International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 10, No.1 (Oct.-2021)

http://dx.doi.org/10.12785/ijcds/100190

A new approach for case acquisition in CBR based on multi-label text categorization: a case study in child's traumatic brain injuries

[1,2]Hichem Benfriha*, [1]Baghdad Atmani, [1]Fatiha Barigou,[3]Belarbi Khemliche, [4]Ali Douah, [4]Zakaria Zoheir Addou and [3]Nabil Tabet Aoul.

[1]Laboratoire d'informatique d'Oran (LIO), Université d'Oran 1 Ahmed Benbella, BP 1524 El M'naouer- 31000, Oran, Algeria.
[2]Département de Tronc Commun Sciences Techniques (TCST), Université Mustapha Stambouli de Mascara BP 305, Mascara, Algeria.
[3]Réanimation médicale, Etablissement Hospitalo-universitaire, Faculté de médecine, Université d'Oran 1 Ahmed Benbella, Oran, Algeria.
[4]Réanimation Polyvalente Pédiatrique EHS Canastel, Faculté de médecine, Université d'Oran 1 Ahmed Benbella, Oran, Algeria.

Received 02 Jun. 2020, Revised 22 Jul. 2021, Accepted 05 Aug. 2021, Published 28 Oct. 2021

Abstract: Case-based reasoning (CBR) is an approach to solving new problems based on those already solved in the past. This means searching in previous cases for one that is similar to the new one and reusing it in this new problem situation. In the literature, there are several CBR developments that have paid particular attention to the stages of the process without paying as much attention to the Case Acquisition (CA) stage. This paper focuses on this task through the use of a Multi-Label Text Categorization (MLTC) approach. The objective of this work, is to automatically complete additional information on cases that were obtained from the Magnetic Resonance Imaging (MRI) scan reports provided by the pediatric intensive care unit of Oran hospital -Algeria. The results suggest that the methodology we have proposed and which we call Multi-Label Text Categorization for Cases Acquisition (MLTC4CA) is a promising way to add automatically values' labels to the case that represents a medical situation related to a child victim of Traumatic Brain Injuries (TBIs).

Keywords: Case Acquisition, Case Based Reasoning, Multi-label Learning, Text Categorization, Traumatic Brain Injuries, Road Accident.

1. INTRODUCTION

CBR is an artificial intelligence approach synthesizing or presenting solutions to problems based on previous experiences [1]. These problems may be of a variety of natures. In fact, many researchers have successfully developed CBR approach to solve problems in different disciplines [2].

The main reason for CBR being of interest is that the methodology provides a computational model based on past experiences that are very close to human reasoning. In CBR methodology, experiences are represented by what are called cases. As defined in [3], cases are a contextualized piece of knowledge representing an experience that teaches a lesson and it is used to help solve future problems. Basically, the cases can be seen as a record of experience and are usually organized in two parts: a problem part and a solution part.

The CBR research community has a great challenge to perform this type of reasoning from cases using a retrieval, reuse, revise, and retrain process [4-7]. For example, when new cases are presented, their similar cases are retrieved before being adapted or directly used to make action, decision or prediction.

However, prior to all this process, case acquisition and representation of cases are necessary and important steps for the proper functioning or prediction of results. For this, a multi-label classification approach proposed in [8] is used in this work for case acquisition. Our objective is to automate the acquisition of knowledge of a medical

E-mail: hichem.benfriha@gmail.com, baghdad.atmani@gmail.com, fatbarigou@gmail.com, khemlichebb@gmail.com, douah-ali@hotmail.fr, addouzakariacanastel@gmail.com, tabetrea@yahoo.fr.

CBR system that can be used in the future by physicians as a support to diagnose new cases with TBI.

This step is in fact very important because the problematic part of the case representing the medical situation of a child suffering from TBI lacks information on head injuries. To overcome this limitation, we propose an approach to automatically complete this information.

The case acquisition in this study is performed in two steps: first the neurological, hydrodynamics and respiratory assessments are manually recorded, second the intracranial lesions are automatically identified from the MRI report by using the MLTC framework.

For the automatic case acquisition, two concepts have been considered. First, through the use of a text categorization process [13,14] because the identification of these lesions is done on the basis of Magnetic Resonance Imaging (MRI) scan reports, second, through the use of multi-label classification [15] because a case may have more than one head lesions.

Knowledge acquisition on cases from textual data has been studied before in many works with different techniques [9-12,16-28] but none of these works have used the Multi-Label (ML) approach for this task.

Multi-label text categorization is an important task in modern text mining applications; it consists in identifying a set of category labels for a text document [15]. A framework for multi-label text categorization has been proposed in a previous work [8] to identify intracranial lesions in child victims of TBI.

The authors performed many experiments using a variety of term weighting schemes (binary, TF, TFIDF), classifiers (such as Naïve Bayes, decision tree, …) and different multi-label problem transformation methods such as Binary relevance (BR), Classifier Chains (CC) and Label Powerset (LP) to determine which combination is the most appropriate to generate the best model for categorizing multi-label clinical reports.

To apply this task on Magnetic Resonance Imaging scan reports, we use in this paper three groups of approaches; data transformation, method adaptation and ensemble methods combined with a variety of classifiers and terms weighting schemes to find the best model for detecting cerebrals lesions (or labels) and automatically add them to new cases.

For the experimental study, we used 174 MRI reports collected from the intensive care unit of Oran hospital in Algeria. The dataset is randomly divided into train set and test set. The training set consists of 120 MRI reports (69%), and the test set includes the remaining data (31%).

We conducted the experiments with CLR transformation approach, BRkNN and MLkNN adaptations approaches and ECC and RAkEL ensemble methods. The results of experiments have shown that the use of the MLTC framework is a promising approach for the automatic case acquisition phase., We found, in fact, that the Ensemble Classifier Chains approach used with the Naïve Bays classifier and the TF-IDF weighting gave the best performance for the different evaluation measures.

This paper is structured as follows: section II first provides an overview of current CBR systems by considering the case acquisition process, and then a basic concept of multi-label learning used in our study for case acquisition is introduced. Section III presents our proposed methodology which we have named: Multi-Label Text Categorization for Case Acquisition (MLTC4CA) with a detailed description of TBI case-base and acquisition mechanism. Section IV presents the experimental and results study. Finally, Section V provides conclusions and future work.

2. BACKGROUND

A. *Works related to case acquisition in CBR*

knowledge is stored as cases. There are two phases in this process:

- The first one, called the problem formulation includes case acquisition and representation, case indexing and case storage.
- The second one is the CBR circle and it includes four processes: retrieve, reuse, revise and retain.

In this paper, we will focus on the first problem and especially in case acquisition. Currently, many works are paying great attention to the automatic case acquisition in the CBR process. In this section two distinct studies are discussed. Case acquisition using only Natural Language Processing (NLP) and case acquisition using variations of NLP techniques and/or machine or deep learning approaches.

In [9], Yang et al. propose a methodology for automatic acquisition and creation of cases for maintenance of complex systems such as trains and aircraft. The operational data collected contains the symptom messages generated by the various sensors installed on the system, which are then transformed into free-text. This elaborated methodology uses NLP in the initial stage of CBR system development. Automatic acquisition is used to reduce the difficulty of case creation and management in CBR diagnostic applications.

In [10], Berghofer et al. propose a system called SCOOBIE for automatic case acquisition from texts. SCOOBIE is an ontology-based information extraction system, which uses symbolic background knowledge for extracting information from text documents to bring additional values to cases.

Bach et el. propose a CBR system applied in a machine diagnosis customer support scenario [11]. Among the techniques used in this system there is automatic text-based case acquisition using NLP techniques.

Dufour et al. use NLP for the automatic knowledge acquisition from the text to support a process-oriented case-based reasoning [12]. In this study, NLP techniques are used to extract verbs, their complements and relevant modifiers from procedural texts in order to build a rich Knowledge base for CBR processes.

In [16], Sizov et al. analyze free text documents to identify knowledge that could be used to generate cases from this text. In this work, NLP techniques are also used for extracting causal relational graphs to avoid the laborious process of manual case acquisition from aircraft incidents.

Cordier et al. develop in [17], a system called Taaable, it provides cooking recipes in response to queries from users. This system contains a combination of various methods and techniques from knowledge-based systems such as CBR, knowledge representation, knowledge acquisition and discovery, knowledge management, and NLP techniques. For the automatic case acquisition, NLP is used to extract terms like named entities or sequences of words and they are explicitly associated with their respective and most specific classes in the ontology.

Sizov et al. in [18], propose a novel case retrieval method called evidence driven retrieval (EDR). In this work, cases are automatically acquired from incident reports from the Transportation Safety Board of Canada. NLP techniques are used during automatic acquisition.

In [19], Dufour and Lieber propose a CBR system called CRAQPOT that retrieves and adapts processes represented as instruction texts. In this system, cases are automatically extracted from texts using NLP techniques.

Reuss et al. propose a framework in the aircraft domain that integrates different algorithms and methods to transform the available data into knowledge for vocabulary, similarity measures, and cases [20]. A semi-automatic case acquisition is applied using NLP techniques on textual data.

