCT reference set for diabetes diagnosis concepts.

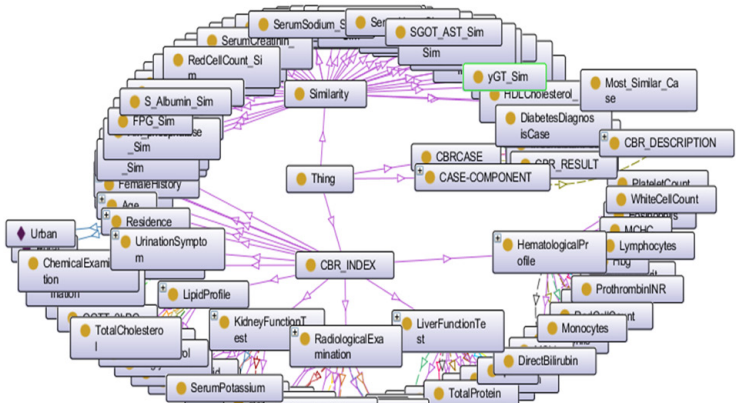

Figure 7: A snapshot of case-Base Fuzzy OWL 2 ontology.

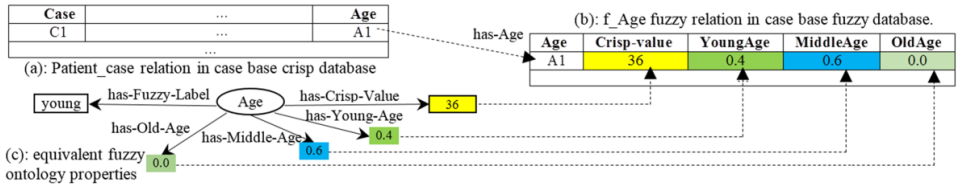

Figure 8: The mapping between elements of fuzzy data and fuzzy ontology.

the same name (ABOX). Moreover, we have represented the selected concepts using its ConceptIDs. Fully specified names, symptoms, and preferred names can be added as annotations with their corresponding names. As shown in Figure 3, this ontology is not user readable. We will resolve this issue in our future work. Each concept name begins with the pattern "C_" to be readable by JCOLIBRI API (http://gaia.fdi.ucm.es/research/colibri/jcolibri) as a concept and differentiate it from instances. The resulting ontology is a Directed Acyclic Graph (DAG), which supports single inheritance only. An ontology has a structured format with relationships between concepts. The "*IS_A*" relationship between a parent and a child is the core relationship, whereas other semantic relationships provide additional associations between terms (such as "*part-of*" or "*active-ingredient-of*"). Our ontology concentrates on the "*IS_A*" relationship only to form a taxonomy of concepts. Enriching the ontology with other relationships and axioms will be considered in future work.

Case-base ontology engineering module

This module has two basic steps: (1) the construction of case base crisp ontology. We propose a diabetes diagnosis ontology engineering methodology [26], (2) the extension of this ontology to a fuzzy ontology. According to our previously created case bases fuzzy database contents and the CBROnto standard case-base ontology of JCOLIBRI 2, we extend the case-base crisp ontology to a fuzzy ontology shown in Figure 7.

The crisp ontology elements that can be fuzzified include datatypes, object properties, and data properties. In other words, the fuzziness of ontology includes modeling of [10]: *(1) Fuzzy concepts:* concepts whose instances may belong to it in certain degrees, such as YoungPatient. Because Young is a vague predicate, the concept is also vague and, therefore, can be represented as a fuzzy one; it allows the fuzzy concept assertions such as Patient X be an instance of YoungPatient to a degree of 0.7.

(2) Fuzzy relations*:* there are two main types, (2.1) (Modified) Fuzzy object relations, which link concept instances at a certain degree, and it allows fuzzy role assertions as Patient X (very) has-Disease Y at a degree of *0.8*. (2.2) (Modified) Fuzzy data type relations, which either assign literal value to concept instances at certain degrees (e.g., Patient X has-Residence "Rural" at a degree of 0.4), which includes the Residence fuzzy predicate, or a fuzzy datatype is assigned to a concept instance (e.g., Patient X has-Age (very) young), which includes the Age fuzzy predicate.

