MOHAMMED VI POLYTECHNIC UNIVERSITY

IST&I INSTITUTE OF SCIENCE TECHNOLOGY & INNOVATION

Master 1 Modelling and Data Science

Project of Semester 2

Identification of Symptoms Based on Natural Language Processing for Disease Diagnosis Based on International Classification of Diseases (ICD-11)

Project Realised By:

AARAB Ilham & TCHOKPONHOUE Aurel Davy

Supervised By:

Pr. Taoufik Rachad

Jury:

Pr. Taoufik RACHAD and Pr. Ali IDRI

ALKHAWARIZMI DEPARTMENT

2021/2022

Acknowledgement

We would like to express our deep gratitude to Professor RACHAD Taoufik, our project supervisor, for their patient guidance, enthusiastic encouragement and useful critiques of this research work.

Our grateful thanks are also extended to Mr.IDRI for teaching us the principles of KDD and machine learning in a sophisticated way.

We would also like to extend our sincere thanks to all our professors of this year for their worthwhile efforts and their explanations.

Finally, we wish to thank our parents for their support and encouragement throughout our study.

Abstract

The field of biosciences have advanced to a larger extent and have generated large amounts of information from Electronic Health Records. This have given rise to the acute need of knowledge generation from this enormous amount of data. Datamining methods and machine learning play a major role in this aspect of biosciences. Furthermore, the automatic diagnosis of different diseases is extremely necessary for doctors, in order to facilitate the consultation process, but also for patients, so that they can have quick access to consultation anywhere, anytime and at low cost.

However, this process is often difficult to perform by computer because of the natural language used and the irregular expressions. For this reason, natural language processing (NLP) techniques are used. In this project, an NLP system was created to identify the symptoms and predict the illness of the patient from the description of their state.

We have used the ICD11 as a terminology of diseases and symptoms and a set of machine learning algorithms. For instance, the Decision Tree, Logistic Regression and Random Forest models, in addition to Gradient Boosting that perform similarly on this task, including 95% accuracy, 93% recall and 93% F1 score. The experimental results show that the decision tree and logistic regression improve the efficiency of the system.

Keywords: NLP, ICD11, Machine Learning, Decision tree, Random Forest, Logistic Regression, Disease, Symptoms

Résumé

Le domaine des biosciences a progressé dans une plus large mesure et a généré de grandes quantités d'informations à partir des dossiers de santé électroniques. Cela a donné lieu au besoin aigu de génération de connaissances à partir de cette énorme quantité de données. Les méthodes de datamining et d'apprentissage automatique jouent un rôle majeur dans cet aspect des biosciences. De plus, le diagnostic automatique de différentes maladies est extrêmement nécessaire pour les médecins, afin de faciliter le processus de consultation, mais aussi pour les patients, afin qu'ils puissent accéder rapidement à une consultation n'importe où, n'importe quand et à moindre coût.

Cependant, ce processus est souvent difficile à réaliser par ordinateur en raison du langage naturel utilisé et des expressions irrégulières. Pour cette raison, des techniques de traitement du langage naturel (NLP) sont utilisées. Dans ce projet, un système NLP a été créé pour identifier les symptômes et prédire la maladie du patient à partir de la description de son état

Nous avons utilisé ICD11 comme terminologie des maladies et des symptômes et un ensemble d'algorithmes d'apprentissage automatique. Par exemple, les modèles d'arbre de décision, de régression logistique et de random forest, en plus du Gradient Boosting qui fonctionnent de manière similaire sur cette tâche, y compris une précision de 95 %, un rappel de 93 % et un score F1 de 93 %. Les résultats expérimentaux montrent que l'arbre de décision et la régression logistique améliorent l'efficacité du système.

Mots-clés : NLP, ICD11, Machine Learning, Arbre de décision, Random Forest, Régression Logistique, Maladie, Symptômes

Contents

Introduction

Health is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity (WHO, 1947). It is a state of well-being essential to the full development of a person and essential to his or her development. Characterized by a very central place in the lives of all human beings, its place is justified within the UN Sustainable Development Goals (SDGs), notably SDG 3 on "Good Health and Well-Being". Highlighting the type of disease or problem that an individual is suffering from requires a diagnosis. Thus, diagnosis is a process that aims to discover the type of disease a patient is suffering from so that measures can be taken for its rapid intervention.

The technique that is commonly adopted to diagnose the disease from which a patient suffers is that of anamnesis, which consists of questions and answers directly or indirectly between patients and health workers (who can diagnose the disease)[1]. Through this technique, the health specialist identifies the symptoms and signs that the patient presents, on the basis of which they identify the disease from which he/she suffers by cross-referencing the information and knowledge that he/she possesses regarding the diseases.

Furthermore, with the new technological advances and innovations in the medical field, it is necessary to adopt medical expert systems that will supervise and control the diagnosis and treatment processes [2]. Thus, thanks to artificial intelligence algorithms, new systems have emerged that provide remote health interventions through a synchronous dialogue system [3,4]. This has many advantages such as rapid decision support for medical diagnosis, treatment, and prevention of diseases. These tools, which can improve the performance of the doctor, also contribute to making consultations accessible to everyone at a lower cost. For this reason, several research works are in the process of setting up efficient solutions for the diagnosis of diseases [5]–[7] and symptoms extraction [8]–[10] . These are based on natural language processing (NLP) techniques that allow knowledge to be extracted from unstructured text[11] .

Our project consists in building a simple disease diagnosis application based on the description of the symptoms by the patient (symptom checker). Thus, through natural language processing (NLP) technics, the device will first extract the symptoms of the diseases which will then be passed to a machine learning model to predict the disease from which the patient is suffering.

The parts of this report is constructed as follows: first we start with a methodology then the second chapter is dedicated for data collection process, then the third chapter includes the workflow and realization of the project, last but not least the 4th chapter represents the results and finally a conclusion of all the work.

Chapter I: Methodology

1-1- Introduction

This chapter describes the methodology adopted in the building of the project, beginning with the definition of ICD11, then revealing the source of used datasets, besides the techniques of NLP, then defining a set of machine learning algorithms with their performance measurements.

1-2- ICD11 API

The International Classification of Diseases and Related Health Problems (ICD-11) is a list of medical classifications made by the World Health Organization (WHO) which contains disease codes, signs and symptoms, and various details about other diseases.