Shen et al. in [21] introduce an integrated system of text mining and case-based reasoning to help designers retrieve the most similar green building cases for references when producing design for a new green building. For the components of automatic case acquisition from textual data, NLP techniques are applied to add some values in the initial case-based.

In [22], Manzoor et al. propose a methodology for automatic case acquisition of a quality case-base using genetic algorithm. This iterative approach has been applied to randomly generating the initial case base and later improve it prior to submission. The performance of the proposed approach has been evaluated and discussed for the examination scheduling problem.

We He propose the use of text mining and web 2.0 technologies to improve and enhance user experience with CBR system [23]. The results of this study showed that applying concept extraction, categorization and clustering to identify terms, concepts and categories are promising ways to automatic knowledge acquisition and can provide additional value to cases.

Bach et al. in [24], propose a methodology for extraction and acquisition of cases from times series in CBR process. The research presented a clustering-based method for automatically detecting and capturing predictive cases originally created by domain experts.

In [25], Nasiri et al. propose a medical CBR approach called DePicT, which combines image classification and text information. In this study, NLP and words associations are applied to build a semantic profile of the textual data record of the cases.

Wienhofen et al. propose in [26], a case acquisition approach to build an initial case-base in CBR process in an industrial domain. Acquisition is done from multiple textual sources like observation, unstructured interview, structured interview and questionnaire. Several methods have been applied for knowledge acquisition and the focus has been on a user-centered iterative process.

Mathisen et al. propose the use of a deep learning approach in [27]to create a model for the automatic case acquisition to be used in a CBR process to predict operational windows for marine operations.

In [28], Amin et al. combine deep learning and big data to automate the case acquisition from text. NLP techniques and word embedding are used for text preprocessing to convert words to vectors. For the automatic CA, three neural network models are used: CNNs, RNNs and LSTMs to explore and compare accuracy results.

Amin et al. propose in [29], a semi-automatic case acquisition approach from text that contain context-free abbreviations, grammatically incorrect text and mixed language. In this study, Manhattan Long short-term memory (MaLSTM) is used for the case acquisition process to minimize the effort from human experts.

Through this series of works, we can see that the acquisition of knowledge about cases from textual data has been studied before in various domains with different techniques [9-12,16-28] but none of these works used the multi-label approach.

Table 1 show the case acquisition has been performed automatically or semi-automatically and that the NLP and machine learning algorithms are used for this task.

TABLE I. RELATED WORKS ON CASE ACQUISITION

Ref	Mechanism of acquisition	Acquisition techniques	Application domain of the proposed CBR
[9]	Automatic	NLP	Train and aircraft maintenance
[10]			Industrial
[11]			Machine diagnosis
[12]			Cooking recipes
[16]			Aircraft incidents
[17]			Cooking recipes
[18]			Transportation Safety
[19]			Cooking recipes
[20]	Semi-automatic		Aircraft incidents
[21]	Automatic		Green building
[22]		Genetic algorithm	Scheduling
[23]		Concept extraction, categorization and clustering	Education
[24]		Clustering	Times series
[25]		NLP and words associations	Disease diagnosis
[26]		NLP and user-centered iterative process	Industrial
[27]		Deep learning	Marine operations
[28]		NLP, CNNs, RNNs and LSTMs	Automobile industry
[29]	Semi-automatic	MaLSTM	Business

B. *Multi-label learning (MLL)*

MLL is a field that has generated a notable interest in recent years, it is becoming increasingly widespread as a predictive data mining task beside its multiple applications to classify different types of knowledge.

As a consequence, MLL is applied in many applications such as multimedia resources labeling, genetics, biology, music classification and text categorization [8,23].

When using text-based MLL, the training corpora contains documents associated with a set of labels. It can classify the label sets of unseen documents on the basis of training documents with known label sets.

In general, one document is present in a ML object and K number of class labels are associated with it as shown in table 2. With the MLL, which applies to text categorization, the categorizer is just a means to assign the right labels [15].

TABLE II. SAMLPE OF MULTI-LABEL CORPORA

	Terms					**Labels**			
Docs	t_1	t_2	t_3	..	t_m	L_1	L_2	..	L_p
d_1	0	1	0	..	1	1	0	0	1
d_2	1	1	0	..	0	1	0	1	1
d_3	1	0	0	..	1	0	1	0	1
..									
d_n	0	1	1	..	1	1	1	1	1

Formally, let D (d_1, d_2, d_3,..,d_n, L_i) is the set of documents, whereas T (t_1, t_2, t_3,.., t_m) represent the terms and Li represents the target class label, for a given instance D_I in T. In single label binary classification task L_i=2 takes only two modalities, D_i will belong to only one value of L_i. In single label multi-class classification, L_i>2 takes more than two values, D_i belongs to one of the values of L_i.

In MLL, there is more than one target class L_i, L_j,.., Lp. In this case, each target label is binary and takes only values 0 or 1; each document will belong to more than one target class.

MLL can be accomplished through three different approaches [15,30], they are data transformation, adaptation methods, and ensemble methods. Fig. 1 summarizes the different approaches and their associated algorithms that will be used in this study.

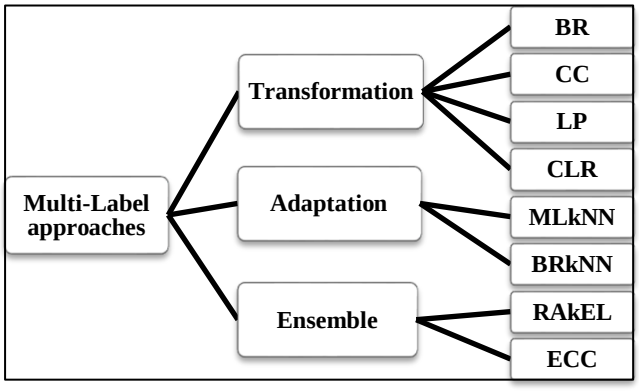


Figure 1. Multi-Label appraoches

1) Transformation Method

Problem transformation methods transform the MLL problem into one or more classification or regression problems with a single label marker. For a small single label problem, there is a plethora of automatic learning algorithms. Transformation methods can be grouped into three categories: Binary Relevance, Classifier Chains and Label Powerset.

- **Binary Relevance (BR)**: is the most popular and simplest method of this class of approaches. It transforms the MLL problem into Q problems of classification or single-label regression. It addresses the MLL problem by learning a classifier for each label, using all examples labeled with this label as positive examples and all other examples as negative [30]. During prediction, each binary classifier predicts whether its label is relevant for the example given or not, which gives at the end, a set of relevant labels. In the classification scenario, the labels are classified according to the probability associated with each label by the respective binary classifier. BR's major advantage is its low learning complexity (relative to a basic classifier) which allows it to easily scale up and therefore be a very good candidate for MLL problems from large data. However, BR is unaware of the existence of potential correlations between labels [15]. In addition, binary classifiers may suffer from the imbalance between classes (1 and 0) if the number of labels is large and the density of labels is low.

- **Classifier Chains (CC)**: is a method closely related to the BR method proposed by [31, 32].CC is an improvement of the BR method that also transforms the MLL problem into Q problems of classification or single-label regression. However, classifiers are trained in a random order defined before the learning phase [1..,j,..Q]where each binary classifier hj learning a label yj adds all the labels associated with the classifiers that precede it in the chain (i.e. y1, ..., yj-1) in its attribute space. Like BR, for a new example, CC returns all the predictions generated by all classifiers. Its advantages are its speed of model learning and its modeling of correlations between labels, but its random definition of the learning order of the models remains a weakness.

- **Label Powerset (LP)**: transforms the ML problem into a single, multi-class, single-label learning problem. LP considers each combination of labels present in the learning set as a class and then learns a multi-class classifier h. For a new example, the classifier returns the most likely class (i.e. combination of labels) [33]. The main advantage of LP is its low complexity of model calculation but also its natural use of correlations between labels. However, some classes may be difficult to learn if the number of labels is large and the number of examples is small. The number of classes is at most equal to min (2Q, N) where Q is the number of labels and N is the number of instances xi. Its other disadvantage is that it does not provide good generalizations: it does not allow predicting new classes (label combinations) that do not exist in the learning set.

- **Calibrated Label Ranking (CLR)**: proposed by [34] is an extension of Ranking by Pairwise Comparison (RPC) approach. In RPC, relevant and irrelevant labels are not distinguished separately. However, CLR augments the original ML dataset set with a virtual label Lv also known as calibration label then it partitioned the labels into relevant and irrelevant class labels and constructs Q(Q-1)/2 binary classifiers as in RPC to represent relationship between each label Li and Lv. Finally, all the labels (Li+Lv) are ranked where relevant labels are clearly separated from irrelevant labels by Lv.

2) Adaptation Method

ML methods that adapt, extend and customize an existing machine learning algorithm for the MLL task are called algorithm adaptation methods. Extended methods are capable of directly managing ML data. Algorithms like MLkNN and BRkNN use this approach.

