Integration of datamining techniques into the CBR cycle to predict the result of immunotherapy treatment

Fatima Saadi
Laboratoire d'Informatique d'Oran (LIO)
University of Oran 1 Ahmed Ben Bella
Oran, 31000, Algeria
saadi_fatima@hotmail.fr

Baghdad Atmani
Laboratoire d'Informatique d'Oran (LIO)
University of Oran 1 Ahmed Ben Bella
Oran, 31000, Algeria
baghdad.atmani@gmail.com

Fouad Henni
Laboratoire d'Informatique d'Oran (LIO)
University of Mostaganem (FSEI)
Mostaganem, 27000, Algeria
fouad.henni@univ-mosta.dz

***Abstract*—The functioning of the medical diagnostic process is very comparable to the pattern of the CBR cycle. The doctor often starts to analyze the whole situation and takes advantage of previous situations resolved successfully in order to efficiently diagnose a new situation. More generally, this mode of operation, based on experience and analogy, can be found practically in all diagnostic fields (medical, industrial, etc.). In this work, we propose the integration of datamining techniques in the CBR cycle for medical decision support. This integration aims to improve the performance of the retrieval phase by reducing the number of attributes considered in the similarity calculation through the selection of the most relevant attributes. To evaluate our approach, we have customized the jCOLIBRI 2.1 framework with a real case base to predict the response to immunotherapy treatment for patients who have plantar and common warts disease.**

Keywords—CBR, Datamining, Decision Support, Immunotherapy, jCOLIBRI, Warts disease.

I. INTRODUCTION

For some doctors it's always difficult to diagnose and give a specific treatment for each patient. They often strive to predict which treatment has a better impact on a particular patient. The reasoning used in medical diagnosis is very similar to the Case Based Reasoning (CBR) methodology for problem resolution, which constitutes an ideal framework for the design of decision support systems.

The process of CBR goes through several steps (Retrieve-Reuse-Revise-Retain). We are interested in the retrieval phase because of its significant impact on the performance of any CBR-based system. This step consists in searching, among previous cases, those with a problem part similar to the new situation to diagnose. The retrieval step is based on the similarity calculus between two problems.

Improving the retrieval step will allow a gain in terms of time which is one of the most important factors in medical diagnosis. Regarding the adaptation phase, in most CBR-based systems this step is left to the domain experts because of the lack of predefined adaptation rules. What's more, medical diagnosis is a delicate area because it is about the patient health. In such domains, a manual adaptation is not considered as a negative aspect. Moreover, in [1], authors state: "when the field of knowledge is not really clear, automatic adaptation is difficult to develop or not recommended".

In this paper, we propose a medical decision support system to predict the response of the patient who has plantar and common warts to the immunotherapy treatment. The objective of this work is to improve the retrieval step through the integration of one of datamining techniques "decision tree". This technique is used as a discriminating tool that eliminates less pertinent attributes and reduces the number of attributes used in the calculation of similarity to accelerate the recuperation of similar cases.

For the similarity measure, we used the k-nearest neighbours, which consist of calculating a distance between the case to be solved and the cases stored in the case base by considering only the relevant attributes deduced from the decision tree.

To implement our approach, we used the Weka platform [14] for the generation of the decision tree and we customized the platform of case-based reasoning jCOLIBRI 2.1 [2] in accordance with the characteristics of the application domain.

The article is organized as follows: in Section 2, we briefly define CBR and Datamining, then we expose some proposed approaches that use a combination of Datamining techniques and the CBR methodology. In Section 3, we present our main contribution and give the proposed approach. Then in section 4, we give a presentation of experimentation and interpretation of results. Finally, section 5 is devoted to the conclusion and some perspectives of this work.

II. STATE OF THE ART

This section is organized in two parts. First, we give the definitions of CBR and Datamining. Second, we expose the main related works:

A. Definitions

- *CBR (Case Based Reasoning)*

CBR is a methodology [3] for problem solving, based on the use of past experiences to solve a new problem.

The resolution cycle in CBR can be described by four steps: Retrieve-Reuse-Revise-Retain.

A. *Retrieve* from the case base the most similar case (or cases) to the new problem to resolve.
B. *Reuse* information and knowledge in this case to solve the problem.
C. *Revise* the proposed solution.
D. *Retain* the parts of this experiment likely to be useful for the resolution of future problems.