Disease classification in the ICD list uses numerical codes, letters, or numeric letter combinations, this aims to homogenize names and classes of diseases, injuries, symptoms and factors that affect health. ICD is an international standard in the medical world in terms of diagnostic classification of diseases. The 10th ICD has been made to the tenth revision (ICD-11) which began in 1983 and was completed in 1992. More than 70,000 codes are found on the ICD-11, which is far from the 14,000 ICD-9 codes. These codes represent each classification of the type of asset along with the index and other detailed information [10].

1-3- Dataset

Experiments are conducted on Symptoms with knowledge database of disease-symptom associations generated by an automated method based on information in textual discharge summaries of patients at New York Presbyterian Hospital admitted during 2004. This dataset contains 3 attributes (including symptoms, Count of Disease Occurrence and diseases) and 1867 instances, this dataset contains 38485 patients that suffer from those diseases and the total number of diseases is 140. The most popular python Programming Data analytics tool has been used to construct the prediction framework.

The second dataset is obtained using web scrapping on ICD11 website, it contains the diseases with their description, with 140 instances.

1-4- Techniques used

1-3-1- Natural Language Processing (NLP)

NLP is a field of computer science that deals with the interaction between computers and humans, such as Indonesian or English. NLP is used to process writing in order to understand what is said by humans.

The main purpose of NLP is to make machines or computers to understand the meaning of human language so that they can provide appropriate responses. NLP application in the medical field is very important as Clinical Decision Support (CDS) which helps health professionals make clinical decisions, deal with medical data about patients or with the knowledge of drugs needed to interpret the data [12].

A- Tokenization

In NLP, tokenizing is the first processes in NLP that identify basic tokens or units for the next process. In simple terms, tokenizing breaks down large text data into smaller shapes to facilitate the analysis process, such as paragraphs being sentences, or sentences being words. In this study, we do not break paragraphs into sentences but sentences into words because the results of speech to text conversion are in the form of a long sentence [13].

B- Stemming

Stemming works by transforming words into their basic forms. the basic form is not always the same as the root word. in general, the basic word in Indonesian has a combination.

Prefix 1 + Prefix 2 + Basic Word + Suffix 3 + Suffix 2 + Suffix 1

C- Stop words Removal

Almost all NLP implementations in the Machine Learning field use the stop words removal method. This method works by removing several conjunctions but does not affect the overall content. Stopwords removal is used to improve system performance in order to effectively process the data needed.

D- Transformers

Sentence Transformers is a Python framework for state-of-the-art sentence, text and image embeddings**,** transformers are deep learning model used in NLP, it's used for encoding sentences which means the vectorial representation of sentences.

E- Semantic similarity

The semantic similarity, or textual semantic similarity, is a task in the domain of natural language processing (NLP) that evaluate the relationship between texts or documents using a defined metric. It determines the degree of similarity between two tetxts.

F- TF-IDF

TF-IDF (term frequency-inverse document frequency) is a statistical measure that evaluates how relevant a word is to a document in a collection of documents. This is done by multiplying two metrics: how many times a word appears in a document, and the inverse document frequency of the word across a set of documents. It has many uses, most importantly in automated text analysis, and is very useful for scoring words in machine learning algorithms for Natural Language Processing (NLP).

1-3-2-Symptoms extraction

A- Description

In the majority of the work done in the literature on the extraction of symptoms or named entities, Name entity recogniton (NER) techniques are used. NER is a technique based on deep learning which is used in the extraction of information from textual documents.

In this work, we have adopted another method to achieve this. We based our work on the extraction of symptoms from similarity calculation. Indeed, each phrase that a patient express alludes to a symptom that he presents. The question would be to identify which symptom it corresponds to, knowing that the phrase or description expressed may be totally outside the medical vocabulary. In order to do this, we used a pre-trained model based on Transformers, which are the most efficient models in NLP in order to have a vector representation of all the symptoms associated with the diseases present in our dataset, as well as the description of the patient. Following this, a metric of similarity calculation between these vector representations, namely the cosine similarity, is used to identify the symptom to which each expression of the patient best corresponds among the list of all the symptoms contained in the dataset. Thus, we obtain the symptoms that the patient presents.

B- Cosine similarity

The formula of cosine similarity metric is:

$$similarity = \cos(\theta) = \frac{< A \, , B >}{|A| * |B|} = \frac{\sum_{i=1}^{n} a_i * b_i}{\sqrt{\sum_{i=1}^{n} a_i^2} * \sqrt{\sum_{i=1}^{n} b_i^2}}$$

$$where\ a_i\ and\ b_i\ represent\ the\ i^{th}\ conponent\ of\ the\ vector\ A\ and\ B$$

1-3-3- Machine Learning

Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy. The process leading to the creation of a machine learning model goes from the pre-processing stage to the deployment of the model and its evaluation.

A- Preprocessing technique

The Binarizer multilabel to indicate the presence or absence of a given symptom for a disease. Label encoding to assign a numerical value to each disease in the database.

B- Algorithms used

- Naive Bayes Classifier

Naïve Bayes classifier is a supervised algorithm. It is a simple classification technique using Bayes theorem. It assumes strong (Naive) independence among attributes. Bayes theorem is a mathematical concept to get the probability. The predictors are neither related to each other nor have correlation to one another. All the attributes independently contribute to the probability to maximize it. It is able to work with Naïve Bayes model and does not use Bayesian methods. Many complex real-world situations use Naive Bayes classifiers [14]:

$$P(X/Y) = \frac{P(Y/X) * P(X)}{P(Y)}$$

$P(X/Y)$ is the posterior probability, $P(X)$ is the posterior probability, $P(Y)$ is the predictor prior probability, $P(Y/X)$ is the likelihood, probability of predictor.

Naïve Bayes is a simple, easy to implement, and efficient classification algorithm that handles non-linear, complicated data. However, there is a loss of accuracy as it is based on assumption and class conditional independence.

- Decision Tree

Decision tree is a classification algorithm that works on categorical as well as numerical data. Decision tree is used for creating tree-like structures. Decision tree is simple and widely used to handle medical dataset. It is easy to implement and analyze the data in tree-shaped graph. The decision tree model makes analysis based on three nodes.

Root node: main node, based on other nodes functions.

Interior node: handles various attributes.

Leaf node: represent the result of each test.

This algorithm splits the data into two or more analogous sets based on the most important indicators. The entropy of each attribute is calculated and then the data are divided, with predictors having maximum information gain or minimum entropy:

$$Entropie(S) = \sum_{i=1}^{c} - P_i * log_2(P_i)$$

$$Gain(S, A) = Entropie(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropie(S_v)$$

The results obtained are easier to read and interpret [15]. This algorithm has higher accuracy in comparison to other algorithms as it analyzes the dataset in the tree-like graph. However, the data may be over classified and only one attribute is tested at a time for decision-making.

