IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 13, 2020, accepted May 28, 2020, date of publication June 4, 2020, date of current version June 17, 2020.

*Digital Object Identifier 10.1109/ACCESS.2020.3000075*

# Application of Machine Learning and Word Embeddings in the Classification of Cancer Diagnosis Using Patient Anamnesis

**ANDRÉS ALEJANDRO RAMOS MAGNA[1], HÉCTOR ALLENDE-CID[1], CARLA TARAMASCO[2], CARLOS BECERRA[2], AND ROSA L. FIGUEROA[3], (Associate Member, IEEE)**

[1]Escuela de Ingeniería en Informática, Pontificia Universidad Católica de Valparaíso, Valparaíso 2374631, Chile
[2]Escuela de Ingeniería Civil en Informática, Universidad de Valparaíso, Valparaíso 2362905, Chile
[3]Departamento de Ingeniería Eléctrica, Universidad de Concepción, Concepción 4070409, Chile

Corresponding author: Andrés Alejandro Ramos Magna (andres.ramos@uv.cl)

The work of Héctor Allende-Cid was supported by the Pontificia Universidad Católica de Valparaíso project under Grant 039.457/2020. The work of Carla Taramasco was supported by Plataforma Tecnológica Nacional para el Registro y Seguimiento del Cáncer en Chile under Grant CORFO 18BPE-93827, and in part by the Centro Nacional en Sistemas de Información en Salud (CENS) under Grant CORFO 16CTTS-66390.

**ABSTRACT** Currently, one of the main challenges for information systems in healthcare is focused on support for health professionals regarding disease classifications. This work presents an innovative method for a recommendation system for the diagnosis of breast cancer using patient medical histories. In this proposal, techniques of natural language processing (NLP) were implemented on real datasets: one comprised 160, 560 medical histories of anonymous patients from a hospital in Chile for the following categories: breast cancer, cysts and nodules, other cancer, breast cancer surgeries and other diagnoses; and the other dataset was obtained from the MIMIC III dataset. With the application of word-embedding techniques, such as word2vec's skip-gram and BERT, and machine learning techniques, a recommendation system as a tool to support the physician's decision-making was implemented. The obtained results demonstrate that using word embeddings can define a good-quality recommendation system. The results of 20 experiments with 5-fold cross-validation for anamnesis written in Spanish yielded an F1 of **0.980** $\pm$ 0.0014 on the classification of 'cancer' versus 'not cancer' and **0.986** $\pm$ 0.0014 for 'breast cancer' versus 'other cancer'. Similar results were obtained with the MIMIC III dataset.

**INDEX TERMS** Natural language processing (NLP), machine learning, deep learning, recommendation system, anamnesis.

## I. INTRODUCTION

Currently, one of the greatest challenges for information systems in healthcare is focused on helping clinicians in disease classification through the proposal of diagnoses through anamnesis. Different clinical registration systems are used as disease classifiers, the most common being the International Classification of Diseases, Version 10[1] (IDC-10), SNOMED CT[2] and International Classification of Primary Care[3] (ICPC-2).

However, the codification process is not trivial, and these classifications do not adequately represent the needs expressed by clinicians, given that, today, IDC-10 has more than 69, 000 types of diagnoses and approximately 72, 000 procedures. This large number of classes poses difficult choices in medical systems, leading to omitted or misguided diagnoses.

For instance, in Chile, according to the statistics maintained by the Ministry of Health (MINSAL) in 2017, approximately 50% of diagnoses admitted to emergency care were classified as "other" [1]. This problem may occur as a result of heavy workloads or because clinicians cannot find a satisfactory diagnosis within the internationally accepted ICD-10 classification adopted by the MINSAL.

To help with this task, two lines of development have been used [2]. The first are those methods based on medical

The associate editor coordinating the review of this manuscript and approving it for publication was Tossapon Boongoen.

[1]https://icd.who.int/browse10/2019/en
[2]https://www.snomed.org/snomed-ct
[3]http://www.ph3c.org/4daction/w3_CatVisu/en/icpc.html?wCatIDAdmin=1106

language processing, for example, approaches that employ clinical ontologies in the healthcare area to identify medical concepts in clinical reports [3], [4] or those that explore the semantic similarities between the diagnoses and names given in the ICD-10. The second line of development uses supervised machine learning (ML) to analyze word frequency and to compare the results in ICD-10 [5].

In the healthcare area, natural language processing (NLP) has a great number of potential uses [6], [7], among which multiple examples of applications through the use of ML can be found [8]. Within the processes of question and response on the web, there are instances of [9] opinion and experience analyses concerning medical treatment or drugs, studies of clinical trials, relationships between symptoms versus lifestyle and the efficacy of the treatment, positive and negative effects of the treatment, and information about the patient's health and psychological state [10].

In this context, NLP is used to investigate and implement computational mechanisms for communication between people and computers through the use of natural language, applying techniques corresponding to models of word, phrase or document representation. Documents or sentences are represented using two main methods: the first is based on word frequency, such as TF-IDF or TF-RFL [11], and the second involves vectorization models, also referred to as word embeddings [12]. The latter method corresponds to a set of techniques where words or phrases are linked to vectors of real numbers.

This proposal aims to represent the patient history or anamnesis using word-embedding representation models, such as skip-gram's word2vec and BERT, to automatically classify them into their corresponding disease, while contrasting them with classic text representations, i.e., TF-IDF. To process a diagnostic corpus and to generate the representation vectors, we processed a corpus of anamnesis and extracted one vector from every clinical history, whereby we applied supervised machine learning techniques to obtain a medical diagnosis classifier. For this approach, we used medical histories to implement a recommendation system for breast cancer and to help define a model for other diseases.

The present work is structured as follows. In Section 2, we briefly address studies related to medical diagnosis classification. In Section 3, we present the proposed method, including the preparation of the data, the vectorization of corpora and the machine learning and deep learning models employed for the classification process. Section 4 describes the results obtained in the classification process, and in Sections 5 and 6, we present a brief discussion regarding the results and conclusions.

## II. RELATED WORK

The anamnesis, or medical history of a patient [13], contains a description of the doctor's interview and analysis of the patient based on their symptoms and medical evaluation. The anamnesis serves to derive a hypothesis by attending physician before a possible pathology or disease is confirmed. Depending on the extent of the clinical domain, the number of hypotheses can vary from a few to thousands, which makes it difficult to diagnose diseases that require early attention [14]. Because much of the analysis that the physician performs is in the record of this medical history, it presents an opportunity for review. With the help of NLP techniques and ML, a system of diagnosis recommendations can be generated as a tool to support the decision-making of the physician.

The use of coding systems, such as ICD-10,[4] provides a tool to support the registration and decision-making of medical diagnoses. However, although the coding systems have significant advantages, as shown in [15] and [16], there are also analyses that address issues that doctors have encountered [17], [18] due to the large number of codes contained within these classification systems; these cases show errors in diagnosis or omission of up to 70%. Due to this, the necessity to help physicians with recommendation systems for diagnosis and medical classification arises. Therefore, the use of ML has been proposed as a support tool in the classification of diagnoses [19].

Most of the classification techniques that have implemented ML in the medical diagnostic process have been centered on the use of quantitative data applied to medical exams and clinical samples [20], [21].

In most of these approaches, algorithms such as naive Bayes [22], artificial neural networks (ANN) [23], CART and C4.5 decision trees [24] are used. The quantitative data that are used to model the clinical histories have been mainly focused on the clinical characteristics of anatomy, type of disease, medical consultations, disorders, procedures, chemistry, and drugs [25].

In the work of [26], we found the first approximation to the use of the history and word structures. However, the approximation, based on the search terms, produces a result that can be considered optimistic.

However, the use of terms is complex in contexts where the texts are written mainly as terms or acronyms, which makes it difficult to find in data dictionaries or data libraries, where in many cases they can be confused with stop words.

In our work, we show that the method for improving the analysis and classification of the diagnosis using the clinical history as input, over natural language processing, is using word embedding and the support of modern methods such as deep learning.

The main problems present in systems based on ML for NLP are centered on the necessity of generating structured data for the specific application. To this end, NLP provides tools that allow the representation of these registers, although most tools require the prior implementation of complex procedures of morphosyntactic analysis or an automatic definition of the sentence to be written [27], [28].

In other studies, NLP has been used to model the corpus to extract characteristics that define clinical experiments. This approximation has been described as a hybrid according to

[4] http://cie10.org/index.html

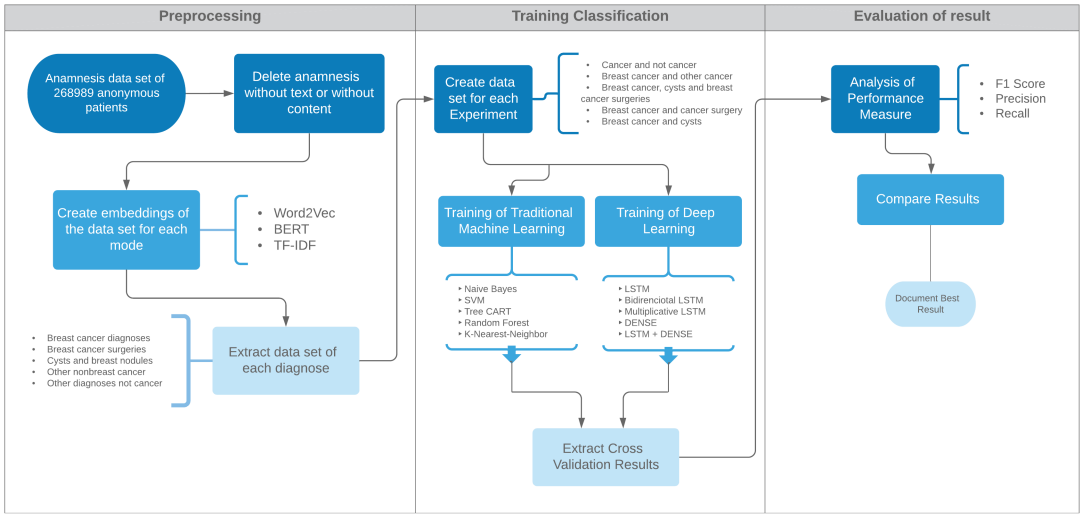


**FIGURE 1.** The process used in work, for the application of automatic learning in the classification of anamnesis and its diagnosis.

the authors of [29], employing NLP to extract six types of features, such as a characteristic of pairs, dependence relations, lexical relations, WordNet relations, predicated argument and discourse relations, performed according to standard criteria. In the classification process, an F1-score of 0.61 was obtained with a support vector machine (SVM). Despite the low value, a useful contribution was identified from these techniques in the text classification process. A similar technique was applied in the work of [30], in which they also used SVMs for the classification process. The task was to assist diagnosis, obtaining an F1 value of 0.94, proving that the proposed approach is a good approximation for the classification task.