- *Datamining*

Datamining is at the heart of the knowledge database discovery (KDD) process. It is a set of techniques for

978-1-5386-8125-1/19/$31.00 ©2019 IEEE

extracting useful and new information from large amounts of data. These techniques reveal predictive models, classification rules and other types of knowledge that will be used for decision support [4]. Among the variety of datamining methods, decision tree and the k-nearest neighbors (KNN) are commonly used in the medical field.

B. Related works

Nowadays, a large number of integrated case-based reasoning (CBR) and datamining (DM) techniques are available for data analysis and prediction in a large variety of domains (medical, industrial, etc.), we cite some works below:

De and Chakraborty [18] proposed a system for car fault diagnosis (CFD) based on case-based reasoning (CBR) methodology. It uses Decision tree and Jaccard Similarity Method. Decision tree is used to store cases into the Case Base (CB), and Jaccard Similarity Method is used to calculate similarity between a new case and stored cases.

Benbelkacem and al. [6] developed a CBR system in the medical context, especially for the treatment of tuberculosis, to facilitate the choice of the appropriate treatment. They used the decision tree as a measure of similarity to recover similar cases and also to optimize the computation time unlike other similarity measures such as the method of the k nearest neighbors which is expensive in computation time. Each case of tuberculosis is described by a set of descriptors that determine the problem part of the case, to which is assigned a treatment representing the solution part of the case. They evaluated their approach on real cases involving patients treated for tuberculosis. Then, they compared with the k nearest neighbors' method and deduced that the retrieval by decision tree makes it possible to optimize the computation time.

Burke and al. [7] proposed an approach for solving timetable problems; In this approach, the case base is represented as a decision tree and cases represent the attributes of the graph. They used the Branch and Bound algorithm to reduce the search space.

Kriegsmann and Barletta [5] proposed a CBR system that uses the decision tree and k nearest neighbors to identify and retrieve cases similar to the target problem. The retrieval is done in two stages. The decision tree is initially used to determine a set of similar cases. Next, k nearest neighbors is used to identify the most relevant cases. This system has been applied to diagnose hardware, software and connectivity issues on various platforms.

Menezes and al. [16] used the decision tree for analysis and diagnosis of incipient failures in power transformers by using the concentration in ppm of the combustible gases present in samples of transformer oils. They built the decision tree using the algorithm C4.5, and for the selection of attributes, they considered that the ratio of the gain is able to extract the maximum information available and to find the attribute that performs the best division.

Guo and al. [17] proposed a hybrid CBR-Bayesian network to improve case retrieval of BN model under big data. They implemented two algorithms: Weighted Index Coefficient of Dirichlet Distribution algorithm (WICDD) to overcome the problem of probability learning under big data, and then enhance the accuracy of case retrieval; and the Within-Cross algorithm (WC) to solve the problem of computation task assignment under the condition of large number of parameters while conducting parallel data processing for parameter independence test.

Clerkin and al. [8] presented a collaborative song recommendation approach in the Web (Smart Radio) based on the user's evaluation of songs. This approach aims to automatically generate the case base by using the K-means algorithm for the segmentation of databases containing the history of the operations carried out in the studied domain.

Mansoul and al. [9] proposed an approach that combinates CBR and clustering to reduce the search space in the retrieval step. They used the K-means algorithm for clustering. The objective is to consider only the most relevant cases and the most interesting solution to allow a better decision support. The proposed approach has been applied to a medical dataset the Vertebral Column Data Set of orthopedic patients.

Mekroud and al. [10] proposed a hybrid approach, CBR and Datamining, applied in industrial diagnostics. The proposed process begins with a fragmentation of the case base into two spaces: Symptoms-Breakdowns & Symptoms-Solutions; followed by a clustering of the two spaces, and a mapping between their clusters. Finally, a CBR cycle is applied for each space.

Tahmasebian and al. [11] extracted the effective information from the summary of medical records of patients with kidney disease and deduced the weights of this information using datamining techniques to measure the similarity rate. Finally, they used a fuzzy system to compare the similarities between the current case and previous cases. The system has been implemented on the Android platform.

