A Disease Identification System using Electronic Medical Records

Youssef Elmir, Meriem Bendida
University Tahri Mohammed of Bechar
Algeria
{elmir.youssef@yahoo.fr}

ABSTRACT: *In the medical field, medical analyses are important to properly diagnose the patient's case by the doctor, especially if there is a history of several analyses of the same patient which are stored in the patient's electronic medical record; this can help the doctor to make the right decision. However, the doctor always needs other techniques and methods in order to make the right decision. In this work, a disease identification system is performed from electronic medical records using the k-nearest neighbours classification algorithm, which classifies different types of diseases (six diseases were studied) according to the values of the medical analysis. The experiment results show that the identification rate (classification) is 43.18% using very small sample as reference data, and this obtained result is acceptable and present the proof of the feasibility of the proposed system.*

Keywords: Electronic Medical Record, Diagnostic, Multi-class Classification, Machine Learning, *K* Nearest Neighbours

Received: 27 August 2020, Revised 4 December 2020, Accepted 11 December 2020

DOI: 10.6025/jic/2021/12/1/8-24

1. Introduction

Man has always sought to reduce his efforts and improve his productivity in all areas, by getting help by machines and discovering technologies capable of performing human functions independently. Today, with computer technologies and artificial intelligence (AI) the medical field has developed and equipped with several computer techniques that can help doctors and patients, such as electronic medical records and automatic diagnostic systems. Thanks to computer technology, the paper medical record has been developed and made into electronic medical record much easier to use compared to the old one. In recent years, most medical analyses laboratories have used electronic records to manage the patient's medical analyses and also to store the necessary information of all related actors.

Often, the patient's medical analyses contain results, but they are generally based on biological standards that do not allow giving a general diagnostic on the condition of the patient's body. In literature, many automatic solutions were proposed to realize intelligent systems that can give a diagnostic on the state of the body from medical analyses. In the field of medical diagnostic and disease identification, the applications of Artificial Intelligence are also very limited, due to the presence of several diseases and even doctors may not make the same diagnostic for same case.

The Watson Project [1]is a computer system for diagnosing patients diseases, created in 2006 by IBM (International Business Machines), it offers treatments to a patient and evaluates its merits by analysing the patient's medical history, each treatment proposal is accompanied by a degree of relevance, it provides doctors with a tool to make a diagnostic as accurate as possible from the data observed by the medical profession and stored in the patient file, in a minimum of time. According to IBM, It is a form of expertise.

A Bayesian network-based classifier was produced by the team of Aline Conseil et al [2] For the Diagnostic of Hypothyroidism. Their goal was to achieve an automatic system to diagnose hypothyroidism, they tried to make several Bayesian network structures to determine the most reliable with a high classification rate and of course looking at other criteria, the obtained classification rate was 99%.

In 2013, Bakhoche Houda and Benglia Rachida from KASDI Merbah University of Ouargla [3] were able to carry out a system to assist in diagnosing diseases in internal medicine based on case-based reasoning, this system based on the principle of artificial intelligence (AI) and case-based reasoning that allows the reuse of past experiments to solve new problems in a specific field.

In July 2016, DeepMind [4] a Google subsidiary specializing in advanced artificial intelligence (AI) research, announced the launch of a partnership with the UK's national health system and Moor Fields Hospital in London to develop an AI capable of detecting eye diseases, the system was capable of successfully detecting more than 50 types of eye diseases by examining 3D images of the retina, with a lower margin of error than ophthalmologists. Thus, the DeepMind system can not only detect pathologies such as diabetes or macular degeneration, but can also recommend the best treatment for patients and suggest treatments to be carried out urgently.

In 2017, another work was carried out by Belhouari Imane and Benabdelkrim Fatima from Abu Bakr Belkaïd University of Tlemcen [5], their project was a medical decision support system for early detection of glaucoma, they used five different techniques for classification in supervised learning on a new database: the classification rate was for K nearest neighbours (79.28%), support vector machines (96.62%), multi-layer perceptron neural networks (96.02%) and Decision Trees (CART, Random Drill) (99.01%) and drill of RF decision trees (98.60%), this algorithms were evaluated using a dataset (1004 data from 514 patients).

In 2016, Simon Kocbek et al [6] presented a text mining system for detecting admissions marked as positive for several diseases. Authors specifically examined the effect of linking multiple data sources on text classification performance. They investigated Support Vector Machine classifiers are built for eight data source combinations. They also explore the impact of feature selection; analyse the learning curve; examine the effect of restricting admissions to only those containing reports from all data sources; and examine the impact of reducing the sub-sampling. These experiments provide better understanding of how to best apply text classification in the context of imbalanced data of variable completeness. The obtained results for Radiology questions plus patient and hospital admission data contribute valuable information for detecting most of the diseases, significantly improving performance when added to radiology reports alone or to the combination of radiology and pathology reports, by consequence, Authors concluded that linking data sources significantly improved classification performance for all the diseases examined. However, there is no single approach that suits all scenarios; the choice of the most effective combination of data sources depends on the specific disease to be classified.

It is important to mention that references in this area are few and studies that use multi-class classification are very rare, while most studies have been done on binary classification (detecting the existence of a disease or not). However, the main objective of this work is to propose a system that can identify more than one disease and give a medical diagnostic similar to a doctor's medical diagnostic from electronic medical records.

The rest of this paper is organized into three sections: section 2 gives a general introduction to the basic concepts of electronic medical records, where section 3 describes the proposed design and implementation of a disease identification system based on electronic medical records. Experiments are presented in section 4 with discussion of the obtained results. In this section, the chosen model chosen is evaluated using a real database collected/acquired from a private medical analyses laboratory.

2. Electronic Medical Records

Electronic medical records remain the most important and best feature of developments in the world of medical records, allowing all doctors from any hospital wherever they are and at any time they want easy access to all health information of all patients, one of the greatest challenges of this medical development is to follow huge amounts of medical information related to patients and protect them from any violation.

2.1. General Information on Electronic Medical Records

2.1.1. Paper Medical Record

The medical record contains the medical, social, professional data concerning a patient and having an interest in the knowledge of his health. It summarizes all the information that is necessary to identify the patient's illness as well as how the therapy will be or was conducted: diagnostic elements, follow-up examinations and then drug or non-drug therapy (work stoppage, nursing, functional reduction, etc.). [7]

While electronic medical records are also widely known as electronic health records and many people confuse between them.