- K-Nearest Neighbor (KNN)

The K-Nearest Neighbors algorithm is a supervised classification algorithm method. It classifies objects dependent on nearest neighbor. It is a type of instance-based learning. The calculation of distance of an attribute from its neighbors is measured using Euclidean distance [15]. It uses a group of named points and uses them on how to mark another point. The data are clustered based on similarity amongst them and is possible to fill the missing values of data using K-NN. Once the missing values are filled, various prediction techniques apply to the data set. It is possible to gain better accuracy by utilizing various combinations of these algorithms.

K-NN algorithm is simple to carry out without creating a model or making other assumptions. This algorithm is versatile and is used for classification, regression, and search. Even though K-NN is the simplest algorithm, noisy and irrelevant features affect its accuracy

- Random Forest Algorithm

Random forest algorithm is a supervised classification algorithmic technique. In this algorithm, several trees create a forest. Each individual tree in random forest lets out a class expectation and the class with most votes turns into a model's forecast. In the random forest classifier, the greater number of trees give higher accuracy. The three common methodologies are: Forest RI (random input choice); Forest RC (random blend); Combination of forest RI and forest RC.

It is used for classification as well as regression task, but can do well with classification task, and can overcome missing values. Besides, being slow to obtain predictions as it requires large data sets and more trees, results are unaccountable.

- Logistic Regression (LR)

LR models have been acquired from the statistics branch. This algorithm has adapted for binary classification problem statements. The main aim of LR is to discover the value of coefficients. The LR converts the value to 0-1. LR model selects the probability of the given data instance of the class to predict as 0 or 1. This technique can be applied for problems when we emerge with multiple reasons for predicting.

The LR standard function is defined as follows:

$$P(Y=1|X)=\frac{e^{\beta_0+\beta_{11}X_{11}+\cdots+\beta_{ki}X_{ki}}}{1+e^{\beta_0+\beta_{11}X_{11}+\cdots+\beta_{ki}X_{ki}}} \quad or \quad logit[P(Y=1|X)]=\beta_0+\beta_{11}X_{11}+\cdots+\beta_{ki}X_{ki}$$

$$where \;\; logit(X)=\ln\left(\frac{X}{1-X}\right)$$

X_{ij}: j^th^ modality of the i^th^ variable

$X=(X_1, X_2, \ldots, X_i)$

β_{ij} : represents the effect of the modality X_{ij} adjusted for the effects of all other variables included in the model

- Support Vector Machine (SVM)

In a Support Vector Machine (SVM), the hyperplane is built between distinct classes or objects in order to classify the data. The hyperplane is generated by calculating the dimensions of the problem space. Additionally, dimensionality reduction is possible in SVM to balance data dimensions. In order to create a gap between the classes, the marginal distance is calculated from the hyperplane’s center using the class corner points and the support vectors. Kernels, C coefficients, and intercepts are some of the parameters used in SVM. Kernels are the most important aspect of SVM. These kernels have been fine-tuned based on the type of data they process.

- Gradient boosting Classifier

Groups of algorithms called Gradient Boosting Classifiers combine numerous poor learning models to achieve an effective prediction. It is usual practice to employ decision trees to increase the gradient. Regression and classification problems can be solved by using gradient boosting, a machine-learning technique that builds a predictive model from a collection of low-quality models. Decision trees with a weak learner are known as gradient trees, and they often outperform Random Forests in comparison. Rather than building the model sequentially, it applies it by minimizing an arbitrarily differentiable loss function, as other techniques do.

C- Performance measurement

To measure the quality of the models created, several performance metrics such as accuracy, recall, precision and f1-score were used.

These different metrics are calculated from a tabular structure called the confusion matrix. It is defined in four parts: the first is the true positive (TP) in which the values are identified as positive and are really positive. The second is the false positive (FP) in which the values are negative but are predicted to be positive. The third is false negative (FN) in which the value was positive but was identified as negative. The fourth is true negative (TN) in which the value was negative and was identified as negative.

- From this we can calculate the accuracy which represents the rate of good predictions:

$$\boldsymbol{Accuracy} = \frac{\mathrm{TP} \; + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

- Then we have the Precision and Recall which respectively allow us to have an idea of the proportion of real positive individuals correctly identified by the model and the proportion of real negative individuals identified by the model. The formulas are respectively:

$$\boldsymbol{Precision} = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \qquad \boldsymbol{Recall} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$$

- Lastly, the F1-score, which combines the precision and the Recall through their harmonic mean, indicates how accurate the model is

$$\boldsymbol{F1\_score} = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}}$$

1-5- Conclusion

In this chapter, we had to describe the methodology we adopted for this work. More specifically, the data source, the NLP techniques used as well as the Machine Learning techniques used and the evaluation methods. Also, we talk about the process of symptoms extractions.

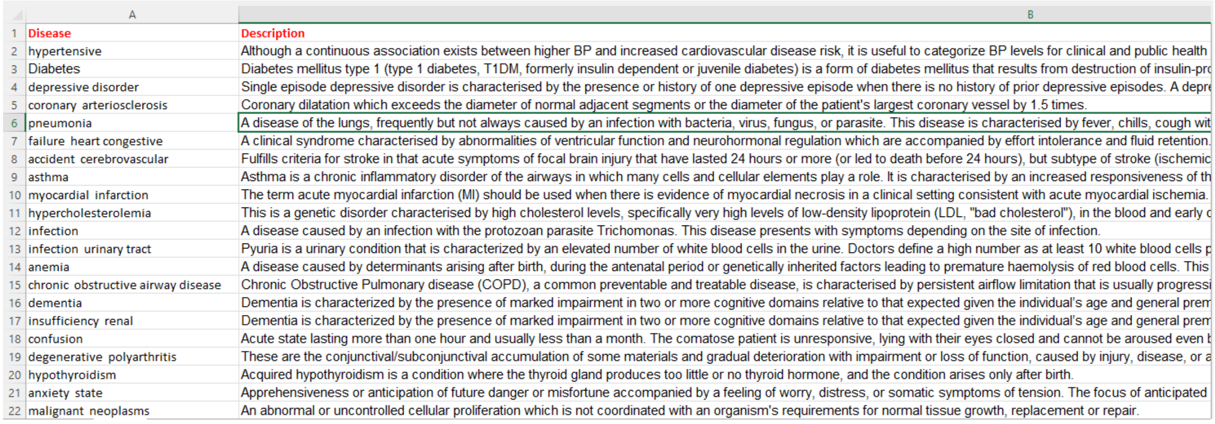
Chapter II: Data Collection

1-1- Introduction

In this project we have used two datasets, one called dataset knowledge which allow us to have the disease and the symptoms which is associate, and the other is ‘Disease/Description’ dataset. In this chapter we will present the whole process used to render them in a structured format.