El-Sappagh and Elmogy [12] proposed a new CBR ontology (CBRDiabOnto) for the fuzzy KI-CBR diagnostic system (diagnosis of diabetes mellitus). The resulting ontology is enriched by several types of data, such as net, semantic, fuzzy reasoning, textual and ordinal. These different types of data facilitate the representation and recovery of cases and support the expression of queries by physicians. In this way, they improved the most critical stage of the CBR system (recall and recovery of the case).

Benamina and al. [13] have developed a fuzzy CBR for diabetes diagnosis that incorporates fuzzy logic and datamining to improve the response time and accuracy of recovery of similar cases. The proposed fuzzy CBR is composed of the part of classification by fuzzy decision tree realized by Fispro and the part of case-based reasoning realized by the platform jCOLIBRI. They used fuzzy logic to reduce the complexity of calculating the degree of similarity that may exist between diabetic patients requiring different surveillance plans.

III. CONTRIBUTION

Among several recently used treatments, immunotherapy with the CANDIDA antigen has proved its effectiveness as a treatment for warts and especially for plantar and common warts that are very painful and contagious for children, adolescents and people with weakened immune systems who are more likely to suffer from these warts.

We propose a medical decision support system whose objective is to help practitioners choosing the right treatment.

A. Case description

Each case consists of a problem part and a solution part. The problem part is described by a set of relevant characteristics called descriptors. The solution part represents the result of the treatment.

1) Problem part

This part represents information about the patient; that is: Identifier, Sex, Age, time elapsed before treatment (months), number of warts, type of warts (common, plantar, both), surface of the largest wart (mm^2), induration diameter of the initial test (mm).

2) Solution part

In the solution part, the result attribute of the processing (Result_of_treatement) is either success or failure.

B. Proposed approach

The "Immunotherapy" database was extracted from the UCI Machine Learning Repository. Data was collected from 90 patients with plantar and common warts who referred to the dermatological clinic. These two types of warts are the most common [14].

Data has been reported in a table to form a CSV (Comma Separated Values) file.

The main steps of our approach are:

- Generation of the decision tree with the WEKA platform from a learning sample (case base) to extract the relevant attributes.
- Calculation of local similarities based on the attributes revealed by the decision tree. These functions are integrated into the jCOLIBRI platform. The similarity between two cases is obtained by the nearest-neighbor technique since it is the most used in systems based on the CBR [3].
- Comparison between the similarity results obtained by the proposed approach and classical approaches which consider all the attributes.

The overall architecture of the proposed approach is illustrated in Fig. 1:

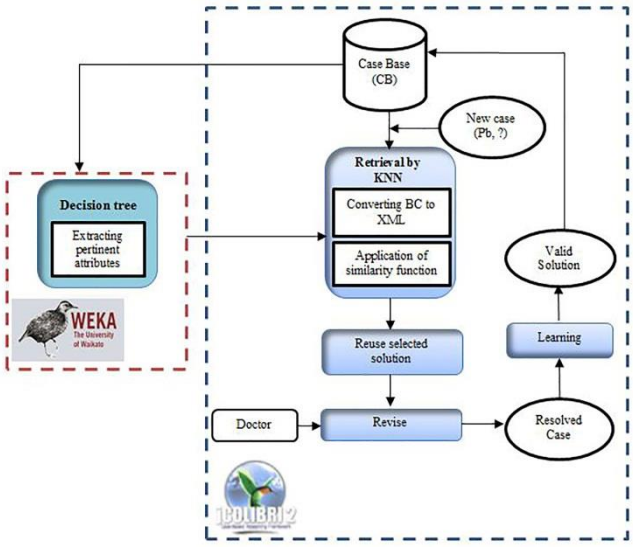

Fig. 1. Architecture of the proposed approach

In order to support doctors in predicting the result of the immunotherapy treatment on a given patient, the collected data has been introduced into the jCOLIBRI framework as the case base. A case is described by 8 attributes mentioned above. Each case is associated with a class that determines the outcome of the immunotherapy treatment.

To generate the decision tree, we used the C4.5 algorithm (J48) implemented in the WEKA platform [15]. This algorithm is an extension of the ID3 algorithm. It constructs the decision tree from a set of training data using the concept of information entropy [17] which measures the amount of information provided by a node. It also uses the gain obtained by measuring the degree of mixing classes for each sample and therefore for any position of the tree under construction.

The information gain can meet the criterion of a good decision tree which at each node must be associated with the non-target attribute that provides the most information relative to the other attributes not yet used in the path from the root.