2.1.2. Electronic Medical Record

The electronic medical record (EMR) is a type of system developed to support the activities of doctors and other health professionals. The patient's file, previously available in paper format, is transformed into an electronic folder containing the same information. During a survey on the future prospects of primary care doctors, argue that electronic medical records are an important way to increase productivity within the practice, improve coordination of care between professionals and patient safety (through minimizing medical errors). [8]

2.1.3. Electronic Health Record

The electronic health record (EHR) is an electronic longitudinal record of patient health information generated by one or more meetings in any care delivery setting. This information includes patient demographics, progress scores, problems, medications, vital signs, medical history, vaccinations, laboratory data and radiology reports. The EHR automates and streamlines the clinician's workflow. The EHR has the ability to generate a complete record of a clinical meeting with a patient, as well as support other care-related activities directly or indirectly through the interface, including evidence-based decision support, quality management and reporting of results.

2.2. The Contents of the Electronic Medical Record

The patient file therefore gathers all the information collected by health professionals (doctors, paramedics, other professionals). It contains:

• The identity of the patient.

• Family and personal history, history of current illness, data from previous visits and hospitalizations.

• The results of clinical, radiological, biological, functional and histopathological examinations (signed by the responsible doctor).

• The opinions of the doctors consulted (signed by these doctors).

• Provisional and definitive diagnoses (signed by the doctor who made the diagnostic).

• The treatment implemented in case of surgery, the operating protocol and the anaesthesia protocol (signed by the treating surgeon and anaesthetist).

• The evolution of the disease.

• Possibly the autopsy protocol (signed by the pathologist).

• An exit report that must contain all necessary information to enable any other doctor consulted by the patient to ensure continuity of care. This report will either be given to the patient or passed on to the treating doctor and any doctor concerned. [8]

3. Design and Implementation of the Disease Identification System from Electronic Medical Records

The design phase is considered to be one of the most important phases in the process of carrying out a computer system; it is

a process that aims to formalize the preliminary stages of the process of development of computer systems. In this work, this phase aims to explain the necessary steps of using learning algorithms to achieve a diagnostic system that can classify diseases automatically.

3.1. Modelling

In this work, *k*-nearest neighbours' algorithm is used as supervised classification method to classify the results of medical analyses and identify the corresponding diseases.

3.1.1. The Method of the K-nearest neighbours

The k-nearest neighbours (KNN) method is a supervised method. It has been used in statistical estimation and model recognition as a non-parametric technique, which means that it makes no assumptions about the distribution of data. [9]

This is one of the non-parametric techniques frequently used in non-linear financial prediction. This preference is mainly due to two reasons:

• First, the algorithmic simplicity of the method compared to other global methods such as neural networks or genetic algorithms.

• Second, KNN has empirically demonstrated an important predictive ability.

The idea of the method is to predict the future of a time series by analysing how it evolved in a similar situation in the past. Thus, to make a prediction the most recent historical data available is taken and among these data, the k nearest instances called also the nearest *k* vectors, are looked for. [10]

The KNN algorithm is one of the simplest of all machine learning algorithms. It is a type of learning based on lazy learning. In other words, there is no explicit or very minimal training phase. This means that the training phase is quite fast. [9]

3.1.2. The Principle of the Method of the *k* Nearest Neighbours

The KNN method assumes that the data is in a space of features. This means that the data points are in a metric space. The data can be scalars or even multidimensional vectors.

The method of KNN is used for classification and regression. In both cases, the input consists of the *k* nearest reference data in the feature space. [11]

To find the class of a new case, this algorithm is based on the following principle: it looks for the *k* nearest neighbours of this new case, and then it chooses from among the candidates found the closest and most frequent result.

To assign a new individual to a class, the algorithm looks for the k nearest neighbours among the individuals already classified. Thus, the individual is assigned to the class that contains the most individuals among the candidates found.

This method mainly uses two parameters:

• A similarity function to compare individuals in the feature space.

• The *k* number that decides how many neighbours influence the classification. [9]

To test the similarity between two vectors, the calculation of a distance is used. It measures the degree of difference between two vectors. There are several types of distance, including:

- The Euclidian distance: Distance that calculates the square root of the sum of square differences between the coordinates of two points.

$$d(x,y) = \sqrt{\sum_i^n (x_i - y_i)^2} \quad (1)$$

Minkowski's distance: or *p*-distance, generalizes the Euclidian, it is the small root of the sum of the absolute values of the deviations to the power *p*.

$$d(x,y) = (\sum_{i=1}^{n}|x_i - y_i|^p)^{\frac{1}{p}} \quad (2)$$

The distance from Manhattan: The distance between two given vectors is the maximum difference between their coordinates on a dimension.

$$d(x,y) = \sum_{i=1}^{n}|x_i - y_i| \quad (3)$$

• Where: *x, y* are vectors.

• ρ : Parameter. [12]

That is the example of Figure 2 with two dimensions corresponding to the e^1 and e^2 attributes, and with k^3. In this example the three nearest neighbours of a are b^4, b^2 and b^5, so a will be assigned to the majority class among these three points. [13]

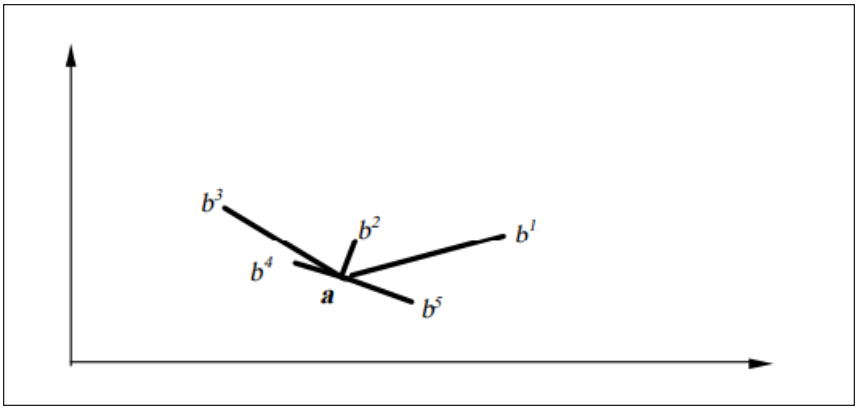


Figure 1. 3-nn example **[13]**

3.1.3. The Advantages of the *k* nearest Neighbours Method

The method of nearby *k* represents advantages such as:

• The algorithm is super simple and easy to implement. [14]

• The KNN algorithm is robust with noise data. [9]

• The method of nearby *k*'s is effective if the data are broad and incomplete.

• There is no need to build a model, adjust multiple parameters or make additional assumptions.

