Enhancing Random Forest Classification with NLP in DAMEH: A system for DAta Management in EHealth Domain

Flora Amato[a], Luigi Coppolino[b], Giovanni Cozzolino[c], Giovanni Mazzeo[d], Francesco Moscato[e], Roberto Nardone[f]

[a]*DIETI, University of Naples Federico II, Naples, Italy*
flora.amato@unina.it
[b]*DI, University of Naples Parthenope, Naples, Italy*
luigi.coppolino@uniparthenope.it
[c]*DIETI, University of Naples Federico II, Naples, Italy*
giovanni.cozzolino@unina.it
[d]*DI, University of Naples Parthenope, Naples, Italy*
giovanni.mazzeo@uniparthenope.it
[e]*DiSciPol, University of Campania Luigi Vanvitelli, Caserta, Italy*
francesco.moscato@unicampania.it
[f]*DIIES, University Mediterranea of Reggio Calabria, Reggio Calabria, Italy*
roberto.nardone@unirc.it

Abstract

The use of pervasive Internet of Things devices in Smart Cities has increased the volume of data produced in many and many fields. Interesting and handy applications have also grown in the eHealth domain, where smart devices are employed to provide innovative services able to collect data and to fill medical records of patients, so managing a considerable amount of data in a highly distributed environment. One of the open problems for these services is the gathering of data to produce records and to analyze medical records based on their contents. Since data gathering in eHealth involves very different and heterogeneous devices (not only wearable medical sensors but also environmental smart devices, like weather, pollution and other sensors), it is challenging to classify data depending their contents to enable better management of the patients' medical records. Indeed, medical records are written in natural language, and the final objective is to populate them automatically with data coming from smart devices. In this paper, we propose DAMEH, a novel and highly scalable system able to determine the best features for

classification, depending on existing medical records. DAMEH could be used to provide innovative functionalities in the eHealth domain, as the suggestion of therapies provided by others doctors based on the presence of the same symptoms in the diagnosis or detecting anomalies between the diagnosis and the prescribed therapy. The implemented workflow is based on a pre-filtering phase that relies on Natural Language Processing. This activity enhances the efficiency of the following step of Machine Learning classification based on Random Forests. The entire workflow and the architecture has been then experimented on a real-world dataset, constituted of about 5000 medical records (adequately anonymized) coming from various health-care organizations in Italy. We evaluated the accuracy of the presented approach in terms of Accuracy-Rejection Curves.

Keywords: Big Data Processing, eHealth, Machine Learning, Random Forests, Multi-Classification Schema

---

The Internet of Things (IoT) paradigm is nowadays inlaid in many and many activities we carry out in everyday life. IoT devices range from objects we deal with every day (e.g. smartphones, fridges, washing machines) to modern industrial equipment (e.g., smart monitoring systems, video surveillance). The massive employment of these devices introduced a kind of "intelligence" in the management of the massive amount of data, so providing complex value-added services to people. This phenomenon is the base for the concept of "Smart Cities".

Even if the pervasiveness of smart devices increases dramatically, the problem of capturing, classifying, indexing, retrieving and using the information in collaborative environments is still one of the open problems in the literature about IoT, Big Data and smart environments [1, 2]. Cloud-based applications may help to solve the storage problem, but many problems are arising in respecting current regulations and laws about privacy. Therefore, approaches that aim at extracting knowledge from data sources and at updating prediction models immediately, without storing data for long times (i.e., data streams), are now a solution to both storage and privacy management. From the other hand, the variety of information, its huge volume and the velocity of data production must be managed by proper techniques, and at state of the art, Machine Learning techniques are the only viable solution to this problem. Some studies have forecast that the global amount of data generated by IoT devices will reach the amount of close to 80 zettabytes (ZB)

per year by 2025[1].

Recent advantages in digital technologies made more easy gathering and managing vast amounts of data. It resulted in increments in volumes of collected data, leading to the necessity of storing and analyzing as much data as possible, and to enact complex statistical and data mining procedures in order to discover new correlations among data [3], even prospecting the scenario of "Big Data as a Service" [4]. It is worth to consider some numerical examples that show the amount of data production in several application areas, ranging from economic transaction to social network data: *i)* more than 144.8 billion e-mail messages are sent every day; *ii)* Twitter, a microblogging source, produces more than 10 TB of data per day; *iii)* a Boeing jet engine can produce 10 TB of operational information for every 30 minutes of flight and sensors networks and smart devices even in small cities can generate more than 1 TB of data per day.

Even the eHealth domain is not free from the influences of new technologies and IoT [5]. eHealth applications benefit from the pervasiveness of IoT devices, of Edge and Cloud Computing, and Big-Data analysis systems. Smart devices and sensors allow for monitoring patients at home, of course, but even the retrieval of pollution, weather and other types of data from sensor networks are useful to retrieve undiscovered correlations in medicine [6]. One of the central and challenging problems in eHealth is that all data collected from smart (medical or not) devices, as well as health records, are unstructured; they can be placed on different electronic medical records, and they can have different sources (both direct and indirect). Furthermore, eHealth data are collected from different media, like results of laboratories, images, medical reports and so on. Accessing to this valuable amount of information and making advanced analytics is decisive, not only to improve the patient care and the outcomes but also to give insights to evidence-based health services and decision making, in which doctors contextualize the best available research evidence by integrating it with their clinical expertise and their patient's values and expectations.

Big Data analysis needs innovative forms of data processing in order to extract information and discovering knowledge from data. Volume and Variety of data (two of the well-known Vs of Big Data analytics [7]) directly

[1]https://www.analyticsinsight.net/billions-iot-devices-produce-79-4-zettabytes-data-2025-says-idc, accessed on Jan. 2020

impact on load. So increasing volumes of data imply that the computational power should increase in order to reduce latency providing actionable intelligence at the right time. Traditional query mechanisms can report undesirable results and disregards significant results, depending on how the queries are implemented [8], while retrieving valuable amount of information and making advanced analytics is crucial, not only to improve patients care and the outcomes but also to advance the evidence-based health services and decision-making procedures.

Innovative data processing procedures are needed in order to extract information and to discover knowledge from big data. At state of the art, ad-hoc architectural solutions have been developed. They range from classical data warehouse techniques to innovative cloud-based architectures that provide potentially infinite resource and power computation [9], or in-memory computing architectures on heterogeneous CPU-GPU clusters for big data[10, 11], that use GPU's massive parallel processing ability, with a Just-In-Time (JIT) compiling schema and a heterogeneous task management strategy to maximize the computation capabilities of big data processing clusters.

The application of such new techniques may have great limitations for the eHealth domain, and the main reason relies on data privacy issues, the second one is related to the needed operational skills to setup/use/manage/audit a private cloud [12, 13]. Nevertheless, special purpose machines may not be suitable to implement high accuracy classification systems, as they are not programmable and many classifications and decision support systems need tuning and reprogramming. Furthermore, the classifier tuning activity is usually manually performed.

To face these two open issues, in this paper, we propose a system for DAta Management in eHealth Domain (namely DAMEH) that is a distributed system for eHealth data gathering and processing. Indeed, DAMEH comprises a workflow and a concrete architecture where all pertaining data are semantically processed to extract structured information. Then, collected data are stored on a central server, so enabling all kind of classification and analysis for decision support. DAMEH could be potentially used to provide innovative functionalities, as the suggestion of therapies provided by others doctors based on the presence of the same symptoms in the diagnosis or detecting anomalies between the diagnosis and the prescribed therapy.

DAMEH architecture includes two main components. The first is devoted to the extraction and the semantic processing of both documents of medical operators and data from smart devices. This component is distributed

on the smart devices, so moving the computational effort of the information extraction and structuring towards the Edge of IoT in a distributed way. The second component is devoted to the collection and storage of structured data on a central server. The classification system is based on multi-classifier schema [14], which combines lexical (Terms), syntactical (Lemmas) and semantic (Synonyms) modules. As will be described in the following, the computational complexity of the second component is considerably reduced by the pre-filtering operations executed by during the first stage. The resulting architecture promotes the scalability thanks to the distribution of the physical components, making affordable the management of a comprehensive data volume and preserving the efficiency of resources.

In this paper, we also demonstrate that the classification performed by DAMEH, obtained by combining the three filters based on Terms, Lemmas and Synonyms, increases the performance of each one of them in terms of Accuracy-Rejection Curves. The classifier in DAMEH relies on Random Forests classifier [15], widely adopted for classification tasks in several application areas. At last, the paper also describes the validation of DAMEH, performed by applying the workflow and the components to a real-world case study coming from various health-care organizations in Italy. In particular, the classifier output demonstrates the significant potential value-added offered by DAMEH and the innovative functionalities it enables in real eHealth context. The main contributions introduced by this paper are in the integration of well-known techniques in DAMEH and their concrete application to the eHealth domain. Moreover, the high scalability of the concrete architecture supporting DAMEH as well as the validation over real-world data are additional contributions of this work.

The rest of the paper is organized as follows. Section 1 discusses the motivations for the design of DAMEH and gives an overview of the systems and of the implemented workflow. Sections 2 and 3 give all the details of the phases of the DAMEH workflow. Section 4 shows and describes experimental results obtained by applying DAMEH to real-world data. Section 5 discusses related works and the innovations introduced by DAMEH in the plethora of work addressing natural language processing. Section 6 ends the paper by drawing conclusions and addressing future work.

