IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 13, 2021, accepted December 27, 2021, date of publication January 11, 2022, date of current version January 20, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3142100

AI-Driven Clinical Decision Support: Enhancing Disease Diagnosis Exploiting Patients Similarity

CARMELA COMITO, DEBORAH FALCONE, AND AGOSTINO FORESTIERO
Institute for High Performance Computing and Networking of the National Research Council of Italy (ICAR-CNR),
87036 Rende, Italy

Corresponding author: Agostino Forestiero (agostino.forestiero@icar.cnr.it)

ABSTRACT Detecting diseases at early stage can help to overcome and treat them accurately. A Clinical Decision Support System (CDS) facilitates the identification of diseases together with the most suitable treatments. In this paper, we propose a CDS framework able to integrate heterogeneous health data from different sources, such as laboratory test results, basic information of patients, health records and social media data. Using the data so collected, innovative machine learning and deep learning approaches can be employed. A neural network model for predicting patients' future health conditions is proposed. The approach employs word embedding to model the semantic relations of hospital admissions, symptoms and diagnosis, and it introduces a mechanism to measure the relationships of different diagnosis in terms of symptoms similarity to exploit for the prediction task. Several CDSs, including diagnostic decision support systems for inferring patient diagnosis, have been proposed in the literature. However, these methods typically focus on a single patient and apply manually or automatically constructed decision rules to produce a diagnosis. Even worst, they consider only a single medical condition, whereas it is not uncommon that a patient has more than one medical condition at the same time. The novelty of the proposed approach is the combination of supervised and unsupervised artificial intelligence methods allowing to combine several and heterogeneous data sources related to a multitude of patients and concerning different medical conditions. Furthermore, with respect to previous approaches, the diagnosis prediction problem is formulated to predict the exact diagnosis in terms of semantic meaning by exploiting Natural Language Processing concepts. Experimental results, performed on a real-world EHR dataset, show that the proposed approach is effective and accurate and provides clinically meaningful interpretations. The obtained outcomes are promising for future extensions of the framework that could be a valuable means for automatic inferring disease diagnosis.

INDEX TERMS Clinical decision support system, digital patient, disease diagnosis prediction, patient similarity, word embedding.

I. INTRODUCTION

Health data from disparate medical sources is collected continuously, leading to the generation of huge amount of information. As an example, patients'medical information is extracted from their personal medical data, such as physiological data, electronic health records (EHRs), 3D images, radiology images, genomic sequencing, clinical and billing data. The availability of such data enables real-time and personalized health services for patients and professionals.

Artificial intelligence (AI) techniques like machine learning and deep learning methods, can be exploited to help doctors in diagnosing and treating their patients more efficiently. Using their experience and knowledge, the physicians classify patients and diagnose their diseases, but in doing so, it could happen that they commit some mistakes, particularly when they lack adequate experience on certain subjects. In such situations, Clinical Decision Support systems (CDS), including systems that provide diagnosis, personalized medical measurement and treatment, could be helpful to the physicians by providing them with specific knowledge, patients' information and intelligent applications, which can improve the efficiency of the decision-making processes [1]. CDS systems focus on extracting characteristics of patients and diseases, based on which they classify patients and provide corresponding clinical suggestions to the physicians.

Several clinical decision support systems, including diagnostic decision support systems for inferring patient

The associate editor coordinating the review of this manuscript and approving it for publication was Ravinesh C. Deo.

diagnosis, have been proposed in the literature. However, these methods typically focus on a single patient and apply manually or automatically constructed decision rules to produce a diagnosis. Even worst, they consider only a single medical condition, whereas it is not uncommon that a patient has more than one medical condition, at the same time, because of the complications of the first disease.

To address the above issues and challenges, in [2] we proposed a CDS framework that integrates heterogeneous health data collected from disparate sources, such as laboratory test results, medical images and electronic health records. The framework implements a set of services to support physicians in diagnosing or treating patients' health issues. To this aim novel methods, exploiting deep learning as well as machine learning tools, are embedded within the framework with the main goal of automate disease diagnosis and treatments.

Leveraging on the CDS framework proposed in [2], the key motivation behind this work is to exploit the disparate artificial intelligence methods (like deep learning and machine learning) to enable automatic disease diagnosis also taking advantage of alternative data sources like social media data and data coming from body sensors networks. Specifically, this paper addresses the challenging issue of inferring patient diagnoses, exploiting electronic health records and medical data upon hospitalization. To this purpose, is proposed a novel framework for diagnostic prediction based on patient-similarity, using basic patient-specific information gathered at hospital admissions, including medical history, blood tests, laboratory results and demographics to identify similar patients and subsequently predicting patient outcomes. Patients similarity is defined as the similarity between patients' diagnoses and symptoms rather than a dichotomous problem stating the absence/presence of just one disease. We learn patients semantic models from the overall collected data by developing an AI method able to generate context-based and rich representation of health related information. In particular, we exploit word embedding models since neural networks shown to have the ability to learn complex feature representations. We apply the word embedding approach to categorize text fragments, at a sentence level, based on the emergent semantics extracted from a corpus of medical text.

In this context a fundamental challenge is how to correctly model such temporal and high dimensional EHR data to significantly improve the performance of prediction. To address this challenge, the paper put emphasis on the implementation within the proposed CDS architecture of an ad-hoc mechanism for searching patient's documents in a distributed health system, based on Natural Language Processing (NLP) concepts.

The diagnosis prediction method proposed has been evaluated over a real-world medical data, the MIMIC III dataset [3]. The results shown that the prediction approach, based on word embedding and exploiting semantic similarity of both symptoms and diagnoses, resulted to be effective and accurate, reaching considerable precision and recall rates. The obtained outcomes are promising for future developments and extensions of the framework that could be a valuable means for automatic inferring disease diagnoses.