• The algorithm is versatile. It can be used for classification, regression and information search. [14]

3.1.4. The disadvantages of the method of the *k* nearestneighbours

• Accuracy depends on the quality of the data.

• With big data, the prediction phase can be slow.

• Sensitive to data scale and irrelevant characteristics.

• Requires a high memory, need to store all the drive data.

• Since it stores all the training, it can be expensive in calculation. [15]

3.2. Classification

The *k* nearest neighbours method is a non-parametric method where a new observation is classified in the belonging class of

the nearest learning sample observation, based on the variable used. Determining their similarity is based on distance measurements.

In this work, the multi-class classification was used using the *k*-nearest neighbours method to model diseases based on medical analyses, this method is chosen based on the type of the problem studied.

The identification process steps are as follow:

• Formally, let L the available data set or reference sample:

$$L = \{(y_i, x_i), i = 1 \ldots, n_l\} \tag{4}$$

• Where $y_i \in \{1, \ldots, c\}$ denotes the class of the individual i, in the case of our project the y_i denotes the classes of the patient's diseases.

• And represents $x_i = x_{i1}, \ldots, x_{ip}$ the predictors of the individual i.

• Formally, let T the available data set or test sample:

$$T = \{(y_i', u_i), i = 1 \ldots, n_t\} \tag{5}$$

- Where $y_i' \in \{1, \ldots, c\}$ denotes the class of the individual i.

- And $u_i = u_{i1}, \ldots, u_{ip}$, represents the predictors of the individual i.

Step 1: Distance Calculation.

The determination of the nearest neighbour is based on an arbitrary distance function $d(.,.)$, the Euclidian distance is used to calculate the distance between the reference vectors and the vectors of the test, this distance is defined by:

$$\mathrm{d}\left((\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_\mathrm{p}), (\mathrm{u}_1, \mathrm{u}_2, \ldots, \mathrm{u}_\mathrm{p})\right) = \sqrt{(\mathrm{x}_1 - \mathrm{u}_1)^2 + (\mathrm{x}_2 - \mathrm{u}_2)^2 + \cdots + \left(\mathrm{x}_\mathrm{p} - \mathrm{u}_\mathrm{p}\right)^2} \tag{6}$$

Step 2: Determining the minimum distance.

For a new observation (y, x) the k nearest neighbours (y_i', u_i) in the reference sample is determined by the:

$$d(x, x_1) = min_i\,(d\,(x_i, u_i)) \tag{7}$$

If $k = 1$, the class of the nearest neighbours $\hat{y} = y_i$ is selected for the prediction of y'. [16]

Otherwise, the nearest k observations are used. Thus the decision is in favour of the class mainly represented by the neighbouring k.

• if k_r The number of observations from the group of nearest neighbours belonging to the R class:

$$\sum_{r=1}^{c} k_r = k \tag{8}$$

Instead of determining a single minimum distance, in the case of the k nearest neighbours, it is determined k minimum distances.

Step 3: Max choice.

Thus a new observation is predicted in class l with: [16]

$$l = max_r\ (k_r) \tag{9}$$

This prevents the predicted class from being determined only from a single observation. The degree of locality of this technique is determined by the *k* parameter:

• The only nearest neighbour's method is used as the maximum local technique.

• For $k \rightarrow n_l$, the majority class is used over the full set of observations (this involves a constant prediction for each new observation to be classified). [17]

3.2.1. Choosing *k*

• The *k* setting must be determined by the user: $k \in N$.

• In binary classification, it is useful to choose odd *k* to avoid egalitarian votes.

• The best choice of *k* depends on the data set. In general, large *k* values reduce the effect of noise on classification and thus the risk of over-learning, but make the boundaries between classes less distinct.

• A good *k* can be selected by various heuristic techniques, for example, validation-cross.

• The *k* value must be chosen in a way that minimizes the misclassification. [18]

3.3. System Performance Evaluation

Once the algorithm is launched, it is always necessary to carry out tests to verify that the algorithm reacts correctly. This phase is very important for testing, measuring and monitoring the performance of a predictive model before and after deploying it in production. The measures to be used to assess this performance must then be defined.

Several experimental regimes have been used in the literature to establish the performance of a classification system such as the classification rate.

The classification rate is a significant parameter for assessing a classifier [2]:

$$CR = \frac{Number\ of\ well\ classified\ patients}{Total\ Number\ of\ patients\ classified} \tag{10}$$

4. Experiments and Results

The evaluation parameter used in this work which is the classification rate is also presented, with different *k* values, after applying the technique of the *k*-nearest neighbours as a classification technique.

4.1. Data Collection and Preparation

Today, most medical analyses laboratories in Algeria use software to store and manage the results of patients' analyses. These results are stored in a database, and this data is private and confidential, and therefore difficult to obtain.

So the first step was the collection of patient analyses, this step took a long time, as mentioned before, the confidentiality of these information made the operation of the acquisition of the data quite difficult. After several attempts, data were obtained from the medical analyses laboratory of Dr. BENCHAIB Sidi Mohamed in the state of Bechar in Algeria, as a backup file of an Interbase database. This database contains patient analyses over a period of ten years period from 2010 to 2020, for each patient a history of analyses during the same period. This type of backup file does not be opened directly, so a restore process was necessary before to get the original file (Interbase extension file). In the case of this study, the database is Interbase of the extension (IB) and the extension backup file (GBK).

To restore the available data from the backup file, an Interbase server is used, this server is provided with libraries that support

the development of integrated SQL and SQL client applications [19]. Also, IBOconsole is used for monitoring and administration of the database and the Interbase server. After restoring, Data consulting was done using RazorSQL which is an SQL query tool, database browser, SQL editor, and database administration tool for Windows, macOS, Mac OS X, Linux, and Solaris [20].

4.1.1. Choosing Analyses

To achieve the proposed system, it was necessary to have a sufficient number of analyses of the same disease, so that it could make a decision or predict a patient's disease correctly, so for each type of disease, the maximum medical analyses were prepared. In this part SQL queries were used to attach the tables of the database and obtain results that contains necessary information (age and sex its information is idea the doctors made a correct decision) and biological analyses of patients.