- **MLkNN**: The nearest k-neighbors in multi-label (ML-kNN) are an extension of the nearest k-neighbors (kNN) algorithm [35]. The recovery of the nearest k-neighbours is the same as in the traditional kNN algorithm.ML-kNN is a BR type method that combines the standard kNN algorithm with Bayesian inference. In the learning phase, ML-kNN estimates the a priori and a posteriori probability of each label from the learning examples. For a new example xi, ML-kNN calculates its nearest k neighbors and then measures the frequency of each label in that neighborhood. This frequency is then combined with the probabilities estimated in the learning phase to determine its set of labels according to the principle of the maximum a posteriori (MAP).

- **BRkNN**: proposed by [36] is an adaptation of the kNN algorithm that is conceptually equivalent to using BR in conjunction with the kNN algorithm. Initially kNN is applied on the multi-label data to obtain k neighbors. Once neighbors are obtained, then BR classifier uses these neighbors independently for prediction of each label.

3) Ensemble methods

Ensemble methods whose basic classifiers are MLL are considered by [37]as a special group of methods because they are developed in addition to problem transformation and algorithm adaptation approaches. The best-known problem transformation sets are the RAkEL system and Sets of Classifiers Chains (ECC).

- **Random k-label sets (RAkEL)**: proposed by [38] is an ensemble of multiple LP classifiers having different combination of all labels referred as a label set. In this approach the complexity of LP is reduced by considering only a group of labels together even if an instance has many labels associated. It draws m random subsets of labels of size k from all the labels L and forms a label calibration classifier using each set of labels. A simple voting process determines the final set of labels for a given example. The advantage of this method is that it allows the prediction of new label combinations that do not exist in the learning set. In addition, it makes it possible to naturally exploit correlations between labels. However, it does not explore all label subsets sufficiently to capture all correlations and focuses only on learning a few k-size subsets.
- **Ensemble of Classifier Chains (ECC)**: are sets of CC that represent a ML classification technique based on classifier [31]. In this approach, m CC classifiers are trained and each one is trained with a random string order of (L) and a random subset of instances, which improves the accuracy of prediction. Therefore, each classifier model is likely to be unique and capable of giving different ML predictions. These predictions are summed per label so that each label receives a certain number of votes.

Table III is a summary of the multi-label methods described above.

TABLE III. MULTI-LABEL APPRAOCHES

Approaches	Method	How it works	Advantages	Weakness
Transformation	Binary Relevance	The MLL problem is transformed into learning a binary classifier for each label	Low learning complexity; It can be easily scaled up	Does not know whether there are any potential correlations between labels
	Classifier Chain	Like BR but classifiers are trained in a random order defined before the learning phase	Speed of model learning; modeling of correlations between labels	Its random definition of the learning order of the models
	Label Powerset	Considers each combination of labels present in the learning set as a class and then learns a multi-class classifier	Low complexity of model calculation; Natural use of correlations between labels.	Classes may be difficult to learn if the number of labels is large and the number of examples is small; It does not allow predicting new classes that do not exist in the learning set.
	Calibrated Label Ranking	Augments the dataset with a virtual calibration label then it partitioned the labels into relevant and irrelevant class labels and constructs Q(Q-1)/2 binary classifiers.	Relevant labels are separated from irrelevant labels.	It only applies to soft classifiers, which are able to provide confidence scores with the prediction.
Adaptation	MLkNN	Is a BR type method that combines the standard kNN algorithm with Bayesian inference.	Simplicity and low computational complexity.	Any potential correlation information among labels is disregarded.
	BRkNN	Initially kNN is applied to obtain *k* neighbors, then BR classifier uses these neighbors independently for prediction of each label.	Applies better in domains with large number of labels and examples; requiring low response times	Low prediction result
Ensemble	RAkEL	It draws *m* random subsets of labels of size *k* from all the labels L and forms a label calibration classifier using each set of labels.	Low computational complexity by considering only a group of labels; it allows the prediction of new label combinations that do not exist in the learning set; exploit correlations between labels.	It does not explore all label subsets sufficiently to capture all correlations and focuses only on learning a few *k*-size subsets.
	ECC	*m* CC classifiers are trained and each one is trained with a random string order of (L) and a random subset of instances.	Improves the accuracy of prediction.	Time complexity.

3. THE PROPOSED METHODOLOGY

Our methodology is mainly based on using the MLTC framework proposed by [8] to the CA process. Before

describing this methodology, we will first describe the Traumatic Brain Injuries (TBI) cases base.

A. *The case base TBI collection*

Childs head injuries are common and have a bimodal incidence before the age of 15 years. Their seriousness lies in the occurrence of intracranial lesions like cerebral edema, fracture, brain hemorrhage and foreign substance.

Initially, and as early as the pre-hospital phase, management consists of maintaining vital functions. The therapeutic and monitoring modalities are determined by altered consciousness, signs of concentration, skull fracture and accident kinetics.

Since there is no predefined representation of the case for CBR systems, we have opted, in this first solution, for a simple representation in vector form. TBI-case base is collected from the pediatric intensive care unit of Oran hospital between 2017 and 2020 and contains 174 child cases.

Taking into account the information collected from the child record and the files made available to us, the TBI-case is composed of 40 descriptors organized in four categories (1- patient information, 2- clinical symptoms, 3- intracranial lesions, 4- Actions to be taken) and split into a problem part and a solution part.

a) Problem part

- Patient Information (PI): sex, age and weight.
- Clinical Symptoms (CS):
  - Hydrodynamics: Temperature, systolic pressure, diastolic pressure, etc.
  - Neurological : Glasgow Score, convulsion, paralysis, etc.
  - Respiratory: respiration rate, SpO2, etc.
- Intracranial lesions (IL): cerebral edema (IL1), fracture (IL2), brain hemorrhage (IL3) and foreign body (IL4).

b) Solution part

The solution part represents the Gestures Performed (GP) or the treatment to be given to the child, it contains 8 actions: O2 therapy, intubation of the child, perfusion, transfusion, etc. In the end, the TBI-case can be represented under the following vector:

TBI-case (PI_1, PI_2, PI_3, CS_1, ..., CS_{25}, IL_1, IL_2, IL_3, IL_4, GP_1,.., GP_8).

B. *MLTC4CA methodology*

During the modeling phase, and as a result of the examination of a set of cases originating from the Pediatric intensive care unit of Oran Hospital, two observations attracted our attention:

The first is that a child victim of a head injury may suffer from more than one intracranial lesion, for example, he/she may have a problem of cerebral edema (IL1), and a fracture (IL2) at the same time.

The second is that some neurological exams were missing from the problem description. This lack of information corresponds to the intracranial lesions caused by the head trauma.

Furthermore, the physicians in this unit expressed a need for a tool that would allow them to address this problem of neurological diagnosis which is frequently absent in some reports.

Hence, considering the doctors' request and also the two observations mentioned above, this paper is motivated by the problem of identifying automatically intracranial lesions of children with head injuries in order to complete the problem part of a case with this missing information.

We formalized this problem as a multi-label text categorization task. Therefore, the automatic acquisition of IL label values (0/1) is carried out using multi-label text categorization of the MRI scan reports.

In summary, the acquisition of the TBI-case in our study is performed in two steps: first the neurological, hydrodynamics and respiratory assessments (PI, CS, GP) are manually recorded (Fig.2-a), second the intracranial lesions (IL1, IL2, IL3, IL4) are automatically identified from the MRI report by using the MLTC framework (Fig.2-b).

Our MLTC4CA methodology consists of the use of the MLTC framework to identify values of IL (IL1, IL2, IL3, IL4) from the MRI scan report. In this methodology TBI-case has two contents: before and after the automatic acquisition.

As illustrated in Fig.2 below, each TBI-case can be assigned to several Intracranial Lesions (IL1, IL2, IL3, and IL4) simultaneously.

For this, we need to build the best MLTC model that can automatically predict the set of labels (IL1, IL2, IL3, IL4) for new cases. A full description of this process is provided in the following section.

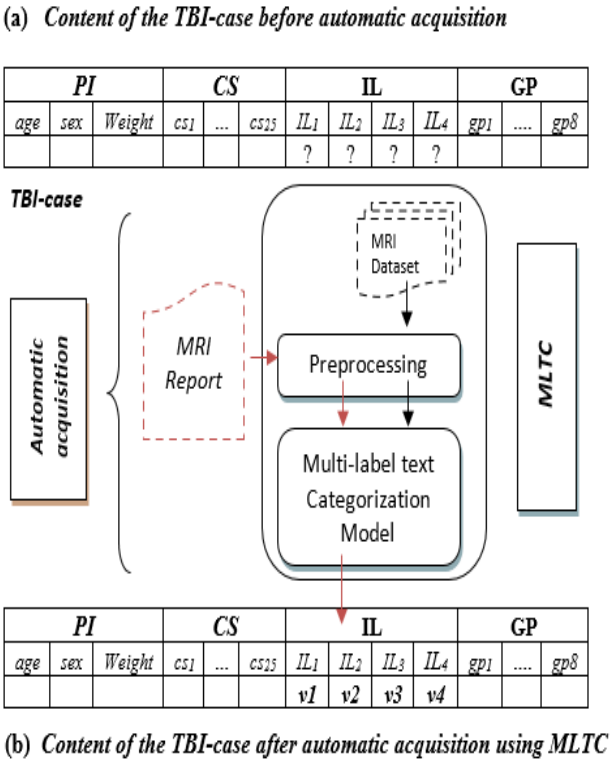


Figure 2. The MLTC4CA methodology.