We apply the procedural steps of IKARUS-Onto [33] methodology, and the resulting ontology is represented by Bobillo and Straccia syntax as OWL 2 ontology using Fuzzy OWL2 2.1.1 plug-in in Protégé 4.1 [31]. The resulting ontology contains 104 classes, 59 (fuzzy) object properties, 141 fuzzy datatype properties, 105 fuzzy datatypes, 1350 axioms, 736 logical axioms, and 2640 concept instances for the 60 real world diabetes-diagnosis patient cases.

Case-base ontology population

We have created a fuzzy database for the proposed fuzzy EER model and filled it with 60 cases of diabetic patients. These data have been collected from the EHR of 60 patients in Mansoura University Hospitals, Mansoura, Egypt [10]. This database is the source for populating our proposed fuzzy case-base ontology. The population process of database tuples to ontology instances is shown in Figure 8.

Case query parser module

For a new patient diagnosis problem, the physician enters the new patient description in the query form; this forms the new case without a solution. Next, the query is fuzzified and coded with the same methods used for the case-base ontology to facilitate similarity and mapping. The new problem structure is transformed into the fuzzy case-base ontology vocabulary by some strategy; then, the semantic query is sent to the *Case Retrieval Engine* to compute the similarity between the query concepts and the concepts of the new semantic-query problem.

For example, by using a small fragment of patient features, let the new patient is described by *Q= <Age=38, Residence= "Rural", Fatigue=*

"++", Gender= "Male", disease= "Malignant tumor involving left ovary by direct extension from endometrium" …>. After fuzzification, *Q= < **(young= 0.2, middleAged=0.8, old= 0, fuzzyLabel= middleAged, Age=38)**, Residence= "Rural", Fatigue= "++", Gender= "Male", disease= "Malignant tumor involving left ovary by direct extension from endometrium"…>*. After encoding, *Q= < (young= 0.2, middleAged=0.8, old= 0, fuzzyLabel= middleAged, Age=38), Residence= "Rural", Fatigue= "++", Gender= "Male", **disease= "369524001"** …>*. The other ordinal and categorical features remain the same. The vector *Q* needs to be transformed into a semantic query.

Case retrieval engine module

In case-base fuzzy ontology, the cases are displayed as concept instances and their features as relations and properties. The fuzzy semantic case retrieval algorithm utilizes the structure and content of the ontology to calculate the semantic similarity between the features and consequently for the cases. This section proposes a case retrieval algorithm. It involves the combination of reasoning capabilities of classical ontologies (e.g., semantic similarity of concept features storing SCT concepts) with fuzzy ontologies (e.g., fuzzy semantic similarity for other features) in order to create a powerful hybrid reasoning mechanism. Cases are commonly expressed as "*case = (problem, solution).*" Consider a query case $C_q = P_q$, ? ,stored cases $C_i = P_i, S_i$ for $i = 1,\ldots,n$

and n is the number of cases in the case-base, and feature weights w_i. Case retrieval module calculates the similarity between C_q and C_i for $i = 1,\ldots,n$ and return cases with highest similarity. Similarity calculation involves calculating local similarity between features and aggregates these similarities using a global similarity function. Local similarity depend on feature types. We propose custom functions for the following feature types:

For nominal features *(e.g., Gender)*, the exact match is used as in Equation 1.

$$sim_{NOM}(a,b) = \begin{cases} 1 & if\ a = b \\ 0 & if\ a \neq b \end{cases} \quad (1)$$

For ordinal features *(e.g., Urination frequency)*, our domain expert proposes a *similarity matrix* for each ordinal feature; $Sim_0(a,b)$ is calculated based on this matrix. Due to space restrictions, we do not show any matrices.

For fuzzy features *(e.g., HbA1c)*, we have two options:

(1) The feature value is numerical. Our proposed fuzzy similarity measure utilizes all of the fuzzy sets of compared features in calculating similarity. As the case-base fuzzy ontology store case with fuzzified features, the input query numerical features is fuzzified using the same fuzzy sets, and a comparison are conducted between stored and fuzzy query values. The normalized Euclidean distances between fuzzy sets of a feature are used to calculate the similarity as in Equation 2.

$$\mathrm{Dist}_F(C_j, Z_j) = \frac{\sqrt{\sum_{k=1}^{n}(\mu_{cj_k} - \mu_{zj_k})^2}}{\sqrt{n}} \quad (2)$$

Where C_j = crisp value of a feature in query, Z_j= crisp value of a feature in a stored case, *n*= number of fuzzy sets for feature f, μ_{cj_k} *and* μ_{zj_k} are k's fuzzy values for query and stored cases' feature, respectively. The similarity is calculated using Equation3.

$$\mathrm{sim}_F(C_j, Z_j) = 1 - \mathrm{Dist}(C_j, Z_j) \quad (3)$$

- ***For crisp numerical features***, the similarity is calculated using Equation 4:

$$sim_N(C_j, Z_j) = 1 - \frac{|D_j - Z_j|}{Max - Min} \quad (4)$$

- ***For semantic features*** (i.e., features SCT store concepts), the similarity is calculated using Equation 5.