The relational model normalization process is based on decomposition and consequently increases the number of tables in a schema. Thus, the majority of queries use joins necessary to be able to extract data from separate tables [21] , this is the SQL query used to collect only necessary and useful data for this study:

```
SELECT DISTINCT DATE_RETRAIT_RESULTAT, NOM,PRENOM,AGE, NUMERO_DDE,
MEDECIN_TRAITANT, LIB_RUBRIQUE, LIB_ABREGE_RUBRIQUE, UNITE_RUBRIQUE,RESULTAT,
LIB_NORME, ID_SEXE, TYPE_AGE_NORME, LIBELLE_AUTOMATE
FROM DEPOTS_ANALYSES dp
INNER JOIN RESULTATS r
ON dp.ID_DEPOT=r.ID_DEPOT
INNER JOIN RUBRIQUES rq
ON rq.ID_RUBRIQUE=r.ID_RUBRIQUE
INNER JOIN LISTE_ANALYSE la
ON la.ID_DEPOT=r.ID_DEPOT
INNER JOIN NORMES noo
ON noo.ID_RUBRIQUE=rq.ID_RUBRIQUE
INNER JOIN AUTOMATE autoo
ON autoo.CODE_AUTOMATE=rq.CODE_AUTOMATE
INNER JOIN ANALYSES ana
ON ana.CODE_ANALYSE=rq.CODE_ANALYSE
WHERE ID_SEXE='1' and TYPE_AGE_NORME='4' and ID_TUBE='6' and AGE_DEBUT='00.00' and
AGE between '40' and '50'ORDER by AGE
```

Figure 2. SQL query of collecting data

4.1.2. Medical Consultations

Since KNN in a supervised classification, consultation with doctors was indispensable task, in order to prepare the reference data and evaluate the test one. In this part, the doctors decide the patient's illness and diseases that has or may have in the future. Three doctors (2 general practitioners and 1 specialist doctor)[1] were consulted to read the collected analyses, and each doctor had to read the patient's analyses and give a conclusion separately.

The first doctor is Dr. ABZZOU Zineb is a general practitioner who has tried to read all the analyses, and give each analysis a conclusion, whether the patient is sick or not, and if the patient is sick she gave the diagnostic. For example in the case in Figure 3, the patient is a man who did two tests of the same type in deferent dates, the first result of the analyses Dr. ABZZOU concluded that the complete blood count (CBC) (FNS in French as the database) is normal, After she found the same thing after 6 years in the second result, and at the end she gives a conclusion that health check is unremarkable.

Dr. MASIKA Meriem is the second general practitioner participated in this study; she also tried to read all the analyses, and gave a conclusion for each analysis and a general conclusion about each patient's probable disease. For example in the case in

[1]The goal of the multiplicity of doctors is to increase the number of possible diseases so that the system gives a correct decision.

DATE_RETRAIT_	N de patien	AGE	MEDECIN_TRAITANT	LIB_RUBRIQUE	LIB_ABREG	UNITE_RUBRIQUE	RESULTAT	LIB_NORME		
12/08/2011	10	41	SALMI	GLOBULES ROUGES	GR	Million/mm3	4,52	3.8 - 5.40		
12/08/2011	10	41	SALMI	GLOBULES BLANCS	GB	Mille/mm3	5 980	4.0 à 12.0		
12/08/2011	10	41	SALMI	V.G.M	VGM	µ3	88	70-100		
12/08/2011	10	41	SALMI	HEMOGLOBINE	Hb	g/100ml	13,3	11 - 14		conclusion:
12/08/2011	10	41	SALMI	C.C.M.H	CCMH	%	34	32 - 37	FNS normale	bilan sans particularité
12/08/2011	10	41	SALMI	HEMATOCRITE	Hte	%	39,6	33 - 44		
12/08/2011	10	41	SALMI	PLAQUETTES	PLQ	Mille/mm3	469	150 - 500		
12/08/2011	10	41	SALMI	T.G.M.H	TGMH	pg	29	27 - 32		
12/08/2011	10	41	SALMI	POLYNUCLEAIRE BASOPHILE#	PN.BASO	/ mm3	00	0 - 100		
12/08/2011	10	41	SALMI	POLYNUCLEAIRE EOSINOPHILE#	PN.EOSI	/ mm3	60	0 - 500		
12/08/2011	10	41	SALMI	POLYNUCLEAIRE NEUTROPHILE#	PN.NEUTR	/ mm3	3 770	1 500 - 4 000		
12/08/2011	10	41	SALMI	LYMPHOCYTES#	LYMPHO	/mm3	1 560	2 000 - 7 000		
12/08/2011	10	41	SALMI	MONOCYTES#	MONO	/ mm3	590	200 - 1 000		
12/08/2011	10	41	SALMI	PHOSPHATASE ALCALINE	PAL	UI/L	77	< 645		
12/08/2011	10	41	SALMI	C- REACTIVE PROTEINE (CRP)	CRP	MG/L	6	inf à 5	tous est normale	
12/08/2011	10	41	SALMI	SGOT/ASAT	GOT	UI/L	19	inf à 60		
12/08/2011	10	41	SALMI	SGPT/ALAT	GPT	UI/L	17	inf à 45		
12/08/2011	10	41	SALMI	ANTISTREPTOLYSINE.O (ASLO)	ASLO	UI/L	inf à 200	< OU =200		
09/18/2017	10	47		GLOBULES ROUGES	GR	Million/mm3	4.39	3.8 - 5.40		
09/18/2017	10	47		GLOBULES BLANCS	GB	Mille/mm3	7.6	4.0 à 12.0		
09/18/2017	10	47		V.G.M	VGM	µ3	84.9	70-100		
09/18/2017	10	47		HEMOGLOBINE	Hb	g/100ml	12.2	11 - 14		
09/18/2017	10	47		C.C.M.H	CCMH	%	32.7	32 - 37		
09/18/2017	10	47		HEMATOCRITE	Htc	%	37.3	33 44		
09/18/2017	10	47		PLAQUETTES	PLQ	Mille/mm3	397	150 - 500	FNS normale	
09/18/2017	10	47		T.G.M.H	TGMH	pg	27.8	27 - 32		
09/18/2017	10	47		POLYNUCLEAIRE EOSINOPHILE%	PN EOS	%	4.80	0 - 5		
09/18/2017	10	47		POLYNUCLEAIRE NEUTROPHILE%	PN.NEUTR	%	64.30	0 - 40		
09/18/2017	10	47		POLYNUCLEAIRE BASOPHILE%	PN BASO	%	0.00	0 - 1		
09/18/2017	10	47		POLYNUCLEAIRE BASOPHILE#	PN.BASO	/ mm3	0	0 - 100		
09/18/2017	10	47		POLYNUCLEAIRE EOSINOPHILE#	PN.EOSI	/ mm3	400	0 - 500		
09/18/2017	10	47		POLYNUCLEAIRE NEUTROPHILE#	PN.NEUTR	/ mm3	4 900	1 500 - 4 000		
09/18/2017	10	47		LYMPHOCYTES#	LYMPHO	/mm3	2 000	2 000 - 7 000		
09/18/2017	10	47		MONOCYTES#	MONO	/ mm3	300	200 - 1 000		
09/18/2017	10	47		LYMPHOCYTES%	LYMPH	%	26.40	40 - 75		
09/18/2017	10	47		MONOCYTES%	MONO	%	4.50	02 - 10		
09/18/2017	10	47		CALCIUM ca+ +	Ca	mmol	1.13	1.10 - 1.40		
09/18/2017	10	47		POTASSIUM SANGUIN (K+)	K+	MEQ/L	4.05	3.5 - 5.5	ionogramme sanguin normale	
09/18/2017	10	47		SODIUM SANGUIN (NA+)	NA+	MEQ/L	135.6	135 - 155		
09/18/2017	10	47		CHLORURES (CL)	cl	MEQ/L	105	95 - 107		
09/18/2017	10	47		GLUCOSE SANGUIN (GLYCEMIE)	GLUC	G/L	0.84	0.70 - 1.20		
09/18/2017	10	47		C- REACTIVE PROTEINE (CRP)	CRP	MG/L	9.81	inf à 5		
09/18/2017	10	47		TSHus	TSH	µUI/ML	1.74	0.24 à 4.70	bilan thyroïdien normale	