The rest of the paper is organized as follows. Section II overviews related work. Section III introduces the clinical decision support architecture together with its main functionalities. Section IV presents as case study the diagnosis prediction method, formulating the problem, the data model and introducing the key aspects of the approach. Section V describes the real-world medical data used for the experimental evaluation. The results of the evaluation are shown in Section VI. Section VII concludes the paper.

II. RELATED WORK

In the healthcare domain, EHRs mining represents one of main research field and topics like disease progression [4]–[7], diagnosis prediction [8]–[11], electronic genotyping and phenotyping [12]–[14], adverse drug event detection [15] were widely investigated. Moreover, designing deep learning models, often, has provided significant improvements to the performance. Unplanned readmissions and risks related to electronic health records management were predicted using convolutional neural networks (CNNs) [16], [17], while multivariate time series health data, fox example, can be profitably modeled through recurrent neural networks (RNNs), also when some values are missing [18], [19].

In [8], an approach to acquire knowledge from health data, named Med2Vec, is proposed with the aim to predict future diagnosis information. The method overrides the long-term dependencies of health codes among diagnosis. A graph-based attention model for representation learning in healthcare is proposed in [20]. In this model, medical ontologies are used to learn robust representations, while patients' visits are modeled exploiting RNN. In order to perform binary prediction tasks, a predictive model implementing a reserve time attention mechanism was proposed in [9]. It employs a location-based approach to predict future possible diagnosis. The attention weights, related to a visit at a given time, are calculated by using latest medical information, and exploited to implement the hidden state of RNN in order to predict the visit at the next time. The relationships between all previous visits and the current one are not taken into account. A diagnosis prediction model, using attention-based bidirectional recurrent neural networks for learning low-dimensional representations of the patient visits, was proposed in [21]. The approach embeds the high dimensional clinical variables into a low dimensional space, and generates the hidden state of a RNN building a code representation through an attention-based bidirectional RNN.

Medical text classification is a special case of a classification task for health data. In this context, natural Language Processing (NLP) and Machine learning algorithms have been applied obtaining promising results. For example, classification of patient record notes were successfully executed by applying Latent Dirichlet Allocation and Support Vector Machines [22], while [23] addresses the classification task of documents in diabetes diseases, also reporting satisfying results. In general, models proposed in the literature, often, exploit neural network approaches for document classification. Mikolov in [24] introduces *Word2Vec*, a two layers neural network that elaborates a text corpus in order to identify correlation among words, also capturing the semantic context. A unique real-valued vector is generated to represent each word included in the corpus. The approach is extended in Le *et al.* in [25], where, starting from *Word2Vec*, the distributed representations of documents or paragraphs is proposed. Kalchbrenner *et al.* [26] proposed an approach to semantically model sentences by exploiting a Dynamic Convolutional Neural Networks. The approach manages sentences of different length and provides a feature graph for each sentence that is able of capturing short and long relations.

Srinivasu *et al.* [27] proposed an approach to classify skin diseases from the image captured from mobile devices, by exploiting a deep learning based method and Long Short Term Memory technique. The model relies on the design of the application through which the image of the affected region of the skin is captured to determine the class of the skin disease. In [28], a cervical cancer prediction model (CCPM) that offers early prediction of cervical cancer using risk factors as inputs, was proposed. In particular, an outlier detection approach and a density-based spatial clustering with noise method, with synthetic minority over sampling technique for data balancing, and random forest for cervical cancer prediction based on risk factors, have been exploited. The novelty is to combine, to improve the prediction performance, data oversampling techniques, an outlier detection methods and random forest classifier for cervical cancer prediction based on risk factors. [29] proposes a personalized healthcare system to monitor diabetic patients by using real-time processing of data from BLE-based sensors. Apache Kafka is utilized to handle incoming sensor information, while MongoDB is exploited for storing unstructured sensor data. The large amount of continuous data (e.g. weight, blood pressure, heart rate, BG,etc.) coming from sensor devices can be managed in real-time through real-time data processing. To classify the diabetes patient, a Multilayer Perceptron classification algorithm is utilized, while the prediction of the BG level relies on Long Short-Term Memory (LSTM) method.

III. THE AI-DRIVEN CLINICAL DECISION SUPPORT SYSTEM

A general and innovative distributed CDS architecture, aiming to support physicians in formulating the diagnosis, was proposed in [2]. The architecture includes a set modules allowing for the collection and management of huge volumes of heterogeneous clinical data originated by distributed sources. Different and non-homogeneous data are involved in the framework including more traditional health data like EHRs, laboratory test data, symptoms of patients, patient health logs, medical images, but also alternative data sources like social media data and data coming from wearable devices. Indeed, with the current revolution of the internet and social media, and the pervasiveness of IoT devices, new sources of medical information are easily accessible. Wearable technologies detect, analyze and transmit information concerning vital signs and/or ambient data, allowing for continuous monitoring of health status. Thanks to advances in artificial intelligence methods, such data can be used in a large set of medical applications, to make a diagnosis or perform a triage.