C. MRI dataset

As noted in the introduction section, MRI dataset for TBI's is collected from the pediatric intensive care unit of Oran hospital.

The source data for our corpus is received in the form of papers written by physicians during the patient's examination in the intensive care unit. The electronic version still does not exist. So, we had to enter these reports by ourselves. Given the unclear and non-readable wording, a doctor had to be present during this data entry operation. This presence is not always possible because the doctor has other occupations. Considering the time needed to enter a large amount of data, and in order to further develop our research, we stopped the data entry at 174 reports.

We have randomly extracted 120 MRI reports (69%) for training, and keep the remaining data (31%) for evaluation. We would point out that this dataset is about children with TBI involving four ILs: cerebral edema, fracture, brain hemorrhage and foreign body.

Cardinality and density [39] are used for evaluating the characteristics of the training dataset. Considering S this dataset of reports with n instances and Yi the set of labels for the i-th instance, cardinality is the average of the number of labels of each report belonging to S, defined by equation (1), and the density is the cardinality of S divided by |L|, defined by equation (2). Characteristics of MRI dataset are summarized in Table IV.

TABLE IV. CHARACTERISTICS OF MRI DATASET

Number of Instances	Size of Vocabulary	Number of labels (|ILs|)	Cardinality	Density
120	1307	04	1.585	0.156

$$Cardinality(s) = \frac{1}{n}\sum_{i=1}^{n}|Y_i| \qquad (1)$$

$$Density(s) = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i|}{|L|} \qquad (2)$$

where L is the set of labels (ILs).

D. Multi-Label Text Categorization (MLTC) framework

As mentioned earlier, MLTC framework is used to automatically obtain from MRI report, information about IL caused in child's road accident to complement the problem part of the case.

But to extract this information we must find and use the best model of knowledge. To this end we have experimented and evaluated a variety of ML approaches and base classifiers.

In [8], authors have proposed a MLTC framework for a best cerebral lesion's identification. The task has been carried out with different multi-label approaches based on problem transformation such as Binary relevance (BR), Classifier Chains (CC) and Label Powerset (LP) with a variety of base classifiers (Sequential minimal optimization, decision tree, naive Bayes, k-nearest neighbours) and different weightings schemes (Binary, term frequency, frequency-inverse document frequency). In this study,

In order to be able to choose the best classifiers to be used in this work, we evaluated 10 different ones that are: NB, SMO, IB1, C4.5, AddaBostM1, Random Forest, BRkNN, MLkNN, IBLR and BPMLL. We then found that the best results for the identification of intracranial lesions can be obtained when using the NB, SMO, C4.5 and KNN classifiers. Thus, these classifiers have been selected for the next experiments.

To decide on the most appropriate multi-label approach, we have, in addition to the algorithms used in [8], experimented with other transformation and adaptation algorithms and also with the ensemble approach.

Fig. 3 illustrates the different components of the developed framework.

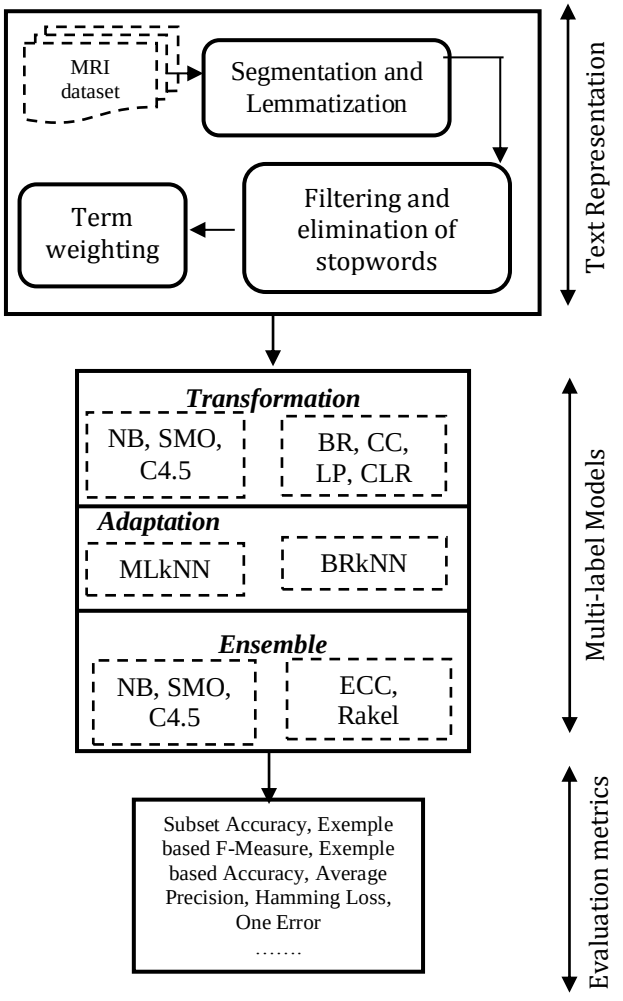


Figure 3. The MLTC Framework

1 *Text representation*

In this section, we will present all the linguistic preprocessing leading to the construction of the numeric representation of the MRI reports obtained from the Pediatric Intensive Care Unit of Oran hospital. This representation is known as bag of words model.

This preprocessing includes segmentation character sequences present in report into distinct word units, normalizing words to provide a canonical form for each word, and filtering to remove the most frequent words.

- **Segmentation and Lemmatization**: This step is a crucial step since poor segmentation and lemmatization has a significant negative impact on the outcome of the categorization [13]. Segmentation of MRI's reports consists in separating a sequence of characters into semantic elements or medical words. A word type is the class of all words having the same sequence of characters and a medical term is a type of word that is kept to form the vocabulary. Lemmatization is an extensive linguistic analysis designed to remove inflectional variants of medical words in order to return them to their lemmatized form. For verbs, this transformation provides the form for the infinitive, for adjectives the masculine singular form and for nouns the singular form.
- **Filtering and elimination of stop words:** after lemmatization, cleaning is performed to remove the non-representative or stop words [14]. These non-representative words are the grammatical words which are irrelevant to report contents, so they need to be removed for more efficiency [40]. In [8], authors prepared a list of stop words with the help of doctors. This list is used to clean up report texts.
- **Term weighting:** reflect the importance of a term in a specific report and it refers to the step of calculating and assigning the weight for each term as its importance degree in order to improve text categorization [41]. Three weighting schemes are used: binary, term frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF) weightings.

2 *Multi-label models*

As mentioned previously, we have performed results obtained in [8] by using other ML transformations, adaptations and ensemble approaches, and others base classifiers to build the best f multi-label model. All ML approaches used in this study are discussed in section II-B. Mulan platform [42] is used as well as cross-validation with ten folds as the dataset is small with only 120 MRI scan report.

3 *Evaluation*

This section presents the various metrics that have been proposed in the literature and used in our study with the aim to compare and evaluate the performance of the proposed framework by varying multi-label approaches, classifiers and weighting schemes to find the best MLTC model that will be used in the acquisition of TBI_case.

A variety of metrics are proposed in [15, 30] to evaluate the performance of the MLL systems. Fig. 4 shows the different metrics evaluation applied in MLTC. They are divided into binary bipartition and label ranking metrics.

a) Binary Bipartition

This group operated on using a binary vector that indicates which of the labels belonging to the MRI dataset is relevant to the processed sample. Metrics using this approach operate using the confusion matrix.

Measurement can be made by example (report) or by label (IL).

- **Example – based measures:** each report R is evaluated and then averaged according to the *n* reports considered. Therefore, the same weight is assigned to every report in the final score. Those metrics are defined by equations 3 to 8, the ⟦ ⟧operator in equation 7 represents the Iverson operator. According to equation 8, *k* is the number of labels and Δ stands for the symmetric difference between L*i*, the exact label set of the ith *R*, and *Pi*, the predicted one.