Figure 3. Example a patient analyses and conclusion made by Dr. ABZZOU Zineb

Figure 20, the same patient with two analyses of the same type in deferent dates, the first conclusion of the analyses made by Dr. MASIKA concluded that there is a suspicion of inflammation and after reading the second analysis, she finds that the patient has an inflammatory syndrome and in the end she gives a conclusion that the health check is unremarkable.

We also requested the help of a specialist doctor Dr. Gritli Youcef, specialist in Otorhinolaryngology (ENT) to confirm the consultations of the two doctors and give a final (diagnosis) decision on the patient's condition.

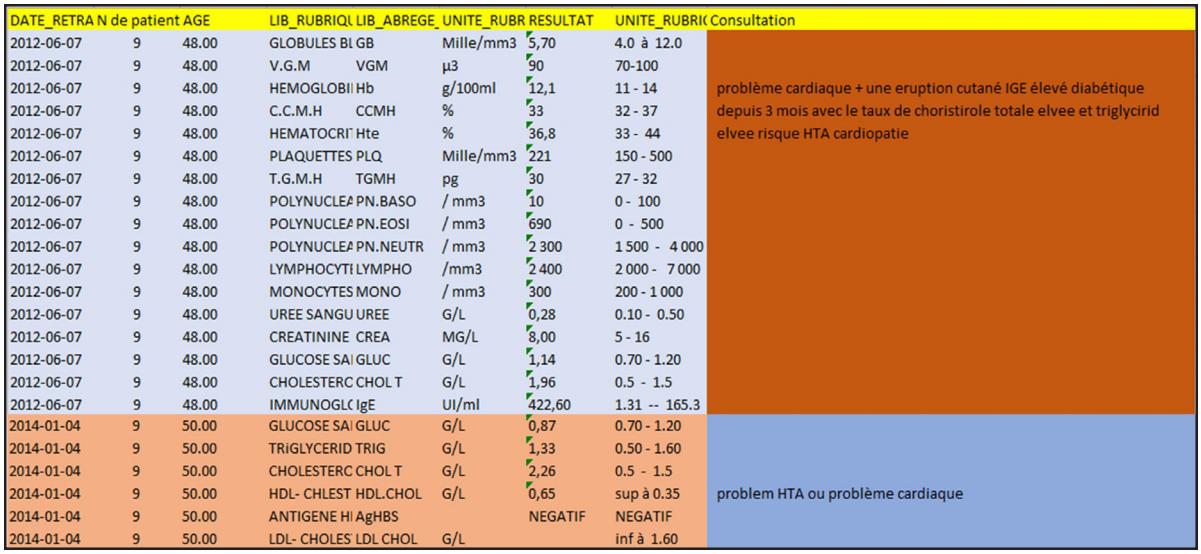
DATE_RETRAIT_	N de patien	AGE	MEDECIN_TRAITANT	LIB_RUBRIQUE	LIB_ABREG	UNITE_RUBRIQUE	RESULTAT	LIB_NORME		
12/08/2011	10	41	SALMI	GLOBULES ROUGES	GR	Million/mm3	4,52	3.8 - 5.40		
12/08/2011	10	41	SALMI	GLOBULES BLANCS	GB	Mille/mm3	5 980	4.0 à 12.0		
12/08/2011	10	41	SALMI	V.G.M	VGM	µ3	88	70-100		
12/08/2011	10	41	SALMI	HEMOGLOBINE	Hb	g/100ml	13,3	11 - 14		
12/08/2011	10	41	SALMI	C.C.M.H	CCMH	%	34	32 - 37	suspicion d'un	
12/08/2011	10	41	SALMI	HEMATOCRITE	Hte	%	39,6	33 - 44	inflammation	
12/08/2011	10	41	SALMI	PLAQUETTES	PLQ	Mille/mm3	469	150 - 500		
12/08/2011	10	41	SALMI	T.G.M.H	TGMH	pg	29	27 - 32		
12/08/2011	10	41	SALMI	POLYNUCLEAIRE BASOPHILE#	PN.BASO	/ mm3	00	0 - 100		
12/08/2011	10	41	SALMI	POLYNUCLEAIRE EOSINOPHILE#	PN.EOSI	/ mm3	60	0 - 500		
12/08/2011	10	41	SALMI	POLYNUCLEAIRE NEUTROPHILE#	PN.NEUTR	/ mm3	3 770	1 500 - 4 000		Conclusion:
12/08/2011	10	41	SALMI	LYMPHOCYTES#	LYMPHO	/mm3	1 560	2 000 - 7 000		
12/08/2011	10	41	SALMI	MONOCYTES#	MONO	/ mm3	590	200 - 1 000		bilan sans
12/08/2011	10	41	SALMI	PHOSPHATASE ALCALINE	PAL	UI/L	77	< 645		particularité
12/08/2011	10	41	SALMI	C- REACTIVE PROTEINE (CRP)	CRP	MG/L	6	inf à 5		
12/08/2011	10	41	SALMI	SGOT/ASAT	GOT	UI/L	19	inf à 60		
12/08/2011	10	41	SALMI	SGPT/ALAT	GPT	UI/L	17	inf à 45		
12/08/2011	10	41	SALMI	ANTISTREPTOLYSINE.O (ASLO)	ASLO	UI/L	inf à 200	< OU =200		
09/18/2017	10	47		GLOBULES ROUGES	GR	Million/mm3	4.39	3.8 - 5.40		
09/18/2017	10	47		GLOBULES BLANCS	GB	Mille/mm3	7.6	4.0 à 12.0		
09/18/2017	10	47		V.G.M	VGM	µ3	84.9	70-100		
09/18/2017	10	47		HEMOGLOBINE	Hb	g/100ml	12.2	11 - 14		
09/18/2017	10	47		C.C.M.H	CCMH	%	32.7	32 - 37		
09/18/2017	10	47		HEMATOCRITE	Hte	%	37.3	33 - 44		
09/18/2017	10	47		PLAQUETTES	PLQ	Mille/mm3	397	150 - 500		
09/18/2017	10	47		T.G.M.H	TGMH	pg	27.8	27 - 32		
09/18/2017	10	47		POLYNUCLEAIRE EOSINOPHILE%	PN EOS	%	4.80	0 - 5		
09/18/2017	10	47		POLYNUCLEAIRE NEUTROPHILE%	PN.NEUTR	%	64.30	0 - 40		
09/18/2017	10	47		POLYNUCLEAIRE BASOPHILE%	PN BASO	%	0.00	0 - 1		
09/18/2017	10	47		POLYNUCLEAIRE BASOPHILE#	PN.BASO	/ mm3	0	0 - 100	syndrome inflammatoire	
09/18/2017	10	47		POLYNUCLEAIRE EOSINOPHILE#	PN.EOSI	/ mm3	400	0 - 500		
09/18/2017	10	47		POLYNUCLEAIRE NEUTROPHILE#	PN.NEUTR	/ mm3	4 900	1 500 - 4 000		
09/18/2017	10	47		LYMPHOCYTES#	LYMPHO	/mm3	2 000	2 000 - 7 000		
09/18/2017	10	47		MONOCYTES#	MONO	/ mm3	300	200 - 1 000		
09/18/2017	10	47		LYMPHOCYTES%	LYMPH	%	26.40	40 - 75		
09/18/2017	10	47		MONOCYTES%	MONO	%	4.50	02 - 10		
09/18/2017	10	47		CALCIUM ca++	Ca	mmol	1.13	1.10 - 1.40		
09/18/2017	10	47		POTASSIUM SANGUIN (K+)	K+	MEQ/L	4.05	3.5 - 5.5		
09/18/2017	10	47		SODIUM SANGUIN (NA+)	NA+	MEQ/L	135.6	135 - 155		
09/18/2017	10	47		CHLORURES (CL)	cl	MEQ/L	105	95 - 107		
09/18/2017	10	47		GLUCOSE SANGUIN (GLYCEMIE)	GLUC	G/L	0.84	0.70 - 1.20		
09/18/2017	10	47		C- REACTIVE PROTEINE (CRP)	CRP	MG/L	9.81	inf à 5		
09/18/2017	10	47		TSHus	TSH	µUI/ML	1.74	0.24 à 4.70		