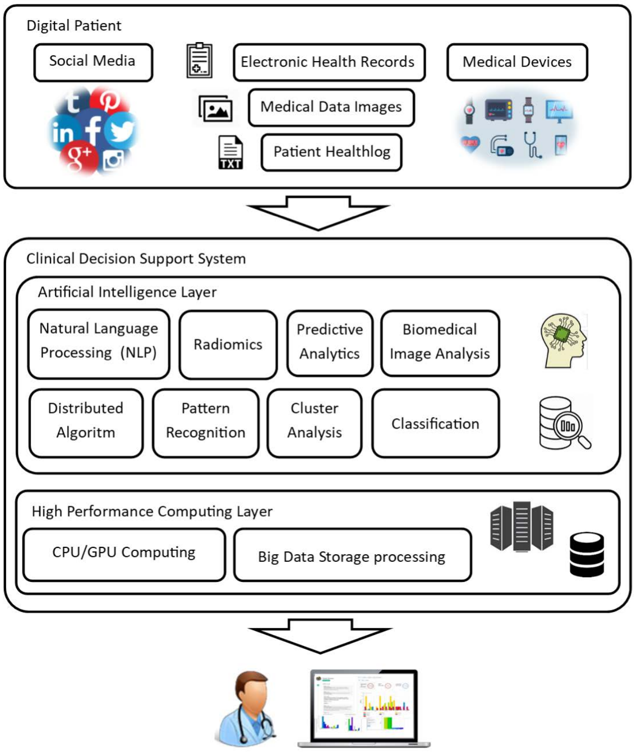

Digital Patient
Social Media
Electronic Health Records
Medical Devices
Medical Data Images
Patient Healthlog
TXT
Clinical Decision Support System
Artificial Intelligence Layer
Natural Language Processing (NLP)
Radiomics
Predictive Analytics
Biomedical Image Analysis
Distributed Algoritm
Pattern Recognition
Cluster Analysis
Classification
High Performance Computing Layer
CPU/GPU Computing
Big Data Storage processing

FIGURE 1. The clinical decision support system architecture.

The proposed Clinical Decision Support framework, whose architecture is shown in Figure 1, relies on a set of cooperative Clinical Data Repository (CDR) hosts, which are geographically distributed healthcare providers (e.g., hospitals or health research centers). Each CDR manages its local information system composed of EHRs, knowledge bases, clinical databases, etc. Essentially a collection of metadata containing information related to the clinical history of each single patient. For the decision making process, the architecture integrates off-line standardized knowledge bases from domain experts and Clinical Practice Guidelines (CPG) with online knowledge that is extracted continuously from EHR databases and social media data. Based on its local knowledge bases and a patient profile, the CDS provides diagnosis suggestions relating to an hospital admission. A patient profile is modeled with consideration of patient medical history and current diagnosis.

The proposed framework relies on the concept of digital patient that is a specialization of the digital twin definition. The concept of digital twin was first used in 2001 at the University of Michigan and described as the virtual and digital equivalent of a physical product. According to [30], a digital twin is a digital replica of a living or non-living physical entity, like processes, people, places, systems and devices that can be used for various purposes. Ideally, it contains all the information of the physical object through a three-dimensional representation of its mechanical, geometric and electronic aspects, i.e. embedded software, micro software, product data, data associated with sensors and actuators, increasingly pervasive. The affirmation of this concept is related to the growing diffusion of the IoT, the Cloud, mobile technologies, and AI that have made the advantages associated with digital twins accessible to many sectors, from smart fabrics to smart agriculture, from smart health to smart cities, and many others aiming to ride digital transformation. In the healthcare industry was originally proposed for equipment prognostics[31]. The patient's digital twin is forged through different health data sources like image records, in-person measurements, laboratory results, and genetic. It is meant to assist during diagnosis and to build personalized models for patients, continuously adjustable based on tracked health and lifestyle parameters. Accordingly, a digital patient simulates the health status of the patient, as captured from available clinical data, and infers the missing parameters from statistical models.

In our approach, the digital patient can benefit also from non-clinical data like the ones from social media and wearable sensors to support physicians in clinical decision-making. In particular, the continuous patient data collection, allows to detect symptoms at early stages, given doctors the capacity to diagnose the patient before getting ill. Besides, during treatment, it will be able to evaluate if the treatment is being effective. Also, it can support therapy planning with individual quantitative optimization of clinical output or predicting disease propagation.

The architecture of the framework is composed of three layers, as can be observed in Figure 1. The first level is responsible for preparation (i.e. cleaning, handling and integration) of medical data generated from multiple heterogeneous sources. The artificial intelligence (AI) layer is in charge to provide integrated health information and intelligent applications, which can improve the efficiency of the decision-making processes. It includes machine learning and deep learning modules, like Natural language Processing (NLP) module, radiomics module, predictive analytics, and so on. Such modules allow to automatically generate context-based and rich representation of health-related information and implement a set of high level clinical services, including systems that provide diagnosis based on patient symptoms similarity, personalized medical measurement, treatment prediction and monitoring.

A High Performance Computing layer, based on CPU/GPU infrastructure, allows the platform to handle large amounts of data in a limited processing time. The module exploits a vector representation of CDR documents, defining a distance/similarity metric, in order to achieve a logical sorting among them. The distance/similarity between two documents can be computed through the cosine distance/similarity between their vector representations. Given two health documents vectors, v_1 and v_2, the cosine measure utilized to compute the similarity between them is reported in formula 1.

$$cos(\vec{v_1}, \vec{v_2}) = \frac{\vec{v_1} \cdot \vec{v_2}}{|\vec{v_1}| \times |\vec{v_2}|} \tag{1}$$

The aim is to obtain a sorted virtual list among CDR hosts, according to which, each server is logically linked to only other two hosts: the CDR host with the vector value immediately lower and the CDR host with the vector value immediately higher of all network. Each CDR host H_{CDR} with vector v_{CDR} performs the following algorithm:

- compute L_{CDR} and H_{CDR} lists containing the currently linked servers with vector value lower and higher than v_{CDR}, respectively;
- if L_{CDR} length is higher than 1, (i) identify in the list the hosts with the minimum and the sub-minimum vector values; (ii) create a virtual link between them;
- if H_{CDR} length is higher than 1, (i) identify the hosts with the maximum and the sub-maximum vector values; (ii) create a virtual link between them;
- notify by message all involved CDR hosts (the host with the minimum/maximum vector value and the host having the sub-minimum/sub-maximum vector value) the information related to their new virtual neighbors; addressed CDR hosts update its lists with the information end considerate these new neighbors in the successive computation.
- remove in L_{CDR} the CDR host with the minimum vector value and in H_{CDR} the CDR host with the maximum vector value;