The dataset has been subdivided into two parts: learning set and test set. We have adopted the partition (70%, 30%) for the learning and test sets respectively. The individuals have been partitioned randomly from the dataset. This experience was repeated several times which produced many decision trees. To choose our decision tree a “majority vote” produced the tree in fig. 2.

The attribute that has the highest value of information gain is used as a decision attribute. From Fig. 2, we can deduce that the most relevant attributes are Time, Type and Age. These three attributes will be used in the similarity calculus between a new case and memorized cases. It is clear that doing so will significantly improve the retrieval step, in particular when the case base becomes very large.

In the experimentations, we used the Euclidean distance as a local similarity measure between two attributes, because we have only numeric attributes. The similarity function between two problems is calculated from the results of local similarities relative to the chosen attributes (Time, Age, Type). In order to provide a result between 0 and 1, we proceeded to the normalization of the Euclidean

distance according to the formula 1 below, then we implemented this function in the jCOLIBRI framework.

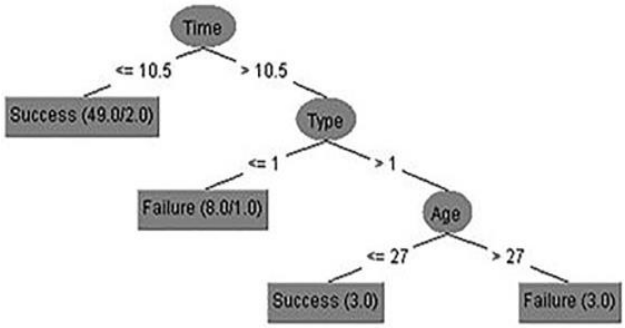

Fig. 2." Decision tree generated with WEKA

$$D(x,y) = \frac{\sqrt{x^2 - y^2}}{x + y} \qquad (1)$$

Once the similarity measures between the attributes are computed, the overall similarity calculates the distance between two cases. It uses a global similarity function based on the nearest-neighbor technique since it is the most widely used in CBR-based systems [3]. In this algorithm, the similarity between a new case and a memorized case is calculated through a weighted summation of similarities between the attributes (two by two). Since the local similarity functions have been normalized, the global similarity value remains between 0 and 1; where 0 represents the total similarity. For our system, the algorithm is defined according to the following equation:

$$Sim(C,S) = \sum_{f=1}^{n} w_f * sim(C_f, S_f) / sum(w_f) \qquad (2)$$

"C" is the new case, "S" is a stored case, "w" is the weight defined by an expert, "n" is the number of attributes for each case, "f" is the index of the attribute and "sim(C_f,S_f)" is the local similarity for attribute "f".

IV." Experimentation And Discussion Of Results

To study the efficiency of our approach, we developed a medical decision support system based on CBR, as mentioned above. In this study we used the immunotherapy dataset.

At the beginning, the system generates the decision tree in order to deduce the relevant attributes. It uses the "immunotherapy" dataset as training set.

The entry in our system is in the form of a simple interface that allows to launch a query. From this interface, the doctor enters the information concerning a new patient with the warts disease.

In addition, our system provides the opportunity for the physician to define the number of cases to retrieve and the possibility to choose one case from those selected. This last possibility forced us to modify the "Cycle" function of the jCOLIBRI main interface "StandardCBRApplication" by adding a parameter to allow the expert to choose the number of cases to retrieve.

The system calculates the local similarity between the deduced relevant attributes and the overall similarity between the cases in the database and shows the best cases resulting from the retrieval step. These cases are displayed in descending order of similarity. The doctor can consult the cases one by one and then choose the case that he/she considers most appropriate (Fig. 3). After this choice, the doctor can make changes (adjustments) to the proposed solution. This solution with the query represents a case that the system has just solved.

For example, consider as a new case a woman aged 48, time elapsed before treatment (8,5 months), number of warts(2), type of warts (plantar), surface of the largest wart (13 mm²), induration diameter of the initial test (8 mm).

The results of the calculation of local and global similarities between the new case and the best cases resulting from the retrieval step are illustrated in Table I.