1. Motivation and overview of DAMEH

The availability of increasingly large amounts of heterogeneous health data represents a significant opportunity to improve patient care as long as current medical information systems are capable of exploiting big data technologies and processing data. Nowadays, health organizations increasingly rely on digital systems for the management of medical records. Electronic medical records usually comply with standards like the Health Level 7 [16] (HL7). HL7 is the internationally recognized standard for managing the exchange of information relating to clinical and administrative health data, which allows health systems to communicate with each other. HL7 defines the interoperability level for the exchange of messages among the various systems and organizations for decision support, the mark-up languages used for defining documents, integration of interfaces and methods for developing messages and the data representation model. It is a roadmap for the presentation and communication of information between two or more parties, in a technological context that provides for an uninterrupted exchange and integration of the transmitted data. HL7 is nowadays present in more than 35 countries in the world.

Even if HL7 was born to contribute to the transition towards an incremental data structuring when a larger volume of loosely structured documents and data have been produced in eHealth, current medical information systems are often designed with relational databases and interfaces, but they still present many unstructured fields. Users are enabled to place and store free text information in the main field of health records. An example of unstructured fields could be a diagnosis and a prognosis of a patient, which are usually represented as free text in electronic medical records. Moreover, things got worst when dealing with data collected by smart devices for eHealth. Even though data is categorized and labelled, and although some alarm conditions automatically arise when dealing with dangerous situations (like with cardiopathic patients), human intervention is still needed in many cases to analyze information. Thus, an eHealth information system should be able to elaborate medical and smart sensors standards, that, like HL7, contains both structured (annotated) and unstructured record fields.

DAMEH has been designed to cope with these issues. It has to deal with the data heterogeneity, so including not only structured data (e.g., in HL7) but also text documents from other sources written in natural language. The main problems of this phase are making the interoperability and design

of a way to *understand* the meaning of data contained in medical records. Besides, information has to be extracted and elaborated from data: a really smart system should be able to correlate medical records and data from different sources automatically. For this reason, in DAMEH, we promote the combined usage of natural language processing (NLP) and unstructured data mining approaches.

DAMEH is a medical record processing system that assists doctors in the medical record composition and analysis, suggesting parts to be inserted in the free text fields, where data sources include smart devices for eHealth applications. The proposed system exploits semantic procedures applied to the medical records to extract and codify concepts that express health information. Furthermore, the analysis step is based on a high-throughput classification system that aims at processing all medical records, creating a data model in order to extract concepts properly.

The architecture of DAMEH is constituted of two main components enacting the following activities: *semantic processing* of unstructured texts from medical records, and *classification* of data, through a centralized node that collects pre-processed data from heterogeneous sources and performs inferences and classification actions on data. Of course, when dealing with large amounts of information, data pre-elaboration and processing represent the two hard task with the highest computational complexity. For this reason, the first component is executed directly on smart (medical) devices, so taking advantage of distribution and enabling scalability. The DAMEH classifier (offering the second functionality), which is instead centralized and may be a bottleneck for the whole system, is designed to be realized on reconfigurable hardware. The reconfigurability of this node is a fundamental property, as we need to train the classifier for this particular domain, in order to have the highest classification efficiency and throughput.

Classification is based on the tree-like model defined by the learner, and the tree visiting algorithm is usually executed sequentially. Each tree node contains a predicate that represents a condition, while leaves are labelled with classes the samples belong to. The implementation on reconfigurable hardware of the component named *Predictor*, i.e. the component performing classification, intrinsically exploits hardware parallelism to reach the best performance. Hence, its implementation can be optimized depending on structures of data models used for classification. In particular, DAMEH can synthesize "on the fly" the Predictor through the generation of proper VHDL code.

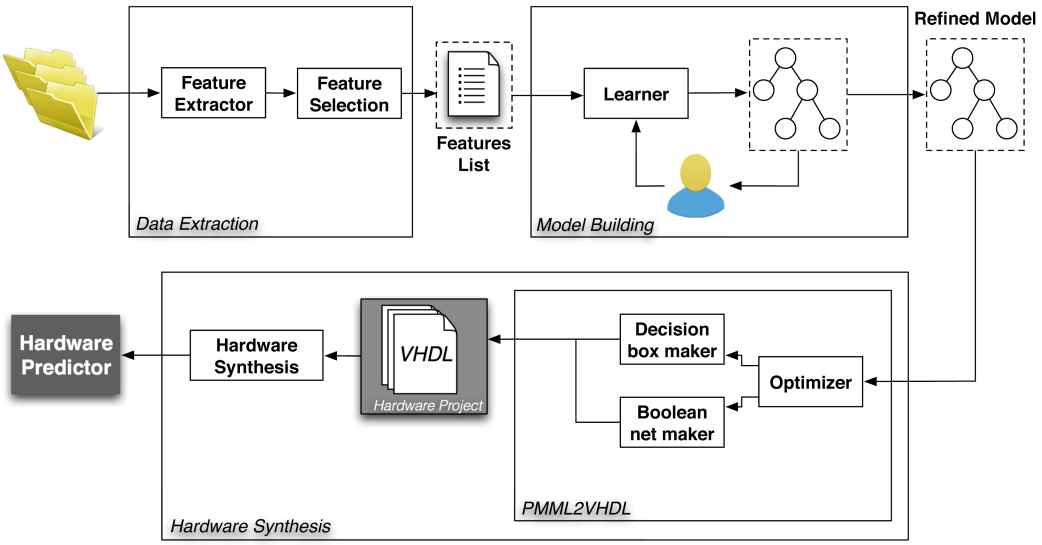


Figure 1: Workflow: from medical records to hardware implementation of the predictor

The workflow implemented by DAMEH is depicted in Figure 1; it is made of three phases: (i) the *Data Extraction* phase; (ii) the *Model Building* phase, and (iii) the *Hardware Synthesis* phase. The first block automatically structures heterogeneous data into a common schema, extracting relevant information by implementing the methodology proposed by Amato et al. in [17, 18]. The output, stored in a tabular format, is given in input to the *Model Building* block that implements the learning phase. The output of this step are parameters for the prediction model (PMML [19]) that is and used for classification.

The predictor model is translated in a hardware description written in VHDL (Very High Speed Integrated Circuit Hardware Description Language) language [20] and synthesized on an FPGA (Field Programmable Gate Array). It is implemented as a set of comparators to obtain the best performance. In the *Hardware Synthesis* step, the VHDL description is synthesized to obtain a predictor working on reconfigurable hardware. Periodically, or when a given rate of misclassification happens, the learner can recompute the parameters based on a new training set in order to refine the behaviour of the classifier (possibly excluding the predictor behaviours that lead to the misclassifications) and generating updated parameters for the prediction model. Again, optimization is performed to generate a new VHDL realization of an

updated and more efficient predictor.

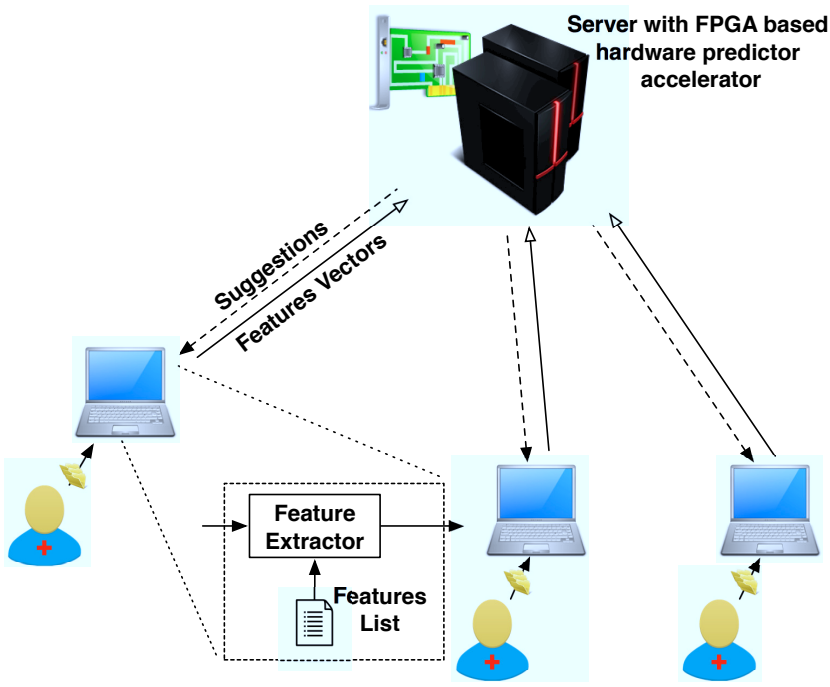


Figure 2: Doctors Medical Records management.

Figure 2 depicts the final architecture. As previously described, the feature extractor is located on smart medical devices and performs the data extraction step. The feature lists, coming from different devices, are sent to the central server equipped with the FPGA implementing the predictor. Suggestions evaluated by the classifier are sent back to smart medical devices. Periodically, the predictor is updated by performing the described optimizations.