At the end of the procedure, the last CDR host contained in the list L_{CDR}, represents the linked CDR host with the highest vector value among all linked CDR hosts with the vector value lower than v_{CDR}; while, the H_{CDR} list contains the linked CDR host with the lowest vector value among all linked CDR hosts with the vector value higher than H_{CDR}. Obviously, the host having the vector with the absolute minimum value and the absolute maximum value of all vectors in the network, will be linked to a unique CDR host. If not exist any linked CDR host with vector higher/lower than v_{CDR}, i.e. H_{CDR}/L_{CDR} is empty, H_{CDR} represents the CDR host with the highest/lowest vector value of the whole network.

IV. DIAGNOSIS PREDICTION CASE STUDY: PROBLEM FORMULATION AND METHOD

Electronic Health Records (EHR), consisting of longitudinal patient health data, including demographics, diagnoses, procedures, and medications, have been utilized successfully in several predictive modeling tasks in healthcare. One critical task is to predict the future diagnoses based on patient's historical EHR data.

In this paper, we address the problem of leveraging EHR patient data to infer the discharge diagnosis of patients. We introduce an automated method that exploits basic patient-specific information gathered at admissions, including medical history, symptoms, and preliminary diagnoses, to identify similar patients and predict patient discharge diagnosis.

The problem addressed in the paper is formulated as follows:

Definition 1: Given a patient p, a set of symptoms s felt by the patient and a set of preliminary diagnosis pd, the objective is to predict the discharge diagnosis d for p by exploiting the similarity in terms of symptoms and preliminary diagnoses with other patients already treated and for which a discharge diagnosis has been already formulated.

In particular, we propose a novel patient-similarity-based framework for diagnostic prediction, where the similarity is defined as the similarity of diagnoses and symptoms among patients. The multilabel classification problem is converted to a single-value regression problem by integrating the pairwise patients' clinical features into a vector and taking the vector as the input and the patient similarity as the output.

It is worth pointing out that, differently to previous approaches, the diagnosis prediction problem is addressed differently since the aim of the work proposed in this paper is to predict the exact diagnosis in terms of semantic meaning. In fact, instead of referring to the diagnosis in terms of the International Classification of Diseases (ICD-9) codes we consider the diagnosis semantics by exploiting Natural Language Processing. This way we are able to implement semantic enhanced diagnosis prediction. In particular, we exploit word embedding models since neural networks shown to have the ability to learn complex semantic feature representations.

A. DATA MODEL

In the proposed model the digital patient is built by combining traditional medical sources like EHR and external knowledge bases like social media and sensors data. For each of such sources relevant data features are identified and extracted. The features from the disparate sources are then represented in ad hoc data structures introduced in the following of the Section. Specifically, in the proposed approach the symptoms, lab tests and preliminary diagnoses are integrated into the feature vectors of patients. The objective of the feature construction effort is to capture sufficient clinical nuances from the heterogeneity of data of the different patients. A major challenge is in data reduction and in summarizing the temporal event sequences in EHR data into features that can differentiate patients. The adopted feature-centered framework serves as the basis for implementing different similarity-based diagnoses prediction algorithms. Essentially, each patient is represented by a feature vector, which serves as the input to the similarity measure. Our objective is to design a similarity measure that operates on patient feature vectors and that it is consistent with physician feedback in terms of whether two patients are clinically similar or not. Accordingly, the semantic similarity metrics is a key aspect in the disease prediction approach.

Even if the architecture shown in Figure 1 shows that we include disparate and heterogeneous medical data sources, in this paper we focus only on classical Electronic Health Records (EHR).

To the aim of this work we refer to the well-known MIMIC-III database [3] and, thus, we rely on its data structures and model, as will be detailed in Section V. In particular, we refer to the database schema and to the data structures maintaining information about the target features that in the faced problem are the diagnosis and the symptoms. Specifically, since the symptoms are not explicitly specified in the database, an ad-hoc extraction module has been implemented and for each admission of a patient a list of symptoms is extracted. Differently, preliminary diagnoses are explicitly modeled in the database; they are established after hospital admissions and recorded in the admission notes. Preliminary diagnoses are tentative hypotheses and may be not satisfactory. Conversely, discharge diagnoses are achieved after treatment, confirmed by the doctors and extracted from the discharge summary notes. They are accurate and definite. Preliminary and discharge diagnoses are categorical and unintelligible for a computer. Several categorical clinical concepts have been classified into hierarchical taxonomies, such as the International Classification of Diseases (ICD-9). In the MIMIC-III database discharge diagnoses are expressed in terms of ICD-9 code, while preliminary diagnoses recorded in the admission notes are not encoded with ICD-9.

The aim of the proposed diagnosis prediction method is to infer the exact semantic meaning of the diagnosis since adopting diagnosis codes like ICD-9 code, an approximation can be introduced due to the categorization step. Accordingly, in our model a diagnosis is a list of words.

We define a patient entry as a data structure storing relevant features about patients, symptoms and diseases:

Definition 2: A patient entry p is defined as a tuple $p = (id, s, pd, ssv)$, where id is the patient identifier, s is a list of symptoms $s = (s_1, s_2, \ldots, s_s)$ where each element s_i corresponds to a symptom and consists of a list of words describing the symptom s_i, pd is the list of preliminary diagnoses that similarly to the symptoms is represented as a list of lists $pd = (pd_1, pd_2, \ldots, pd_d)$, and ssv is the semantic symptom vector corresponding to the list of symptoms s and preliminary diagnoses pd.