TABLE I. " Results Of The Calculation Of Local And Global Similarity

	Age	Time	Type	Global similarity
Case N°53	49	9	2	0,1146
Local similarity	0,1015	0,1690	0,0	
Case N°76	47	9,25	2	0,1239
Local similarity	0,1025	0,2055	0,0	
Case N°48	51	8,75	2	0,1285
Local similarity	0,2581	0,1203	0,0	

Fig.3 shows the best cases resulting from the retrieval step.

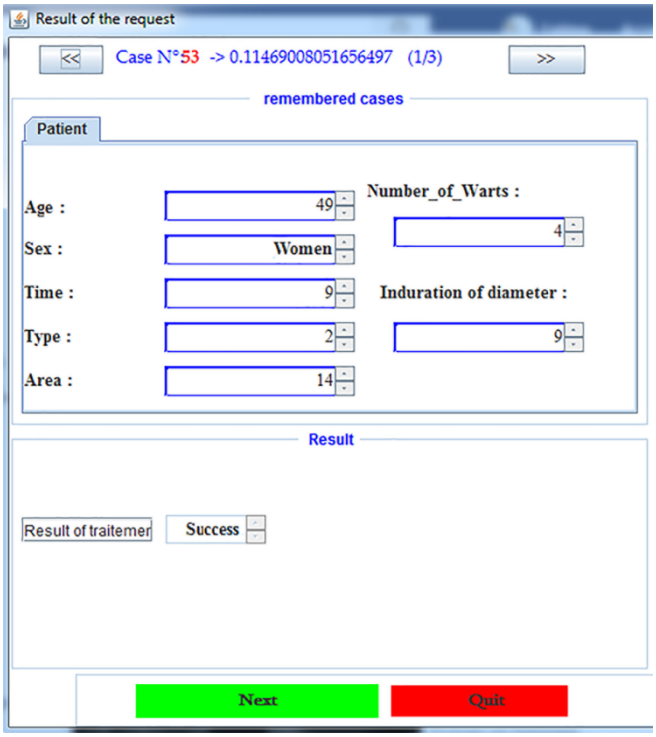

Fig. 3." Result interface

Once the solution is adjusted (or adapted) by the doctor, the retention interface gives the possibility to retain this new case or to leave it in a temporary case base that we added in our application. If the case is retained, it is integrated in the case base (learning). The doctor may decide to take an extra time before deciding on the retention of this new case. In addition, the doctor may, from time to time, consult the

temporary case base to complete cases and/or decide to select a few cases for retention.

To verify our approach, we performed two main experiences. In the first experience, the global similarity used the KNN with all the attributes of the case (method 1). In the second experience, the global similarity function also used KNN with only the attributes revealed by the decision tree (method 2); that is (Time, Age, Type). In both experiences, cases have been randomly chosen from the test set and introduced as new cases. For each new case, we have chosen to retrieve 7, 5 and 3 similar cases (TABLE II). Then we considered the intersection between the sets of retrieved cases obtained by the two methods. Table 2 reports the number of identical cases obtained from the two methods.

TABLE II. INTERSECTION OF RESULTS OBTAINED BY METHOD 1 AND METHOD 2

	7 cases	**5 cases**	**3 cases**
New case 1	4	3	1
New case 2	2	2	2
New case 3	3	1	1
New case 4	3	2	0
New case 5	2	2	2

From TABLE II., we can observe that in most held experiences, the intersection between the two sets of retrieved cases is not empty. This means that "method 2" (reduction of the attributes according to the generated decision tree) can retrieve pertinent cases in general. We cannot consider this as a validation. Further experiences with the same case base, and other case bases from other domains can confirm this result

V. CONCLUSION

In this article, we presented our approach that integrates data mining techniques into the CBR cycle, and more precisely into the retrieval phase for the similarity calculus. Our objective is to improve the retrieval step which is known as time consuming, in particular in a very large case base.

The designed solution is intended to help the doctor in the prediction of the outcome of the immunotherapy treatment on a new patient with plantar or common warts disease.

To that end, we proposed to use a decision tree as a discriminating tool to reveal the most relevant attributes in the case. The reduction of the attributes taken into account in the similarity calculus can significantly speed up the retrieval of similar cases.

We used the WEKA tool for the generation of the decision tree and the jCOLIBRI 2.1 framework that we customized according to the field of application.

The experiences performed are very encouraging since they revealed that the results obtained when we consider only the attributes deduced from the decision tree are comparable to those obtained when we use all the attributes.