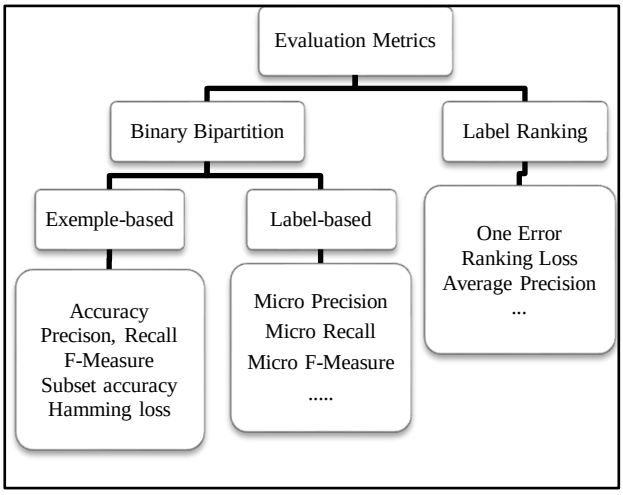


Figure 4. Evaluation metrics used in MLL.

$$Accuracy = \frac{1}{n}\sum_{i=1}^{n}\frac{|Li \cap Pi|}{|Li \cup Pi|} \quad (3)$$

$$Precision = \frac{1}{n}\sum_{i=1}^{n}\frac{|Li \cap Pi|}{|Pi|} \quad (4)$$

$$Recall = \frac{1}{n}\sum_{i=1}^{n}\frac{|Li \cap Pi|}{|Pi|} \quad (5)$$

$$F_{Measure} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (6)$$

$$Subset\ accuracy = \frac{1}{n}\sum_{i=1}^{n}[\![Li = Pi]\!] \quad (7)$$

$$HammingLoss = \frac{1}{n}\frac{1}{k}\sum_{i=1}^{n}|Li\ \Delta\ Pi| \quad (8)$$

- **Label-based measures**: each label is computed independently before it is averaged. Two methods can be used: Micro and Macro averaging. In this study we have used only the micro- averaging measures given in equations 9, 10 and 11. This form aggregates firstly counters hits and misses for each label then the metric is computed only once. Let TP_j, FP_j, FN_jdenote the true-positives, false-positives and false-negatives for the class-label j, k is the number of labels.

$$Micro\ Precision\ (MP) = \frac{\sum_{j=1}^{k} TP_j}{\sum_{j=1}^{k} TP_j + \sum_{j=1}^{k} FP_j} \quad (9)$$

$$Micro\ Recall\ (MR) = \frac{\sum_{j=1}^{k} TP_j}{\sum_{j=1}^{k} TP_j + \sum_{j=1}^{k} FN_j} \quad (10)$$

$$Micro\ \mathrm{F-measure} = 2 \times \frac{MP \times MR}{\mathrm{MP + MR}} \quad (11)$$

b) Label Ranking

This group operated on using a ranking of labels, so a confidence degree or belonging probability of each label is calculated. In the equations 12, 13 end 14, the rank (x_i, l) is defined as a function allowing to calculate, for the instance xi and its corresponding label l ∈L, with a known position, the degree of confidence of l concerning the prediction P_i returned by the classifier.

$$Average\ Precision = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|Li|}\sum_{l \in Li}\frac{|\{l' : rank(x_i, l') \leq rank(x_i, l), l' \in L_i\}|}{rank(x_i, l)} \quad (12)$$

$$Ranking\ Loss = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|L_i|.|\overline{L_i}|}|l_a, l_b : rank(x_i, l_a) > rank(x_i, l_b), (l_i, l_b) \in L_i \times \overline{L_i}| \quad (13)$$

$$One\ Error = \frac{1}{n}\sum_{i=1}^{n}[\![\mathrm{argmax}\ \langle rang(x_i, l)\rangle \neq L_i]\!] \quad (14)$$

4. EXPERIMENTATION AND RESULTS

In this section we will present the results of the experimentation carried out in our study. We have divided these experiments into two parts. The first one shows the results obtained by the MLTC system.

This comparative study has been carried out in order to find the best learning algorithm, the best multi-label approach and the best term weighting measure in order to obtain the best MLTC model. The second part shows the results obtained by using the best model previously obtained in the process of case acquisition.

A. MLTC Results

Fig.5 to Fig.12 depict the comparative analysis and the performance of the MLTC approach to find the best model using the MRI dataset.

For a more accurate classification, it is important to note that measures like hamming loss, Ranking loss and One error should be less or equal to zero for the best classifier. The remaining measures, however should be less or equal to one. Experimental results on the MRI dataset for TBI's indicate that the model based on NB classifier, ECC approach and TF-IDF weighting gave the best performance for all metrics.

We divided the discussion of results according to the classification of measures defined in section III-C-4 into two classes: binary bipartition and label ranking.

As illustrated in Fig. 5, the results obtained show that model based on NB classifier, ECC approach and TF-IDF weighting gave the best performance in term of subset accuracy with 55,2% and surpasses all the other models. The lowest values are obtained by the model based on C4.5 classifier, BR and CC approaches and binary weighting with 39,3%.

According to Fig. 6 the model based on NB classifier, ECC approach and TF-IDF weighting outperforms all the other models in term of F-measure with 72,4 %. The same applies to the accuracy where the model based on NB classifier, ECC approach and TF-IDF weighting gave the best performance with 67,2%. Also, the results are 2.1% better than the other models based on NB classifier, BR, CC and Rakel approaches and TF-IDF weighting. The lowest values are obtained by the model based on NB classifier, LP approach and binary weighting with 47,4%.

In term of accuracy and as show in Fig. 7, the result surpasses the model based on NB classifier, CLR approach and TF-IDF weighting by 2,3%. The lowest values are obtained by the model based on MLkNN approach and TF weighting with 47,7%.

In term on hamming loss and as show in Fig. 8, model based on NB classifier, ECC approach and TF-IDF weighting gave the best performance with 12,1% and outperforms those models based on NB classifier, CLR approach and binary weighting by 3,7%. The lowest values are obtained by the models based on: MLkNN approach, TF weighting, and the model based on C4.5 classifier, CLR approach and TF weighting with 24,6%.

Still in binary bipartition and now considering label-based measures and as show in Fig. 9, we notice that in term of micro-average F-Measure, model based on NB classifier, ECC approach and TF-IDF weighting gave the best performance with 79,9% and surpasses the models based on NB classifier, CLR approach and binary weighting by 12%. The lowest values are obtained by the model based on BRkNN approach and TF-IDF weighting with 26,2%.

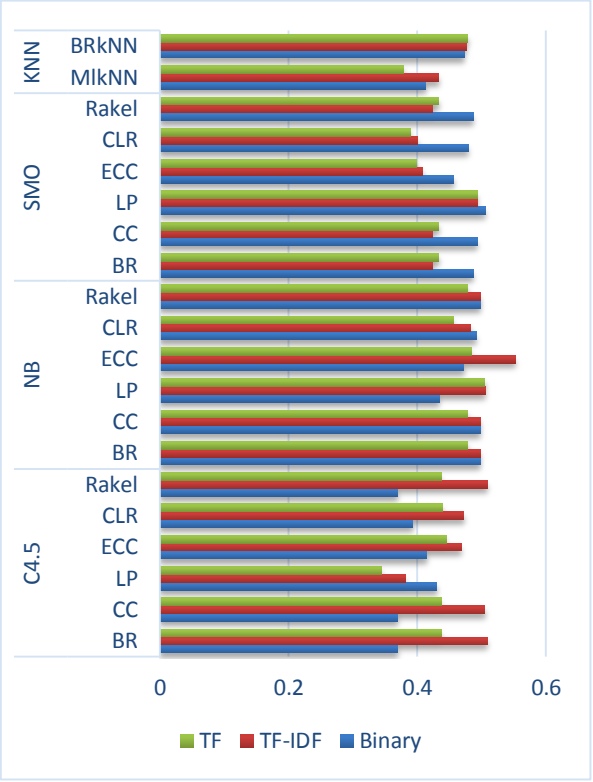


Figure 5. Subset accuracy (↑).

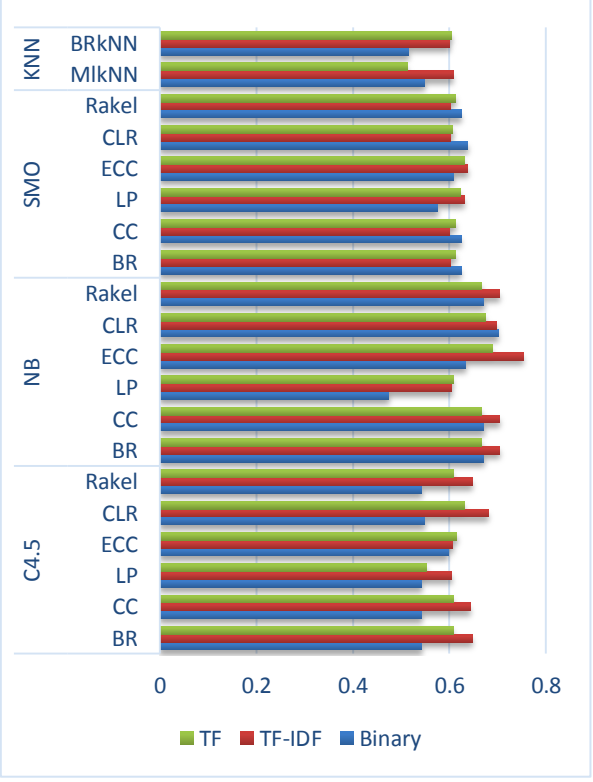


Figure 6. F-Measure (↑).