Definition 3: A semantic symptom vector $ssv = (ssv_1, ssv_2, \ldots, ssv_s)$ is a vector where each element in turn is an n-dimensional vector in a semantic space obtained through word embedding of each single symptom and preliminary diagnosis. Each feature item is represented as a real-valued vector, and each dimension contains a certain amount of semantic information.

We learn patient representation from EHR by developing a supervised machine learning model that allow an automatically generated context-based and rich representation

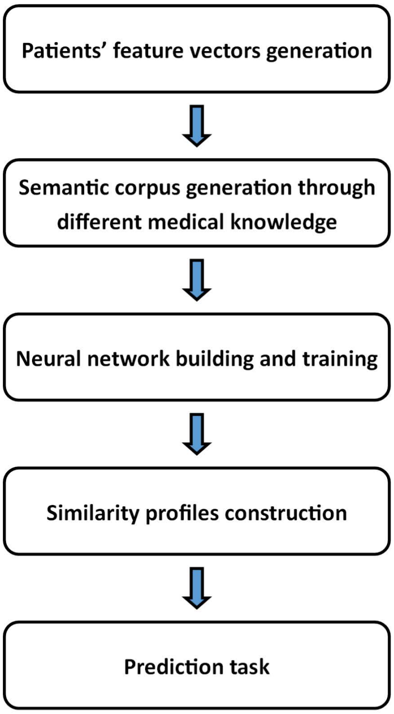

Patients' feature vectors generation
Semantic corpus generation through different medical knowledge
Neural network building and training
Similarity profiles construction
Prediction task

FIGURE 2. The flowchart of the prediction method proposed.

of health related information. In particular, we apply a word embedding approach to categorize text fragments, at a sentence level, based on the emergent semantics extracted from a corpus of medical text. With the embedding approach, symptoms and diagnoses are represented as vectors in real space and trained using artificial neural networks. Usually, the vector representations of words are computed using their contexts so that words with similar meanings will have similar vector representations. Sentence embedding maps sentence to fixed length dense vectors and can be generated from the aggregation of embeddings of words in the sentence.

As sentence embedding model we used Sent2Vec [32], an unsupervised model allowing to compose sentence embeddings using word vectors along with n-gram embeddings, simultaneously training composition and the embedding vectors themselves. Conceptually, the model can be interpreted as a natural extension of the word-contexts from the C-BOW approach [25] to a larger sentence context, with the words being specifically optimized towards additive combination over the sentence, by means of the unsupervised objective function.

B. METHOD

The supervised prediction method proposed to compute discharge diagnosis similarity of patients is formulated as a sequence of steps, as showed in Figure 2. The input of the method are the feature vectors of each patient, which are built by integrating two parts: symptoms and preliminary diagnosis data. Accordingly, the first step is the construction of the patient feature vectors. The second step is the development of the semantic corpus by integrating the different medical knowledge. The third step is the building of the neural network and its training. The forth step is the construction of the similarity profiles, while the final step is the prediction task.

We propose a distance-based similarity approach exploiting symptoms and preliminary diagnoses similarity. Given a set of historical admissions, the goal of the proposed algorithm is to predict the diagnosis of a new patient admission based on the similarity of symptoms and preliminary diagnoses with those of the historical admissions.

The semantic similarity metrics is formulated by exploiting the cosine similarity between the semantic symptoms vectors of the target patient p and that of another patient $p_k$in the historic EHR. The similarity is computed by taking for each item in the semantic symptom vector of p the maximum similarity value with the semantic feature of patient p_k. In other words, for each item in the ssv^p the semantically most similar term in the $ssv_i^{p_x}$ of patient p_k is determined.

Let $n =| ssv^p |$, $m =| ssv^{p_k} |$ be the size of the semantic symptom vectors of the target patient p and of the patient p_k, respectively. Then, the semantic similarity between them is computed through the following function:

$$sim_{sem}(p, p_k) = \frac{\sum_{x=1}^{n} \underset{y\in\{1,\ldots,m\}}{\overset{x}{\arg\max}}\; cos(ssv^p(x), ssv^{p_k}(y))}{\arg\max(n, m)} \quad (2)$$

where the arg max function gives the highest cosine similarity value computed among the $n \times m$ couples (x, y) of features of ssv^p and ssv^{p_k}.

1) COMPLEXITY

Two main trends in NLP have emerged. Recurrent neural networks (RNNs), LSTMs, attention models are widely used for NLP applications. However, even if they exhibit extreme strong expressiveness, the increased model complexity makes such models much slower to train on larger datasets. On the other end, simpler approaches like matrix factorizations or bilinear models can be successfully trained on much larger datasets, which is an important advantage, in the unsupervised setting considered. In particular, is important to note that for constructing sentence embeddings naively using averaged word vectors was shown to outperform LSTMs (see Wieting *et al.* [33] for plain averaging, and Arora *et al.* [34] for weighted averaging). This example shows potential in exploiting the trade-off between model complexity and ability to process huge amounts of text using scalable algorithms. In view of this tradeoff, the adopted approach of Pagliardini et al [35]. further advances unsupervised learning of sentence embeddings. The Sent2Vec model proposed by Paglairdini *et al.* exploits, a simple unsupervised model allowing to compose sentence embeddings using word vectors along with n-gram embeddings, simultaneously training composition and the embedding vectors themselves. Pagliardini et al demonstrated that the empirical performance

TABLE 1. Summary statistics of MIMIC-III dataset.