Semantic similarity between our proposed SCT ontology instances measures the similarity in meaning between these instances.

$$SIM_{Semantic}(u,v) = w_1\, sim_{Path}(u,v) + w_2 sim_{feature}(u,v) \quad (5)$$

Where $w_1, w_2 \epsilon (0,1]$ are weights for $w_1 + w_2 = 1$, and $Sim_{path}(u,v)$ (Equation 6) is an adapted version of Wu and Palmer [34] (Equation 7) because $Sim_{wu\ and\ palmer}(u,u) < 1$ which violates the Identity Of the Indiscernibles property (IOI) [27].

$$sim_{Path}(u,v) = \begin{cases} 1 & if\ u = v \\ sim_{Wu\ and\ palmer} & otherwise \end{cases} \quad (6)$$

$$sim_{Wu\ and\ palmer}(u,v) = \frac{2 * depth(lca(u,v))}{shortest\_path(u, lca(u,v)) + shortest\_path(v, lca(u,v)) + 2 * depth(lca(u,v))} \quad (7)$$

In addition, $sim_{Feature}(u,v)$ is based on Batet et al., [34], Equation 8 and Equation 9:

$$sim_{Feature}(u,v) = 1 - Dist_{Batet}(u,v) \quad (8)$$

$$Dist_{Batet}(u,v) = log_2\left(1 + \frac{|A(u) \setminus A(v)| + |A(v) \setminus A(u)|}{|A(u) \setminus A(v)| + |A(v) \setminus A(u)| + |A(u) \cap A(v)|}\right) \quad (9)$$

Where $A(u)$ is the set of ancestors of *u*, i.e., $A(u) = \{v \mid u \leq v\}$ $A(u) \setminus A(v)$ is specificity of *u*, and $A(u) \cap A(v)$ is the commonality between *u* and *v*. Our proposed measure calculates the clinical similarity between two concepts rather than the semantic distance. Global similarity is calculated using the Euclidean distance function.

Evaluation of the P0roposed CBR System

Each module of the proposed system is separately evaluated upon completion, and the completely integrated system is evaluated. The proposed framework is the first to integrate the capabilities of standard medical ontologies (i.e., SCT), fuzzy logic, ontology, and CBR in a hybrid system. Our proposed ontologies (i.e., SCT refset ontology, case base crisp ontology, and case base fuzzy ontology) have been tested to check their consistency using several reasoners including Pellet, FaCT++, HermiT, and fuzzyDL. Moreover, for testing the ontology correctness, we have used the online tool *OOPS!* (http://oops.linkeddata.es/) *Pitfall Scanner* to detect potential modeling errors; results indicated no critical errors. Content coverage has been checked for each ontology with the domain experts.

The system has been tested using *Leave-One-In* technique. It measures the accuracy of the system to retrieve an existing patient case. Our system has an accuracy of 100% in this regard. Next, the system's decisions are compared with expert domain decisions. We have applied this study using a case-base containing 60 cases from EHR of Mansoura University Hospitals. Our method shows promising results. We used the *leave-one-out* technique to measure the performance for non-existing cases.

Namely, cases are taken out from the case-base one by one, and we have computed the similarity of this case with all the remaining cases in the case-base. It is a particular case of cross-validation. The domain experts evaluate the performance of the implemented framework by

System decision		Domain expert decision	
		Positive	**Negative**
	Positive	TP	FP
	Negative	FN	TN

TP = the *CBR system* decides the diabetic case, and *domain expert* decides a diabetic case.
FP = the *CBR system* decides a diabetic case, but the *domain expert* do not.
FN= the *CBR system* decides not a diabetic case, but the *domain expert* decides it be diabetic.
TN= the *CBR system* decides not a diabetic case and the *expert decides* not a diabetic case.