Additionally, word embeddings have been extensively used in the document classification process, and in particular cases, they have shown good results in medical texts by creating word vectors with the word2vec CBOW algorithm and employing traditional ML and deep learning, obtaining F1 scores greater than 0.9 in the reported experiments [31]–[33].

## III. METHOD

To carry out this research, 2 datasets were used: the records of the clinical histories of 268, 989 patients (from which only 160, 560 have their corresponding label), written in Spanish, of the Dr. Guillermo Grant Benavente Regional Clinical Hospital in the City of Concepcion, Chile, which was authorized by the hospital ethics committee; and 46, 500 anamnesis from the well-known MIMIC III dataset, written in English.

To conduct the experiments, we applied the process presented in Figure 1. The process is based on the 7 steps of machine learning [34], comprised of data collection, data preparation, model choice, model training, model evaluation, parameter tuning, the prediction process and results evaluation.

This paper is an initial proposal for a new diagnostic recommendation system that uses clinical histories written by the medical practitioner, and it focuses on the classification of the anamnesis of patients with cancer, with breast cancer being the most common case. Currently, cancer is one of the major causes of death in the population. The number of records available in the dataset that are classified with this diagnosis and the importance of its evaluation is highlighted by a previous work that reports that breast cancer is among the first five causes of cancer death [35], primarily affecting women.

The dataset was separated into five categories, as presented in Table 1.

**TABLE 1.** Quantity of anamnesis by type of diagnosis.

| Dataset | Number of Clinical History |
|---|---|
| Breast cancer diagnoses | 4,396 |
| Breast cancer surgeries | 2,679 |
| Cysts and breast nodules | 4,061 |
| Other nonbreast cancer | 3,452 |
| Other diagnoses not cancer | 145,972 |

The proposed method was developed in multiple stages: first, the unstructured information of the patients was preprocessed, followed by vectorization of the corpus, representation of the clinical histories employing a representation vector, carrying out experiments with machine learning and finally reporting the results with several performance measures.

### A. PREPROCESSING OF UNSTRUCTURED INFORMATION

As mentioned above, the dataset in Spanish contains 268, 989 clinical registers of the Dr. Guillermo Grant Benavente Regional Clinical Hospital in the City of Concepcion, Chile and was anonymized due to the requirements of the ethics committee.

The registers contain data on gender, clinical histories, habits, and medical diagnosis, out of which only 60% such data include written clinical histories, while the rest of the cases are empty registers or symbols identified as scripts or any other character. In addition, within the cases with clinical histories, the gender distribution of the diagnoses was 61.437% female cases, 38.559% male cases and 0.004% cases identified as 'A'.

Regarding the diagnoses classified by the clinical histories, the most significant number of cases registered occur in registries with the diagnosis of breast cancer, followed by cholelithiasis. In addition, as shown in Figure 2, cancer cases are among the five most frequent in the sample; thus, they were selected as the objective of this work.

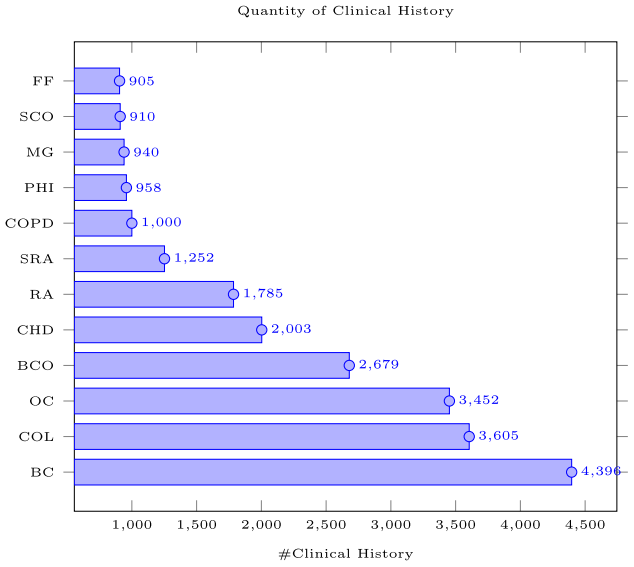


**FIGURE 2.** The number of cases with a complete clinical history by a registered diagnosis corresponds to the cases with the largest number of records and is defined by the following: BC:Breast cancer, COL:Colelithiasis, OC:Other Cancer,BCO:Breast Cancer Operated, CHD:Congenital Hip Dysplasia, RA:Rheumatoid arthritis, SRA:Seropositive Rheumatoid Arthritis, COPD: Chronic obstructive pulmonary disease, PHI:Phimosis, MG:Multinodular goiter, SCO:Scoliosis, FF:Flatfoot.

For the present work, the representations of other types of cancer were grouped as a single class and compared with the target class corresponding to breast cancer. It is worth noting that the number of diagnoses registered in other types of cancers was lower overall than the number existing in the target class. An example of this can be seen in Figure 3, where the number of cancer records per type that are different from breast cancer does not exceed 1, 209, as is the case with gastric cancer, with 1, 209 records versus 4, 396 cases of breast cancer.

In an analysis of the morphosyntactic structure of the clinical histories, the high complexity of linguistic structure must be emphasized due to the way doctors write stories with a significant amount of description with very low-value information. For this reason, the first task was to discard all

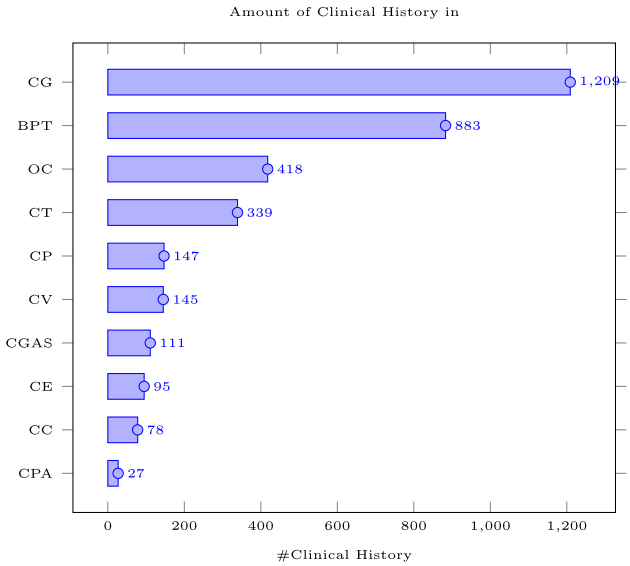


**FIGURE 3.** The number of cases with anamnesis with a diagnosis of other types of cancer other than breast cancer is defined as follows: BPT: papillary thyroid cancer, CP: pulmonary cancer, CP: gastric cancer, CT: thyroid cancer, CGAS: gastroesophageal cancer, CE: esophageal cancer, CV: vesicula's cancer, CC: colon cancer, CPA: pancreas cancer, OC: other cancer.

records with fewer than ten characters since it was not feasible to identify content in the description of such clinical histories.

In addition to the above, the analysis shows great complexity in the writing methodology of the descriptions registered by the doctors. In particular, there is significant use of acronyms, and in many cases, they represent the entire record of the clinical histories.

These cases make any analysis difficult by means of standard disease dictionaries, such as those conducted in previous work [36]. An example of such complexity is indicated in a clinical history shown in Table 2, where the original record is compared with acronyms and their meaning in the text without acronyms.

**TABLE 2.** Comparison of registered anamnesis versus its meaning in natural language.

| **It says in register with acronyms** |
|---|
| MP VAX IZQUIERDA RT QT TX COMPLETÓ 5 AÑOS |
| **Meaning Text Without Acronyms** |
| Mastectomía parcial se realizó el vaciamiento axilar izquierdo, se plica radio terapia, quimioterapia y no se puede evaluar el tumor primario, completó 5 años |

As a consequence of the abbreviation problems, it is necessary to use techniques that do not depend on the dictionaries of the morphosyntactic structures for analysis; therefore, this task is based on a corpus to which word-embedding techniques were applied, which present multiple benefits for the description of clinical information [37].

As explained before, we also used a dataset written in English: the MIMIC III (Multiparameter Intelligent

Monitoring in Intensive Care) database [38], which corresponds to an update to the widely-used MIMIC-II [39]. MIMIC III is an openly available dataset developed by the MIT Lab for Computational Physiology, comprising deidentified health data associated with ∼60,000 intensive care unit admissions [40].

The data correspond to the coronary care unit (CCU), cardiac surgery recovery unit (CSRU), medical intensive care unit (MICU), surgical intensive care unit (SICU) and trauma surgical (TSICU). They have been collected since 2001 and correspond to records of medical care, surgery, exams and others [38]. The database contains data associated with 53, 423 distinct hospital admissions for adult patients (aged 16 years or above) admitted to critical care units between 2001 and 2012 from Beth Israel Deaconess Medical Center of Boston.

The data cover 38,597 distinct adult patients and 49,785 hospital admissions. The median age of adult patients is 65.8 years, from which 55.9% patients are male, and a mean of 4,579 charted observations with (''chartevents'') 380 laboratory measurements (''labevents'') are available for each hospital admission.

For this research, we used a patient subset from cancer diagnosis. To do this, we extracted patients who were registered with a diagnosis of malignant neoplasm of the nipple and areola of the female breast or malignant neoplasm of the central portion of the female breast, among others.

To carry out the tests, events from different categories were selected, from which we found mainly for the case of breast cancer, 265 physician category, 1,003 radiology, 247 nursing notes and 178 discharge summaries, among others. These notes correspond to 98% of female patients, and only two cases were male patients.

Each category was selected according to the length of note description, which gave us a richer text, therefore allowing us to obtain better results in the classification process. The mean length of all description notes in the breast cancer dataset was 1, 085 words, a maximum length of 5, 541 words, and a minimal length of 23 words, resulting in a corpus of 126, 960 distinct, unique words.

The records contained data on gender, age, admission, and performed care. It also included the medical records of each patient, along with the associated diagnoses in ICD9 coding, which allowed us to identify the cases required to make an analysis similar for the texts in Spanish. Figure 4 shows the top 11 diagnoses, in terms of their quantity, present in the dataset.

The database contains the description of more than 600, 000 medical diagnoses, from which it can be noted that 39.1% of the cases correspond to diseases of the circulatory system, 10.2% to trauma, 9.7% to diseases of the digestive system, 9% to lung diseases, 7% to infectious diseases and 6.8% to malignant or benign neoplasms. For this work, we extracted the following categories (see Table 3):

Unlike the data in Spanish, the medical records in English were described with a more formal language, where acronyms are avoided. With this level of detail in writing, the medical records exceeded 5, 000 words in some cases, which makes the word relationships richer.