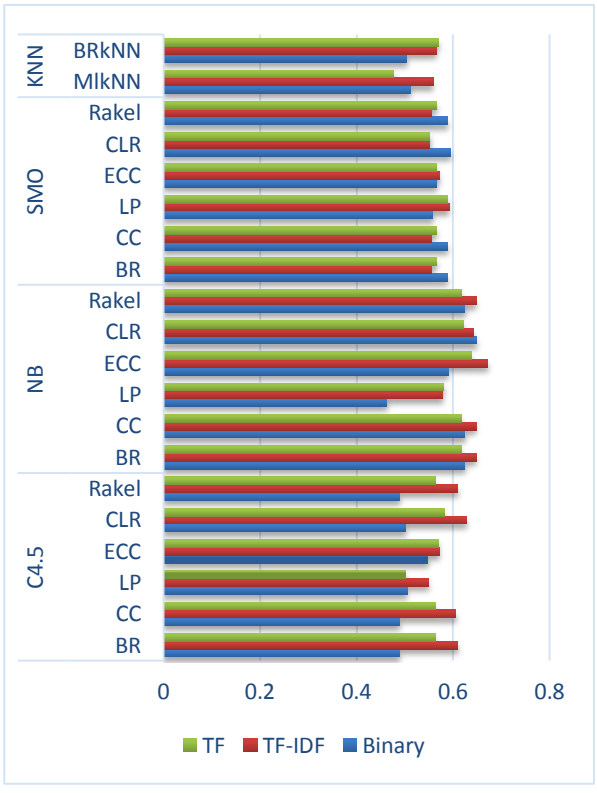


Figure 7. Accuracy(↑).

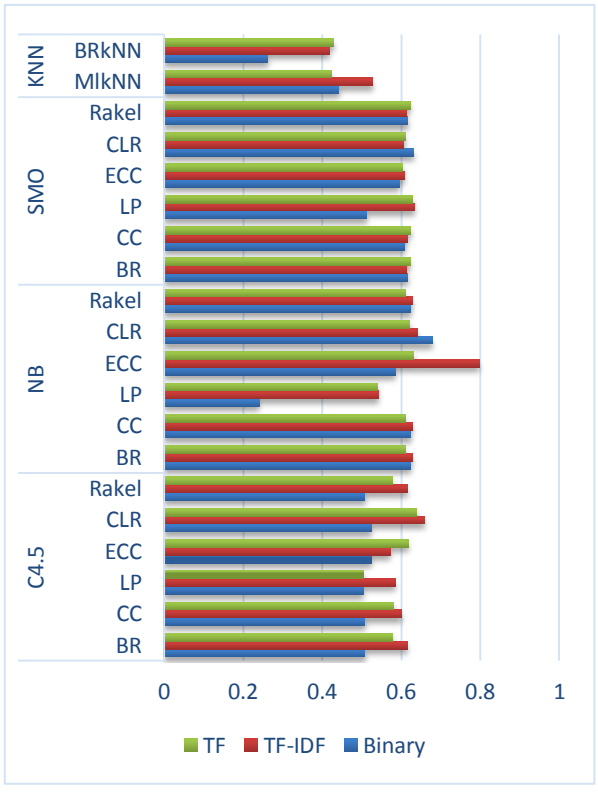


Figure 9. Micro F-Measure (↑).

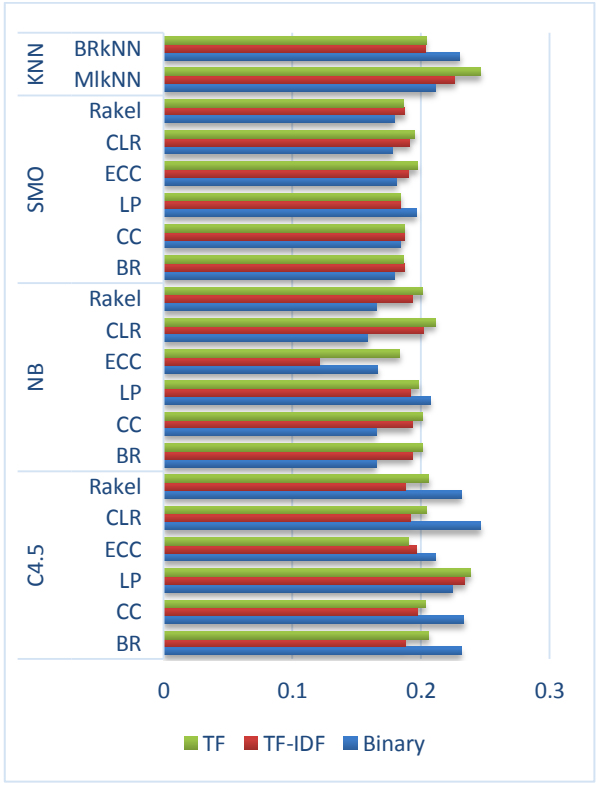


Figure 8. Hamming loss (↓).

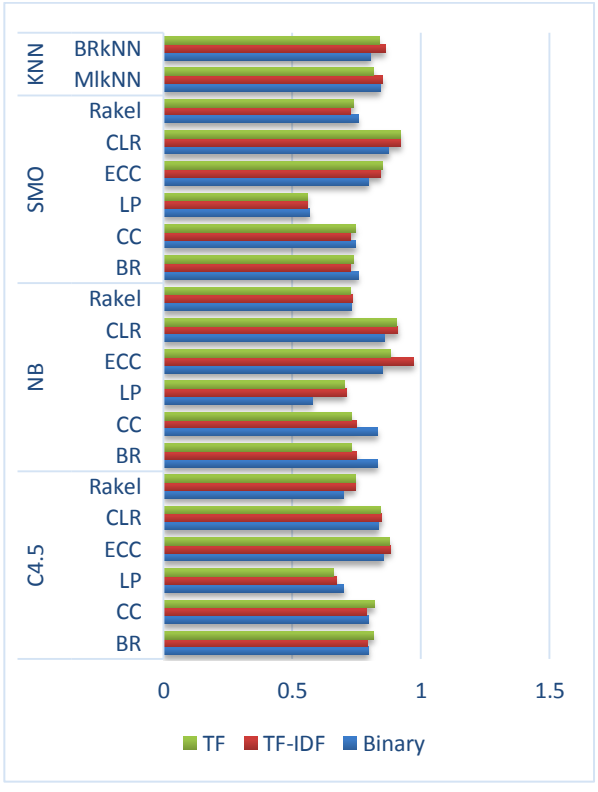


Figure 10. Average Precision (↑).

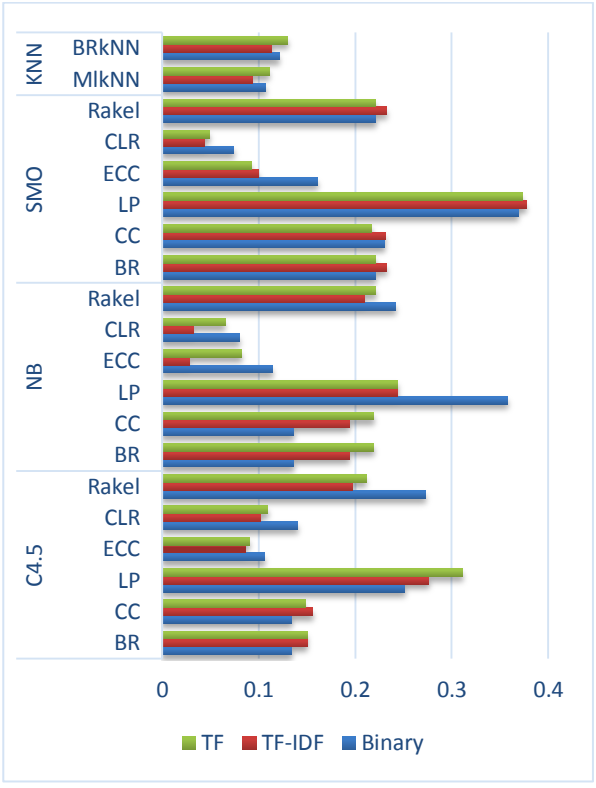


Figure 11. Ranking loss (↓).

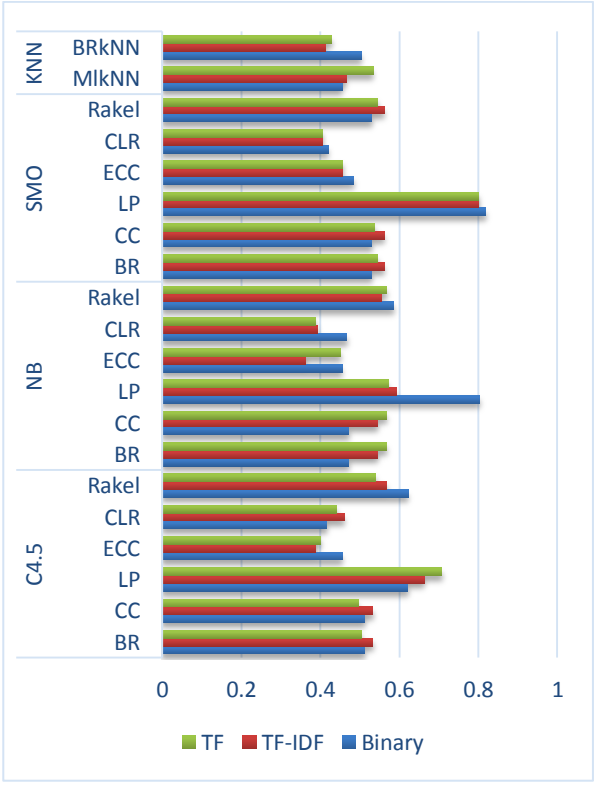


Figure 12. One Error (↓).