Data	Total
# admissions in the MIMIC-III (v1.4) database	58,576
# admissions which are the first admission of the patient	46,283
# admissions which are the first admission of an adult patient (> 15 years old)	38,425
# admissions where adult patient died 24 h after the first admission	35,627

of their proposed general-purpose sentence embeddings very significantly exceeds the state of the art, while keeping the model simplicity as well as training and inference complexity exactly as low as in averaging methods ([33], [34]). Accordingly, the computational complexity of our embeddings is only O(1) vector operations per word processed, both during training and inference of the sentence embeddings. This strongly contrasts all neural network based approaches, and allows our model to learn from extremely large datasets, in a streaming fashion, which is a crucial advantage in the unsupervised setting.

Fast inference is a key benefit in downstream tasks and industry applications. In contrast to more complex neural network based models, one of the core advantages of the proposed technique is the low computational cost for both inference and training. Given a sentence S and a trained model, computing the sentence representation vS only requires |S| · h floating point operations (or |R(S)| · h to be precise for the ngram case), where h is the embedding dimension.

V. DATASET AND SEMANTIC KNOWLEDGE CORPUS

A. MIMIC-III DATASET

MIMIC III [3], Multiparameter Intelligent Monitoring in Intensive Care, is a publicly available critical care dataset developed by the MIT Lab. This database integrates deidentified, comprehensive clinical data of patients admitted to an Intensive Care Unit (ICU) at the Beth Israel Deaconess Medical Center (BIDMC) in Boston, Massachusetts during 2001 to 2012. It includes demographics, vital signs, laboratory tests, medications, and more. MIMIC-III contains data associated with 53,423 distinct hospital admissions for adult patients (aged 15 years or above) and 7870 neonates admitted to an ICU at the BIDMC. The data covers 38,597 distinct adult patients with 49,785 hospital admissions. Table 1 shows the statistics of the dataset.

The dataset version used in this research is MIMIC-III is v1.4, released on 2 September 2016 [3].

The tools used in this study were PostgreSQL and Python. PostgreSQL was used as the database management system which allowed SQL-based queries though PgAdmin4 to extract and select data from MIMIC-III database. Python was used to create more structured way of data processing.

B. SEMANTIC CORPUS MODEL: BioSentVec

We trained the word embedding on the BioSentVec corpus, the first open set of sentence embeddings trained with over 30 million documents from both scholarly articles in PubMed and clinical notes in the MIMIC-III Clinical Database. BioSentVec is publicly available at https://github.com/ncbi-nlp/BioSentVec.

TABLE 2. Summary statistics of BioSentVec corpus.

Source	Documents	Sentences	Tokens
PubMed	28,714,373	181,634,210	4,354,171,148
MIMIC III Clinical notes	2,083,180	41,674,775	539,006,967

BioSentVec implements the sentence to vector model (sent2vec) It is actually an unsupervised version of FastText, and an extension of word2vec (CBOW) to sentences. Accordingly, BioSentVec is a pre-trained model for readily generating sentence embeddings given as input any arbitrary sentence. In benchmarking, BioSentVec shows superior performance on two public datasets for computing sentence similarity, compared to the current state of the art.

Specifically, BioSentVec is created by applying sent2vec to both biological and clinical texts at a large scale, to compute the 700-dimensional sentence embeddings, using the bigram model and set window size to be 20 and negative examples 10.

VI. EXPERIMENTAL EVALUATION

In this section we present the results of the experimental evaluation performed with a twofold goal: one was assessing the effectiveness and accuracy of the proposed diagnosis prediction approach; the other goal was to asses the scalability of the distributed discovery services. We implemented the Disease Diagnosis method exploiting Cython,[1] an optimizing static compiler for Python programming. The source code is available at http://staff.icar.cnr.it/diseaseDiagnosis.zip.

A. DIAGNOSIS PREDICTION METHOD PERFORMANCE

In this Section we show the performance of the proposed approach in predicting patient diagnoses based on symptoms similarities. We consider the following performance metrics:

- *Precision:* $P = TP/(TP + FP)$
  It is the fraction of the number of successfully detected ground-truth diagnoses (TP: true positive), out of the total number of detected diagnoses (TP + FP). FP (false positive) is the number of mistakenly detected ground-truth diagnoses. A ground-truth diagnosis is considered successfully detected if exists a predicted diagnosis that matches it.
- *Recall:* $R = TP/(TP + TN)$
  It is the percentage of ground truth diagnosed successfully detected, where TP + TN (TN: True Negative) is the total number of ground-truth diagnoses.
- *F-Measure:* $FM = 2RP/(R + P)$
  It is the harmonic mean of precision and recall metrics, that reaches its best value at 1 and worst at 0.

In our experiments we are interested in calculating precision and recall at k. In particular, we take the top k predictions with higher similarity. If one of them matches with the ground-truth diagnosis, the method classifies the prediction as correct. Top-1 accuracy is a special case, in which only the highest similarity is taken into account for the prediction. The data consists of 50.870 hospital admissions. Different types of admissions are considered: 'ELECTIVE', 'URGENT',