Table 2: The 2 × 2 confusion matrix.

System decision		Domain expert decision	
		Positive	**Negative**
	Positive	27	0
	Negative	1	15

Table 3: Diabetic decision confusion matrix.

organizing a set of 43 experiments. The test cases are selected in a manner that allowed them to span the majority of topics and content represented in the case base. Each test query is fed into the system, and the corresponding response was recorded. The proposed system's decisions are compared with the expert domain ones.

The semantic performance of the system is 97.67%, compared to 66% using Node Distance (ND) metrics only, 79% using IC similarity metric only, and 82% using a combination of both IC and ND . Table 2 is a 2 × 2 confusion matrix to calculate the evaluation metrics of our system. For *Diabetic* decisions only, the values of TP, FP, FN, TN can be interpreted as shown in Table 2.

The above parameters can be evaluated for Pre-diabetic and Normal as well. For space restrictions, we calculate Precision (P), Recall (R), Accuracy (A), Sensitivity (S), Effectiveness (E), and Negative Prediction Value (NPV) for *Diabetic* decisions only as follows. The metrics E and NPV are calculated using Equations. 10, 11:

$$Effectiveness(E) = F - Measure(Score) = \frac{1}{(1/2P)+(1/2R)} \quad (10)$$

$$Negative\ Prediction\ Value(NPV) = \frac{TN}{TN+FN} \quad (11)$$

From the performed experiments, we have calculated the values in Table 3 for the proposed system.

The P, R, A, S, E regarding diabetic diagnosis are:

$P = \frac{27}{27+0} = 100\%$, $R = \frac{27}{27+1} = 96.43\%$, $A = (27+15)/(27+15+0+1) = 97.67\%$, $S = \frac{15}{15+0} = 100\%$, $E = \frac{1}{(1/2*(1))+(1/2*(0.9643))} = 98.18\%$, and $NPV = \frac{15}{15+1} = 93.75\%$.

Although, the pre-diabetic and normal patients from less than half of the case-base, the proposed system accuracy for predicting them is 100%. The performance of our proposed system is enhanced because its similarity measures take into account the nature of all features.

Another type of comparison has been done with a set of machine learning classifiers. Techniques such as artificial neural networks (ANN), support vector machines (SVMs), neuro-fuzzy systems and expert systems that developed by different authors have been discussed. Firstly, all these studies have lower performance than ours. However, these systems mostly depend on Pima Indians Dataset (https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes). To compare our system with these techniques, it is better to run these algorithms on our dataset. This dataset has been prepared before, and all noise and missing data have been handled [21]. For the comparing purpose, we apply some machine learning classifiers including C4.5, k-NN, SVM, Bayesian classifier, and ANN on our dataset and measure their performance. We use the 2-fold, 3-fold, 4-fold…10-fold. The cross-validation technique is our evaluation process. Cross-validation is a statistical technique useful in determining the robustness of a model. The n-fold cross validation divides the whole data set into n folds. The n-1 folds are used for training, and one fold is used for testing. This process is continued until each fold from n is used for testing. The overall performance of these algorithms is presented in Table 4. For the k-NN algorithm, we select k=3 as done in our system; however, its performance is low. C4.5 achieves the best performance (about 89.19%) among machine-learning techniques; however, our system outperforms it. After testing the machine learning algorithms using from 2-fold to 10-fold cross-validation techniques, we calculate the average performance of each fold, and we make a comparison of different folds' results. Figure 9 shows that the best performance is achieved with 5-fold cross validation.

We calculate the average precision, recall, accuracy, f-measure, and specificity for all folds. These averages are compared with the proposed system, the 5-fold cross validation, and the traditional (i.e. not fuzzy and not semantic) system, as shown in Figure 10. Our findings show that the fuzzy KI-CBR can classify data more accurately than the other machine learning techniques and conventional CBR.

It can be seen in Figure 10 that the machine learning classifiers have better performances than conventional CBR systems. This means that our study makes a high improvement in the CBR performance. The average accuracies of C4.5, conventional CBR, and proposed system are 88.88%, 57.14%, and 98.18% respectively. The proposed approach demonstrates a major improvement than machine learning techniques and conventional CBR system.