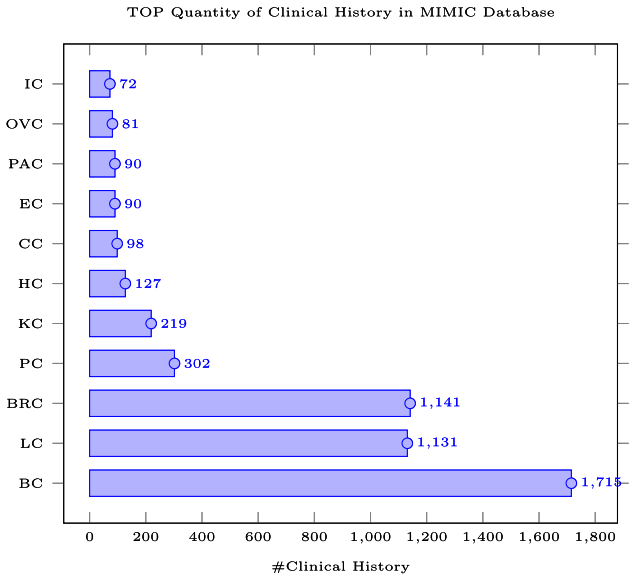


**FIGURE 4.** Number of cases with anamnesis with a diagnosis of cancer in MIMIC Database: BC:breast cancer, LC:lavier cancer, BRC:bronchus or lung cancer, PC:prostate cancer of prostate, KC:kidney, except pelvis cancer, HC:head of pancreas cancer, CC:cardia cancer, EC:esophagus cancer, PAC:pancreas cancer, OVC:ovary cancer, IC:Intrahepatic bile ducts cancer.

**TABLE 3.** Quantity of anamnesis by type of diagnosis in MIMIC III Dataset.

| Dataset | Quantity of Clinical History |
|---|---|
| Breast cancer diagnoses | 1,715 |
| Cysts and breast nodules | 3,436 |
| Other nonbreast cancer | 26,918 |
| Other diagnoses not cancer | 23,866 |

### B. CORPUS VECTORIZATION

For this research, one of the main activities of natural language processing was focused on the classification of documents. Multiple techniques are applied to these tasks. Modern methods are mainly aimed at automatic classification with the use of ML, using supervised, unsupervised and semisupervised techniques [41].

To perform the classification task, it was necessary to transform the unstructured data into a vectorized representation. For this, it is feasible to use representations based on unsupervised techniques, which allow for the vectorization of words, such as word2vec,[5] Glove,[6] FastTEXT,[7] BERT[8] or similar approaches. For this research, Google's word2vec and BERT algorithms were applied, and we compared the obtained

[5]https://code.google.com/archive/p/word2vec/
[6]https://nlp.stanford.edu/projects/glove/
[7]https://research.fb.com/fasttext/
[8]https://github.com/google-research/bert

results with a traditional technique called term frequency-inverse document frequency TF-IDF.

Word2Vec is a framework or group of related models that are used to produce word vectors. Word2Vec was proposed by Google and defines two main implementation algorithms for the definition of embedding: CBOW and Skip-gram [12].

To build the word vectors, windows of size 5 were used, based on the recommended size by the authors of the method, as well as 300 characteristics per word, a size that is common in this type of work and supported in different systems of vectorization [42].

### C. DEEP BIDIRECTIONAL TRANSFORMERS WITH BERT

BERT is a language representation model that stands for bidirectional encoder representations from transformers [43]. BERT is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both the left and right context in all layers. Consequently, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without the need for substantial task-specific architecture modifications. It has been recently applied to clinical text data in English [44]. The authors address the need for exploring and releasing BERT models for clinical text. They create embeddings for generic clinical text and for discharge summaries specifically. They demonstrate that using a domain-specific model yields performance improvement on three common clinical NLP tasks compared to nonspecific embeddings. In [43], the authors show that this representation outperforms other word-embedding techniques in several applications.

### D. DATASET CONSTRUCTION

Once a vocabulary is built with its respective representation vector using word2vec and BERT, for each clinical history description, a representation vector is created using the arithmetic mean of the vectors of every word in the clinical history as follows:

$$V_I = \frac{\sum_{j=1}^{n} v'_{wj}}{n} \tag{1}$$

where $V_I$ corresponds to the vector of the instance, $j$ corresponds to each word found in the vocabulary $W$ in the case represented by the output vector $v'_w$, and $n$ represents the number of words that the instance contains. For each word in the clinical history, the average of the word vectors that compose it is extracted. Every averaged vector is a new representative vector of the clinical history.

Through this process, the final dataset that will be used for the classification algorithm training process is extracted. The datasets for each of the experiments performed are described in Table 4 and 5.

### E. TRADITIONAL TECHNIQUES TF-IDF

To compare the results obtained with the vectorization proposed by word2vec and BERT, we also use a traditional technique that provides an alternative method for representing the clinical histories by employing term frequency-inverse document frequency (TF-IDF) representation; for this, we used Boolean frequencies:

$$tf(t, d) = 1 \text{if } t \text{ occurs in } d, \text{ and } 0 \text{ otherwise;} \tag{2}$$

In conjunction with the above, we applied the inverse document frequency, which is a measure to represent whether the term is typical in the collection of documents. This measure is obtained as a result of the logarithm of the division of the total number of records by the number of documents that contain the term. It is shown in the following:

$$idf(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|} \tag{3}$$

where $|D|$ is the cardinality of D, or the number of documents in the collection. $|\{d \in D : t \in d\}|$ is the number of documents where the term t appears. If the term is not in the collection, a division by zero will occur. Therefore, it is common to adjust this formula to $1 + |\{d \in D : t \in d\}|$.

$$tfidf(t, d, D) = \text{tf}(t, d) \times \text{idf}(t, D) \tag{4}$$

However, a significant problem that occurs during the dataset processing of the clinical histories should be addressed, namely, the necessity in increasing the capacity of the server to allow for processing large vectors for each clinical history.

Because each clinical history is represented in principle by vectors with a size of 130, 152 characteristics, it was necessary to use a reduction dimensionality technique that allows for obtaining vectors with an appropriate size. Thus, the principal component analysis (PCA) algorithm was employed to extract the same-sized vectors as those of the previous experiment with the skip-gram algorithm, with vectors of size 300.

The previous definition was necessary due to the considerable computational cost of processing the vectors, those that have a size of 130, 152 and for this purpose, we used one of the most common machine learning tools that allow a reduction in dimensionality through the main components or characteristics of the vector, which is known as PCA [45].

**TABLE 4.** List of datasets used in each of the experiences (Spanish anamnesis).

| Data Set | Size |
|---|---|
| Cancer and not cancer | 21,054 |
| Breast cancer and other cancer | 6,904 |
| Breast cancer, cysts and breast cancer surgeries | 8,037 |
| Breast cancer and cancer surgery | 5,358 |
| Breast cancer and cysts | 8,122 |

**TABLE 5.** List of datasets used in each of the experiences (MIMIC III anamnesis).

| Data Set | Size |
|---|---|
| Cancer and not cancer | 54,604 |
| Breast cancer and other cancer | 3,430 |
| Breast cancer and cysts | 3,430 |

Together with the aforementioned dimensionality reduction, and to make a better comparison based on the same vector dimension, it was decided to reduce the dimension to 300 characteristics as well as the other embedding models.

### F. CLASSIFICATION OF CLINICAL HISTORY WITH ML AND DEEP LEARNING

To carry out the classification process of the clinical histories, our work implemented different ML algorithms. ML corresponds to a branch of artificial intelligence and a subfield of the study of computer science [46] that has seen rapid growth in recent years.

The ML techniques are divided into three broad groups of algorithms that are defined according to the data and task (supervised, unsupervised and reinforcement learning). In this case, we used supervised techniques that correspond to classification algorithms to identify the class of each of the instances present in the dataset.

In our experiments, the classes used are described in Figure 2, where each of the classes and the number of instances contained in them are depicted. To carry out the classification training, we employed a process that allowed us to obtain corroborated results through cross-validation, in conjunction with several iterations of the same experiments. For each iteration, the data are randomly mixed to avoid biased results. In general, the steps of the applied experiment are shown in Algorithm 1.

**Algorithm 1** Training Machine Learning Algorithm

**Require:** Balanced data set by class of each word embedding.

```
for iteration = 0 to 20 do
    Shuffle instances randomly.
    for ML = 1 to alltraining do
        Train the learning machine using cross-validation with cv = 5.
        Validate result with performance measure using test dataset.
    end for
end for
```

Because the dataset has defined class labels, it allows us to use supervised techniques; therefore, we defined the use of 6 different ML classification models.

#### 1) TRADITIONAL MACHINE LEARNING

To implement each of the models, we used the scikit-learn [47] library, available for the Python language. The algorithms and their parameters are presented as follows:

- **Support Vector Machines:** A support vector machine is a supervised machine learning algorithm that permits classifying the data by a separating hyperplane. Specifically, given a labeled dataset, the algorithm obtains an optimal separating hyperplane with new example categories. The main parameters for this method are the kernel, regularization, gamma, and margin.
  There are several SVM implementation algorithms, and Python provides three implementations of this algorithm; for the case in our experiment, we used C-support vector classification (SVC),[9] which corresponds to an implementation based on libsvm [48]. Although the cost of execution is very high, the results in different tasks have shown good results with the use of these learning machines. We used SVC with two different kernels: linear and radial basis functions (RBFs).
- **Decision Trees:** A decision tree classifier is a simple and widely-used supervised classification technique that poses a series of questions about the attributes of the test record. In general, the decision tree is constructed from the attributes of the dataset. Finding the optimal tree is computationally infeasible because of the exponential size of the search space. As a solution to this problem, several algorithms are proposed, such as Hunt's algorithm, ID3, C4.5, CART, and SPRINT. In this work, we used the CART version.
- **Naive Bayes:** Naive Bayes is a family of probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. Even with this characteristic, it is a robust algorithm used for real-time prediction, text classification/spam filtering, recommendation systems, etc.
  Scikit-learn has implemented three naive Bayes models in its libraries: Gaussian, multinomial, and Bernoulli. For the purpose of this research, Gaussian naive Bayes implementation was used, which does not require a significant number of parameters for the experiments; therefore, we used the default parameters.
- **K-nearest Neighbors:** K-nearest neighbors is a non-parametric method used for classification and regression. In both cases, the input consists of the $k$ closest training examples in the feature space. In the case of classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the most common class among its K-nearest neighbors. In this work, we compared the result of using 1 to 20 neighbors employing the weighted method, and we selected the Euclidean distance for every vector instance.
- **Random Forest:** Random forest is an ensemble learning method for classification and regression that operates by constructing several decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. We used the Gini index as a quality estimator and searched from 1 to 20 trees to obtain the best results.

#### 2) DEEP LEARNING

In this work, we also applied different types of neural network architectures. In these architectures, the neurons can

[9] http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm. SVC

**TABLE 6.** Architectures used for the Middle Layer.

| Name | Description Layer | size |
|---|---|---|
| LSTM | One layer of LSTM | 300 Units |
| BILSTM | One layer of Bidirectional LSTM | 600 Units (300 per direction) |
| LSTM + Dense | One layer of LSTM and one dense fully connected perceptron layer | 300 + 300 |
| 3LSTM | Three layers of LSTMs stacked | 300 + 300 + 300 |
| MultiLSTM | One layer of Multiplicative LSTM | 300 Unit |
| DENSE | Two layer of DENSE | 300 + 300 |

be stacked in different layers, where the inputs of each layer are the outputs of the other. The structure used for the experiments in this work consists of a three-layer split, which includes the input layer, middle layer, and output layer [49].