Figure 4. Example of conclusions of Dr. MASIKA Meriem on the analyses of a patient

Detailed diagnosis for each patient is used to validate the results of the proposed system.

4.2. Classification Process

After the data collection and preparation phase, comes the division phase of the database into two parts, where 70% of it was used as reference dataset and 30% as test dataset. The reference dataset is a set of patient analyses results, which contains



DATE_RETRA	N de patient	AGE	LIB_RUBRIQU	LIB_ABREGE_	UNITE_RUBR	RESULTAT	UNITE_RUBRI	Consultation
2012-06-07	9	48.00	GLOBULES BI	GB	Mille/mm3	5,70	4.0 à 12.0	
2012-06-07	9	48.00	V.G.M	VGM	µ3	90	70-100	
2012-06-07	9	48.00	HEMOGLOBI	Hb	g/100ml	12,1	11 - 14	problème cardiaque + une eruption cutané IGE élevé diabétique
2012-06-07	9	48.00	C.C.M.H	CCMH	%	33	32 - 37	depuis 3 mois avec le taux de choristirole totale elvee et triglycirid
2012-06-07	9	48.00	HEMATOCRI	Hte	%	36,8	33 - 44	elvee risque HTA cardiopatie
2012-06-07	9	48.00	PLAQUETTES	PLQ	Mille/mm3	221	150 - 500	
2012-06-07	9	48.00	T.G.M.H	TGMH	pg	30	27 - 32	
2012-06-07	9	48.00	POLYNUCLEA	PN.BASO	/ mm3	10	0 - 100	
2012-06-07	9	48.00	POLYNUCLEA	PN.EOSI	/ mm3	690	0 - 500	
2012-06-07	9	48.00	POLYNUCLEA	PN.NEUTR	/ mm3	2 300	1 500 - 4 000	
2012-06-07	9	48.00	LYMPHOCYTI	LYMPHO	/mm3	2 400	2 000 - 7 000	
2012-06-07	9	48.00	MONOCYTES	MONO	/ mm3	300	200 - 1 000	
2012-06-07	9	48.00	UREE SANGU	UREE	G/L	0,28	0.10 - 0.50	
2012-06-07	9	48.00	CREATININE	CREA	MG/L	8,00	5 - 16	
2012-06-07	9	48.00	GLUCOSE SAI	GLUC	G/L	1,14	0.70 - 1.20	
2012-06-07	9	48.00	CHOLESTERC	CHOL T	G/L	1,96	0.5 - 1.5	
2012-06-07	9	48.00	IMMUNOGLC	IgE	UI/ml	422,60	1.31 -- 165.3	
2014-01-04	9	50.00	GLUCOSE SAI	GLUC	G/L	0,87	0.70 - 1.20	
2014-01-04	9	50.00	TRIGLYCERID	TRIG	G/L	1,33	0.50 - 1.60	
2014-01-04	9	50.00	CHOLESTERC	CHOL T	G/L	2,26	0.5 - 1.5	
2014-01-04	9	50.00	HDL- CHLEST	HDL.CHOL	G/L	0,65	sup à 0.35	problem HTA ou problème cardiaque
2014-01-04	9	50.00	ANTIGENE HI	AgHBS		NEGATIF	NEGATIF	
2014-01-04	9	50.00	LDL- CHOLES	LDL CHOL	G/L		inf à 1.60	

Figure 5. Sample of the medical analyses and consultations made by Dr. Gritli Youcef

blood components like red blood cells and white blood cells, blood glucose etc. that are needed to diagnose the patient's disease by the doctor. This reference data also includes consultations with doctors who have correctly diagnosed the tests.