The results of this study clearly indicate that the hybridization of CBR with fuzzy ontology and medical ontologies is the most suitable technique for solving medical diagnosis problems. The enhanced performance of our system is a result of a couple of reasons. Firstly, the proposed CBR framework is integrated and complete. All components have been fully implemented and tested. The knowledge representation formalism using fuzzy ontology integrates the reasoning capabilities of fuzzy logic, description logic, and CBR. There are many studies, which use each of these reasoning mechanisms individually, but they have not achieved high accuracy. The second reason is the preparation of case-base data. These data have been pre-processed, fuzzified, and encoded before populated into the case-base knowledge. As a result, accurate data will produce accurate decisions. The third reason is the usage of a suitable weight vector for the used case features; the global similarity function has produced suitable similarities. The fourth reason is the proposed semantic retrieval algorithm. We have handled most of the possible datatypes, which appear in the medical domain. The fuzzy types support the reasoning using linguistic terms and enhance the similarity calculation. Ordinal features' similarity is based on the expert domain knowledge in the form of similarity matrixes. Semantic features support the calculation of clinical similarities between SCT concepts.

In addition to its enhanced performance, the proposed system is tested for problems that are complex and cannot be solved by traditional systems. For example, If the case base contains a case C1= (age=20, disease= "Acute proliferative", urination frequency= "++" …) and the query case is (age=young, disease= "Idiopathic crescentic", urination frequency= "Nil"…); in the traditional CBR systems, these cases are

	Fold	Algorithm	Precision (%)	Recall (%)	Accuracy (%)	F-Measure (%)	Specificity (%)
Machine learning algorithms	2-Fold	C4.5	93.1	93.1	93.33	93.1	93.54
		k-NN (k=3)	63.3	63.3	63.33	63.3	64.5
		SVM	6.3	58.6	63.33	60.7	67.74
		Naive Bayes	81.8	62.1	75	70.6	87.09
		ANN	65.5	65.5	66.66	65.5	67.74
	3-Fold	C4.5	90	93.1	91.66	91.5	90.32
		k-NN (k=3)	60	60	60	59.9	64.51
		SVM	71	75.9	73.33	73.3	70.96
		Naive Bayes	65.4	58.6	65	61.8	70.96
		ANN	72.4	72.4	73.33	72.4	74.19
	4-Fold	C4.5	89.7	89.7	90	89.7	90.32
		k-NN (k=3)	68.7	68.3	68.33	68	77.41
		SVM	69	69	70	69	70.96
		Naive Bayes	77.3	58.6	71.66	66.7	83.87
		ANN	75.9	75.9	76.66	75.9	77.41
	5-Fold	C4.5	92.9	89.7	91.66	91.2	93.54
		k-NN (k=3)	68.3	68.3	68.33	68.3	70.96
		SVM	78.6	75.9	78.33	77.2	80.64
		Naive Bayes	77.3	58.6	71.66	66.7	83.87
		ANN	78.6	75.9	78.33	77.2	80.64
	6-Fold	C4.5	89.3	86.2	88.33	87.7	90.32
		k-NN (k=3)	61.7	61.7	61.66	61.5	67.74
		SVM	67.7	72.4	70	70	67.74
		Naive Bayes	61.5	55.2	61.66	58.2	67.74
		ANN	73.3	75.9	75	74.6	74.19
	7-Fold	C4.5	89.7	89.7	90	89.7	90.32
		k-NN (k=3)	73.6	73.3	73.33	73.2	80.64
		SVM	69.7	79.3	73.33	74.2	67.74
		Naive Bayes	70.4	65.5	70	67.9	74.19
		ANN	71.9	79.3	75	75.4	70.96
	8-Fold	C4.5	89.7	89.7	90	89.7	93
		k-NN (k=3)	68.7	68.3	68.33	68	77.41
		SVM	74.2	79.3	76.66	76.7	74.19
		Naive Bayes	82.6	65.5	76.66	73.1	87.09
		ANN	70	72.4	71.66	71.2	70.96
	9-Fold	C4.5	89.3	86.2	88.33	87.7	90.32
		k-NN (k=3)	66.8	66.7	66.66	66.4	74.19
		SVM	75	82.8	78.33	78.7	74.19
		Naive Bayes	79.2	65.5	75	71.7	83.87
		ANN	77.4	82.8	80	80	77.41
	10-Fold	C4.5	74.2	79.3	76.66	76.7	90.32
		k-NN (k=3)	73.1	65.5	71.66	69.1	70.96
		SVM	77.4	82.8	80	80	77.41
		Naive Bayes	79.2	65.5	75	71.7	83.87
		ANN	74.2	79.3	76.66	76.7	74.19
	Average (%)		**73.88**	**73.39**	**75.1**	**74.04**	**78.04**
Conventional CBR system			**85.7**	**42.85**	**57.14**	**57.13**	**85.7**
Proposed fuzzy KI-CBR system			**100**	**96.43**	**97.67**	**98.18**	**100**