Even if DAMEH is centred on eHealth application, the whole methodology is general enough to be applied to different domains. However, in a domain like eHealth, where the availability of support systems for fast and quick decision-making actions is crucial, DAMEH is appealing since designed to reduce time complexity and has the aims of automatically classifying data. Of course, another important innovation introduced by DAMEH resides in its distributed architecture, which exploits the parallelism of proper software and hardware solutions.

Next sections will illustrate in details every single step of the workflow

and will describe the internal of each module.

2. Data Extraction and Model Building

In this section we describe the details of the phases *Data Extraction* and *Model Building* of the workflow depicted in Figure 1.

The final goal of these phases is to create the classification model for the classification in DAMEH, so to apply edge filters to data retrieving only relevant information for classification, and to define a data model to organize retrieved information.

The *Data Extraction* phase aims at extracting all relevant tokens from the input elements that are both data streams from IoT devices and other medical records. It selects relevant features by two sub-modules: *Feature Extraction* and *Feature Selection*.

Notice that this phase addresses both XML-annotated data from IoT devices (e.g., in HL7), as well as text documents from other sources. The main problem of this phase is to extract information from text, to realize interoperability among different formats, standards, and documents in natural languages as well.

The *Feature Extraction* module performs the following activities: *(i)* it breaks up a data stream into a list of words, *(ii)* it marks up tokens related to parts of speech, *(iii)* it associates useful lemma to parts of speech, *(iv)* it filters token list obtaining the most relevant ones to build a features list and *(v)* it selects the most relevant features for the domain we are considering.

In order to extract the relevant tokens from input data streams and document corpora, the *Feature Extractor* submodule realizes a text processing pipeline implementing the sequential steps: *Text Tokenization*, *Text Normalization*, *Part-Of-Speech Tagging*, *Lemmatization* and *Synset Recognition*. Figure 3 shows the activities of *Feature Extraction* sub-module.

The main goal of these procedures is the extraction of relevant terms used to recognize concepts in the text. *Text Tokenization* and *Text Normalization* procedures perform a first grouping of the extracted terms [21], introducing a partitioning scheme that establishes an equivalence class on terms. In particular, *Text Tokenization* removes any punctuation, splitting the document up by spaces to get our tokens, making everything lowercase and removing stop words. *Text Normalization* aims at reducing the "entropy" of the input data applying techniques that eliminate numbers or non-letter characters,

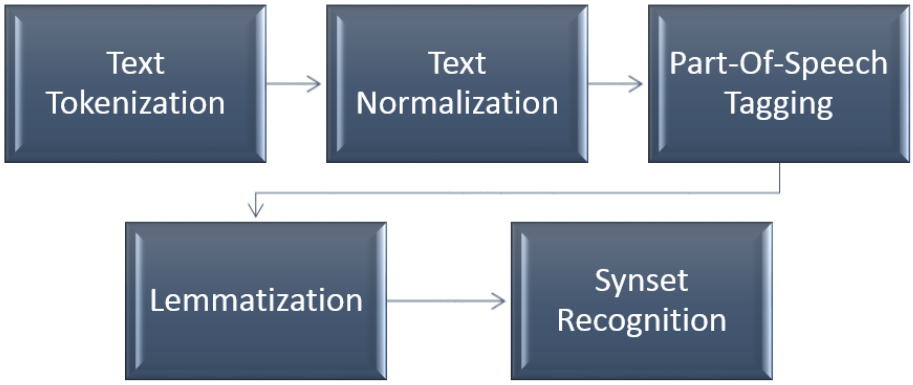


Figure 3: *Feature Extraction* sub-module

unifying special characters, disambiguate sentence boundary and identifying abbreviations or acronyms.

The *Part-Of-Speech Tagging* and *Lemmatization* steps aim at enriching text with meta-information about syntactical aspects associated to the extracted tokens, aiming at performing the second type of grouping of the words, based on reduction of terms in a basic form, independently from the conjugations or declination.

In natural language, many words may have different meanings, (this characteristic is called "polysemy"). We can identify the correct meaning of an ambiguous word through the context where it appears. For this reason, we apply on the text a word-sense disambiguation technique, that assigns the most suitable meaning to content words [22], such as nouns, verbs, adverbs or adjectives, based on probabilistic analysis of the word context [23].

The *Lemmatization* phase is performed on the list of disambiguated terms, to reduce all the inflected forms to related lemmas, or citation form, coinciding with the singular male/female form for nouns, the singular male form for adjectives and the infinitive form for verbs. *Lemmatization* introduces a second partitioning scheme on the set of extracted terms, creating new equivalence classes: it produces a more compact dataset because more terms are grouped.

Once relevant terms are detected, the *Synset Recognition* procedure proceeds in clustering lemmas in synsets. This procedure allows for associating proper concepts to the list of terms chosen to address them. In this way, it is possible to refer a concept independently from the particular term used for

denoting it. The clustering has been performed by integrating two external resources: medical ontologies by "UMLS" and "Mesh" [24], and several thesauri of medical terms. The use of specialized external resources has a double purpose: the first is "endogenous" (different terms can refer same concepts); the second is "exogenous" (in a natural language query, a given concept can be denoted by using terms which are different from those occurring in documents).

All these operations are performed to detect relevant words, by skimming the text of grammatical words not carrying useful information. These procedures are language-dependent, consisting of several sub-steps, and are implemented by using the state of the art NLP modules [25].

At this point, a list of tokens is obtained from the text. The *Features Extraction* phase produces a vector-based representation of inputs documents that lead to very high-dimensional problem feature spaces. In the adopted Bag of Word model [26], each distinct word corresponds to a feature. In this way, thousands of different terms are produced even for a moderate-sized data collection. Even if it is desirable to use as many features as possible to classify various documents so that a feature selection algorithm has more choices to find the optimal subset obtaining a high accuracy, having many dimensions leads to very long computing time because it can be expensive to run algorithms on huge vectors.

Moreover, it is not better to use more features but to use the right features. Many classifiers are known not to scale very well to high problem sizes [27]. Moreover, high dimension feature vectors could also cause overfitting, which is the trend of learning models to classify better objects they were trained with.

To avoid these problems, the list of extracted terms is processed by a dimensionality reduction stage, i.e. the *Feature Selection*, that reduces the size of vectors, to reduce the probability of overfitting and to make the vectors more manageable for the learning module. In the *Feature Selection* stage, each term is scored using a function that is indicative of its degree of correlation with the document class, and only terms with the highest score are used as features in document representation, so we select only the ones that best discriminate between classes. We applied several well-known feature selection approaches [28] to cut down the features by using various evaluation measures such as word frequency, document frequency, Correlation, Information Gain and Information Gain Ratio [29]. These approaches produce a ranking of features. We consider for further processing only the first Top_k

features of the ordered list.

To evaluate the effectiveness of the applied selection, we evaluate the accuracy of the classification results (performed on the data built on these features). For this reason, in the evaluation phase described in the following of this paper, we set different thresholds and evaluate the accuracy classification results for the three selection methods while varying the threshold k and, consequently, the number of selected features. Once the filtered list of features is identified, we proceed to evaluate the semantic relevance of the selected features in the document corpus by the TF-IDF matrix (Term Frequency - Inverse Document Frequency) [30], computed over the corpus vocabulary on the base of the term frequency and term distribution within the corpus. This matrix gives an indication of the semantic relevance of the features for each document in the corpus.

To build the model, we need to provide to the *Learner* module, the TF-IDF matrix with, also, a column specifying the class assigned for each row/-document.

2.1. Application of Data Extraction

After the description of the first two phases of the DAMEH workflow, that are *Data Extraction* and *Model Building*, we give in this subsection the first proof of concept of the proposed framework for the pre-processing of medical records. The goal is to show through a concrete application to real-world data all the steps to pre-process data and give proper input to the learner to set up the parameters for the classifier. Notice that the learning phase is executed off-line and just once. In this concrete application, we chose a fragment of a clinical record coming from Italian Hospitals that uses IoT smart devices in order to collect data from both remote and hospitalized patients. The data were anonymized, and we provided the proper English translation of each reported sentence (that were given to us in the Italian language). We use a set of over 67.000 medical diagnoses coming from various health care organizations, with 600 smart devices for patients monitoring.

The whole dataset has been classified in the following macro-categories: consulting, doppler, ecoc (for eco cardio), echography, endoscopy, operation, radiology, and synthesis. The data stream from IoT devices is clear: it contains information that identifies patients, the detected medical information and a timestamp (like for doppler data from portable ultrasounds).

The most complicated part is the management of medical records associated with retrieved data. They are in natural language and contain a de-

scription of patient health care on which we performed the analysis process, and an Exam code, containing the numerical ID of the exam.

Consider, as an example, the health report fragment reported below:

> *The patient reports the occurrence of paresthesias and cold sensation in the lower limbs during night. TAC L/S and doppler exams are required. Successively, elettromiografia exam to lower limbs.*[2]

We extracted relevant terms of this fragment and the associated concepts according to the described *Feature Extraction* step.