[1] https://cython.org/

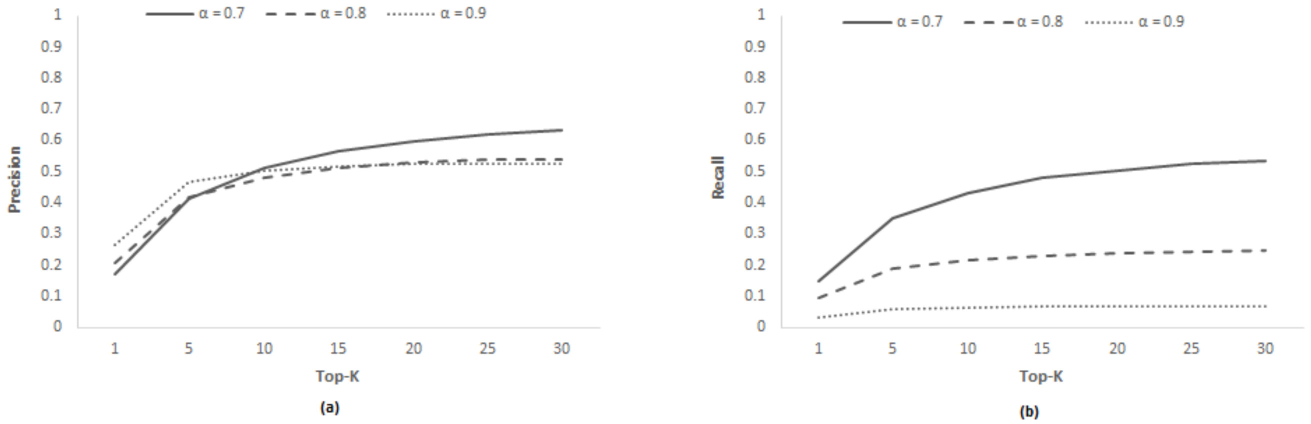


FIGURE 3. Precision (a) and Recall (b) with respect to symptoms similarity threshold α on 2000 predictions.

or 'EMERGENCY'. Emergency/urgent indicate unplanned medical care, while elective indicates a previously planned hospital admission. Each set of experiments is performed considering 5-fold cross validation. For each fold, the dataset was randomly splitted into training (80%) and test (20%) sets, then a model is trained using a training data and the resulting model is validated on a test set. The performance measures reported by the 5-fold cross validation are then the average of the values computed at each fold. In a first set of experiments we aimed to identify the best parameter settings for the prediction algorithm. The parameters are α, the symptoms similarity threshold, and β, the diagnoses similarity threshold. Thus, we considered different values of α and β and computed the precision and recall metrics.

Figure 3 shows precision and recall at $k = \{5, 10, 15, 20, 25, 30\}$ obtained with respect to the symptoms similarity threshold $\alpha = \{0.7, 0.8, 0.9\}$. The graph shows that, as expected, both precision and recall have an increasing trend as the number of top predicted diagnoses increases. From the graph one can note that the recall got lower values for $\alpha = 0.8$ and $\alpha = 0.9$. This happens because by increasing the similarity threshold between the symptoms, the similarity constraint becomes more stringent, and consequently the number of successfully predicted diagnoses decreases. According to these results, the best configuration for symptoms similarity threshold is $\alpha = 0.7$. Configuration that will be used throughout the paper. For this first experiment is considered the semantics similarity threshold $\beta = 1$, setting that implies a semantic correspondence at 100% between the predicted and ground truth diagnosis, that in a real-life scenario is unlikely. By slightly lowering the semantic similarity between the predicted and ground truth diagnosis (as will be shown in Figure 5) the performance of the approach greatly increases.

In Figure 4 is shown the number of correct predictions at $k = \{5, 10, 15, 20, 25, 30\}$ with respect to the dataset size. The graph exhibits how the fraction of correctly detected ground truth diagnoses increases with the size of the dataset and with the number of top predicted diagnoses.

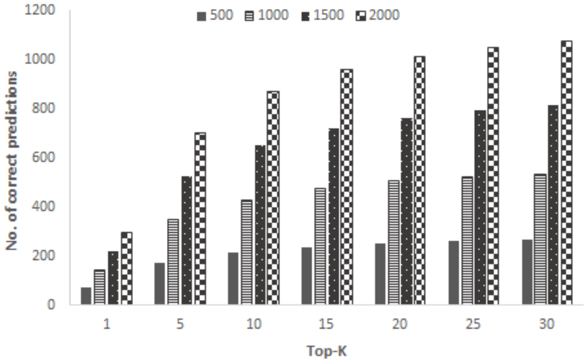
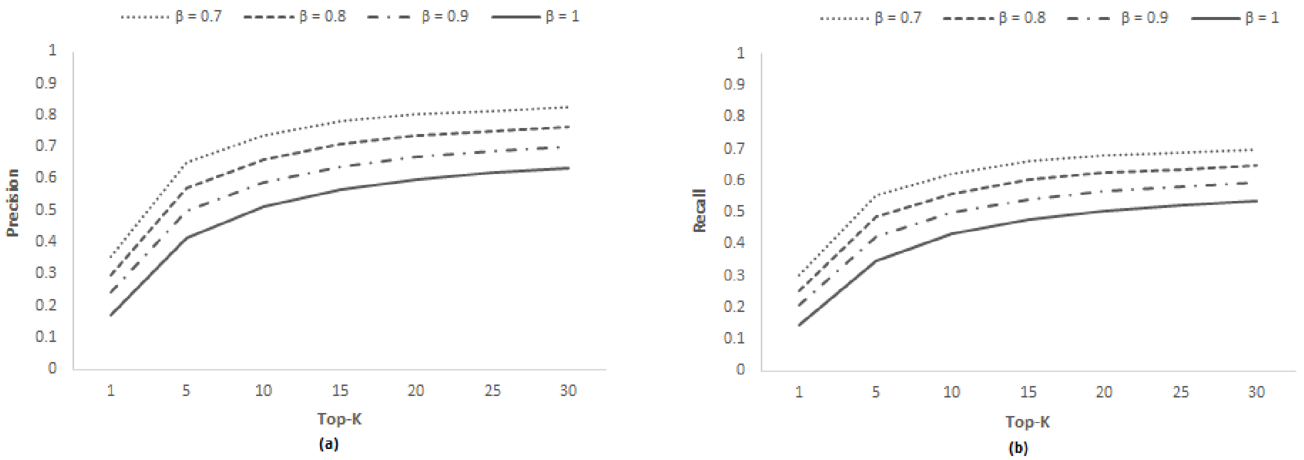


FIGURE 4. Number of correct predictions with respect to dataset size.