Table 4: Performance of machine learning algorithms on our dataset.

not similar, and C1 will not be returned. For fuzzy systems, the age is matched right as age=20 is the same as age=young(i.e., $\mu_{Young}(20)=1$). However, the comparison of semantic and ordinal features fails to get the similarity. In semantic CBR systems, they fail to get the similarity of fuzzy and ordinal features. Due to these conditions, the results of these systems might prove to be not accurate. In our proposed system, we have proposed algorithms to handle all of these types.

Conclusion

In this paper, we propose a fuzzy ontology-based semantic-CBR system. This framework has been implemented for diabetes diagnosis as a case study. The proposed approach has many contributions and novelties. Our implemented fuzzy ontology has followed a formal methodology, and it has represented using fuzzy OWL2 language. The proposed fuzzy-semantic retrieval algorithm outweighs all of the JCOLIBRI algorithms, and it covers their limitations. Our system has achieved a performance of 97.67%. These results show that the proposed system has a high accuracy, and physicians can consult it when diagnosing patients. However, the proposed study has some limitations including the ability to handle termoral data, diabetes treatment, and

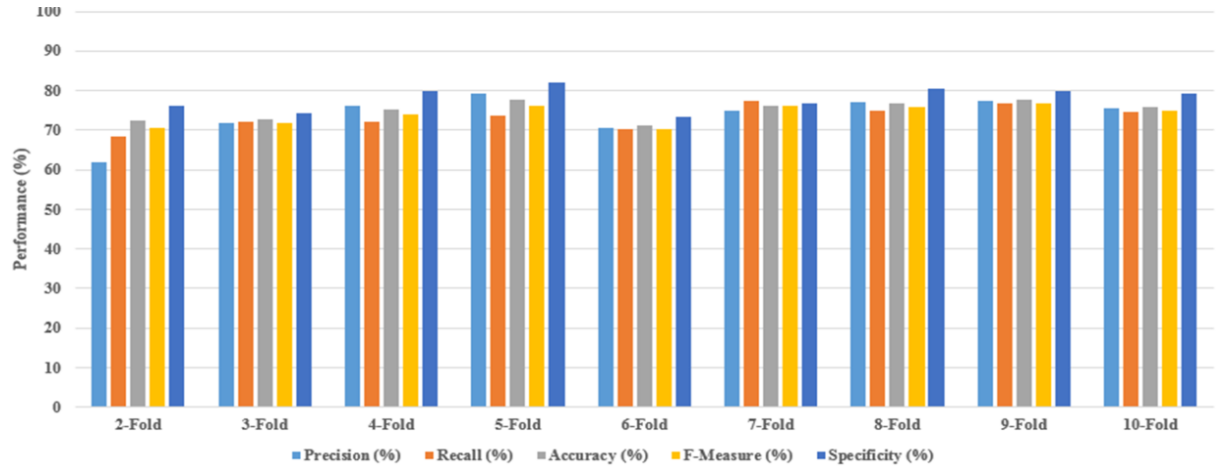

Figure 9. A comparison between the n-folds cross validation results.