Table 1 shows all transformations performed on the list of tokens in the Features Extraction module. Notice that transformations have been performed on the Italian version; in Table 1, we report the English translation. In the Part of Speech phase, the acronyms tagging the words indicates the grammar category, for example, VER is for the verb, NOUN is for the name, NPR is for Proper noun, ADJ is for the adjective, and ADV is for adverb. The interested reader can find further details and examples on the method adopted for NLP in [17].

The extracted tokens are then filtered by the *Feature Selection* stage, whose goal is to reduce the number of features to improve the classification learning phase from a performance point of view but without losing accuracy.

To this aim, the feature vector is processed by different well-known feature selection approaches, in order to choose an appropriate subset of words for useful classification automatically. The features are processed with the following three methods, and the results are evaluated based on obtained accuracy. In this way, no manual selection is required. The chosen methods are well-known selection techniques, highly suggested and adopted in the current literature [31]. As it will be shown in the following, the obtained results are satisfying.

We used the following three techniques:

- Information Gain method, where the Information Gain of a feature is calculated by measuring the difference in uncertainty (i.e., entropy),

[2]The original Italian version is: *La paziente riferisce la comparsa notturna di parestesie e di sensazione di freddo agli arti inferiori. Si richiede TAC L/S e doppler arteriovenoso arti inferiori. Successivamente elettromiografia agli arti inferiori.*

Phase	**Processed Fragment of Clinical Record**
Original	The patient reports the occurrence of paresthesias and cold sensation in the lower limbs during night. TAC L/S and doppler exams are required. Successively, elettromiografia exam to lower limbs.
Tokenization	The-patient-reports-the-occurrence-of-paresthesias- and-cold-sensation-in-the-lower-limbs-during-night.-TAC-L/S-and-doppler-exams-are-required.-Successively,-elettromiografia-exam-to-lower-limbs.
Normalization	The-patient-reports-the-occurrence-of-paresthesias- and-cold-sensation-in-the-lower-limbs-during-night-TAC-L/S-and-doppler-exams-are-required-successively-elettromiografia-exam-to-lower-limbs
Part of Speech Tagging	The (ART) – patient (NOUN) – reports (VER) – the (ART) – occurrence (NOUN) – of (PRE) – paresthesias (NOUN) – and (CON) – cold (ADJ) – sensation (NOUN) – to (PRE) – the (ART) – lower (ADJ) – limbs (NOUN) – during (ADV) – night (NOUN) – TAC (NOUN) - L/S (NPR)- and (CON) - doppler(NPR) - exams (NOUN) - are required (VER) - successively (ADV) - elettromiografia (NOUN) - exam (NOUN) - to (PRE) - lower (ADJ) – limbs (NOUN)
Lemmatization	The (ART) – patient (NOUN) – report (VER) - the (ART) - occurrence (NOUN)- of (PRE) – paresthesias (NOUN) - and (CON) – cold (ADJ) – sensation (NOUN) – to (PRE) - the (ART) – lower (ADJ) - limb (NOUN) - during (ADV) - night (NOUN) - TAC (NOUN) - L/S (NPR) - and (CON) - doppler (NPR)- exam (NOUN) - require (VER) - successively (ADV) - elettromiografia (NOUN) - exam (NOUN) - to (PRE) - lower (ADJ) - limb (NOUN)

Table 1: Example of Processing of Clinical Record Fragment

Information Gain	Gain Ratio	Correlation
sclerosis	rate	cardiac
normal	cholecyst	sclerosis
instruction	thickening	gall-bladder
education	ECG	biliary bladder
formation	echocardiogram	cholecyst
constitution	gizzard	spleen
assistance	sonogram	splene
operation	belly	liver
presence	abdomen	pancreas
frequency	sinuses	courage

Table 2: Fragment of Top feature lists extracted by the three considered method

between the cases without and with knowledge of the value of that attribute;

- Information Gain Ratio method is the normalized version of Information Gain, as it evaluates Information Gain divided by the entropy of the attribute. It can be used when the Information Gain approach overestimates the importance of features with large numbers of values;

- Correlation-based method, which evaluates the worth of features by measuring the Pearson's correlation between it and the class[28].

The top-ranked features extracted with the selected methods, on a small dataset of five documents, are shown in Table 2, adequately translated from Italian to English.

Based on empirically derived thresholds, we just considered the Top_k of the list of the ranked features, outputted by each method. Furthermore, based on the evaluation scores for the features selected with the different methods, shown in Table 3, we choose as feature list the first Top_k ranked features outputted by the Correlation method (with better accuracy). As said previously, the results in Table 3 confirm that a selection of features leads to better results in terms of accuracy, in addition to better performance in the learning state due to a lower number of features to consider.

At this point, the semantic relevance of the features in the document corpus is evaluated by the TF-IDF matrix, based on terms frequencies and

Selection	Feat.	Accuracy	FMeasure	Precision	Recall
No Selection	2967	68.00%	66.88%	74.81%	68.00%
Correlation	303	88.00%	85.57%	83.86%	88.00%
GainRatio	436	83.00%	81.22%	80.94%	83.00%
InfoGain	363	87.00%	86.03%	85.95%	87.00%

Table 3: Accuracy results evaluated on the filtered features, for the considered dataset

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}
Doc1	1.91	0	0	1.91	0	0	0	1.91	0	3.31	0
Doc2	0	0	0	0	1.22	0	1.91	0	1.91	0	0
Doc3	2.13	0	1.51	0	1.12	1.91	0	0	0	0	0
Doc4	0	0	1.31	0	1.22	0	0	0	0	0	1.91
Doc5	0	1.91	0	0	0	0	0	0	0	0	0

Table 4: TF-IDF Matrix for the considered dataset

distributions within the corpus. In Table 4, we report a matrix portion with TF-IDF computed for the five medical reports fragments. The first row is related to the fragment discussed above; columns represent the features, i.e., the TF-IDF value of the following concepts (translated adequately from Italian to English): f_1 =arteryvenous, f_2 =articular, f_3 =biopsy, f_4 =doppler, f_5 =echodriven, f_6 =echostructure, f_7 =lower, f_8 =paresthesia, f_9 =patient, f_{10} =request, f_{11} =CAT.

This matrix with the addition of the column of the class (assigned and refined by a domain expert) is the input for the adopted C4.5 learner module[32], which is responsible for building the model of the Predictor that will be synthesized in hardware and used online on the whole set of medical records to classify. In our experiments, for the learning phase, we used a training set of 1000 labelled clinical records.

3. From model to hardware synthesis

This section continues the description of the workflow by illustrating the phases from model building to the hardware synthesis. As depicted in Figure 4, the preliminary phases are composed by *Feature Extraction* and *Feature Reduction*. As described in the previous section, the first steps aimed at

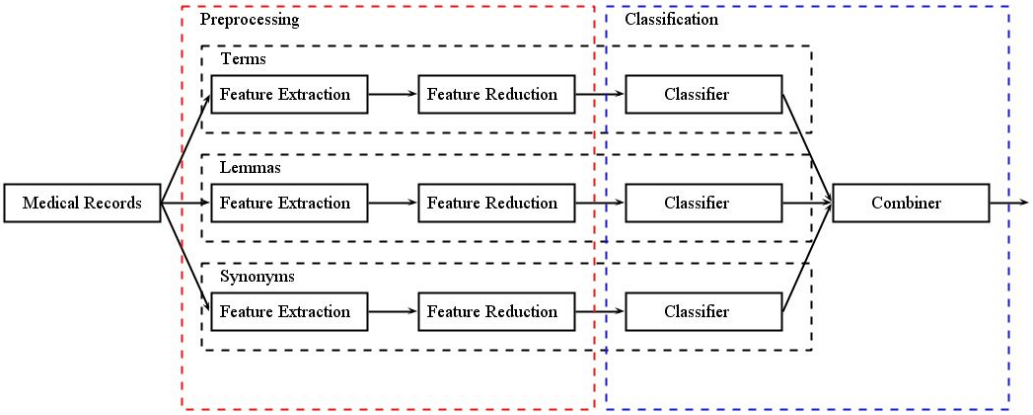


Figure 4: Classification System Architecture

analysing the medical records to extract useful information, called *Features*, to represent them in the classification process. In our framework we take into account three different features types: *Terms*, *Lemmas* and *Synonyms*. Each input document has been represented using a feature vector, which is a row of the TF-IDF matrix.

Since the number of features extracted from medical records is very high, the size of this matrix is very large. On the one hand, it is desirable to use as many features as possible to classify various documents and to obtain a high classification accuracy; on the other hand, it can be expensive to run algorithms on a very large vector, and it could also introduce "rumour" (confusion) during the learning phase of Classification Stage. This phenomenon is also known as *curse of dimensionality*, first introduced by Bellman [33]. For these reasons, in the second phase of Preprocessing, a step of Feature Reduction has been applied to TF-IDF matrix [34].

In the second phase, a classifier exploits the feature matrix to build the classification model and to assign a class to new unseen documents. Our classification approach is based on a *Multiple (Parallel) Topology* which is the most common implementation of a multi-classifier system that combines the results of some base classifiers [35]. As shown in Figure 4, we combine the results of three classifiers, one for each feature type.