Patient ID	RED CELLS	WHITE CELLS	BLOOD GLUCOSE	TSHus	Diagnostic
01	4,5	7.6	0,97	9,36	Normal
02	4,59	5.1	1	0,041	Anaemia
03	4,68	7.1	—	22,99	Hypothyroid
04	4,51	5.2	—	—	Normal
05	4,93	9.2	1,61	—	Diabetes
06	4,4	9.5	1,07	—	Normal
07	5,5	10.6	—	0,05	Hypothyroid
08	4,2	6.4	1,1	—	Anaemia
09	4,8	8.6	—	11,4	Hypothyroid
10	5,6	8.2	—	0,002	Hypothyroid
11	5,06	6.3	0,95	—	Normal
12	4,1	7.3	0,8	—	Anaemia
13	5,1	8	1,36	—	Diabetes
14	5,1	7.4	1,32	—	Diabetes
15	4	5.6	1,23	—	Diabetes
16	3,68	7.3	—	—	Anaemia
17	5,39	8.8	—	—	Diabetes
18	4	5.6	1,2	—	Anaemia
19	5,5	10.9	—	11,4	Hypothyroid
20	5,2	7.8	1,8	—	Normal

Table 1. A set *L* as reference sample

- Let L the dataset available or reference sample:

$$L = \{(y_i, x_i), i = 1, ..., n_l\} \quad (10)$$

Diseases
Normal
Anaemia
Hypothyroid
Diabetes
Hypertriglyceridemia

Table 2. Some patients'possible disease classes

Patient ID	RED CELLS	WHITE CELLS	BLOOD GLUCOSE	TSHus
01	4,5	7.6	0,97	9,36
02	4,59	5.1	1	0,041
03	4,68	7.1	—	22,99
04	4,51	5.2	—	—
05	4,93	9.2	1,61	—
06	4,4	9.5	1,07	—
07	5,5	10.6	—	0,05
08	4,2	6.4	1,1	—
09	4,8	8.6	—	11,4
10	5,6	8.2	—	0,002
11	5,06	6.3	0,95	—
12	4,1	7.3	0,8	—
13	5,1	8	1,36	—
14	5,1	7.4	1,32	—
15	4	5.6	1,23	—
16	3,68	7.3	—	—
17	5,39	8.8	—	—
18	4	5.6	1,2	—
19	5,5	10.9	—	11,4
20	5,2	7.8	1,8	—

Table 3. Some results of the patients' analyses

• Where $y_i \in \{1, ..., c\}$ denotes the class of individual i, in the case of our project the y_i denotes the classes of the patient's diseases.

And the vector $x_i = x_{i1}, ..., x_{ip}$ represented the features of the individual *i*, in the case of this study the x_i represents the results of some analyses of a patient in one occurrence.

- Let *T* the dataset available or test sample:

$$T = \{(y'_i, u_i), i = 1, ..., n_t\} \quad (11)$$

Patient ID	RED CELLS	WHITE CELLS	BLOOD GLUCOSE	TSHus	Diagnostic
01	4,33	9.7	0,71	0,05	???

Table 4. Sample from *T* test dataset

• Where $y'_i \in \{1, ..., c\}$ denotes the class of the individual *i*, in the case of our project y'_i denote the patient's unknown disease class.

• And the vector $u_i, u_{i1},..., u_{ip}$, represents the predictors of the individual *i*, in the case of this project the u_i represent results of the patient's analyses.

Step 1: Calculates the Euclidian distance between the u_i vector and the vector x_i:

$$d\left((u_1, u_2, \dots, u_p), (x_1, x_2, \dots, x_p)\right) = \sqrt{(u_1 - x_1)^2 + (u_2 - x_2)^2 + \cdots + (u_p - x_p)^2} \quad (12)$$

Step 2: Determining an observation (*y, x*) the nearest neighbour by:

$$d(x, x_1)_i = min_i (d(x_i, u_i)) \quad (13)$$

And $\hat{y} = y_i$, the class of the nearest neighbour is selected for the prediction of y'.

Patient ID	RED CELLS	WHITE CELLS	BLOOD GLUCOSE	TSHus	Diagnostic
01	4,33	9.7	0,71	0,05	Normal

Table 5. Example of the obtained result when $k = 1$

The principle of the algorithm of the *k* nearest neighbours is the same principle of the nearest algorithm, but the only difference is instead of taking a single minimum distance between reference vectors and testing, the *k* nearest neighbours' algorithm takes *k* minimum distances depending on the choice of *K* made by the user.

If $k = 3$

Patient ID	Euclidian Distance
01	9.54
02	2.12
03	23.10
04	4.55
05	1.42

Patient ID	Euclidian Distance
06	0.16
07	1.63
08	3.76
09	1.95
10	2.12
11	3.48
12	2.42
13	1.98
14	2.5
15	4.14
16	2.58
17	1.55
18	4.19
19	11.49
20	2.35

Table 6. Euclidian distance between the u_i vector and the vector x_i

Decisions of three minimum distances are:

Patient ID	Diagnostic
05	Diabetes
06	Normal
17	Diabetes

Table 7. The three nearest neighbours.

Step 3: Concluded that the number of 'Diabetes' class is repeated more than 'Normal' class, according to the principle of the algorithm, the *k* nearest neighbours choose the max.

The final result according to $k = 3$:

Patient ID	RED CELLS	WHITE CELLS	BLOOD GLUCOSE	TSHus	Diagnostic
01	4,33	9.7	0,71	0,05	Diabetes

Table 8. The final result of the KNN algorithm when $k = 3$

4.3. Obtained Results

4.3.1. Scenarios

In this section, the results of the application of the proposed classifier on the database, are presented, the database has been divided into two parts: 2/3 of the individuals for learning (references) and 1/3 for tests. The objective of experiments carried out on the database is on the one hand, to evaluate the performance of the used classification algorithm (KNN), and on the other hand, to test the effectiveness of the system in diagnosing diseases.

The table shows the results and performance of the proposed system using the classification rate over the entire test dataset.

k	Classification Rate (%)
1	43.18
3	27.27
5	25
7	25
9	25

Table 9. Corresponding classification rates to the chosen values of *k*

The top results were obtained by testing the system on 44 random data samples selected from the test data, several k values were tested in order to adopt a better result, and $k = 1$ provides a better classification rate 43.18%, and error rate 15.90%. It is important to know that this is a test to see how the model learns from the training data. Based on the results obtained, it can be concluded that this model does not suffer from a problem of under-learning, except the learning base is not very representative. However, to have reliable results on expected performance, new data that has not been used in the learning of our classification model must be used.