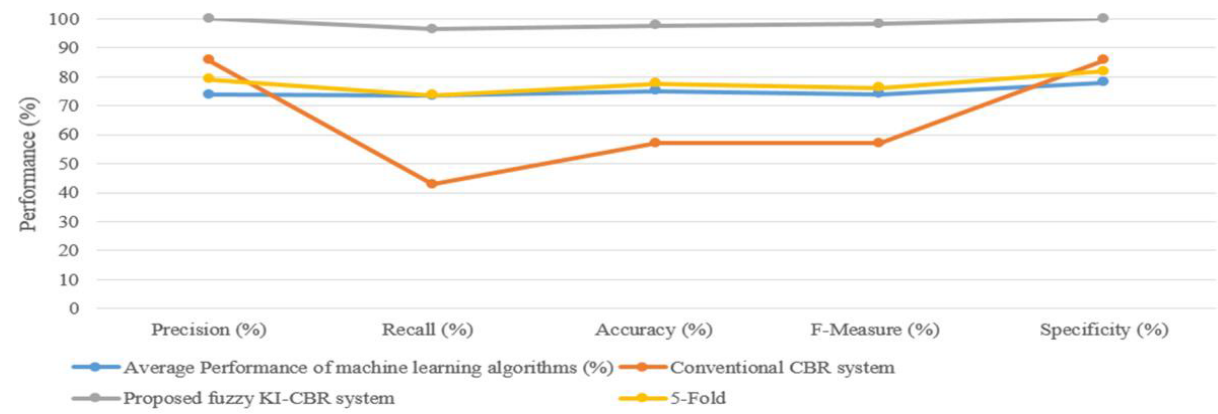

Figure 10: Classification results comparison.

the ability to adapt the proposed solutions. Many studies have solved parts of these problems including case adaptation, but the applicability of these approaches with fuzzy ontology has not been studied yet. In the future, we will implement the rest of the CBR steps especially the case adaptation process. We will utilize fuzzy ontology in the other steps of CBR as case adaptation, retention, and case-base maintenance. Moreover, we will try to integrate multiple medical ontologies in our system because SCT has a limitation in many aspects as lab tests and genes representation. Fortunately, there are many standard medical ontologies for theses domains such as LOINC for lab tests and GO for genes representation. The integration of CBR with EHR environment will enhance the automation of the decision support process and building of distributed CDSS systems.

References

1. International Diabetes Federation. Diabetes Atlas. 5th ed. Brussels, Belgium: IDF Publications. (2011) The Global Burden of Diabetes; pp. 7–13. Available from http://www.idf.org/diabetesatlas/news/fifth-edition-release. Accessed 25 May 2015.
2. Roche M, Wang P (2014) Factors associated with a diabetes diagnosis and late diabetes diagnosis for males and females. Journal of Clinical & Translational Endocrinology 1: 77-84
3. Başçiftçi F, Hatay Ö (2011) Reduced-rule based expert system by the simplification of logic functions for the diagnosis of diabetes. Computers in Biology and Medicine 41: 350–356.
4. Tripathi B, Srivastava A (2006) Diabetes mellitus: Complications and therapeutics. Med Sci Monit 12: RA130-147.
5. Konga G, Dong-Ling Xu, Richard Body, Jian-Bo Yang, Kevin Mackway-Jones, et al. (2012) A belief rule-based decision support system for clinical risk assessment of cardiac chest pain. European Journal of Operational Research 219: 564–573.
6. Alves V, Novais P, Nelas L, Maia M, Ribeiro V (2003) Case-based reasoning versus artificial neural networks in medical diagnosis. Proceedings of IASTED International Conference Artificial Intelligence and Applications 1-5.
7. Blanco X, Rodríguez S, Corchado J, Zato C (2013) Case-Based Reasoning Applied to Medical Diagnosis and Treatment. Distributed Computing and Artificial Intelligence 217: 137-146.
8. Chen J, Su S, Chang C (2010) Diabetes care decision support system. 2nd IEEE International Conference on Industrial and Information Systems (IIS) 1: 323 – 326.
9. Hidalgo J, Maqueda E, Risco-Martín JL, Cuesta-Infante A, et al. (2014) glUCModel: A monitoring and modeling system for chronic diseases applied to diabetes. Journal of Biomedical Informatics 48: 183–192.
10. El-Sappagh S, Elmogy M, Riad A, Zaghloul H, Badria F (2014) A proposed SNOMED CT ontology-based encoding methodology for diabetes diagnosis case-base. The 9th IEEE International Conference on Computer Engineering and Systems (ICCES 2014) 184-191.
11. El-Sappagh S, Elmogy M, El-Masri S, Riad A (2014) A Diabetes Diagnostic Domain Ontology for CBR System from the Conceptual Model of SNOMED CT. The IEEE second International Conference on Engineering and Technology (ICET 2014) 1 – 7.

12. El-Sappagh S, Elmogy M, Riad A (2015) A CBR system for diabetes mellitus diagnosis: case-base standard data model. International Journal of Medical Engineering and Informatics 7: 191-208.

13. Zadeh L (2003) From search engines to question-answering systems the need for new tools. Fuzzy Systems. The 12th IEEE International Conference 2: 1107-1109.