1-2- ICD 11

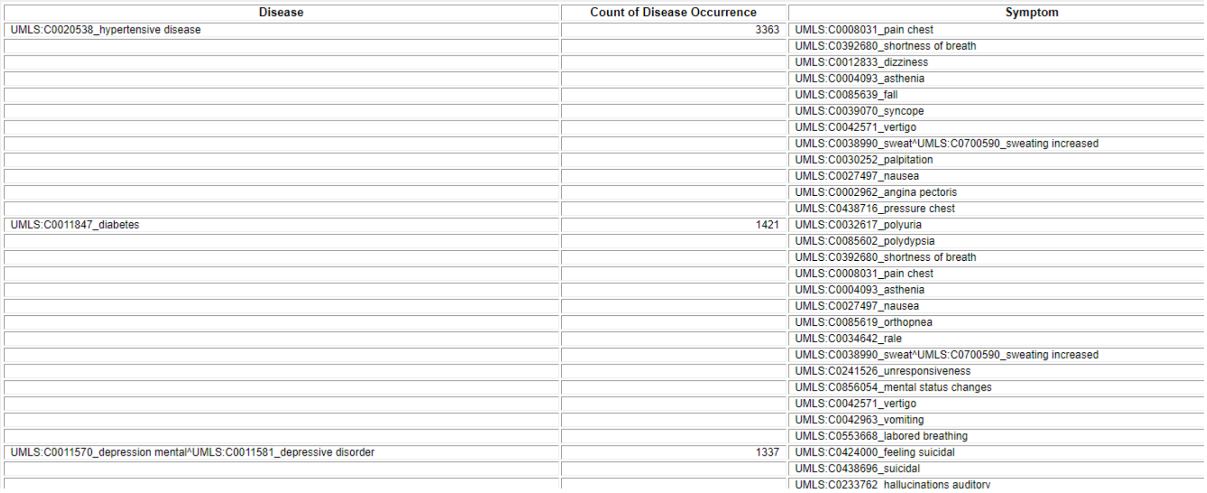
ICD11 was the basic website from which we have used web scraping for retrieving 140 diseases with their descriptions based on WHO classification.



	A	B
1	Disease	Description
2	hypertensive	Although a continuous association exists between higher BP and increased cardiovascular disease risk, it is useful to categorize BP levels for clinical and public health
3	Diabetes	Diabetes mellitus type 1 (type 1 diabetes, T1DM, formerly insulin dependent or juvenile diabetes) is a form of diabetes mellitus that results from destruction of insulin-pro
4	depressive disorder	Single episode depressive disorder is characterised by the presence or history of one depressive episode when there is no history of prior depressive episodes. A depre
5	coronary arteriosclerosis	Coronary dilatation which exceeds the diameter of normal adjacent segments or the diameter of the patient's largest coronary vessel by 1.5 times.
6	pneumonia	A disease of the lungs, frequently but not always caused by an infection with bacteria, virus, fungus, or parasite. This disease is characterised by fever, chills, cough wit
7	failure heart congestive	A clinical syndrome characterised by abnormalities of ventricular function and neurohormonal regulation which are accompanied by effort intolerance and fluid retention.
8	accident cerebrovascular	Fulfills criteria for stroke in that acute symptoms of focal brain injury that have lasted 24 hours or more (or led to death before 24 hours), but subtype of stroke (ischemic
9	asthma	Asthma is a chronic inflammatory disorder of the airways in which many cells and cellular elements play a role. It is characterised by an increased responsiveness of th
10	myocardial infarction	The term acute myocardial infarction (MI) should be used when there is evidence of myocardial necrosis in a clinical setting consistent with acute myocardial ischemia.
11	hypercholesterolemia	This is a genetic disorder characterised by high cholesterol levels, specifically very high levels of low-density lipoprotein (LDL, "bad cholesterol"), in the blood and early c
12	infection	A disease caused by an infection with the protozoan parasite Trichomonas. This disease presents with symptoms depending on the site of infection.
13	infection urinary tract	Pyuria is a urinary condition that is characterized by an elevated number of white blood cells in the urine. Doctors define a high number as at least 10 white blood cells p
14	anemia	A disease caused by determinants arising after birth, during the antenatal period or genetically inherited factors leading to premature haemolysis of red blood cells. This
15	chronic obstructive airway disease	Chronic Obstructive Pulmonary disease (COPD), a common preventable and treatable disease, is characterised by persistent airflow limitation that is usually progressi
16	dementia	Dementia is characterized by the presence of marked impairment in two or more cognitive domains relative to that expected given the individual’s age and general prem
17	insufficiency renal	Dementia is characterized by the presence of marked impairment in two or more cognitive domains relative to that expected given the individual’s age and general prem
18	confusion	Acute state lasting more than one hour and usually less than a month. The comatose patient is unresponsive, lying with their eyes closed and cannot be aroused even b
19	degenerative polyarthritis	These are the conjunctival/subconjunctival accumulation of some materials and gradual deterioration with impairment or loss of function, caused by injury, disease, or a
20	hypothyroidism	Acquired hypothyroidism is a condition where the thyroid gland produces too little or no thyroid hormone, and the condition arises only after birth.
21	anxiety state	Apprehensiveness or anticipation of future danger or misfortune accompanied by a feeling of worry, distress, or somatic symptoms of tension. The focus of anticipated
22	malignant neoplasms	An abnormal or uncontrolled cellular proliferation which is not coordinated with an organism's requirements for normal tissue growth, replacement or repair.

1-3- Database Knowledge

Knowledge database of disease-symptom associations generated by an automated method based on information in textual discharge summaries of patients at New York Presbyterian Hospital admitted during 2004. This dataset contains 3 attributes (including symptoms, Count of Disease Occurrence and diseases) and 1867 instances, this dataset contains 38485 patients that suffer from those diseases.