3.1. Preprocessing

The Preprocessing stage extracts the most *discriminant* features, based on *lexical, syntactic and semantic analysis*, to build the TF-IDF matrix of

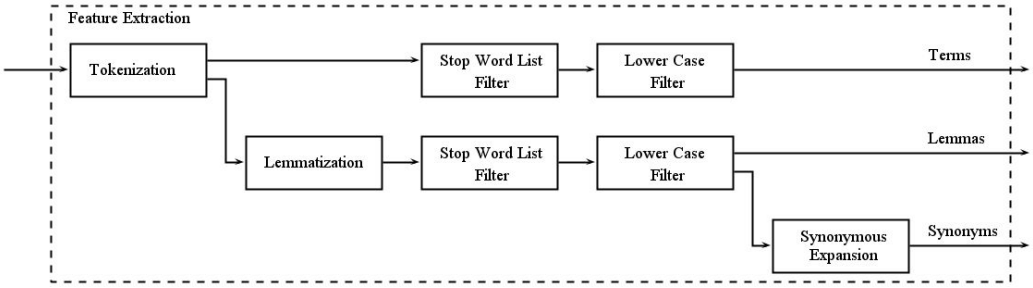


Figure 5: Feature Extraction Module

the input medical records.

This stage is composed of two steps: feature extraction and feature reduction. The first step builds the *document Index (Indexing)* while the second step applies Feature Reduction methods to remove non-discriminant information from TF-IDF matrix in order to reduce computational complexity and, more importantly, to improve classification performance.

3.1.1. Features Extraction

The Feature extraction step extracts from documents: *Terms*, *Lemmas*, *Synonyms*. As depicted in Figure 5 we define a different pipeline for each of three groups of features.

To obtain the **Terms** we apply *lexical analysis* based on NLP techniques to documents: Tokenisation, Deleting Stop Words and Lower Casing. Tokenisation is based on text segmentation. The input of this process is the plain text and the output is a stream of text composed by words (terms) or other significant items, called tokens. Stop words (as for examples English determiners "the", "a" or "an", or Italian conjunctions such as "e" or "o".) are text element useless to understand documents meanings and then they can be ignored to improve the information extraction. Lower Case filter normalises input tokens to lower case.

In order to extract **Lemmas** from documents, we add the Lemmatization module to the pipeline composed by the modules described above. Lemmatisation is the activity of grouping together some inflected forms of a given token so that they can be analysed as a single element: for instance, the lemma of the tokens "have", "has", "had" is "have". To determine a lemma for a given token it is necessary the context understanding and the labeling of a token with its own *Part-of-Speech* (PoS) tag. The PoS tagging is a process

to tag each token with its tag which represents the syntactic category of own token (i.e. the – DT, DeTerminer for "an" or "the"). The Italian PoS tagger employs in the implementation is TreeTagger[3]. Applying Tokenisation, Lemmatization, Deleting Stop Words and Lower Casing to each document, we analyse syntactically the input of our system, as in the second pipeline shown in Figure 5.

Finally, in order to extract **Synonyms**, we apply a Synonymous Expansion block, that exploits a multilingual lexical database, namely MultiWordNet[4], to retrieve the synset starting from given lemmas and its PoS tags. A synset is a set of all synonyms of predetermined lemma, and its items represent a concept expressed by the lemma. For example, the synset of the lemma "home" contains the followings words : "dwelling", "home", "domicile", "abode", "habitation", etc. These words are handy to describe the meaning (concept) of token "home". Synonymous Expansion is the main part of our semantic analysis of the input medical records.

The implementation of lexical, syntactic and semantic analyses are based on Apache Lucene[5]. Within the Lucene framework, these modules are called filters because they *filter* from the documents only the required information. The schema of the implemented semantic-based classifier is depicted in Figure 6.

The values of features identified according to previous analysis were evaluated considering their relevance both, *locally* within each document and *globally* considering their impact within the corpora. We considered TF-IDF matrix [36] to compute the values of the features. The TF-IDF is a $m \times n$ matrix where m is the number of documents in the collection, and n is the number of features. Each row represents a document, and it contains the values of the corresponding features (term or lemma or synonymous). Each value represents how important an element is for a *specific* medical record and for the *whole* corpora[37]. The values of the TF-IDF matrix, named $(tf - idf)$, are defined as follows: $tf - idf = tf \times idf$, where tf represents the local parameter, and it is the number of occurrences of the feature within the specific document; whereas the idf represents the global parameter and

[3]Software is freely available here `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`

[4]See the following web page `http://multiwordnet.fbk.eu/english/home.php` for a more detailed overview of MultiWordNet.

[5]Software is freely available here `http://lucene.apache.org/`

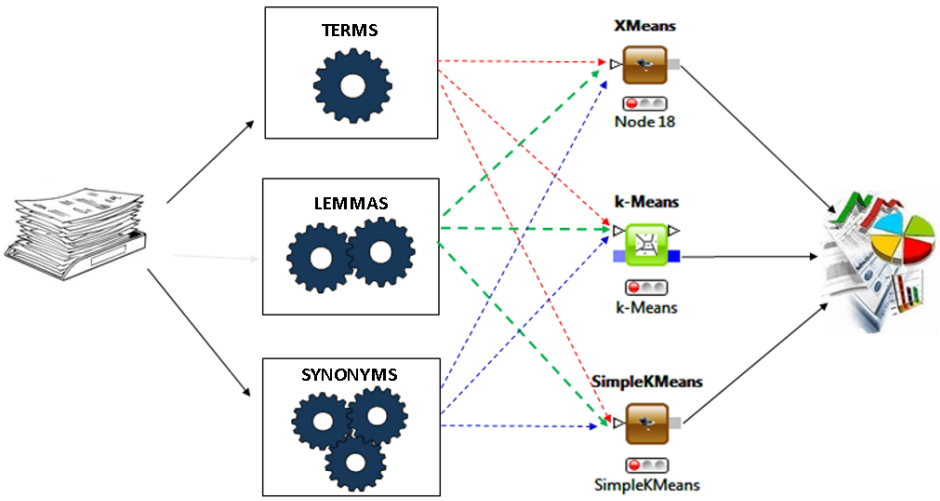


Figure 6: Schema of the Semantic Based Classifier

it is the measure of whether the feature is common or rare across the whole collection of documents.

3.1.2. Features Reduction

The main motivation to reduce the original feature space is the curse of dimensionality introduced the first time by Bellman [33] in 1957. More generally, the curse of dimensionality is the expression of all phenomena that appear with high-dimensional data, and that often has unfortunate consequences on the behaviour and performances of learning algorithms.

In this work, we considered as feature reduction method the *selection* strategy, and we applied it to the $TF-IDF$ matrix. In details, we choose an algorithm based on the *Gain Ratio* (GR) [38]. GR selects the most discriminant features considering as *fitness function* the Information Gain divided by the entropy of the attribute, as summarized in the following formula:

$$\mathbf{GR}(Class, Attribute) = \frac{H(Class) - H(Class \backslash Attribute)}{H(Attribute)} \quad (1)$$

We selected this method because it is a good trade-off between the discriminant power of each attribute and the minimisation of the degree of redundancy from each other.

Classification technique	Accuracy
Random Forest	0,885
Random Forest (regression)	0,993
Fuzzy rule	0,899
Naive Bayes	0,774
PNN	0,836
MLP	0,823

Table 5: Accuracy values of different classification techniques on a preliminary dataset

3.2. Classification

We adopted a multi-classification schema based on the *parallel* topology which is one of the most common implementations of a multi-classifier system [35]. The concept of combining classifiers has been intensely exploited in literature, and it is well known that it improves the performances of the base classifiers if they are *independent* enough [39]. The input of this combination consists of results of the individual classifiers, and the output is a unique combined decision.

We have evaluated various types of classifiers to assess which one was the most suitable for our classification module. In particular, we considered the following classification techniques: Random forest, Naïve Bayes, Fuzzy rule, Machine Learning Neural Network, Probabilistic Neural Network.

We used a smaller dataset, compared to the one that will be shown in Section 4, made by 1523 of medical records, appropriately anonymised and randomly chosen, that we wanted to classify in order to identify patients suffering from certain pathologies. The classification has been done according to the Early Warning Score (EWS), a patient instability assessment scale, used in hospital by nurses to monitor clinical conditions. Five physiological parameters are evaluated; for each of them a score from 0 to 3 is given, then added to the others for a total ranging from a minimum of 0 to a maximum of 14. The higher the value, the greater will be the patient's health risk.

The metric chosen methodologies' comparison is the *accuracy*: the ratio between the sum of *TruePositive* and *TrueNegative* and the total of the analysed elements. Results of preliminary evaluation session are reported in Table 5. The preliminary data, confirmed by the full experimental session reported in Section 4, showed that Random Forest has higher accuracy than the other techniques.

Regarding the classification module implementation, as shown in Figure 4, we considered three base classifiers, one for each feature type (Terms, Lemmas, and Synonyms). Each one of them is based on *Random Forest* (RF) [40].