It can be noted that the results obtained are not comparable to the results obtained previously, this is because of the small size of database used, compared to the database used previously in the study [5] they used 1004 analysis data from 514 patients where just 159 analysis data from 22 patients were used in this study for the reasons mentioned before.

Also, among the factors that affected these results is the presence of many medical analyses that do not contain all the biochemical components that are replaced by zero value. This makes the classification of several diseases very difficult, because there are often biochemical components on diabetes analysis does not exist on the analysis of anaemia, for example the component “BLOOD GLUCOSE (GLYCEMIA)” exists on diabetes analysis but does not exist on the analysis of anaemia, and the component “PLATELET COUNT” is a very essential element in the analysis of anaemia but it is not important for diabetes.

In addition, the difference in the unit of components can cause problems during the process of calculating distance, for example there are biochemistry values with the unit of the “Million” and the other with “Millilitre”.

Indeed, this study encountered many difficulties during the database collection stage, among them:

- Most medical laboratories in the local area use the paper version, but the electronic version is often confidential.

- The electronic database was not well structured and doctors could not easily understand it.

- Also, it was difficult to find volunteer doctors to read and diagnose medical analyses.

All of these problems consumed a lot of time of this study and directly affected the collection of a very large database. Also, it must be taken into account that the multi-class classification between six classes of diseases - normal, hypothyroidism, hypertriglyceridemia, diabetes, high cholesterol, anaemia - is a complicated and difficult task compared to other tasks, as the majority of diagnostic studies used the binary classification of a very specific disease, and as this method requires a very large database to function properly.

5. Conclusion

In this paper an overview is provided of electronic medical records and their significant improvements in the medical field, to help doctors monitor the condition of their patient's body remotely and help patients, to know their body condition, as well as electronic medical records play an important role in the management of most biomedical laboratories.

Through this work, some questions are answered and discussed in the issue. This was accomplished by performing a disease identification system from electronic medical records using an intelligent algorithm (the algorithm of the *k* nearest neighbours.

In summary, the main contributions of this work are:

• Create a disease identification system that provides a medical diagnostic similar to a doctor's medical diagnostic.

• Develop an algorithm capable of classifying several diseases, unlike other algorithms that use binary classification.

As perspectives:

• Increase the amount data of the learning database to improve the performance of the proposed system.

• Investigate other classification algorithms that can be more accurate than the one currently used.

• Link the proposed system directly to the electronic medical records.

• Applying reduction methods like principal component analyses (PCA) to reduce the number of variables and make information less redundant.

Acknowledgement

This work has been carried out as part of the socio-economic impact research project "Application of Semantic Web Services in the field of Citizen Health, Electronic Health Management -Cloud Solution-" backed by Smart Grids & Renewable Energies laboratory.

Authors would like to thank Dr. BENCHAIB Sidi Mohamed, owner of the medical analyses laboratory that offers data about medical analyses. They would also like to thank Dr. ABZOU Zineb, Dr. MASIKA Meriem for reading patients' analyses and giving diagnoses, and Dr. Gritli Youcef, specialist in Oto-Rhino-Laryngology (ORL) for his efforts in the validation of the obtained results.

References

[1] Jollien, Nathalie. (2016, September) Le temps. [Online]. https://www.letemps.ch/sciences/un-outil-diagnostic-medical-nomme-watson

[2] BELAIDI, Asma., BASSAID, Imane. (2015). Classification de l'hypothyroidie par approche, mono classifieur et multi classifieurs, Tlemcen.

[3] BAKHOUCHE, Houda., BENGLIA, Rachida. (2013). Système d'aide au Diagnostique des Maladies dans, Ouargle.

[4] jean philippe louis. (2018, août) Les echos. [Online]. https://www.lesechos.fr/tech-medias/intelligence-artificielle/maladies-de-loeil-lintelligence-artificielle-meilleure-que-les-medecins-136694

[5] BELHOUARI, Imane., BENABDELKRIM, Fatima. (2017). Un système d'aide à la décision médicale pour la détection précoce du glaucome, Tlemcen.

[6] Kocbek, Simon. (2016). Text mining electronic hospital records to automatically classify admissions against disease: measuring the impact of linking data sources., *Journal of Biomedical Informatics*, no. 64, p 158-167.

[7] REZZIK, Fatima., HABECHE, Amel. (2016). La tenue du dossier médical du patient: un impératif de la qualité des soins., Tizi-Ouzou, 2016.

[8] OTMANI, NADA. (2016, juillet) Doctinews. [Online]. https://www.doctinews.com/index.php/doctinews/institutionnel/item/5004-dossier-medical-informatise

[9] LABIAD Ali, Sélection des mots clés basés sur la classification l'extraction des regles d'association, Québec, 2017.

[10] MIFDAL Rachid. (2019). Application des techniques d'apprentissage automatique pour la prédiction de la tendance des titres financiers, Québec, 2019.

[11] Saravanan Thirumuruganathan. (2010, mai) A Detailed Introduction to K-Nearest Neighbor (KNN) Algorithm. [Online]. https://saravananthirumuruganathan.wordpress.com/about/

[12] CHOUAIB Hassan, Sélection de caractéristiques: méthodes et applications, Paris, 2011.

[13] Belacel Nabil, Méthodologie et Applications à l'Aide au Diagnostic Médical, Bruxelles, 1999-2000.

[14] Moncoachdata. Moncoachdata. [Online]. https://moncoachdata.com/blog/algorithme-k-plus-proches-voisins/

[15] Marina Chatterjee. (2020, février) Mygreatlearning. [Online]. https://www.mygreatlearning.com/blog/knn-algorithm-introduction/

[16] ikram Tourqui and fatma Zohra Gueniaa. (2016). Comparaison de méthodes de classification appliquées à la détection d'objets, EL-OUED, 2016.

[17] Eve Mathieu-Dupas, Algorithme des k plus proches voisins pondérés et application en diagnostic, in 42èmes Journées de Statistique, Marseille, France, 2010.

[18] HOUAM Lotfi. (2013). • Contribution à l'analyse de textures de radiographies osseuses pour le diagnostic précoce de l'ostéoporose, Guelma, 2013.

[19] Borland/Inprise, InterBase 6, Operations Guide. Scotts Valley, CA, 1999.

[20] Margaret Rouse. (2020). TechTarget. [Online]. https://whatis.techtarget.com/fr/definition/Backup-sauvegarde

[21] Chloé-Agathe Azencott, introduction au machine learning. Malakoff: Dunod, 2018.