To validate our approach, further experiences must be held. These experiences can concern other types of treatments of the warts disease (e.g. the cryotherapy treatment), and also case bases from other domains. Some attempts are actually in progress.

On another hand, we are working on the integration of many other datamining techniques and their impact on the retrieve phase of the CBR cycle.

REFERENCES

[1] Nasrullah, and M. Hassan, "Evaluation of jCOLIBRI", Rapport de thèse de master, Malardalen University, Vasagatan 44, 72123 Vasteras, Sweden., 2006.

[2] J. Bello-Tomas, P. Gonzalez-Calero, and B. Diaz-Agudo, "jCOLIBRI: Anobject-oriented framework for building cbr systems," IEEE Advancesin Case-Based Reasoning, 7th European Conference on Case-Based Reasoning (ECCBR 2004), pp. 32–46, 2004.

[3] I.Watson, "Case-based reasoning is a methodology not a technology", Knowledge-Based Systems 12, p 303–308, 1999.

[4] M. Hanifi, "Extraction de caractéristiques de texture pour la classification d'images satellites", Thèse de Doctorat, Université de Toulouse, 2009.

[5] M. Kriegsmann and R. Barletta, "Building a case-based help desk application," IEEE Expert, vol. 8, pp. 18–26, 1993.

[6] E. K. Burke, B. L. MacCarthy, S. Petrovic, and R. Qu, "Multipleretrieval case based reasoning for course timetabling problems" , Journal of the Operational Research Society, vol. 57, pp. 148–162, 2006.

[7] P. Clerkin, P.Cunningham, and C. Hayes, "Concept discovery in collaborative recommender systems", Trinity College Dublin, Department of Computer Science, 2003.

[8] A. Mansoul, and B. Atmani, "Clustering to Enhance Case-Based Reasoning", In Modelling and Implementation of Complex Systems (pp. 137-151). Springer, Cham, 2016.

[9] N. Mekroud, and A. Moussaoui, "Approche Hybride pour le Diagnostic Industriel basée sur le RàPC et le Datamining Utilisation de la Plateforme JCOLIBRI 2.1." , In CIIA,2009.

[10] S. Tahmasebian, M. Langarizadeh, M. Ghazisaeidi, and M. Mahdavi- Mazdeh, "Designing and implementation of fuzzy case-based reasoning system on android platform using electronic discharge summary of patients with chronic kidney diseases," Acta Informatica Medica, vol. 24, no 4, pp. 266-270, 2016.

[11] S. El-Sappagh and M. Elmogy, "A fuzzy ontology modeling for case base knowledge in diabetes mellitus domain," Engineering Science and Technology, an International Journal, vol. 20, no 3, pp. 1025-1040, 2017.

[12] M. Benamina, B. Atmani, and S. Benbelkacem, "Diabetes Diagnosis by Case-Based Reasoning and Fuzzy Logic", International Journal of Interactive Multimedia and Artificial Intelligence, (In Press).

[13] F.Khozeimeh, R. Alizadehsani, M. Roshanzamir, A. Khosravi, P. Layegh, and S.Nahavandi, "An expert system for selecting wart treatment method", Computers in biology and medicine, 81, 167-175, 2017.

[14] S. R. Garner, "Weka: The waikato environment for knowledge analysis,"In Proc. of the New Zealand Computer Science Research Students Conference, pp. 57–64, 1995.

[15] A. G. C.Menezes, O. M. Almeida, and F. R. Barbosa, "Use of decision tree algorithms to diagnose incipient faults in power transformers". In 2018 Simposio Brasileiro de Sistemas Eletricos (SBSE) (pp. 1-6). IEEE, (2018, May).

[16] S. A. Nabi, S. Rasool, and P. Premchand, "Detection and extraction of videos using decision trees",(IJACSA) International Journal of Advanced Computer Science and Applications, 2(12), 147-151. , 2011.

[17] Y.Guo, and K.Wu, "Research on case retrieval of Bayesian network under big data", Data & Knowledge Engineering, 118, 1-13, 2018.

[18] S.De, and B.Chakraborty, "Case Based Reasoning (CBR) Methodology for Car Fault Diagnosis System (CFDS) Using Decision Tree and Jaccard Similarity Method", In 2018 3rd International Conference for Convergence in Technology (I2CT)(pp. 1-6). IEEE, (2018, April).