RF is a classifier based on an *ensemble learning* method. It builds a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees. Each of the decision tree models is learned on a different set of rows (records) and a different set of columns (attributes). The records for each decision tree are created by *bagging* - bootstrap aggregation - and have the same size as the input table. This technique starts from the original *training set*: N smaller subsets are created using random sampling with a replacement called *bootstrap*. For each node in a decision tree, a new set of attributes is determined by taking a random sample of $sqrt(m)$ size - square root - where m is the total number of attributes. The output model describes a random forest and is applied to the corresponding predictor node.

The base classifiers outputs are then combined using a *Weighted Majority Voting* (WMV) [41]. If we want to formalize this method, we can assume that the outputs of each classifier will be denoted with a binary vector of size M, $[d_{i,1}, \ldots, d_{i,M}]^T \in \{0,1\}^M, i = \{1, \ldots B\}$, where B is the number of classifiers involved into the ensemble, M is the number of the possible classes, and where $d_{i,j} = 1$ if the i^{th} classifier votes the class C_j for the actual sample, while $d_{i,j} = 0$ otherwise. To consider the *weighted* version we have to define another coefficient vector b, where b_i represents the weights associated with the i^{th} classifier. So the system will decide for the class C_k if:

$$\sum_{i=1}^{B} b_i d_{i,k} = \max_{j=1}^{M} \sum_{i=1}^{B} b_i d_{i,j} \tag{2}$$

In our work, the coefficient vector b_i contains the probability associated with the aftermath of classes predicted by the RF classifier.

For the parallel implementation of the random forest classifier on FPGA, we followed the approach described in [42], that is FastRF, a Fast Implementation of Random Forest Classifier. This approach improves both speed and memory usage, by an average speed-up factor of 2.3. The choice of this implementation method was justified by the fact that it is compliant with the

memory constraints of the Nexys board we used in the experimental session. Note that random forests parallel implementations have been deeply studied in literature [43, 44], but all the other analysed solutions require the usage of specific hardware, not available on the board.

4. Experimental Results

This section reports results of the classification system proposed for each base classifier built for (Terms, Lemma and Synonyms) and for the combination of them using the Weighted Majority Voting.

Here we present results in terms of *Accuracy-Rejection Curves* (ARC) obtained using the ten-fold cross-validation methodology. The points of the curves represent the accuracy as a function of the rejection rate. We used an experimental dataset that includes original medical records and data collected from smart devices, coming from various Italian healthcare organizations.

4.1. Data Set

In our experimental assessment, we used a subset of our corpus, which contain over than 5.000 medical records coming from various healthcare organizations located in Campania (Italy). In the rest of the paper, we call this data set "medical records". Since each fragment is a real document or an acquired record from a smart device that reports patients' ID, they have been properly anonymized, by deleting any sensitive personal information about patients. Each medical record is characterized by three columns, as shown in Table 6. We report the original record in the Italian language between parenthesis.

4.2. Evaluation Measures

We compare classifier performances using the Accuracy-Rejection Curves. These curves represent the classifier accuracy as a function of the rejection rate. An ARC is therefore produced by plotting the accuracy of a classifier against its rejection rate, varying it in the range $[0\%, 100\%]$. The main characteristics of the ARCs are: i) all ARCs have an accuracy of 100% for a rejection rate of 100%, and therefore they converge on the point (100%, 100%); ii) they always start from a point (0%, a%), where a% is the best accuracy obtainable without rejection. The reject option is based on the idea that

Table 6: Examples of Medical Records

Class	Report	Exam Code
Consultation	Visit Urological. The prostate appears in the standard size. (Visita Urologica. La prostata appare di dimensioni nella norma.)	8406
Doppler	Absence of hemodynamically significant stenosis. (Assenza di stenosi emodinamicamente significative.)	8041
Ecoc	Mild dilatation of the left atrium. Aortic sclerosis. (Lieve dilatazione dell'atrio sinistro. Sclerosi aortica.)	3201
Ecographic	In the normal thyroid volume in regular contours. (Tiroide in sede di normale volume a contorni regolari.)	8021
Endoscopy	Esophagus normal. Normoconformed stomach. (Esofago nella norma.Stomaco normoconformato.)	8061
Intervention	Right inguinal hernia. Neo suprapubic region. (Ernia inguinale destra. Neo regione sovrapubica.)	2623
Radiology	Marked bilateral knee OA, in pre ankylosis, more marked on the right. (Marcata gonartrosi bilaterale,in preanchilosi, piu' marcata a destra.)	8128

the documents, for which the classification reliability is less than a threshold value, will be rejected to reduce the likelihood of error. Observing the value of the reject rate computed when the accuracy is equal to 100%, we can establish if the classification system is efficient on not. If this value is very low, the number of a not classified element is meaningless (very reliable classification system).

4.3. Results

In our experiments, we varied the size of the rejection rate from 0% to 100%, by increments of 5%. These rates are plotted against the classification accuracies obtaining several ARCs, as shown in Figure 7. To improve presentation and visualization of our results, classification accuracy is included in the interval $[95\%; 100\%]$; this range is justified because our worst classifier achieves the accuracy of 96.88%.

In the graph shown in Figure 7(a), the three base classifiers built on Terms, Lemmas and Synonyms have been compared. Each of them is built using 10-fold Cross-Validation. The cross-validation has been applied after the preprocessing phase, described in Section 3.1, to validate the model.

As we can see in Figure 7(a), the best of three classifiers is the one based on the Terms because it obtains the accuracy value equal to 100% with a rejection rate of 31.90%. On the contrary, other two classifiers, based on Lemmas and Synonyms respectably, achieve the same accuracy with rejection rates of 33.50% and 100%. The ARC points of Term classifiers are always

greater than the same points related to two other systems (based on Lemmas and Synonyms) because the last two feature extraction methods introduce noise in the classification model.

In Figure 7(b), we can see that the use of the combination of Terms and Lemmas allows having greater performance than the one based on respectively the Terms and the Lemmas. It is worth noting that when we consider the rejection rate equal to 0%, the multi-classifier schema reached an accuracy of 98.60%, while for the two base classifiers, Terms and Lemmas, the accuracies are respectively of 98.22% and 98.36%.

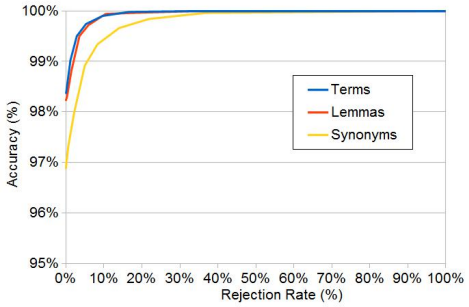
In Figure 7(c), we compare the performance, obtained combining with WMV only the Term and Lemma base classifiers, to the one computed using all the proposed classifiers. As we can observe, the WMV classifier based on a combination of Terms and Lemmas reaches the accuracy at 100% with a rejection rate of 14.32%, while for the combination of all the three classifiers we reached 100% only for a rejection rate of 100%, these results are obtained derive from the Synonymous extraction in which all synonyms of a given lemma have been introduced in the feature set. This choice amplifies classifier errors, and it does not reduce them.

Finally analyzing the experimental results in terms of accuracy without rejecting, we can conclude that the classification process which obtains the best performance is the combination of Terms, Lemmas, and Synonyms. On the contrary, considering the ARCs, the combination of Terms and Lemmas is the best choice. This is due, as just discussed before, to the synonyms issues.

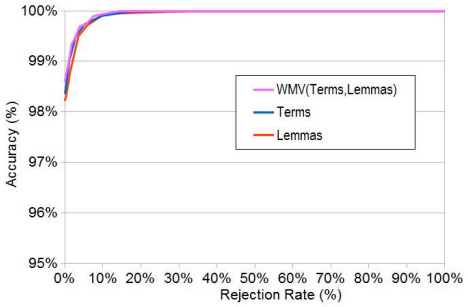
Before concluding, it is worth to report that the time needed by the classifier is not critical with respect to the concrete application of DAMEH in a real-world context. The time consumption in this kind of e-Health applications is not a crucial feature since it is possible to wait also for some minutes before having a result. This would lead in a new configuration of the classifier, that can be synthesized on the board, so updating the classifier itself. However, with the applied dataset, we experimented a maximum classification time of a single sample on the Nexys board lower than 20 seconds. Obviously, this value can be optimized by using a more efficient FPGA.