For the input layer, we used the embedding representation created with skip-gram word2vec, BERT, and the vectorized representation created with TF-IDF in all experiments. In this case, just as in traditional machine learning, the input vector corresponds to the average of the sentence of the anamnesis or clinical histories.

For the middle layer, multiple architectural units were used: dense units, LSTMs, bidirectional LSTMs, and multiplicative LSTM units. The use and number of units are described in Table 6. Additionally, a dropout [50] of 50% was applied to prevent overfitting.

- **Long Short-term Memory (LSTM):** LSTM is a variation in recurrent neural networks that was created as a solution to the problem of short-term memory. These networks have internal mechanisms called gates that can regulate the flow of information, which helps preserve the error that can be backpropagated through time and layers.
  The LSTM units were proposed by [51]; this unit is characterized by a loop that allows the transfer of information between neurons of the layer, which ensures that the information persists.
  Although it can be thought that models based on neural networks such as the case of LSTM present a very high cost compared to the work that they need to do, they have shown that excellent results are obtained when working in complex documents or corpora contexts in NLP analysis [52], [53]. The above incorporates into our experiments the evaluation of the results that may be obtained in this way.
- **Dense Neural Network:** As the name states, layers are fully connected (dense). Each neuron in a hidden layer receives input from all the neurons in the previous layer. The dense layer is known as a perceptron layer, corresponding to the most straightforward units in the neural networks [54]. Given an input ($x_i$) and a weight value ($w_i$), it produces an output ($y$) through the dot product between the inputs and weights, which is then passed to an activation function $f$. Formally, the perceptron is defined as:

$$y = f\left(\sum_{i=1}^{n} x_i w_i\right) \tag{5}$$

  This model is also known as the multilayer perceptron (MLP).

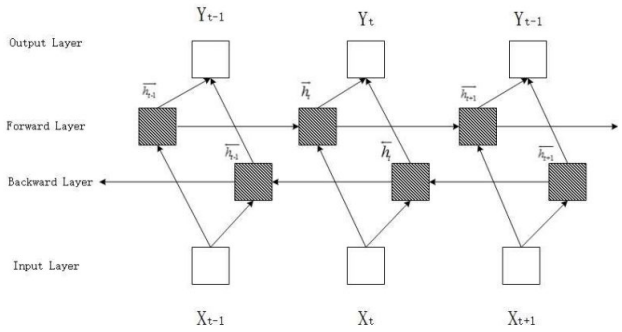


**FIGURE 5.** Bidirectional LSTMs units description.

- **Bidirectional LSTMs:** The bidirectional LSTM [55] is an adaptation of the LSTM units, which follows the behaviors of bidirectional recurrent neural networks (BRNN). The main idea is to use the components of the LSTM to train two separate units: one trains the sequence in a forward direction, and the other trains the sequence in the backward direction. Finally, these two networks connect to the same output layer via a concatenation of the hidden activation $h_t$ for each LSTM layer. Figure 5 shows the described behavior.
  As in the case of LSTM, bidirectional LSTM has been used with good results in NLP, especially in complex document structure analysis processes such as sentiment analysis on Twitter [56]. Along with this, it should be understood that there is a great completeness in the description that is made in the patient's medical history, and in many cases, the level of detail is low with a high use of acronyms that can further complicate the analysis of the texts for classification.
  The task of classification in texts has been presented with optimistic results in different contexts [57], [58], which makes it interesting and is expected to obtain similar results in the classification of the history with the diagnosis.
- **Multiplictive LSTM:** The multiplicative LSTM (mLSTM) is a hybrid architecture that combines the factorized transition of multiplicative recurrent neural networks (mRNNs) in hidden layers with the gating framework from LSTMs [59]. The architecture combines mRNN's structures, adding connections from the mRNN's intermediate state $m_t$ to each LSTM unit. This state is incorporated in each gate of the LSTM architecture, and the dimensionality of $m_t$ and $h_t$ is the same for all our experiments.

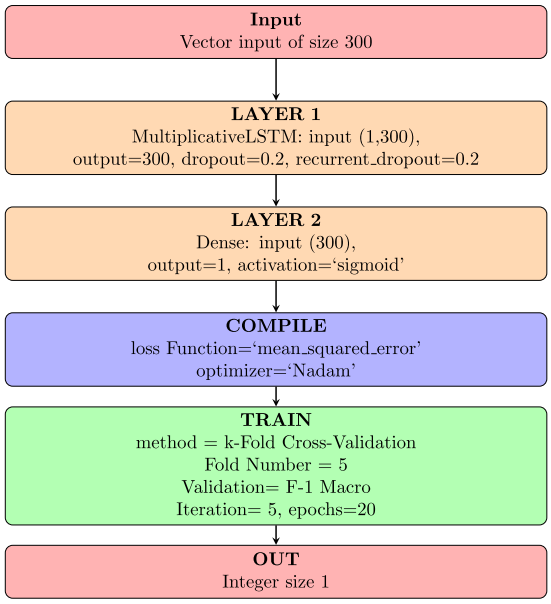


**FIGURE 6.** Multiplicative LSTM more dense layer.

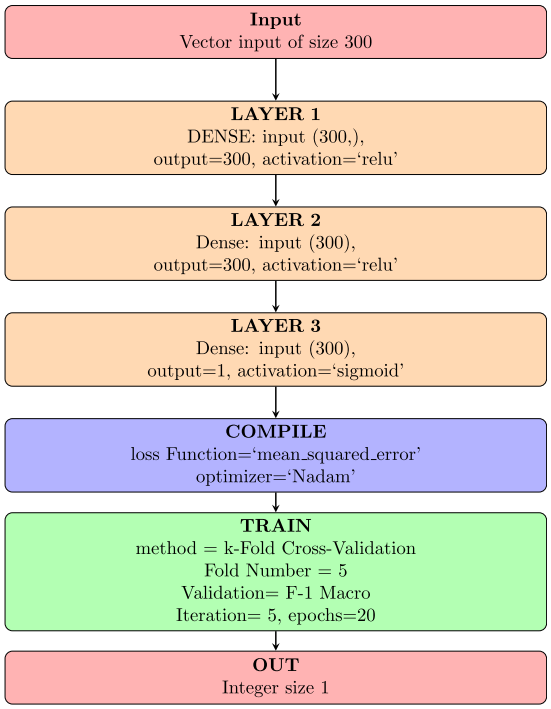


**FIGURE 7.** Three dense layer.

To implement each of the experiments with deep learning, we used the Keras library over TensorFlow [60], available for the Python language. For each experiment, 20 runs with each model were performed, and the results were validated by applying 5-fold cross-validation. As an activation function, we used sigmoid and hyperbolic tangent, and Nadam was used as the optimization algorithm.

#### 3) DEEP LEARNING USED ARCHITECTURE

Several deep learning techniques were applied in different experimental architectures. In each case, we used the same input, which was generated with the arithmetic mean of each word vector, omitting the stop words. Each vector had a size of 300 features and was generated with different word representation models, such as word2vec, BERT, and TF-IDF.

For the implementation of each architecture, we used the Keras library on TensorFlow 2.0 with Python 3 [61]. After multiple tests, the best result in each iteration was obtained by compiling the layer architecture using the objective function (loss) "mean squared error" and Nesterov Adam optimizer or "Nadam" optimizer [62]. Below are each of the implementations used in the home case:

- **MultiplicativeLSTM + Dense Layer:** For this implementation, an input layer of multiplicativeLSTM type and an output layer of dense type were used, where 300 characteristic vectors were received in the input layer and 300 size vectors were output. Each iteration used a 0.2 dropout for the linear transformation of the inputs and a recurrent dropout of 0.2 in the recurrent state.
  For the dense layer, an input of vectors of size 300 and output of size 1 was used. As an activation function, we used the sigmoid function, as shown in Figure 6.
- **Three Dense Layers:** In this implementation, we used three dense layers: the input vector, as in the previous case, had a size of 300, and the initial layer and hidden layer had the activation function ReLU. This function was selected by the obtained results over multiple tests. Similar to the previous case, the output layer corresponds to the dense type, the activation function is sigmoid, the input vector size was 300, and the output vector was 1; see Figure7.
- **LSTM Other Cases:** For other cases, the architectures are implemented with an initial LSTM layer or bidirectional LSTM. In each case, we used a standard configuration and an input vector size of 300 characteristics, and the same dimensions were configured for output.
  For all cases, the output layer corresponded to the dense type, and the activation function was configured in sigmoid or tanh. The input vector size was 300, and the output vector was 1; see Figure 8.

### G. PERFORMANCE MEASURES

To evaluate the classifier performance, we validated the results with several performance measures. We used the macrotest score obtained by the following performance measures:

- **Precision (P):** The precision is defined by the number of true positives over the number of true positives plus the number of false positives.

$$P = \frac{T_p}{T_p + F_p} \tag{6}$$

**TABLE 7.** Averages of F1 macro results after 20 iterations using Word2Vec skip-gram with cross-validation (Anamnesis dataset written in Spanish), SVM(L): support vector machine with linear kernel, SVM (R): support vector machine with RBF kernel, DT: decision tree, NB: naive Bayes, KNN: K-nearest neighbor, RF: random forest.

| Analysis | Cancer vs not cancer | Breast cancer vs other cancer | Breast cancer vs cysts vs cancer surgeries | Breast cancer vs breast cysts | Breast cancer vs breast cancer surgeries |
|---|---|---|---|---|---|
| SVM(L) | **0.918** ± 0.0004 | **0.930** ± 0.0008 | 0.736 ± 0.001 | **0.906** ± 0.001 | 0.707 ± 0.002 |
| SVM(R) | 0.820 ± 0.0003 | 0.702 ± 0.001 | 0.523 ± 0.0009 | 0.672 ± 0.001 | 0.610 ± 0.001 |
| DT | 0.816 ± 0.002 | 0.822 ± 0.003 | 0.652 ± 0.004 | 0.811 ± 0.005 | 0.660 ± 0.006 |
| NB | 0.866 ± 0.0004 | 0.844 ± 0.0008 | 0.656 ± 0.001 | 0.831 ± 0.001 | 0.657 ± 0.001 |
| KNN | 0.896 ± 0.0007 | 0.918 ± 0.002 | **0.757** ± 0.003 | 0.894 ± 0.003 | **0.740** ± 0,005 |
| RF | 0.892 ± 0.002 | 0.710 ± 0.0042 | 0.72 ± 0.042 | 0.870 ± 0.038 | 0.722 ± 0.030 |
| Multiplicative LSTM | 0.971 ± 0.021 | 0.972 ± 0.019 | 0.468 ± 0.040 | 0.961 ± 0.020 | 0.843 ± 0.079 |
| Dense | 0.976 ± 0.011 | 0.985 ± 0.008 | **0.509** ± 0.022 | **0.978** ± 0.010 | **0.931** ± 0.038 |
| 3-LSTM | 0.974 ± 0.019 | 0.977 ± 0.019 | 0.475 ± 0.043 | 0.967 ± 0.025 | 0.865 ± 0.080 |
| Bidirectional LSTM (sigmoid) | 0.974 ± 0.015 | 0.974 ± 0.015 | 0.491 ± 0.040 | 0.965 ± 0.017 | 0.884 ± 0.082 |
| Bidirectional LSTM (tanh) | **0.980** ± 0.014 | **0.986** ± 0.014 | 0.501 ± 0.041 | 0.975 ± 0.020 | 0.905 ± 0.082 |
| LSTM + Dense | 0.977 ± 0.007 | 0.984 ± 0.007 | 0.509 ± 0.025 | 0.974 ± 0.010 | 0.928 ± 0.046 |