Disease	Count of Disease Occurrence	Symptom
UMLS:C0020538_hypertensive disease	3363	UMLS:C0008031_pain chest
		UMLS:C0392680_shortness of breath
		UMLS:C0012833_dizziness
		UMLS:C0004093_asthenia
		UMLS:C0085639_fall
		UMLS:C0039070_syncope
		UMLS:C0042571_vertigo
		UMLS:C0038990_sweat^UMLS:C0700590_sweating increased
		UMLS:C0030252_palpitation
		UMLS:C0027497_nausea
		UMLS:C0002962_angina pectoris
		UMLS:C0438716_pressure chest
UMLS:C0011847_diabetes	1421	UMLS:C0032617_polyuria
		UMLS:C0085602_polydypsia
		UMLS:C0392680_shortness of breath
		UMLS:C0008031_pain chest
		UMLS:C0004093_asthenia
		UMLS:C0027497_nausea
		UMLS:C0085619_orthopnea
		UMLS:C0034642_rale
		UMLS:C0038990_sweat^UMLS:C0700590_sweating increased
		UMLS:C0241526_unresponsiveness
		UMLS:C0856054_mental status changes
		UMLS:C0042571_vertigo
		UMLS:C0042963_vomiting
		UMLS:C0553668_labored breathing
UMLS:C0011570_depression mental^UMLS:C0011581_depressive disorder	1337	UMLS:C0424000_feeling suicidal
		UMLS:C0438696_suicidal
		UMLS:C0233762_hallucinations auditory

1-4- Dataset Preparation

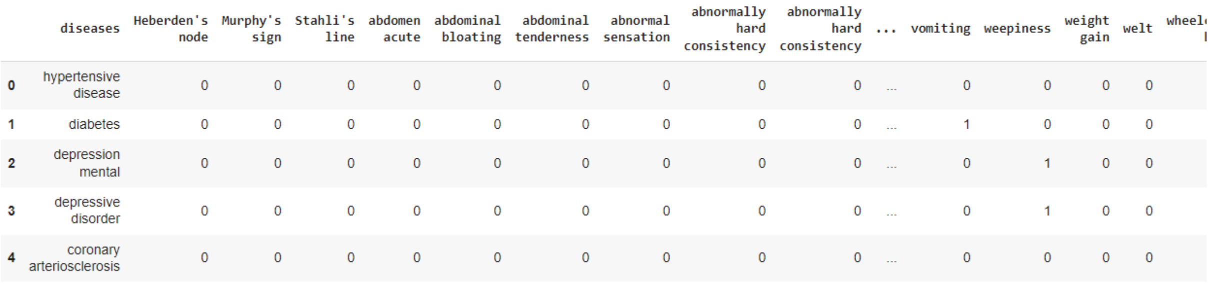
1-4-1- Dataset of description from ICD

For this dataset, to make it structured a TF-IDF was used while taking care to remove the stop words. Before that, the description column was cleaned up by tokenizing and removing punctuation, numbers and unwanted characters.

1-4-2- Dataset of symptoms and disease

For this dataset, we first imputed the missing values resulting from the format of the dataset by the name of the corresponding diseases. Then, we performed a cleaning of the undesirable characters as well as the UMLS codes associated with the symptoms and diseases. This allowed us to create a first dataset in graph format.

In addition, a second dataset containing the diseases and all associated symptoms stored in a list was created. On this dataset was then applied a multilabel binarizer in order to obtain a new dataset containing columns associated with each symptom containing either 1 or 0 where 1 means the presence of the symptom for the disease and 0 its absence. The machine learning algorithms were applied to this dataset.



	diseases	Heberden's node	Murphy's sign	Stahli's line	abdomen acute	abdominal bloating	abdominal tenderness	abnormal sensation	abnormally hard consistency	abnormally hard consistency	...	vomiting	weepiness	weight gain	welt	wheel
0	hypertensive disease	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
1	diabetes	0	0	0	0	0	0	0	0	0	...	1	0	0	0	
2	depression mental	0	0	0	0	0	0	0	0	0	...	0	1	0	0	
3	depressive disorder	0	0	0	0	0	0	0	0	0	...	0	1	0	0	
4	coronary arteriosclerosis	0	0	0	0	0	0	0	0	0	...	0	0	0	0	

1-5- Conclusion

The reason behind processing on two datasets is to apply our ML algorithms and observe which dataset that gives the most convenient and well performed model.