In label ranking and as show in Fig. 10, Fig. 11 and Fig. 12, we notice that in term of average precision, model based on NB classifier, ECC approach and TF-IDF weighting gave the best performance with 97,1% and surpasses the models based on SMO classifier, CLR approach and TF-IDF weighting by 4,9%. The lowest values are obtained by the model based on SMO classifier, LP approach and TF weighting with 56%.

In term of ranking loss, model based on NB classifier, ECC approach and TF-IDF weighting gave the best performance with 2,8% and surpasses the models based on NB classifier, CLR approach and TF-IDF weighting by 0,4 %. The lowest values are obtained by the model based on SMO classifier, LP approach and TF-IDF weighting with 36,9%.

In term of one error, model based on NB classifier, ECC approach and TF-IDF weighting gave the best performance with 36,3% and surpasses the models based on NB classifier, CLR approach and TF weighting by 2,4%. The lowest values are obtained by the model based on SMO classifier, LP approach and binary weighting with 81,8%.

Before concluding this evaluation, it will be necessary to quote some remarks on these results:

- The Model based on NB classifier, ECC approach and TF-IDF weighting achieves high performances for all evaluation measures;
- The CLR approach achieves good performances after ECC;
- The LP approach gave poor results for all measures, the reason could be that the dataset size and number of models used with LP may be too small to properly consider label correlation.
- In general, models based on TF-IDF weighting perform better that those based on TF or binary weighting;
- The C4.5 classifier gives second best performance to the NB classifier;

B. Case Acquisition Results

As defined in section IV and according to the results found in section IV-A, we can now use the best model based on NB classifier, ECC approach and TF-IDF weighting for the case acquisition process to identify IL's (cerebral edema, fracture, brain hemorrhage and foreign body) values in the TBI-case vectors which form our TBI case. Also, we have acquired manually information about children (PI, CS and GP) defined in section III.B.

For the process of validation of our MLTC4CA approach, we used the remaining 31% reports (equivalent to 54 test reports) and we have used the same preprocessing module to generate the numeric representation for each MRI report.

For each test report, the MLTC4CA used its numeric representation as input to automatically identify the corresponding label values and insert them into its corresponding case vector.

Table V summarizes the obtained results of the proposed MLTC4CA approach. It shows the values for all multi-label measures calculated to evaluate the case acquisition using the test TBI collection.

As seen in table V below, we have done the case acquisition process and we have obtained a good performance to acquire values or labels from the fifty-four MRI scan report of test. All of these results demonstrate that our approach of using MLTC framework for the acquisition of cases in CBR process is very satisfactory.

TABLE V. CASE ACQUISITION RESULT OBTAINED BY USING MLTC FRAMEWORK

Measure	Values (%)
Subset accuracy	85,80
F-measure	91,65
Accuracy	89,72
Hamming Loss	04,17
Micro-avg F-measure	93,00
Avg precision	97,96
Ranking Loss	02,07
One Error	19,17

5. CONCLUSION AND FUTURE WORKS

This study presents MLTC4CA, a framework that uses MLTC tool to assist in automatic case acquisition from text. MLTC4CA is able to automate the case acquisition process without exhaustive doctors to complete case information from the MRI scan reports provided by the pediatric intensive care unit of Oran hospital - Algeria.

In this paper, we first conducted numerous experiments focused on finding the appropriate combination of classifiers, multi-label approaches and terms weightings schemes to build the best multi-label categorization model. For this, multi-label approaches (see Figure 1) including a variety of transformations, adaptation and ensemble methods have been used and compared.

According to the results, Ensemble of Classifier Chains (ECC) approach achieves the best performances for the identification of intracranial lesions.

In a second step, we used the best model, which is of course based on the ECC approach, in the case acquisition phase to complete the problem part of the case.

This model, which is supposed to predict the different possible labels representing intracranial lesions of a child victim of TBI, will automatically complete the problem part of the case with the values of these labels.

Our Results, suggest that the proposed approach is a promising way to automatic case acquisition. However, the dataset used is small which makes it more difficult to generalize the results obtained. In future work, we will address this limitation by collecting more clinical reports.

For this first investigation we used magnetic resonance imaging (MRI) reports provided by the pediatric intensive care unit of the Oran-Algeria hospital. These reports provided in paper format are written in French (the language used by doctors in Algeria). Unfortunately, we were unable to include other dataset for comparison because we did not find a public dataset in the form of medical reports written in French. For further experimentation, children's MRI reports can be collected from other hospitals in Algeria, which will allow us to make a deeper and more conclusive analysis. We also suggest that the case acquisition can also be done from the MRI scan image.

We also have in mind to improve the step of retrieving relevant cases by reducing the search space because it is a hard and time-consuming stage. Because the case-based system we are developing should be used to help medical doctors in the care of children who are victims of TBI. In such a situation, to save lives, they have to react quickly.

A final perspective to consider in the future is to apply the MLTC approach proposed in this paper to other tasks such as information retrieval.

ACKNOWLEDGMENT

Authors would like to express their gratitude to Doctor Nesserine Benfriha who has significantly and decisively contributed to the labeling of the MRI scan collection.

REFERENCES

[1] Aamodt, A., Plaza, E.: Case-based reasoning: foundational issues, methodological variations and system approaches. AI Commun. 7(1), 39‑59 (1994).

[2] M. M. Richter and R. O. Weber," Introduction" in Case-Based Reasoning: A Textbook. Springer Berlin, 2013, pp. 3–15.

[3] Kolodner JL (1993) Case-based reasoning. Morgan Kaufmann, San Mateo.

[4] Sabri, Q.U., Bayer, J., Ayzenshtadt, V., Bukhari, S.S., Althoff, K.D., Dengel, A.: Semantic pattern-based retrieval of architectural floor plans with case-based and graph-based searching techniques and their evaluation and visualization. In: 6th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2017), Porto, Portugal, 24–26 February 2017 (2017).

[5] Malburg, L., M¨unster, N., Zeyen, C., Bergmann, R.: Query model and similarity-based retrieval for workflow reuse in the digital humanities. In: Proceedings of LWDA 2018, Mannheim, vol. 2191, pp. 251‑262 (2018). CEUR-WS.org

[6] Skjold, K., Øynes, M.S.: Case-based reasoning and computational creativity in a recipe recommender system. Master's thesis, NTNU (2017)

[7] Finnie, G., Sun, Z.: R5 model for case-based reasoning. Knowl.-Based Syst. 16(1), 59–65 (2003).

[8] Benfriha H., Atmani B., Khemliche B., Aoul N.T., Douah A. (2019) A Multi-labels Text Categorization Framework for Cerebral Lesion's Identification. In: Alfaries A., Mengash H., Yasar A., Shakshuki E. (eds) Advances in Data Science, Cyber Security and IT Applications. ICC 2019. Communications in Computer and Information Science, vol 1098. Springer, Cham.

[9] Yang, C., Farley, B., Orchard, B.: Automated case creation and management fordiagnostic CBR systems. Appl. Intell. 28(1), 17–28 (2008).

[10] Roth-Berghofer, T., Adrian, B., Dengel, A.: Case acquisition from text: ontology-based information extraction with SCOOBIE for myCBR. In: Bichindaritz, I., Montani, S. (eds.) ICCBR 2010. LNCS, vol. 6176, pp. 451–464. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-14274-1 33.

[11] Bach, K., Althoff, K.D., Newo, R., Stahl, A.: A case-based reasoning approach for providing machine diagnosis from service reports. In: Ram, A.,Wiratunga, N.(eds.) Case-Based Reasoning Research and Development, ICCBR. Lecture Notes in Computer Science, vol. 6880. Springer, Heidelberg (2011).

[12] Dufour-Lussier, V., Le Ber, F., Lieber, J., Nauer, E.: Automatic case acquisition from texts for process-oriented case-based reasoning. Inf. Syst. 40, 153–167 (2014)

[13] Sebastiani, F. Machine learning in automated text categorization. ACM computing surveys, 34 (1), 1-47. (2002).

[14] Benfriha, H., Barigou, F., & Atmani, B. (2016). A text categorization framework based on concept lattice and cellular automata. International Journal of Data Science (IJDS), 1(3), 227-246.

[15] Herrera, F., Charte, F., Rivera, A.J., Del Jesus, M.J.: Multilabel Classification: Problem Analysis, Metrics and Techniques. Springer, Heidelberg. (2016).

[16] Sizov, G., Öztürk, P., Styrak, J.: Acquisition and reuse of reasoning knowledge from textual cases for automated analysis. In: Lamontagne, L., Plaza, E. (eds.) ICCBR 2014. LNCS, vol. 8765, pp. 465–479. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11209-1 33.

[17] Cordier, A., et al.: Taaable: a Case-Based System for personalized Cooking. In: Montani, S., Jain, L.C. (eds.) Successful Case-based Reasoning Applications-2. SCI, vol. 494, pp. 121–162. Springer, Heidelberg (2014).

[18] Sizov, G., Ozturk, P., & Aamodt, A. 2015. Evidence-driven retrieval in textual CBR: Bridging the gap between retrieval and reuse. In Procs. ICCBR-2015, pp. 351–365.