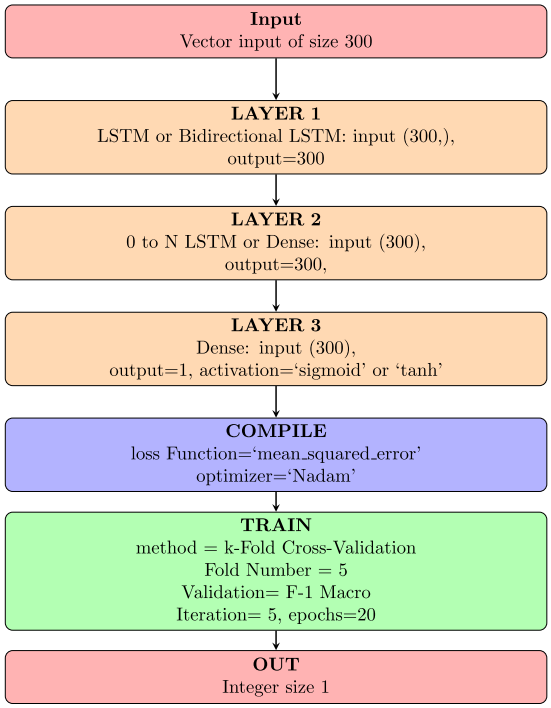


**FIGURE 8.** LSTM and other cases.

- **Recall (R):** The recall is defined by the number of true positives over the number of true positives plus the number of false negatives.

$$R = \frac{T_p}{T_p + F_n} \tag{7}$$

- **F1 score:** [63] It is defined as a weighted average between precision (P) and recall (R), described by the following equation:

$$F1 = \frac{2PR}{P + R} \tag{8}$$

where $T_p$ corresponds to true positive, $F_p$ corresponds to false positive and $F_n$ is false negative.

In summary, the results for the F1 score correspond to each of the 20 iterations using 5-fold cross-validation.

## IV. RESULTS

In Table 7, Table 8 and Table 9, we present the results obtained from 20 experiments with 5-fold cross-validation with each of the text representation techniques (word2vec, BERT and TF-IDF) in the Spanish Anamnesis dataset. First, we performed the task of balancing the classes and normalizing the vectors of the dataset. After that, we trained the ML models using 5-fold cross-validation. In each of the experiments, the hyperparameters were optimized using a grid search with 5-fold cross-validation.

When observing the results with the word2vec representation presented in Table 7, the best results of the classic ML models were obtained by the SVM with linear kernel and K-nearest neighbor. In the case of the deep learning models, the best results were obtained with the bidirectional LSTM with the *tanh* activation function and the dense network. Overall, the deep learning models outperformed the classic machine learning algorithms in four of the five tasks.

In the case of the BERT representation, the results can be seen in Table 8. Within the classic machine learning algorithms, SVMs (linear and RBF kernel) were the methods with better performance. In the case of deep learning models, in all cases, the bidirectional LSTM (sigmoid) was the model with the best performance in all the different problems. To train the vectors, we used the Spanish corpus. This was because the anamnesis was mainly written with acronyms. In the case of the English dataset, we used a pretrained version of the BERT model with Wikipedia since the clinical history was written more formally.

In Table 9, we see the results of the TF-IDF representation. In the case of the classic machine learning models, the best results were obtained with decision trees, K-nearest

**TABLE 8.** Averages of F1 macro results after 20 iterations using BERT with cross-validation (Anamnesis dataset written in Spanish), SVM(L): support vector machine with linear kernel, SVM (R): support vector machine with RBF kernel, DT: decision tree, NB: naive Bayes, KNN: K-nearest neighbor, RF: random forest.

| Analysis | Cancer vs not cancer | Breast cancer vs other cancer | Breast cancer vs cysts vs cancer surgeries | Breast cancer vs breast cysts | Breast cancer vs breast cancer surgeries |
|---|---|---|---|---|---|
| SVM(L) | 0.818 ± 0.0004 | **0.888** ± 0.0007 | 0.576 ± 0.001 | 0.822 ± 0.003 | 0.525 ± 0.005 |
| SVM(R) | **0.819** ± 0.0003 | 0.886 ± 0.003 | **0.577** ± 0.002 | **0.823** ± 0.001 | **0.533** ± 0.004 |
| DT | 0.747 ± 0.002 | 0.812 ± 0.005 | 0.498 ± 0.004 | 0.811 ± 0.005 | 0.514 ± 0.007 |
| NB | 0.802 ± 0.0004 | 0.854 ± 0.003 | 0.529 ± 0.002 | 0.722 ± 0.001 | 0.504 ± 0.001 |
| KNN | 0.807 ± 0.0005 | 0.874 ± 0.007 | 0.548 ± 0.003 | 0.791 ± 0.003 | 0.515 ± 0,005 |
| RF | 0.799 ± 0.003 | 0.863 ± 0.002 | 0.527 ± 0.025 | 0.780 ± 0.001 | 0.513 ± 0.007 |
| Multiplicative LSTM | 0.939 ± 0.033 | 0.973 ± 0.019 | 0.476 ± 0.003 | 0.910 ± 0.033 | 0.817 ± 0.095 |
| DENSE | 0.931 ± 0.030 | 0.962 ± 0.020 | 0.167 ± 0.000 | 0.874 ± 0.018 | 0.711 ± 0.069 |
| 3-LSTM | 0.913 ± 0.037 | 0.964 ± 0.026 | 0.472 ± 0.003 | 0.883 ± 0.029 | 0.759 ± 0.108 |
| Bidirectional LSTM (sigmoid) | **0.953** ± 0.019 | **0.974** ± 0.016 | **0.478** ± 0.004 | **0.915** ± 0.034 | **0.831** ± 0.096 |
| Bidirectional LSTM (tanh) | 0.323 ± 0.001 | 0.333 ± 0.007 | 0.166 ± 0.000 | 0.913 ± 0.038 | 0.333 ± 0.000 |
| LSTM + Dense | 0.915 ± 0.023 | 0.957 ± 0.015 | 0.467 ± 0.001 | 0.851 ± 0.009 | 0.607 ± 0.030 |

**TABLE 9.** Averages of F1 results after 20 iterations using TF-IDF with cross-validation (Anamnesis dataset written in Spanish), SVM(L): support vector machine with linear kernel, SVM (R): support vector machine with RBF kernel, DT: decision tree, NB: naive Bayes, KNN: K-nearest neighbor, RF: random forest.

| Analysis | Cancer vs not cancer | Breast cancer vs other cancer | Breast cancer vs cysts vs cancer surgeries | Breast cancer vs breast cysts | Breast cancer vs breast cancer surgeries |
|---|---|---|---|---|---|
| SVM(L) | 0.428 ± 0.0087 | 0.889 ± 0.0025 | 0.838 ± 0.001 | 0.894 ± 0.001 | 0.851 ± 0.001 |
| SVM(R) | 0.458 ± 0.0073 | 0.867 ± 0.0004 | 0.800 ± 0.0002 | 0.871 ± 0.0001 | 0.817 ± 0.0004 |
| DT | **0.470** ± 0.003 | 0.854 ± 0.009 | 0.507 ± 0.0005 | 0.869 ± 0.009 | 0.808 ± 0.015 |
| NB | 0.440 ± 0.006 | 0.881 ± 0.0013 | 0.823 ± 0.001 | 0.887 ± 0.001 | 0.847 ± 0.001 |
| KNN | 0.124 ± 0.0011 | **0.893** ± 0.010 | 0.848 ± 0.001 | 0.911 ± 0.0013 | 0.883 ± 0,003 |
| RF | 0.334 ± 0.001 | 0.742 ± 0.0075 | 0.305 ± 0.003 | **0.924** ± 0.006 | **0.902** ± 0.007 |
| Multiplicative LSTM | 0.955 ± 0.002 | 0.923 ± 0.005 | 0.487 ± 0.007 | 0.927 ± 0.005 | 0.903 ± 0.011 |
| DENSE | 0.404 ± 0.042 | 0.333 ± 0.000 | 0.242 ± 0.001 | 0.333 ± 0.000 | 0.333 ± 0.001 |
| 3-LSTM | **0.962** ± 0.006 | **0.972** ± 0.014 | **0.527** ± 0.014 | **0.972** ± 0.015 | **0.971** ± 0.022 |
| Bidirectional LSTM (sigmoid) | 0.968 ± 0.008 | 0.958 ± 0.018 | 0.511 ± 0.013 | 0.953 ± 0.017 | 0.947 ± 0.023 |
| Bidirectional LSTM (tanh) | 0.945 ± 0.001 | 0.924 ± 0.004 | 0.491 ± 0.006 | 0.915 ± 0.003 | 0.893 ± 0.002 |
| LSTM + Dense | 0.930 ± 0.001 | 0.886 ± 0.004 | 0.457 ± 0.002 | 0.903 ± 0.003 | 0.847 ± 0.005 |

neighbors and random forest, while with deep learning models, the best results in all cases were obtained with the 3-LSTM model.

Now, if we compare the three text representations with the traditional ML methods, as seen in Figure 9, we observe that in two cases, the representation with word2vec obtained the best results, and in the other three cases, the best results were obtained with TF-IDF. In Figure 10, we present the results of the deep learning models with the three text representations. For this model, in three cases, the word2vec representation obtained the best results, while in the two remaining cases, the best results were obtained with TF-IDF. In both sets of experiments, BERT was outperformed by word2vec and TF-IDF.

It is important to highlight the quality of the results, wherein the best case ('breast cancer' versus 'other cancer') has an average F1 result of 0.980 and a standard deviation of 0.0014. These initial results allow us to infer that these methods can be an excellent tool for supporting decision-making by physicians based on the use of the natural language with which diseases are described in the clinical history of the patient.

In contrast, in the cases in which the classes are not separated in their context, for example, in the 'breast cancer vs. cysts vs. cancer surgeries' dataset, the results have a slightly lower quality in comparison with the other tasks. It should be noted that when using a dataset with a shared context, as was the case of 'breast cancer vs. cysts vs. cancer surgeries', the best results were obtained with the TF-IDF representation, with an average F1 of 0.848 and standard deviation of 0.001. These results ed because the contexts were very similar, such as in clinical histories where the patient was diagnosed with 'breast cancer', cases that were concerned with surgeries, and those that were mainly observations or consultations for this cancer.