14. Jaya A, Uma G (2009) Role of Ontology in Case-Based Reasoning (CBR) for Diagnosing Diabetes. Journal of Information Technology 5: 17-23.

15. Lee C, Wang M (2011) A Fuzzy Expert System for Diabetes Decision Support Application. IEEE transactions on systems, man, and cybernetics—part b: cybernetics 41: 139-153.

16. Ahmadian L, van Engen-Verheul M, Bakhshi-Raiez F, Peek N, Cornet R, et al. (2011) The role of standardized data and terminological systems in computerized clinical decision support systems: Literature review and survey. INT J MED INFORM 80: 81–93.

17. Paterson G, Abidi S, Soroka S (2005) HealthInfoCDA: Case Composition Using Electronic Health Record Data Sources. Studies in Health Technology and Informatics 116: 137-142.

18. Abidi S, Manickam S (2002) Leveraging XML-based electronic medical records to extract experiential clinical knowledge an automated approach to generate cases for medical case-based reasoning systems. International Journal of Medical Informatics 68: 187-203.

19. Borges K, Aquino R, Barcelos T, Simoes J (2012) A methodology for preprocessing data for application of case based reasoning. IEEE Informatica (CLEI), 2012 XXXVIII Conferencia Latinoamericana En., 1-8.

20. Lee D, Lau F, Quan H (2010) A method for encoding clinical datasets with SNOMED CT. BMC Medical Informatics and Decision Making 10:53.

21. El-Sappagh S, Elmogy M, Riad A, Badria F, Zaghlol M (2014) EHR Data Preparation for Case Based Reasoning Construction. Advanced Machine Learning Technologies and Applications Communications in Computer and Information Science, Springer International Publishing 488: 483-497.

22. Chen R, Huang Y, Bau C, Chen S (2012) A recommendation system based on domain ontology and SWRL for anti-diabetic drugs selection. Expert Systems with Applications 39: 3995–4006.

23. Rahimi A, Liaw S, Taggart J, Ray P, Yu H (2014) Validating an ontology-based algorithm to identify patients with Type 2 Diabetes Mellitus in Electronic Health Records. International Journal of Medical Informatics 83: 768–778.

24. Sherimon P, Vinu P, Krishnan R, Takroni Y, AlKaabi Y, et al. (2014) Adaptive Questionnaire Ontology in Gathering Patient Medical History in Diabetes Domain. Proceedings of the First International Conference on Advanced Data and Information Engineering 285: 453-460.

25. Hayuhardhika W, Putra N, Sugiyanto, Sarno R, Sidiq M (2013) Weighted Ontology and Weighted Tree Similarity Algorithm for Diagnosing Diabetes Mellitus. IEEE International Conference on Computer, Control, Informatics and Its Applications pp. 267-272.

26. El-Sappagh S, El-Masri S, Elmogy M, Riad A, Saddik B (2014) An Ontological Case Base Engineering Methodology for Diabetes Management. J Med Syst 38:67.

27. Alexopoulos P, Wallace M, Kafentzis K, Askounis D (2010) Utilizing Imprecise Knowledge in Ontology-based CBR Systems by Means of Fuzzy Algebra. International Journal of Fuzzy Systems 12.

28. Abdul M, Muhammad A, Mustapha N, Muhammad S, Ahmad N (2014) Database workload management through CBR and fuzzy based characterization. Applied Soft Computing 22: 605–621.

29. Khanum A, Mufti M, Javed M, Shafiq M (2009) Fuzzy case-based reasoning for facial expression recognition. Fuzzy Sets and Systems 160: 231–250.

30. Sohn M, Jeong S, Lee H (2014) Case-based context ontology construction using fuzzy set theory for personalized service in a smart home environment. Soft Comput 18: 1715–1728.

31. Bobillo F, Straccia U (2011) Fuzzy ontology representation using OWL 2. INT J APPROX REASON 52: 1073–1094.

32. Aamodt A, Plaza E (1994) Case-based reasoning foundational issues, methodological variations, and system approaches. Journal of AI Communications 7: 39–59.

33. Alexopoulos P, Wallace M, Kafentzis K, Askounis D (2012) IKARUS-Onto: a methodology to develop fuzzy ontologies from crisp ones. Knowledge and information systems 32: 667-695.

34. Harispe S, Sanchez D, Ranwez S, Janaqi S, Montmain J (2014) A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain. J Biomed Inform 48: 38–53.