5. Related Work

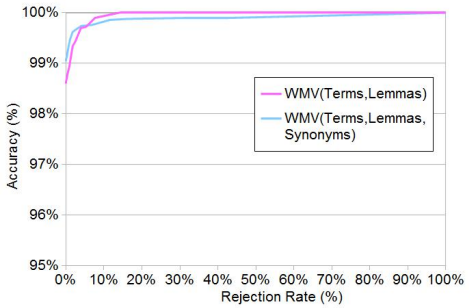
Knowledge management systems should provide instruments for building, maintaining, and development of a knowledge base which represent the cen-



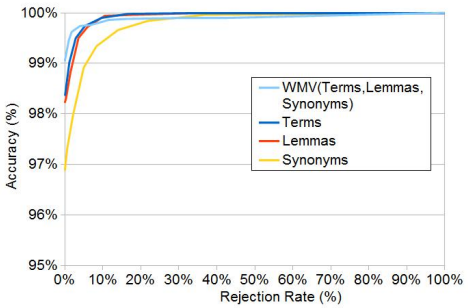
(a) ARCs for the base classifiers obtained for Terms, Lemmas, Synonyms



(b) Comparison among the best combination approach (Terms-Lemmas) and each base classifier involved into the ensemble



(c) ARCs obtained for the multiclassification schema based on Term-Lemmas and Terms-Lemmas-Synonyms



(d) Comparison among the combination of all the three base classifiers and the performance of each of them

Figure 7: Area Rejection Curves to compare our classifiers

tral unit of any knowledge-based intelligent system. The first step involved in the knowledge base development consists in the activity of knowledge acquisition that involves information extraction and machine learning techniques. The increasing availability of on-line sources of information in the form of natural-language texts increased the accessibility of textual information. The overwhelming quantity of available information has led to a strong interest in technology for processing text automatically in order to extract task-relevant information [45, 46]. The main task of information extraction is to automatically extract structured information from unstructured and/or semi-structured documents, exploiting different kinds of text analysis.

Once the text has been extracted, the aim of a document management

system is to correctly group the information according to a specific task. The task of grouping information and assigning a document into a class or category is known as information categorization. It involves classification and clustering techniques that are respectively commonly known as supervised and unsupervised learning techniques for automatic document organization. In many specialist domains, as the healthcare one, the introduction of automatic tools for Managing Information is extremely appealing. Many projects have exploited Natural Language Processing (NLP) approaches to extract the information from the free text part of medical reports [47]. For instance, in [48] these NLP methods have been used to generate structured patient records and then supervised machine learning techniques (classification) have been applied to code the presence of clinically important injuries. Their results showed that the use of structured information (built using NLP technique) improves raw text classification. In [49] an initial exploratory work on the development of an NL parser for medical reports is presented that attempts to assign a stability metric of a reference word within a given sentence and proposed parse configuration. Moreover, in [50] an automatic system, based on NLP methods to add annotations of clinical documents using a Clinical Document Architecture (CDA), is proposed. In [51] the authors studied, evaluated and proposed different swarm intelligence techniques for mining information from loosely structured medical textual records with no a-priori knowledge. The output of this task is a set of ordered/nominal attributes suitable for rule discovery mining and automated processing.

Some text categorization approaches exploit lexical, syntactic and semantic features extracted with NLP methods (as for example Part-of-Speech tagging, Tokenization, etc.), but in other methods, as for example [52, 53], semantic features (synonym or hypernym or hyponym, etc.) have been employed to improve the performance [54]. Moreover, the main contribution of [52] is that semantic features lead in text categorization rather than feature selection and machine learning techniques.

Moreover in [55], a methodology for automatic document categorization, based on the adoption of unsupervised learning techniques, is proposed. Experiments were performed on a corpus of real medical records written in Italian. The objective of their work was both to extract semantic (Concepts) and syntactic (Lemmas and Terms) features in order to define the vector space models. In [56], the authors discusses the importance of Accuracy in classifiers for medical applications. High dimensional dataset degrades the classification accuracy, therefore, they used the Feature Selection (FS), a pro-

cess used to select the most informative features from the given medical data sets, as in our classification framework. In the [57] a method of Chinese text feature selection and weighting based on Synonym merge has been presented. They use synonymous concepts to extract feature values in text based using a Thesaurus called "TongYiCi CiLin".

The need to properly manage and preserve records for quality assurance implies that specific processes for electronic document and records management are required that should be able to support the users in the management of information by means of on-line and/or off-line functionalities. Moreover, to deal with big data, innovative forms of data processing are required in order to enable process optimization and enhance decision making tasks while the volume and variety of data directly impact computational load, the velocity, in most of the data mining methods[58, 59]. In high performance applications, the data increasing implies that the computational power should increase in order to reduce latencies providing actionable intelligence at the right time. Thus hardware implementation of machine learning algorithm can improve the computational power.

Random forests is a widely used method for machine learning classification. Thanks to their inherent concurrent memory accesses and computational parallelism FPGAs are a good platform for performance boost of random forests. In [60] authors demonstrate the trade-offs between area utilization and context switch time between different architectures, and show how each architecture maps well to a different design scenario. The paper [61] compares and contrasts the effectiveness of FPGAs, GP-GPUs, and multicore CPUs for accelerating classification using models generated by compact random forest machine learning classifiers and shows that FPGAs provide the highest performance solution, but require a multi-chip / multi-board system to execute even modest sized forests, while GP-GPUs offer a more flexible solution with reasonably high performance that scales with forest size.

6. Conclusion and Future work

In this paper, we addressed the problem of data analysis in e-health domain with a semantic approach. We presented DAMEH, a classification system based on multi-classifier schema which combines lexical (Terms), syntactical (Lemmas) and semantic (Synonyms) modules using Random Forest method. We demonstrated that the classification results obtained with the combination of three classifiers (based on respectively Terms, Lemmas and

Synonyms) increase the performance of each one of them (in terms of ARC). As discussed in the experimental results, the use of all synonyms related to the single lemma does not contribute to the improvement of the classification performance. To solve this problem in future works, we will consider using a Word Disambiguation approach. We will also deal with to select, in synonymous extraction module, only the synonyms of a synset belonging to a healthcare/medical domain.

Furthermore, as we can observe analyzing the ARC of the combination of Terms and Lemmas in experimental assessment, the accuracy at 100% has been reached with a rejection rate of 14.32%. In general, it means that our proposed classification system is enough reliable because of the low number of rejected medical records with the highest accuracy. Finally, we studied the impact of other feature reduction methods within our classification framework.

As for the semantic, medical records are complex technical semi-structured text documents that need a pre-processing phase. We coped this issue implementing an automatic semantic structuring flow that overcomes the complexity of the documents. The heterogeneous data are transformed into a standard schema, and the only relevant information from the medical records are extracted. In particular, we adopted a sequence of operations for features extraction and selection.

After selecting relevant features for classification, we used a random forest algorithm.

We have shown the validity of our methodology by explaining the suitability and benefits of the introduced techniques.

As for the semantic, we provided results for several techniques in the feature extraction to maximizes accuracy with a minimum cardinality of the set feature to configure parameters of the classifier. A case study on a real medical records set puts in evidence the trade-off between accuracy reached by the data extraction phase and throughput of the hardware accelerator.

As future work, we aim at working to both aspects addressed in this paper to evaluate different classification algorithms and their hardware implementations and to provide a more complex refinement process for pre-processing of data at the user level, in order to fasten IoT data management directly on devices and on the Edge.

References

[1] J. Hu, C. Liu, K. Li, K. Li, Game-based multi-md with qos computation offloading for mobile edge computing of limited computation capacity, in: IFIP International Conference on Network and Parallel Computing, Springer, 2019, pp. 16–27.

[2] J. Mei, K. Li, Z. Tong, Q. Li, K. Li, Profit maximization for cloud brokers in cloud computing, IEEE Transactions on Parallel and Distributed Systems 30 (1) (2018) 190–203.

[3] V. Mayer-Schönberger, E. Ingelsson, Big data and medicine: a big deal?, Journal of internal medicine 283 (5) (2018) 418–429.

[4] X. Wang, L. T. Yang, H. Liu, M. J. Deen, A big data-as-a-service framework: State-of-the-art and perspectives, IEEE Transactions on Big Data 4 (3) (2017) 325–340.

[5] K. Kuan, M. Ravaut, G. Manek, H. Chen, J. Lin, B. Nazir, C. Chen, T. C. Howe, Z. Zeng, V. Chandrasekhar, Deep learning for lung cancer detection: tackling the kaggle data science bowl 2017 challenge, arXiv preprint arXiv:1705.09435 (2017).

[6] K. Hu, Y. Wang, A. Rahman, V. Sivaraman, Personalising pollution exposure estimates using wearable activity sensors, in: 2014 IEEE ninth international conference on intelligent sensors, sensor networks and information processing (ISSNIP), IEEE, 2014, pp. 1–6.

[7] P. Russom, et al., Big data analytics, TDWI best practices report, fourth quarter 19 (4) (2011) 1–34.

[8] X. Zhou, K. Li, G. Xiao, Y. Zhou, K. Li, Top k favorite probabilistic products queries, IEEE Transactions on Knowledge and Data Engineering 28 (10) (2016) 2808–2821.

[9] L. Zhang, K. Li, Y. Xu, J. Mei, F. Zhang, K. Li, Maximizing reliability with energy conservation for parallel task scheduling in a heterogeneous cluster, Information Sciences 319 (2015) 113–131.

[10] C. Chen, K. Li, A. Ouyang, K. Li, Flinkcl: an opencl-based in-memory computing architecture on heterogeneous cpu-gpu clusters for big data, IEEE Transactions on Computers 67 (12) (2018) 1765–1779.

[11] C. Chen, K. Li, A. Ouyang, Z. Zeng, K. Li, Gflink: An in-memory computing architecture on heterogeneous cpu-gpu clusters for big data, IEEE Transactions on Parallel and Distributed Systems 29 (6) (2018) 1275–1288.