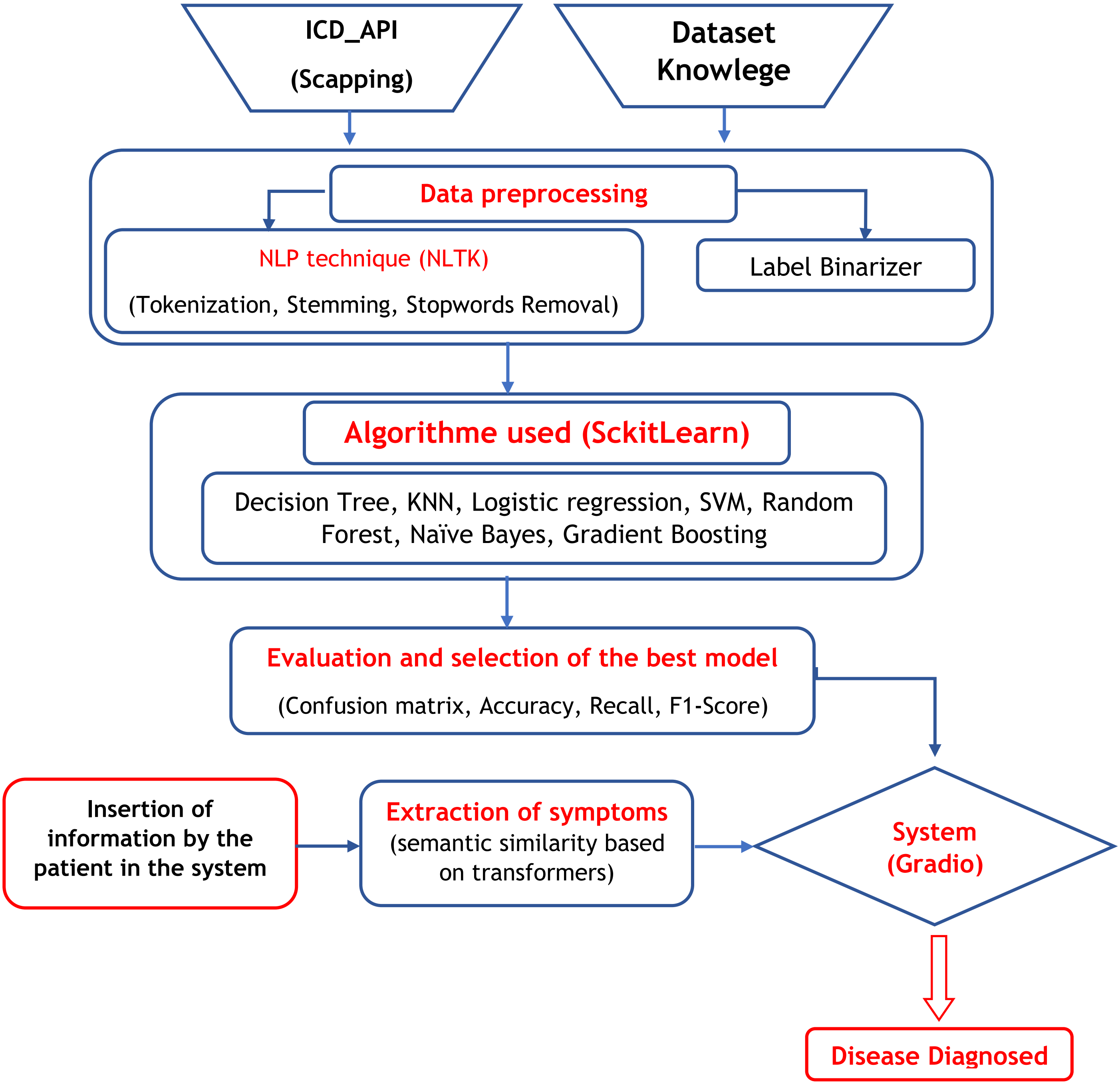
Chapter III: Realization

1-1- Introduction

In this chapter we mention the details about project's workflow, including scrapping from ICD11 and python libraries used for programming, furthermore the framework used for implementing the web interface.

1-2- Experimental Design

Figure 1: Workflow project


ICD_API
(Scapping)
Dataset Knowlege
Data preprocessing
NLP technique (NLTK)
(Tokenization, Stemming, Stopwords Removal)
Label Binarizer
Algorithme used (SckitLearn)
Decision Tree, KNN, Logistic regression, SVM, Random Forest, Naïve Bayes, Gradient Boosting
Evaluation and selection of the best model
(Confusion matrix, Accuracy, Recall, F1-Score)
Insertion of information by the patient in the system
Extraction of symptoms (semantic similarity based on transformers)
System (Gradio)
Disease Diagnosed

Source: Our own work

The figure above shows the workflow of our project. Indeed, following the various tasks carried out for the recovery of the dataset. Some manipulations have been done to render it in a structured format. Regarding the description dataset that was scrapped from the ICD11, NLP pre-processing techniques were applied on it in order to be able to use it for learning. On the other hand, a Multilabel Binarizer was applied to the disease-symptom association dataset.

Machine learning algorithms such as Decision Tree, Logistic Regression, SVM, Random Forest, Naive Bayes, KNN, Gradient boosting were then applied to both datasets to determine which approach would give the best performance. The model deployed in the system was selected on this basis.

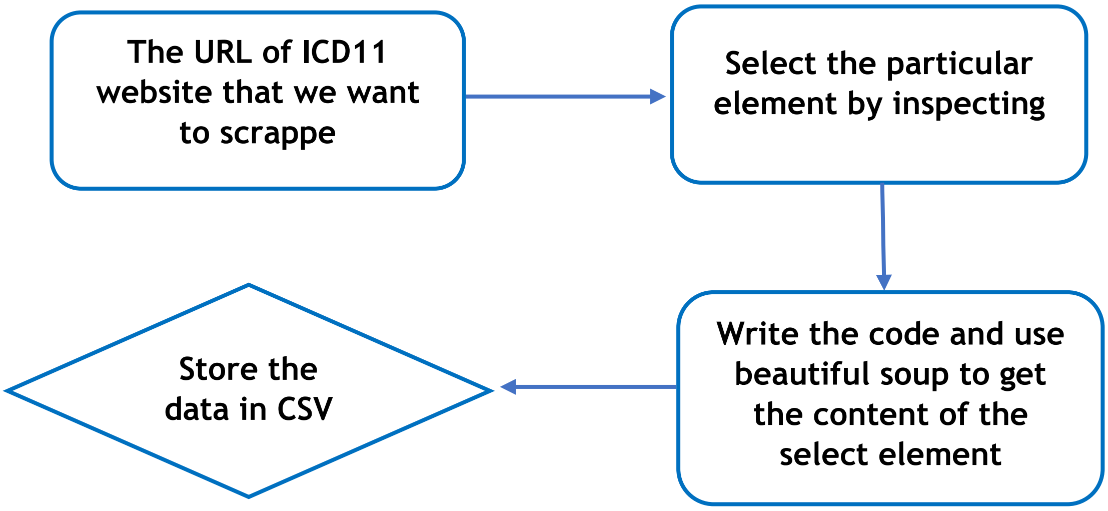
At the outset, a function to extract symptoms from the description of the effects they experience was constructed based on a similarity function using a pre-trained model of Transformers which served to provide the vector representation necessary for the calculation of similarities.

Following this, the web interface serving as a gateway for querying the model for diagnostics was set up.

1-3- Web Scraping

1-3-1- Definition

Web scraping is the process of using bots to extract content and data from a website. Unlike screen scraping, which only copies pixels displayed onscreen, web scraping extracts underlying HTML code and, with it, data stored in a database. The scraper can then replicate entire website content elsewhere.


The URL of ICD11 website that we want to scrappe
Select the particular element by inspecting
Write the code and use beautiful soup to get the content of the select element
Store the data in CSV

1-3-2- Tools: Beautiful Soup

Beautiful Soup is a Python package for parsing HTML and XML documents . It creates a parse tree for parsed pages that can be used to extract data from HTML, which is useful for web scraping.

1-4- Python Libraries

- **NLTK**

The Natural Language Toolkit (NLTK) is a platform used for building Python programs that work with human language data for applying in statistical natural language processing (NLP). It contains text processing libraries for tokenization, parsing, classification, stemming, tagging and semantic reasoning.

- **Scikit-Learn**

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.

- **Gradio**

Gradio is an open-source Python library that is used to build machine learning and data science demos and web applications .With Gradio, you can quickly create a beautiful user interface around your machine learning models or data science workflow and let people "try it out" by dragging-and-dropping in their own images, pasting text, recording their own voice, and interacting with your demo, all through the browser.

1-5- Conclusion

In this chapter, we have presented the different tools and libraries we have used as well as the experimental design. As tools and library, they were beautiful soup, NLTK, scikit-Learn and Gradio.