In Tables 10, 11 and 12, we present the results of the three text representations for the MIMIC III dataset. When validating the results obtained in the classification process

**TABLE 10.** Averages of F1 results after 20 iterations using Word2Vec with cross-validation (MIMIC II dataset), SVM(L): support vector machine with linear kernel, SVM (R): support vector machine with RBF kernel, DT: decision tree, NB: naive Bayes, KNN: K-nearest neighbor, RF: random forest.

| Analysis | Cancer vs not cancer | Breast cancer vs other cancer | Breast cancer vs breast cysts |
|---|---|---|---|
| SVM(L) | **0.869** $\pm$ 0.0001 | 0.734 $\pm$ 0.001 | 0.852 $\pm$ 0.001 |
| SVM(R) | 0.676 $\pm$ 0.005 | 0.670 $\pm$ 0.0005 | 0.601 $\pm$ 0.0003 |
| DT | 0.699 $\pm$ 0.0001 | 0.642 $\pm$ 0.008 | 0.740 $\pm$ 0.015 |
| NB | 0.664 $\pm$ 0.0003 | 0.645 $\pm$ 0.001 | 0.725 $\pm$ 0.002 |
| KNN | 0.789 $\pm$ 0.0011 | **0.851** $\pm$ 0.0015 | **0.883** $\pm$ 0,003 |
| RF | 0.802 $\pm$ 0.002 | 0.696 $\pm$ 0.005 | 0.809 $\pm$ 0.008 |
| Multiplicative LSTM | 0.922 $\pm$ 0.022 | 0.867 $\pm$ 0.040 | 0.927 $\pm$ 0.020 |
| DENSE | 0.961 $\pm$ 0.021 | **0.952** $\pm$ 0.033 | **0.984** $\pm$ 0.015 |
| 3-LSTM | 0.938 $\pm$ 0.025 | 0.891 $\pm$ 0.049 | 0.946 $\pm$ 0.030 |
| Bidirectional LSTM (sigmoid) | 0.968 $\pm$ 0.030 | 0.928 $\pm$ 0.051 | 0.961 $\pm$ 0.024 |
| Bidirectional LSTM (tanh) | **0.980** $\pm$ 0.020 | 0.954 $\pm$ 0.054 | 0.976 $\pm$ 0.031 |
| LSTM + Dense | 0.959 $\pm$ 0.017 | 0.947 $\pm$ 0.037 | 0.976 $\pm$ 0.019 |

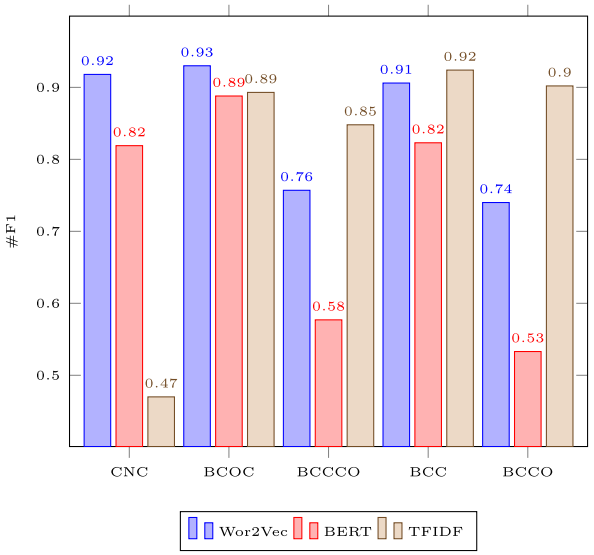


**FIGURE 9.** Comparative of the best result between Word2Vec, BERT, and TFIDF using traditional ML in each case for Spanish Dataset, CNC: cancer vs not cancer, BCOC: breast cancer vs other cancer, BCCCO: breast cancer vs cysts vs cancer surgeries, BCC: breast cancer vs breast cysts, BCCO: breast cancer vs breast cancer surgeries.

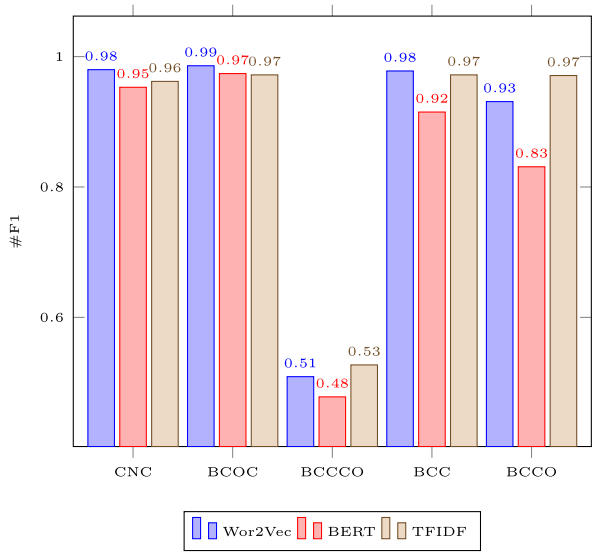


**FIGURE 10.** Comparative of the best result between Word2Vec, BERT, and TFIDF using Deep Learning in each case for Spanish Dataset, CNC: cancer vs not cancer, BCOC: breast cancer vs other cancer, BCCCO: breast cancer vs cysts vs cancer surgeries, BCC: breast cancer vs breast cysts, BCCO: breast cancer vs breast cancer surgeries.

for the data in English, we observe that the best results were obtained with the TF-IDF word vectors. In the case of traditional ML methods, the best results were obtained with the KNN method.

In the case of DL, although the best results were obtained with the TF-IDF representation, the results obtained with BERT and word2vec were very close. The worst result was obtained with the BERT representation in the case of 'breast cancer vs. other cancers' with 0.92 and 0.99 with TF-IF in the same case.

## V. DISCUSSION

The results presented in Table 7 demonstrate the usefulness of NLP and ML for the classification task, which could be convenient in recommendation systems in healthcare. This would allow the diagnosis process based on medical histories and help with an early diagnosis of complex diseases such as breast cancer and other diseases.

Moreover, it is important to highlight the contribution of word2vec, which considers specific contexts. In other application areas, the context can be harmful [64], but in the case in this study, due to the writing style that included a large number of acronyms, the context is very useful since this form of writing prevents applying other types of techniques for language analysis.

It is essential to emphasize that the use of traditional techniques such as TF-IDF provides a better result when the words used in the description of the clinical histories are very similar, such as the case of cancer diagnosed versus cancer surgeries, where the word representations, generated with word embeddings, have similar distances.

**TABLE 11.** Averages of F1 results after 20 iterations using BERT with cross-validation (MIMIC II dataset), SVM(L): support vector machine with linear kernel, SVM (R): support vector machine with RBF kernel, DT: decision tree, NB: naive Bayes, KNN: K-nearest neighbor, RF: random forest.

| Analysis | Cancer vs not cancer | Breast cancer vs other cancer | Breast cancer vs breast cysts |
|---|---|---|---|
| SVM(L) | **0.822** ± 0.0001 | 0.740 ± 0.0001 | 0.780 ± 0.001 |
| SVM(R) | 0.591 ± 0.003 | 0.530 ± 0.0001 | 0.610 ± 0.002 |
| DT | 0.660 ± 0.002 | 0.643 ± 0.007 | 0.702 ± 0.001 |
| NB | 0.621 ± 0.0002 | 0.564 ± 0.0009 | 0.685 ± 0.001 |
| KNN | 0.787 ± 0.003 | **0.699** ± 0.001 | **0.786** ± 0,003 |
| RF | 0.756 ± 0.005 | 0.670 ± 0.003 | 0.740 ± 0.006 |
| DENSE | 0.91 ± 0.021 | **0.923** ± 0.0.36 | **0.974** ± 0.022 |
| 3-LSTM | 0.912 ± 0.041 | 0.892 ± 0.369 | 0.940 ± 0.406 |
| Bidirectional LSTM (sigmoid) | **0.968** ± 0.031 | 0.970 ± 0.037 | 0.981 ± 0.021 |
| Bidirectional LSTM (tanh) | 0.333 ± 2.38 | 0.333 ± 0.2 | 0.333 ± 3.41 |
| LSTM + Dense | 0.951 ± 0.033 | 0.932 ± 0.039 | 0.972 ± 0.023 |

**TABLE 12.** Averages of F1 results after 20 iterations using TFIDF with cross-validation (MIMIC II dataset), SVM(L): support vector machine with linear kernel, SVM (R): support vector machine with RBF kernel, DT: decision tree, NB: naive Bayes, KNN: K-nearest neighbor, RF: random forest.

| Analysis | Cancer vs not cancer | Breast cancer vs other cancer | Breast cancer vs breast cysts |
|---|---|---|---|
| SVM(L) | 0.611 ± 0.0021 | 0.665 ± 0.004 | 0.724 ± 0.004 |
| SVM(R) | 0.532 ± 0.0012 | 0.529 ± 0.002 | 0.622 ± 0.005 |
| DT | 0.929 ± 0.003 | 0.953 ± 0.007 | 0.971 ± 0.003 |
| NB | 0.672 ± 0.001 | 0.722 ± 0.0005 | 0.777 ± 0.002 |
| KNN | **0.975** ± 0.003 | **0.983** ± 0.001 | **0.974** ± 0,003 |
| RF | 0.921 ± 0.004 | 0.900 ± 0.005 | 0.926 ± 0.009 |
| DENSE | 0.916 ± 0.004 | 0.942 ± 0.002 | 0.953 ± 0.002 |
| 3-LSTM | **0.996** ± 0.005 | **0.997** ± 0.004 | 0.995 ± 0.008 |
| Bidirectional LSTM (sigmoid) | 0.993 ± 0.010 | 0.995 ± 0.007 | **0.996** ± 0.007 |
| Bidirectional LSTM (tanh) | 0.995 ± 0.007 | 0.996 ± 0.004 | 0.995 ± 0.007 |
| LSTM + Dense | 0.952 ± 0.003 | 0.959 ± 0.039 | 0.971 ± 0.003 |

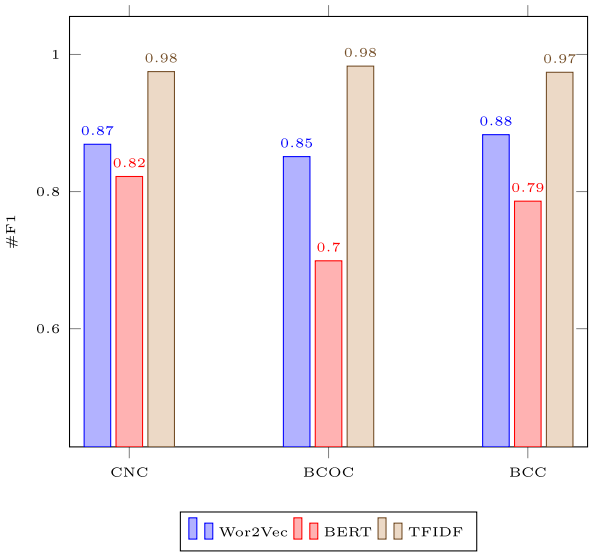


**FIGURE 11.** Comparative of the best result between Word2Vec, BERT, and TFIDF using traditional machine learning in each case for English dataset, CNC: cancer vs not cancer, BCOC: breast cancer vs other cancer, BCC: breast cancer vs breast cysts.