[19] Dufour-Lussier V., Lieber J. (2015). Evaluation a Textual Adaptation System. In: Hüllermier E., Minor M.(eds) Case-Based Reasoning Research and Development. ICCBR 2015. Lecture Notes in Computer Sciences, vol 9343. Springer, Cham.

[20] Reuss P. et al. (2016) FEATURE-TAK - Framework for Extraction, Analysis, and Transformation of Unstructured Textual Aircraft Knowledge. In: Goel A., Díaz-Agudo M., Roth-Berghofer T. (eds) Case-Based Reasoning Research and Development. ICCBR 2016. Lecture Notes in Computer Science, vol 9969. Springer, Cham.

[21] Shen, L.Y.; Yan, H.; Fan, H.Q.; Wu, Y.; Zhang, Y. An integrated system of text mining technique and case-based reasoning (TM-CBR) for supporting green building design. Build. Environ. 2017, 124, 388–401.

[22] Manzoor, J., Asif, S., Masud, M., Khan, M.J.: Automatic case generation for case-based reasoning systems using genetic algorithms. In: 2012 Third Global Congress on Intelligent Systems (GCIS), pp. 311–314. IEEE (2012).

[23] Wu, He. Improving user experience with case-based reasoning systems using text mining and web 2.0. Expert System with Applications, 40 (2) (2013), pp. 500-507

[24] Bach, K., Gundersen, O. E., Knappskog, C., & Öztürk, P. (2014, September). Automatic case capturing for problematic drilling situations. In International Conference on Case-Based Reasoning (pp. 48-62). Springer, Cham.

[25] Nasiri S., Zenkert J., Fathi M. (2015) A Medical Case-Based Reasoning Approach Using Image Classification and Text Information for Recommendation. In: Rojas I., Joya G., Catala A. (eds) Advances in Computational Intelligence. IWANN 2015. Lecture Notes in Computer Science, vol 9095. Springer, Cham.

[26] Wienhofen, L. W., & Mathisen, B. M. (2016, October). Defining the initial case-base for a CBR operator support system in digital finishing. In International Conference on Case-Based Reasoning (pp. 430-444). Springer, Cham.

[27] Mathisen, B. M., Aamodt, A., &Langseth, H. (2017). Data driven case base construction for prediction of success of marine operations. CEUR Workshop Proceedings.

[28] Amin, K., Kapetanakis, S., Althoff, K. D., Dengel, A., & Petridis, M. (2018, July). Answering with cases: a CBR approach to deep learning. In International Conference on Case-Based Reasoning (pp. 15-27). Springer, Cham.

[29] Amin, K., Lancaster, G., Kapetanakis, S., Althoff, K. D., Dengel, A., & Petridis, M. (2019, September). Advanced similarity measures using word embeddings and siamese networks in CBR. In Proceedings of SAI Intelligent Systems Conference (pp. 449-462). Springer, Cham.

[30] Tsoumakas, G., Katakis, I., Vlahavas, I.: Mining multi-label data. In: Data Mining and Knowledge Discovery Handbook, pp. 667–685. Springer .2010.

[31] Read, J., Pfahringer, B., Holmes, G., & Frank, E. (2009, September). Classifier chains for multi-label classification. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 254-269). Springer, Berlin, Heidelberg.

[32] Read, J., Pfahringer, B., Holmes, G., & Frank, E. (2011). Classifier chains for multi-label classification. Machine learning, 85(3), 333.

[33] Boutell M, Luo J, Shen X, Brown C. Learning multi-label scene classification. Pattern Recogn 2004,37:1757–1771.

[34] Fürnkranz, J., Hüllermeier, E., LozaMencía, E., Brinker, K.: Multilabel classification via calibrated label ranking. Mach. Learn. 73, 133–153 (2008).

[35] Zhang ML, Zhou ZH. ML-KNN: A lazy learning approach to multi-label learning. Pattern Recognition; 2007, 40(7):2038–2048.

[36] Spyromitros E, Tsoumakas G, Vlahavas I. An empirical study of lazy multi-label classification algorithms.In: SETN'08: Proceedings of the 5th Hellenic Conference on Artificial Intelligence, Berlin, Heidelberg;2008, 401–406.

[37] Madjarov, G., Kocev, D., Gjorgjevikj, D., &Džeroski, S. (2012). An extensive experimental comparison of methods for multi-label learning. Pattern recognition, 45(9), 3084-3104.

[38] Tsoumakas, Grigorios, and Ioannis Vlahavas, Random k-labelsets: An ensemble method for Multilabel classification, Machine learning: ECML 2007. Springer Berlin Heidelberg, 2007, 406-417.

[39] Tsoumakas, G., Katakis, I.: Multi-label classification: An overview. Int. J. Data Warehouse. Min. 3(3), 1–13 .2007.

[40] Hotho, A., Nürnberger, A., &Paaß, G. (2005, May). A brief survey of text mining. In Ldv Forum (Vol. 20, No. 1, pp. 19-62).

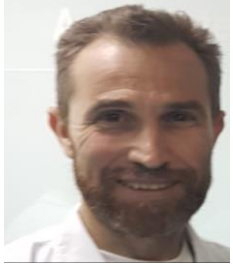
[41] Jo, Taeho. Text Mining: Concepts, Implementation, and Big Data Challenge. Vol. 45. Springer, 2018.

[42] Tsoumakas, G., Spyromitros-Xioufis, E., Vilcek, J., & Vlahavas, I. (2011). Mulan: A java library for multi-label learning. Journal of Machine Learning Research, 12(Jul), 2411-2414.

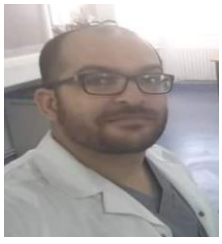
HICHEM BENFRIHA is a computer science teacher in the Department of Technical Sciences, University of Mascara Mustapha Stambouli, Algeria. He is currently a Research Member of Laboratory of Computer Science of Oran. He is currently a PhD candidate in the Computer Science Department of Oran 1 University (Algeria). He received his Master of Science degree in 2012 from the same university. His research interests focus on CBR, data Mining, text mining, information extraction, information retrieval, natural language processing, machine learning and multi-label classification areas.

BAGHDAD ATMANI is currently a Professor in Computer Science. His interest fields are artificial intelligence and machine learning. His research interests focus on knowledge representation, knowledge-based systems, CBR, data mining, expert systems, decision support systems and fuzzy logic. His researches are guided and evaluated through various applications in the field of control systems, scheduling, production, maintenance, information retrieval, simulation, data integration and spatial data mining.

FATIHA BARIGOU graduated from Department of Computer Science, University of Oran 1, Algeria. In 2012, she received his PhD degrees in Computer Science from the University of Oran.
Currently, she is a university lecturer at Computer Science Department of University of Ahmed Benbella Oran 1. She is a research member of the AIR team in the LIO laboratory. She does research in Text Data Mining, Big data and Artificial Intelligence. Her current projects are Sentiment Analysis, AI and Cloud Computing in healthcare.



KHEMLICHE BELARBI MD, PhD. Doctorate in Medicine October 1998. Oran's college of medicine, medical studies diploma with an anesthesiology and intensive care specialty 2002 Algiers's college of medicine. Vice valedictorian in competitive exam of university hospital assistant's professor's recruitment 2003 Algiers's college of medicine. MD-PhD in medical sciences, with an anesthesiology and intensive care specialty. 2012. Oran's college of medicine. Thesis subject: « noninvasive ventilation in acute lung injuries in pediatric critical care unit » with honors and board's congratulations. Level « A » conferences master in medical sciences, with an anesthesiology and intensive care specialty 2014. Professor and Chair of medical intensive care, 2019. Head by interim of pediatric resuscitation department university hospital of Oran, Algeria (EHU Oran) 2016. Head of medical resuscitation department university hospital of Oran, Algeria (EHU Oran) 2017. Head of Unit in the Pedagogy laboratory research in medical science university Oran 1 Ahmed Benbella. Vice president of the regional pedagogic committee in medical resuscitation specialty (CPRS) 2017. Head of the regional pedagogic committee in medical resuscitation specialty (CPRS) 2020.



ALI DOUAH is a pediatric anesthesiologist and critical care physician since 2016 at Canastel Pediatric Hospital in Oran, Algeria. He prepared his PhD at the university of Ahmed Benbella Oran 1 in medical science. He is also a member of Pediatric Injury Laboratory in the University of Oran 1. His field of research is lung ultrasound.

Zakaria Zoheir Addou is pediatric anesthesiologist and critical care physician since 2002 at pediatric hospital in Oran, Algeria. He received his PhD in 2017 at the university of Ahmed Benbella 1 Oran in medical science. He is also a member of laboratory of pediatric injury in the University of Oran. His field of research is safety of anesthesia outside operating room in children.

NABIL TABET AOUL is anesthesiologist and critical care physician since 2002 at pediatric hospital in Oran, Algeria. He received his PhD in 2014 at the university of Ahmed Benbella 1 Oran in medical science. He is a member of laboratory of pediatric injury in the University of Oran. His field of research is sedation and analgesia in the pediatric intensive care.