Chapter IV: Results

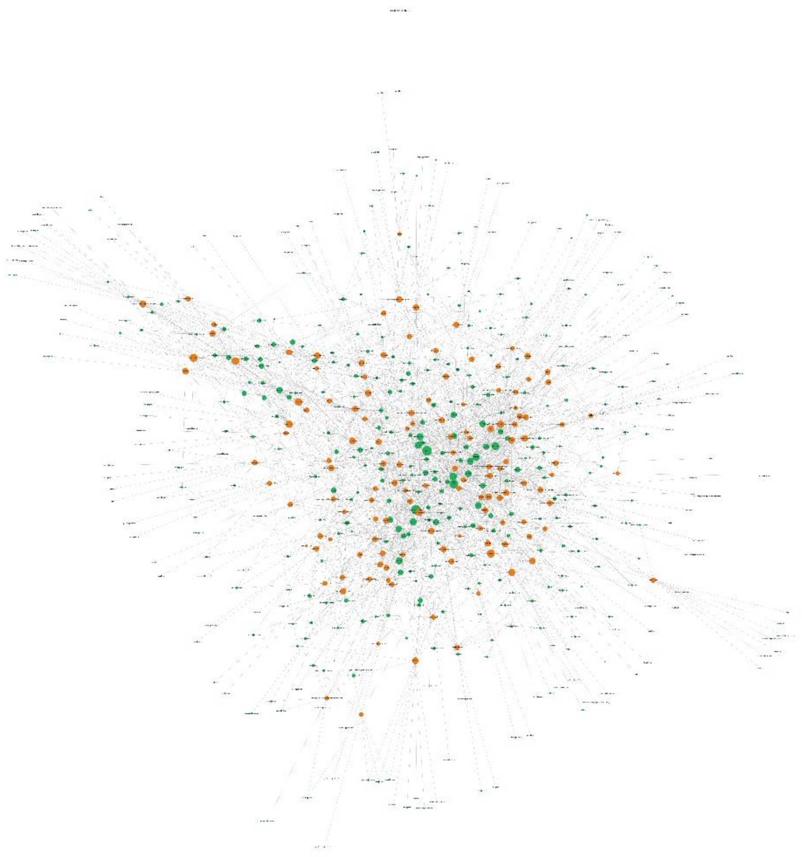
1-1- Introduction

In this chapter, we show the graph of symptoms and diseases, besides the results of ML algorithms applied on datasets 1 and 2 and a short presentation of the web interface that we have created.

1-2- Exploration of the graph of diseases and symptoms

The figure below shows the graph of associations between diseases and symptoms

Figure 2: Diseases and symptoms association graph



Source: Our own work

The analysis of this network through centrality has allowed us to identify the most recurrent symptoms that occur in many diseases. The most important ones are among others 'pain', 'shortness of breath', 'fever', 'bipolar disorder', 'upper respiratory infection', 'abdominal pain', 'psychotic disorder', 'vomiting', 'diarrhea', 'anxiety state'.

1-3- Comparison of the performance of the ML techniques used

The table below shows the performance of each algorithm on dataset 1 (containing the symptoms and associated diseases).

Table 1: Performance by model for dataset 1

Algorithme	Acurracy	Recall	F1-score
Decision Tree	0.95	0.93	0.93
Regression Logistique	0.95	0.93	0.93
SVM	0.94	0.92	0.93
KNN	0.95	0.93	0.93
Naive Bayes	0.95	0.93	0.93
Random Forest	0.94	0.92	0.93
Gradient Boosting	0.95	0.93	0.93

Source: Our own work

The analysis of this table shows that the models with the best performances for our task of predicting diseases from the extracted symptoms are the Decision Tree, Logistic Regression, KNN, Naive Bayes and Gradient Boosting models with similar performance (Accuracy of 95%, recall of 93% and F1-Score of 93%). Furthermore, an in-depth analysis of the error made by the different algorithms employed, reveals that the Decision Tree and Logistic Regression models performed better. Indeed, the errors made by these two models generally concerned diseases of the same family as the real disease from which the patients suffered. Since we also want to have an idea of the probabilities associated with the predicted diseases, and since the logistic model is intrinsically based on probabilities, the logistic regression model was chosen as the prediction model for our system.

On the other hand, there is no significant difference between the performances obtained for the different models tested on the dataset containing the descriptions, notably an Accuracy of

92.9%, a recall of 91% and an F1-score of 91% (Table 2). These performances are lower compared to those obtained for the symptoms and diseases dataset

Table 2: Performance by model for dataset 2 (description dataset)

Algorithme	Acurracy	Recall	F1-score
Decision Tree	0.929	0.93	0.91
Regression Logistique	0.929	0.93	0.91
SVM	0.929	0.93	0.91
KNN	0.929	0.90	0.91
Naive Bayes	0.929	0.90	0.91
Random Forest	0.929	0.90	0.91
Gradient Boosting	0.929	0.90	0.91

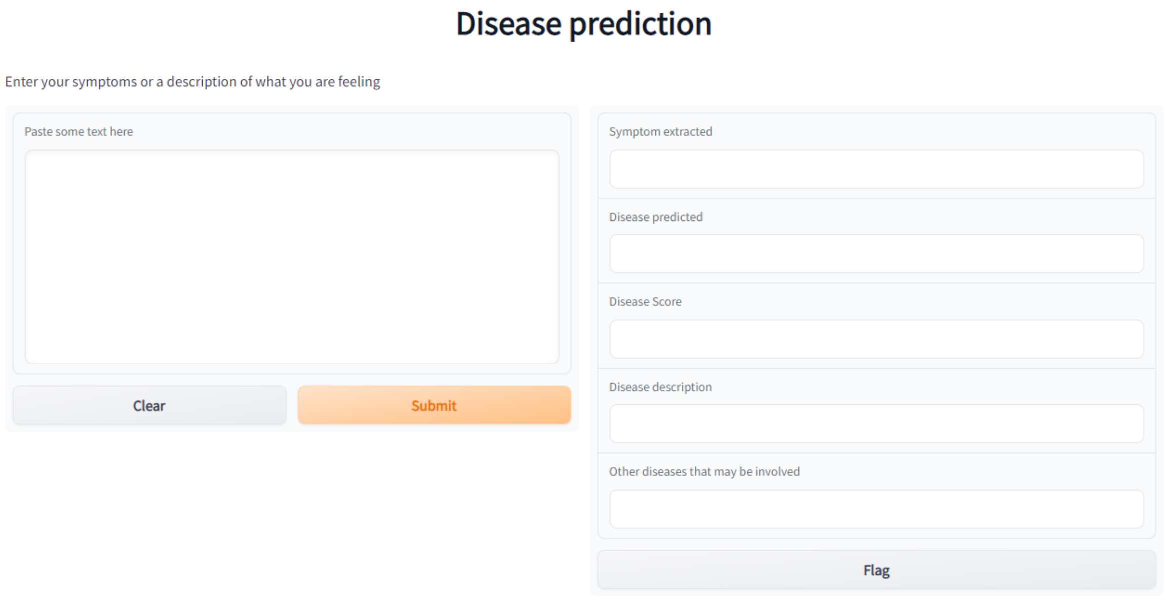
Source: Our own work

From all these analyses, the model retained for our system is the Logistic regression model built on the basis of the dataset containing the diseases and symptoms

1-4- System interface

The figure below shows the interface of the proposed system for the diagnosis of the patient's disease.