However, generating word embeddings requires decidedly fewer computational resources than TF-IDF by obtaining the representation vectors in a simple way and with limited resources. In our case, we used a computer with 32 GB of RAM and an Intel I7 processor to obtain vectors with 300 characteristics, recommended as the average length.

In contrast, when using TF-IDF, it was necessary to use external support tools because the size of the vectors exceeded the processing capacity of the available equipment, which complicated the preprocessing of the data due to the high demand for resources.

It should also be noted that despite using state-of-the-art models for the word-embedding process, the results obtained with word2vec were in all cases better than those obtained with BERT. This could be because BERT requires more data and is more well suited for formally written text.

Based on the results obtained in this work and with a larger dataset, we think that it is possible to develop a support tool for decision-making that could be of great help to physicians. This would allow improvement in the primary diagnoses and reduce the time for a patient to start treatment. More importantly, this would allow a reduction in the task time for searching through diagnoses in the coding systems. It would make the process more effective and less complex.

We repeated the same classification process on the MIMIC III dataset. For this, the breast cancer, other cancer, breast cysts, and other disease datasets were extracted.

With these datasets, the classification tests were carried out under the same conditions applied to the classification process used for the Spanish dataset. As observed in the results presented in Tables 10, 11 and 12, the TF-IDF model for the results in English obtained the best results in traditional ML and deep learning.

Although the results obtained when using the TF-IDF embedding were better than those obtained by word2vec and BERT, the computational cost of creating these vectors was very high compared to the other models. In addition, this model has the problem that it must train the entire sample when a new word is registered.

The word2vec and BERT models, for their part, are independent of the registration of new words because they do not work through the frequency of these words. Additionally, the manufacturing cost is minor compared to the completeness of going through all the texts and generating the necessary dictionaries for TF- IDF.

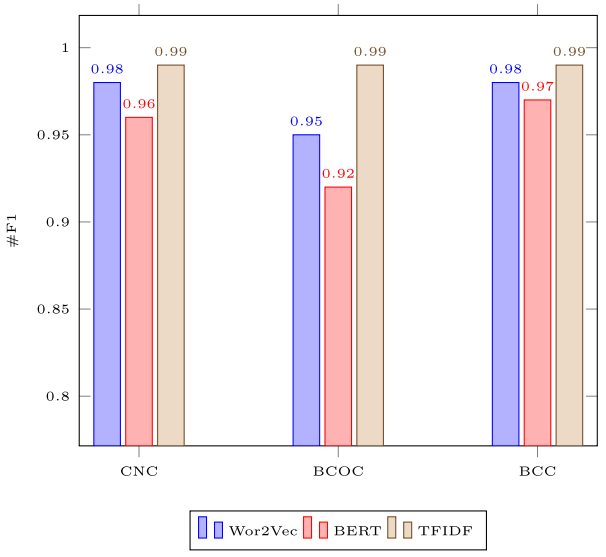


**FIGURE 12. Comparative of the best result between Word2Vec, BERT, and TFIDF using deep learning in each case for English dataset, CNC: cancer vs not cancer, BCOC: breast cancer vs other cancer, BCC: breast cancer vs breast cysts.**

As seen in Figure 12, the F1 results are not distant, where word2vec obtained an F1 score of 0.98 and TF-IDF obtained 0.99. This indicates that it is also possible to use these vectorization models to perform the clinical text classification process.

## VI. CONCLUSION

The objective of this work was to create a classification algorithm that could be used in a support tool for the recommendation of patient diagnosis. In this first approach, the results obtained illustrate that NLP, together with word embedding and machine learning, whether traditional or in deep learning, allows excellent results to be obtained in the classification of breast cancer, both in Spanish and English.

This type of tool, according to experts in the area, would be beneficial for doctors, especially for those who are starting their careers and have the direct responsibility for guaranteeing the health of patients, in many cases in rural hospitals.

The study presented in this document demonstrates that modern tools based on artificial intelligence, with the use of information processing, allow us to create and define algorithms for referral systems that can be useful support tools for the registration of doctors when they face complex decision-making tasks.

The study also shows that in these cases, it is vital to use a corpus based on the context of use for word embedding because many specific terms that are used can be misinterpreted in generic contexts. By considering this, we can obtain a better relationship of words around their distance and distribution, which does not happen with other representation models such as TF-IDF.

It is important to note that when the context is not considered, some analyses can be degraded if the diagnoses are distant in their description, such as ''cancer versus non-cancer'' or ''breast cancer versus other types of cancer''. In addition, it is worth emphasizing the good results of the word2vec model in texts where the syntax is mainly composed of the acronyms of the specialties.

Second, with a well-defined corpus, it is feasible to use automatic learning to classify diagnoses, obtaining excellent results with F1-macro values above 0.9.

However, in cases where the diseases are related, the results using a word2vec model are not as good as in the previous case, where the F1 scores are over 0.75. However, for these cases, it is feasible to use traditional methodologies to obtain better results, i.e., TF-IDF. Therefore, with very similar clinical histories, such as the case of breast cancer and breast cancer surgeries, these cases can be classified with excellent results by using traditional machine learning algorithms, such as random forest and K-nearest neighbor.

The opportunities that this study has afforded have motivated us to consider future work that entails processing a more significant number of clinical histories and defining a process that will enable us to describe other types of diagnoses, thus expanding the number of evaluated classes or the set of support tools to make decisions in highly complex diseases.

Finally, the use of unstructured data, such as the definitions written by physicians in the clinical histories, allows us to define the feasibility of applying large-volume information processing tools, which would make it possible to complement the long time required for preprocessing the information and thus to expand the number of resources with less time and fewer resources.

## REFERENCES

[1] Ministerio de Salud Chile. *Centro Nacional de Referencia de la FCI—DEIS*. Accessed: Aug. 13, 2018. [Online]. Available: http://www.deis.cl/centro-nacional-de-referencia-de-la-fci/

[2] M. Almagro, R. M. Unanue, V. F. Fernández, and S. M. Herranz, ''Estudio preliminar de la anotación automática de códigos cie-10 en informes de alta hospitalarios,'' *Sepln*, vol. 60, no. 1, pp. 45–52, Mar. 2018.