[12] L. Fei, E. Rijnboutt, D. Routsis, N. Venekamp, H. Fulgencio, M. Rezai, A. van der Helm, What challenges have to be faced when using the cloud for e-health services?, in: e-Health Networking, Applications & Services (Healthcom), 2013 IEEE 15th International Conference on, IEEE, 2013, pp. 465–470.

[13] R. Chauhan, A. Kumar, Cloud computing for improved healthcare: Techniques, potential and challenges, in: E-Health and Bioengineering Conference (EHB), 2013, IEEE, 2013, pp. 1–4.

[14] A. F. Kamara, E. Chen, Q. Liu, Z. Pan, Combining contextual neural networks for time series classification, Neurocomputing 384 (2020) 57 – 66.

[15] M. Pal, Random forest classifier for remote sensing classification, International Journal of Remote Sensing 26 (1) (2005) 217–222.

[16] R. H. Dolin, L. Alschuler, S. Boyer, C. Beebe, F. M. Behlen, P. V. Biron, A. Shabo, Hl7 clinical document architecture, release 2, Journal of the American Medical Informatics Association 13 (1) (2006) 30–39.

[17] F. Amato, V. Casola, A. Mazzeo, S. Romano, A semantic based methodology to classify and protect sensitive data in medical records, in: Information Assurance and Security (IAS), 2010 Sixth International Conference on, IEEE, 2010, pp. 240–246.

[18] F. Amato, V. Casola, N. Mazzocca, S. Romano, A semantic-based document processing framework: A security perspective, in: International Conference on Complex, Intelligent and Software Intensive Systems, 2011, pp. 197–202.

[19] A. Guazzelli, M. Zeller, W.-C. Lin, G. Williams, Pmml: An open standard for sharing models, The R Journal 1 (1) (2009) 60–65.

[20] J. Bhasker, A Vhdl Primer, Prentice-Hall, 1999.

[21] E. Hatcher, O. Gospodnetic, M. McCandless, Lucene in action (2004).

[22] S. Stevenson, E. Joanis, Semi-supervised verb class discovery using noisy features, in: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4, Association for Computational Linguistics, 2003, pp. 71–78.

[23] J. Boyd-Graber, Linguistic resource creation in a web 2.0 world, in: NSF Workshop on Collaborative Annotation.

[24] O. Bodenreider, The unified medical language system (umls): integrating biomedical terminology, Nucleic acids research 32 (suppl 1) (2004) D267–D270.

[25] D. Klein, C. D. Manning, Conditional structure versus conditional estimation in nlp models, in: Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, Association for Computational Linguistics, 2002, pp. 9–16.

[26] H. M. Wallach, Topic modeling: beyond bag-of-words, in: Proceedings of the 23rd international conference on Machine learning, ACM, 2006, pp. 977–984.

[27] Z. Yang, Paml: a program package for phylogenetic analysis by maximum likelihood, Computer applications in the biosciences: CABIOS 13 (5) (1997) 555–556.

[28] G. Forman, An extensive empirical study of feature selection metrics for text classification, The Journal of machine learning research 3 (2003) 1289–1305.

[29] T. Li, C. Zhang, M. Ogihara, A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression, Bioinformatics 20 (15) (2004) 2429–2437.

[30] T. Joachims, A probabilistic analysis of the rocchio algorithm with tfidf for text categorization., Tech. rep., DTIC Document (1996).

[31] D. Roobaert, G. Karakoulas, N. V. Chawla, Information gain, correlation and support vector machines, in: Feature extraction, Springer, 2006, pp. 463–470.

[32] W. Sun, J. Chen, J. Li, Decision tree and pca based fault diagnosis of rotating machinery, Mechanical Systems and Signal Processing 21 (3) (2007) 1300–1317.

[33] R. Bellman, Dynamic programming: Princeton univ. press, N J 95 (1957).

[34] J. R. Finkel, A. Kleeman, C. D. Manning, Efficient, feature-based, conditional random field parsing, in: Proceedings of ACL-08: HLT, 2008, pp. 959–967.

[35] M. Woźniak, M. Graña, E. Corchado, A survey of multiple classifier systems as hybrid systems, Information Fusion 16 (2014) 3–17.

[36] C. D. Manning, P. Raghavan, H. Schütze, Scoring, term weighting and the vector space model, Introduction to information retrieval 100 (2008) 2–4.

[37] F. Amato, R. Canonico, A. Mazzeo, A. Picariello, Statistical and lexical analysis for semi-automatic extraction of relevant information from legal documents, Journal of Applied Sciences 11 (4) (2011) 639–46.

[38] A. G. Karegowda, A. Manjunath, M. Jayaram, Comparative study of attribute selection using gain ratio and correlation based feature selection, International Journal of Information Technology and Knowledge Management 2 (2) (2010) 271–277.

[39] L. I. Kuncheva, Combining pattern classifiers: methods and algorithms, John Wiley & Sons, 2014.

[40] L. Breiman, Random forests, Machine learning 45 (1) (2001) 5–32.

[41] C. De Stefano, A. Della Cioppa, A. Marcelli, An adaptive weighted majority vote rule for combining multiple classifiers, in: Object recognition supported by user interaction for service robots, Vol. 2, IEEE, 2002, pp. 192–195.

[42] V. Y. Kulkarni, Effective learning and classification using random forest algorithm (2017).

[43] A. Jinguji, S. Sato, H. Nakahara, An fpga realization of a random forest with k-means clustering using a high-level synthesis design, IEICE TRANSACTIONS on Information and Systems 101 (2) (2018) 354–362.

[44] A. A. Montillo, Random forests, Lecture in Statistical Foundations of Data Analysis (2009).

[45] R. Grishman, Information extraction: Capabilities and challenges (2012).

[46] M. F. Moens, Information Extraction: Algorithms and Prospects in a Retrieval Context, The Springer international series on information retrieval, Springer, 2006.

[47] G. Hripcsak, C. Friedman, P. O. Alderson, W. DuMouchel, S. B. Johnson, P. D. Clayton, Unlocking clinical data from narrative reports: a study of natural language processing, Annals of internal medicine 122 (9) (1995) 681–688.

[48] E. Sarioglu, H.-A. Choi, K. Yadav, Clinical report classification using natural language processing and topic modeling, in: 2012 11th International Conference on Machine Learning and Applications, Vol. 2, IEEE, 2012, pp. 204–209.

[49] R. K. Taira, V. Bashyam, H. Kangarloo, A field theoretical approach to medical natural language processing, IEEE Transactions on Information Technology in Biomedicine 11 (4) (2007) 364–375.

[50] S. L. DuVall, K. W. Boone, A. Gundlapalli, B. R. South, S. Shen, J. R. Nebeker, L. W. D'Avolio, M. H. Samore, Creating reusable annotated corpora with the clinical document architecture, in: 2011 44th Hawaii International Conference on System Sciences, IEEE, 2011, pp. 1–10.

[51] M. Bursa, L. Lhotska, V. Chudacek, M. Huptych, J. Spilka, P. Janku, M. Huser, Ant inspired techniques in textual information retrieval from a hospital information system, in: 2011 Third World Congress on Nature and Biologically Inspired Computing, 2011, pp. 421–426.

[52] D. Çelik, A. Elçi, A broker-based semantic agent for discovering semantic web services through process similarity matching and equivalence considering quality of service, Science China Information Sciences 56 (1) (2013) 1–24.

[53] J. J. Messerly, G. E. Heidorn, S. D. Richardson, W. B. Dolan, K. Jensen, Information retrieval utilizing semantic representation of text, uS Patent 6,076,051 (Jun. 13 2000).

[54] F. Amato, M. Barbareschi, V. Casola, A. Mazzeo, An fpga-based smart classifier for decision support systems, in: Intelligent Distributed Computing VII, Springer, 2014, pp. 289–299.

[55] F. Amato, F. Gargiulo, A. Mazzeo, S. Romano, C. Sansone, Combining syntactic and semantic vector space models in the health domain by using a clustering ensemble., in: HEALTHINF, 2013, pp. 382–385.

[56] H. H. Inbarani, A. T. Azar, G. Jothi, Supervised hybrid feature selection based on pso and rough sets for medical diagnosis, Computer methods and programs in biomedicine 113 (1) (2014) 175–185.

[57] Z. Lu, Y. Liu, S. Zhao, X. Chen, Study on feature selection and weighting based on synonym merge in text categorization, in: 2010 Second International Conference on Future Networks, IEEE, 2010, pp. 105–109.

[58] C. Dobre, F. Xhafa, Intelligent services for big data science, Future Generation Computer Systems (2013).

[59] M. Pallikonda Rajasekaran, S. Radhakrishnan, P. Subbaraj, Sensor grid applications in patient monitoring, Future Generation Computer Systems 26 (4) (2010) 569–575.

[60] X. Lin, R. S. Blanton, D. E. Thomas, Random forest architectures on fpga for multiple applications, in: Proceedings of the on Great Lakes Symposium on VLSI 2017, 2017, pp. 415–418.

[61] B. Van Essen, C. Macaraeg, M. Gokhale, R. Prenger, Accelerating a random forest classifier: Multi-core, gp-gpu, or fpga?, in: 2012 IEEE 20th International Symposium on Field-Programmable Custom Computing Machines, IEEE, 2012, pp. 232–239.