Figure 3: Web interface built for the test



Source: Our own work

The annotated field of (1) is reserved for the insertion of symptoms when the patient is able or the description of his condition.

The annotated field of (2), presents the symptoms extracted from the patient's description, which he/she can take care of adjusting when it does not match the symptoms, he/she expressed.

The annotated field in (3) contains the disease predicted by the model based on the symptoms that were retrieved.

The annotated field in (4) shows the probability associated with the predicted disease.

The field of (5), presents the description of the predicted disease based on the ICD 11

The annotated field in (6), on the other hand, provides information on other possible diseases to which the patient's symptoms could correspond.

1-5- Conclusion

The main goal of this chapter was about showing results of ML algorithms and comparing their performances so that we can pick up the best one and incorporate in our system.

Conclusion

Diseases have long been the greatest sources of death and disability in the world. Thus, their early diagnosis could allow their effective and cost-effective management. In this sense, ML algorithms are considered valuable for disease diagnosis. This study aims to implement a disease diagnosis system based on symptom extraction. For this purpose, several machine learning algorithms were trained on the dataset composed of diseases and symptoms and associated and the one composed of the description of diseases in order to determine the one that performed the best. These included the Decision Tree and Logistic Regression with 95% accuracy. The limitations of our study are that there is no data to train a NER for symptom detection. Future studies will focus on exploiting deep learning algorithms and more suitable data to improve the performance of the system.

Bibliographic references

[1] D. Chambers *et al.*, ‘Digital and online symptom checkers and health assessment/triage services for urgent health problems: systematic review’, *BMJ Open*, vol. 9, no. 8, p. e027743, Aug. 2019, doi: 10.1136/bmjopen-2018-027743.

[2] S. Hoermann, K. L. McCabe, D. N. Milne, and R. A. Calvo, ‘Application of Synchronous Text-Based Dialogue Systems in Mental Health Interventions: Systematic Review’, *Journal of Medical Internet Research*, vol. 19, no. 8, p. e7023, Aug. 2017, doi: 10.2196/jmir.7023.

[3] A. Greene, C. C. Greene, and C. Greene, ‘Artificial intelligence, chatbots, and the future of medicine’, *The Lancet Oncology*, vol. 20, no. 4, pp. 481–482, Apr. 2019, doi: 10.1016/S1470-2045(19)30142-1.

[4] A. KELEŞ, ‘Expert Doctor Verdis: Integrated medical expert system’, *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 22, no. 4, pp. 1032–1043, Jan. 2014, doi: 10.3906/elk-1210-64.

[5] ‘Automatic prediction of coronary artery disease from clinical narratives - ScienceDirect’. https://www.sciencedirect.com/science/article/pii/S1532046417301466 (accessed Jul. 23, 2022).

[6] H. Q. Yu, ‘Experimental Disease Prediction Research on Combining Natural Language Processing and Machine Learning’, in *2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*, Oct. 2019, pp. 145–150. doi: 10.1109/ICCSNT47585.2019.8962507.

[7] ‘Text Messaging-Based Medical Diagnosis Using Natural Language Processing and Fuzzy Logic’. https://www.hindawi.com/journals/jhe/2020/8839524/ (accessed Jul. 23, 2022).

[8] X. Luo, P. Gandhi, S. Storey, and K. Huang, ‘A Deep Language Model for Symptom Extraction From Clinical Text and its Application to Extract COVID-19 Symptoms From Social Media’, *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 4, pp. 1737–1748, Apr. 2022, doi: 10.1109/JBHI.2021.3123192.

[9] C. Saligram, D. B. K, G. K. S, J. A. Kumar, and D. J. R. Munavalli, ‘Symptoms Extraction from a Voice Input using Natural Language Processing’, *International Journal of Engineering Research & Technology*, vol. 9, no. 4, May 2020, doi: 10.17577/IJERTV9IS040645.

[10] F. B. Putra *et al.*, ‘Identification of Symptoms Based on Natural Language Processing (NLP) for Disease Diagnosis Based on International Classification of Diseases and Related Health Problems (ICD-11)’, *2019 International Electronics Symposium (IES)*, 2019, doi: 10.1109/ELECSYM.2019.8901644.

[11] S. Doan *et al.*, ‘Building a Natural Language Processing Tool to Identify Patients With High Clinical Suspicion for Kawasaki Disease from Emergency Department Notes’, *Academic Emergency Medicine*, vol. 23, no. 5, pp. 628–636, 2016, doi: 10.1111/acem.12925.

[12] M. E. Peters *et al.*, ‘Deep Contextualized Word Representations’, 2018. doi: 10.18653/v1/N18-1202.

[13] J. J. Webster and C. Kit, ‘Tokenization as the Initial Phase in NLP’, presented at the COLING 1992, 1992. Accessed: Jul. 23, 2022. [Online]. Available: https://aclanthology.org/C92-4173

[14] M. Fatima and M. Pasha, ‘Survey of Machine Learning Algorithms for Disease Diagnostic’, *Journal of Intelligent Learning Systems and Applications*, vol. 9, no. 1, Art. no. 1, Jan. 2017, doi: 10.4236/jilsa.2017.91001.

[15] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, ‘Can machine-learning improve cardiovascular risk prediction using routine clinical data?’, *PLOS ONE*, vol. 12, no. 4, p. e0174944, Apr. 2017, doi: 10.1371/journal.pone.0174944.

[16] Z. Mushtaq, M. F. Ramzan, S. Ali, S. Baseer, A. Samad, and M. Husnain, ‘Voting Classification-Based Diabetes Mellitus Prediction Using Hypertuned Machine-Learning Techniques’, *Mobile Information Systems*, vol. 2022, p. e6521532, Mar. 2022, doi: 10.1155/2022/6521532.