[3] W. Ning, M. Yu, and R. Zhang, "A hierarchical method to automatically encode chinese diagnoses through semantic similarity estimation," *BMC Med. Informat. Decis. Making*, vol. 16, no. 1, p. 30, Dec. 2016.
[4] S. Chen, P. Lai, Y. Tsai, J. Chung, S. Hsiao, and R. Tsai, "NCU IISRSystem for NTCIR-11 MedNLP-2 task," in *Proc. NTCIR*, 2014, pp. 1–4.
[5] P. Jatunarapit, K. Piromsopa, and C. Charoeanlap, "Development of thai text-mining model for classifying ICD-10 TM," in *Proc. 8th Int. Conf. Electron., Comput. Artif. Intell. (ECAI)*, Jun. 2016, pp. 1–6.
[6] P. Spyns, "Natural language processing in medicine: An overview," *Methods Inf. Med.*, vol. 35, nos. 4–5, pp. 285–301, 1996.
[7] R. H. Baud, A.-M. Rassinoux, and J.-R. Scherrer, "Natural language processing and semantical representation of medical texts," *Methods Inf. Med.*, vol. 31, no. 2, pp. 117–125, 1992.
[8] J. Pestian, H. Nasrallah, P. Matykiewicz, A. Bennett, and A. Leenaars, "Suicide note classification using natural language processing: A content analysis," *Biomed. Informat. Insights*, vol. 3, Jan. 2010, Art. no. BII.S4706.
[9] A. Ben Abacha and P. Zweigenbaum, "MEANS: A medical question-answering system combining NLP techniques and semantic Web technologies," *Inf. Process. Manage.*, vol. 51, no. 5, pp. 570–594, Sep. 2015.
[10] H. Wang, Q. Zhang, and J. Yuan, "Semantically enhanced medical information retrieval system: A tensor factorization based approach," *IEEE Access*, vol. 5, pp. 7584–7593, 2017.
[11] R. Alfaro and H. Allende, "Text representation in multi-label classification: Two new input representations," in *Adapt. Natural Comput. Algorithms*, A. Dobnikar, U. Lotrič, and B. Šter, Eds. Berlin, Germany: Springer, 2011, pp. 61–70.
[12] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*. [Online]. Available: http://arxiv.org/abs/1301.3781
[13] P. L. R. García and L. R. Pupo, "Principios técnicos para realizar la anamnesis en el paciente adulto," *Revista Cubana de Medicina Gen. Integral*, vol. 15, no. 4, pp. 409–414, 1999.
[14] P. Szolovits, R. S. Patil, and W. B. Schwartz, "Artificial intelligence in medical diagnosis," *Ann. Internal Med.*, vol. 108, no. 1, pp. 80–87, 1988, doi: 10.7326/0003-4819-108-1-80.
[15] C. De Coster, H. Quan, A. Finlayson, M. Gao, P. Halfon, K. H. Humphries, H. Johansen, L. M. Lix, J.-C. Luthi, J. Ma, P. S. Romano, L. Roos, V. Sundararajan, J. V. Tu, G. Webster, and W. A. Ghali, "Identifying priorities in methodological research using ICD-9-CM and ICD-10 administrative data: Report from an international consortium," *BMC Health Services Res.*, vol. 6, no. 1, p. 77, Dec. 2006.
[16] C. Benesch, D. M. Witter, A. L. Wilder, P. W. Duncan, G. P. Samsa, and D. B. Matchar, "Inaccuracy of the international classification of diseases (ICD-9-CM) in identifying the diagnosis of ischemic cerebrovascular disease," *Neurology*, vol. 49, no. 3, pp. 660–664, Sep. 1997.
[17] K. J. O'Malley, K. F. Cook, M. D. Price, K. R. Wildes, J. F. Hurdle, and C. M. Ashton, "Measuring diagnoses: ICD code accuracy," *Health Services Res.*, vol. 40, no. 5p2, pp. 1620–1639, Oct. 2005.
[18] L. B. Goldstein, "Accuracy of ICD-9-CM coding for the identification of patients with acute ischemic stroke: Effect of modifier codes," *Stroke*, vol. 29, no. 8, pp. 1602–1604, Aug. 1998.
[19] I. Kononenko, "Machine learning for medical diagnosis: History, state of the art and perspective," *Artif. Intell. Med.*, vol. 23, no. 1, pp. 89–109, Aug. 2001.
[20] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. Van Calster, "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models," *J. Clin. Epidemiology*, vol. 110, pp. 12–22, Jun. 2019.
[21] D. Chen, S. Liu, P. Kingsbury, S. Sohn, C. B. Storlie, E. B. Habermann, J. M. Naessens, D. W. Larson, and H. Liu, "Deep learning and alternative learning strategies for retrospective real-world clinical data," *NPJ Digit. Med.*, vol. 2, no. 1, pp. 1–5, Dec. 2019.
[22] K. Igor, "Experiments in automatic learning of medical diagnostic rules," Jozef Stefan Inst., Ljubljana, Slovenia, Tech. Rep. 1, 1984.
[23] F. Amato, A. López, E. M. Peña-Méndez, P. Vaňhara, A. Hampl, and J. Havel, "Artificial neural networks in medical diagnosis," *J. Appl. Biomed.*, vol. 11, no. 2, pp. 47–58, 2013.
[24] L. Mena and J. A. Gonzalez, "Symbolic one-class learning from imbalanced datasets: Application in medical diagnosis," *Int. J. Artif. Intell. Tools*, vol. 18, no. 2, pp. 273–309, Apr. 2009.
[25] W.-H. Weng, K. B. Wagholikar, A. T. McCray, P. Szolovits, and H. C. Chueh, "Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach," *BMC Med. Informat. Decis. Making*, vol. 17, no. 1, p. 155, Dec. 2017.
[26] M. Kindblom, "Diagnostic prediction on anamnesis in digital primary health care," School Elect. Eng. Comput. Sci., KTH Roy. Inst. Technol., Stockholm, Sweden, Tech. Rep. TRITA-EECS-EX, 2018, p. 523.
[27] V. Laippala, T. Viljanen, A. Airola, J. Kanerva, S. Salanterä, T. Salakoski, and F. Ginter, "Statistical parsing of varieties of clinical finnish," *Artif. Intell. Med.*, vol. 61, no. 3, pp. 131–136, Jul. 2014.
[28] P. Ruch, J. Gobeill, I. Tbahriti, and A. Geissbühler, "From episodes of care to diagnosis codes: Automatic text categorization for medico-economic encoding," in *Proc. AMIA Annu. Symp.*, 2008, pp. 636–640.
[29] J. D'Souza and V. Ng, "Knowledge-rich temporal relation identification and classification in clinical notes," *Database*, vol. 2014, Nov. 2014, Art. no. bau109.
[30] Y. Kim, E. Riloff, and S. Meystre, "Improving classification of medical assertions in clinical notes," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol., Short Papers*. vol. 2, 2011, pp. 311–316.
[31] Y. Li, S. Lipsky Gorman, and N. Elhadad, "Section classification in clinical notes using supervised hidden Markov model," in *Proc. ACM Int. Conf. Health Informat. IHI*, 2010, pp. 744–750.
[32] M. Hughes, I. Li, S. Kotoulas, and T. Suzumura, "Medical text classification using convolutional neural networks," *Stud. Health Technol. Inf.*, vol. 235, pp. 246–250, May 2017.
[33] X. Shi, Y. Hu, Y. Zhang, W. Li, Y. Hao, A. Alelaiwi, S. M. M. Rahman, and M. S. Hossain, "Multiple disease risk assessment with uniform model based on medical clinical notes," *IEEE Access*, vol. 4, pp. 7074–7083, 2016.
[34] G. Yufeng. (Aug. 2017). *The 7 Steps of Machine Learning*. [Online]. Available: https://towardsdatascience.com/the-7-steps-of-machine-learning-2877d7e5548e
[35] B. Koopman, G. Zuccon, A. Nguyen, A. Bergheim, and N. Grayson, "Extracting cancer mortality statistics from death certificates: A hybrid machine learning and rule-based approach for common and rare cancers," *Artif. Intell. Med.*, vol. 89, pp. 1–9, Jul. 2018.
[36] J. Cogley, N. Stokes, J. Carthy, and J. Dunnion, "Analyzing patient records to establish if and when a patient suffered from a medical condition," in *Proc. Workshop Biomed. Natural Lang. Process.*, Jun. 2012, pp. 38–46.
[37] M. Kholghi, L. De Vine, L. Sitbon, G. Zuccon, and A. Nguyen, "The benefits of word embeddings features for active learning in clinical information extraction," in *Proc. ALTA*, 2016, pp. 1–10.
[38] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-W.-H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Sci. Data*, vol. 3, no. 1, Dec. 2016, Art. no. 160035.
[39] M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, T. Heldt, T. H. Kyaw, B. Moody, and R. G. Mark, "Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access intensive care unit database," *Critical Care Med.*, vol. 39, no. 5, p. 952, 2011.
[40] M. L. for Computational Physiology. *Mimic Critical Care Database*. Accessed: Oct. 10, 2019. [Online]. Available: https://mimic.physionet.org/
[41] C. Robert, "Machine learning, a probabilistic perspective," *Chance*, vol. 27, no. 2, pp. 62–63, Apr. 2014, doi: 10.1080/09332480.2014.914768.
[42] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Oct. 2014, pp. 1532–1543.
[43] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: http://arxiv.org/abs/1810.04805
[44] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. B. A. McDermott, "Publicly available clinical BERT embeddings," 2019, *arXiv:1904.03323*. [Online]. Available: http://arxiv.org/abs/1904.03323
[45] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analyzers," *Neural Comput.*, vol. 11, no. 2, pp. 443–482, Feb. 1999.
[46] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2014.
[47] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and J. Vanderplas, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.
[48] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, p. 27, 2011.
[49] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[50] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[51] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[52] S. Ghosh, O. Vinyals, B. Strope, S. Roy, T. Dean, and L. Heck, "Contextual LSTM (CLSTM) models for large scale NLP tasks," 2016, *arXiv:1602.06291*. [Online]. Available: https://arxiv.org/abs/1602.06291
[53] R. Johnson and T. Zhang, "Supervised and semi-supervised text categorization using LSTM for region embeddings," 2016, *arXiv:1602.02373*. [Online]. Available: https://arxiv.org/abs/1602.02373
[54] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychol. Rev.*, vol. 65, no. 6, p. 386, 1958.
[55] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Netw.*, vol. 18, nos. 5–6, pp. 602–610, Jul. 2005.
[56] R. ALRashdi and S. O'Keefe, "Deep learning and word embeddings for Tweet classification for crisis response," 2019, *arXiv:1903.11024*. [Online]. Available: https://arxiv.org/abs/1903.11024
[57] D. S. Sachan, M. Zaheer, and R. Salakhutdinov, "Revisiting LSTM networks for semi-supervised text classification via mixed objective function," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 6940–6948.
[58] G. Liu and J. Guo, "Bidirectional LSTM with attention mechanism and convolutional layer for text classification," *Neurocomputing*, vol. 337, pp. 325–338, Apr. 2019.
[59] B. Krause, L. Lu, I. Murray, and S. Renals, "Multiplicative LSTM for sequence modelling," 2016, *arXiv:1609.07959*. [Online]. Available: http://arxiv.org/abs/1609.07959
[60] A. Gulli and S. Pal, *Deep Learning with Keras*. Birmingham, U.K.: Packt Publishing Ltd, 2017.
[61] S. Raschka and V. Mirjalili, *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2*. Birmingham, U.K.: Packt Publishing Ltd, 2019.
[62] T. Dozat, "Incorporating nesterov momentum into adam," in *Proc. Int. Conf. Learn. Represent.*, PR, USA, 2016.
[63] Y. Sasaki, "The truth of the F-measure," *Teach Tutor Mater*, vol. 1, no. 5, pp. 1–5, 2007.
[64] R. Speer. (Jul. 2017). *How to Make a Racist Ai Without Really Trying | Conceptnet Blog*. Accessed: Aug. 10, 2018. [Online]. Available: http://blog.conceptnet.io/posts/2017/how-to-make-a-racist-ai-without-really-trying/

**ANDRÉS ALEJANDRO RAMOS MAGNA** received the degree in computer science (engineer) from the Universidad de Valparaíso, in 2007, and the M.Sc. degree in computer science, in 2012. He is currently pursuing the Ph.D. degree in computer engineering with the Pontificia Universidad Católica de Valparaíso. He currently works as the Head of the Academic Systems Development Department and as a part-time Professor with the Universidad de Valparaíso in data mining and project management. His research interests include machine learning, natural language processing, and deep learning applications.

**HÉCTOR ALLENDE-CID** received the degree in computer science engineering, in 2009, the Master of Science degree in computer engineering, in 2009, and the Ph.D. degree in computer engineering from the Universidad Técnica Federico Santa María, in 2015. He is currently a full-time Professor with the School of Computer Engineering, Pontificia Universidad Católica de Valparaíso. He is also the President with the Chilean Association of Pattern Recognition. His research interests include machine learning, predictive analysis, and statistical computing.

**CARLA TARAMASCO** received the B.Eng. degree in computer engineering from the Universidad de Valparaíso, Chile, in 2001, the M.Sc. degree in cognitive science from the École Normale Supérieure, in 2006, and the Ph.D. degree (summa cum laude) from the École Polytechnique, France, in 2011. Her Ph.D. thesis was on Obesity and Social Structures. From 2011 to 2013, she was a Postdoctoral Fellow with CNRS. She is currently a Researcher and also a Professor with the Computer Science Department, Universidad de Valparaíso. She also teaches both the undergraduate and graduate levels, along with scientific divulgation. She has acted as the coordinator for over 20 national and international projects. She has scientific publications in books, journals, and conference proceedings. She has organized over ten international workshops/sessions. She was involved in the development of networks for scientific collaboration between Africa, South America, and Europe. She has been involved in the investigation and development of technological solutions for health-monitoring software and hardware. Her main research interests include (i) health, which includes mHealth, ambient assisted living for elderly persons, e-health, telemedicine and telerehabilitation, and monitoring of chronic diseases, and (ii) complex social systems, including dynamic networks, socio-semantic networks, analysis of trajectories both individual and collective, among others. She was the Coordinator of the Latino America Committee of Complex Systems Society for a period of five years.

**CARLOS BECERRA** is currently pursuing the Ph.D. degree with the Computing Engineering School, Universidad de Valparaíso, Chile. He is also an Associate Professor and the Director of the Computing Engineering School, Universidad de Valparaíso. His research interests include software architecture, empirical software engineering, and learning objects.

**ROSA L. FIGUEROA** (Associate Member, IEEE) received the B.Eng. degree and the Ph.D. degree in electrical engineering from the University of Concepción, in 2004 and 2012, respectively. Her Ph.D. Thesis explored different methods to obtain useful information from free text. She is currently a Faculty Member and a Researcher in biomedical engineering degree part with the Electrical Engineering Department, University of Concepción. She has scientific publications in journals and conference proceedings. She is also working in research projects related to secondary use of medical data and text classification. Her research interests include medical informatics area mainly machine learning and text mining.