As discussed above, another key aspect of the proposed prediction method is the diagnoses similarity threshold β, that exploits diagnosis semantics and not just textual similarities between predicted and ground truth diagnoses. In particular, we use, similarly to the symptoms, the sentence embedding techniques to manage the diagnosis semantic information as vectors in the semantic embedding space.

The MIMIC-III dataset maintains diagnosis related groups (DRG) codes for each admission. More precisely, a final discharge diagnosis can consist of several terms based on the number of DRG codes associated with admission. DRG codes represent the diagnoses billed for by the hospital. There are three types of DRG codes in the database which have overlapping ranges but distinct definitions for the codes. The three types of DRG codes are 'HCFA' (Health Care Financing Administration), 'MS' (Medicare), and 'APR' (All Payers Registry). HCFA-DRG and MS-DRG codes have multiple descriptions as they have changed over time. Sometimes these descriptions are similar, but sometimes they are completely different diagnoses. So we need to consider both the type and the description for a certain diagnosis. All admissions have an HCFA-DRG or MS-DRG code, but not all admissions have an APR-DRG code. Note that APR-DRG is believed to be an alternative,



(a)

(b)

FIGURE 5. Precision (a) and Recall (b) with respect to diagnoses similarity threshold β.

more specific, code which could be used in conjunction with the HCFA codes. Consequently, each term is a pair $< DRG_{TYPE}, DRG_{DESCRIPTION} >$, and each diagnosis is a set of terms. Given a ground truth diagnosis gtd and a predicted diagnosis pd, first, we extract their vectors through the sentence embedding method, then we calculate the similarity between them.

Figure 5 shows precision and recall at $k = \{5, 10, 15, 20, 25, 30\}$ obtained with respect to the diagnoses similarity threshold $\beta = \{0.7, 0.8, 0.9, 1\}$. The graph displays that, precision and recall increase with k, while decrease with the increasing of β. This happens because $\beta = 1$ indicates that pd and gtd are identical. By decreasing beta, the similarity constraint is loosened and the prediction is considered correct even if the diagnoses are not the same but are similar at the β level. This approach allows us to make a greater number of predictions with a high level of similarity between diagnoses. We can conclude that the proposed diagnosis prediction method based on word embedding and exploiting semantic similarity of both symptoms and diagnoses, resulted to be effective and accurate reaching considerable precision and recall rates. The obtained outcomes are promising for future developments and extensions of the framework that could be a valuable means for automatic inferring disease diagnosis.

B. DISTRIBUTED DISCOVERY SERVICE PERFORMANCE

With the aim to test the effectiveness of the system we implemented a Java simulator in which the characteristics of real networks of CDR hosts were careful considered. Firstly, the mean number of messages exchanged by each server, i.e. the traffic generated by the algorithm to obtain a stable and ordered situation, was evaluated. Figure 6 shows, for different network size, the average number of messages handled by each server, that is for different number of clinical serves involved in the logical reorganization. Different mean number of connections/neighbors of each server were considered in the experiments. We can note that, the algorithm achieve a stable overlay using a limited number of messages and therefore generating a low value of network traffic.

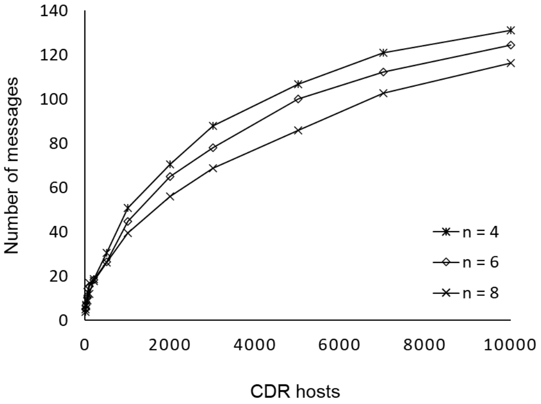


FIGURE 6. The average number of messages handled by each CDR host to obtain the organization, for different network size.

Figure 7 reports the total number of messages exchanged by all servers per each step to reaches a stable situation. The network size is set to 5000. Notice that the algorithm converges in a finite number of steps and the number of messages decreases exponentially.

In order to evaluate the worst case for the organization, i.e. the maximum number of steps required to each CDR host to create the final links to its neighbors in the overlay, a set of experiments was performed and the results are reported in Figure 8. Even in this case, the simulations were executed for different network sizes and for different mean values of neighbors. It can possible to highlight that the maximum number of steps required to a server to identify its neighbors in the overlay is, in any case, very low. Moreover, we can highlight that the maximum number of steps to reach the sorting would be necessary only for the first running because none previous order exists and the system is completely

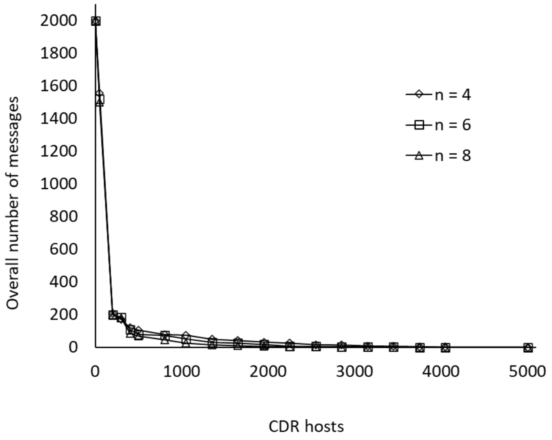


FIGURE 7. The overall number of messages handled by the network for each step. The number of CDR hosts is set to 5000.

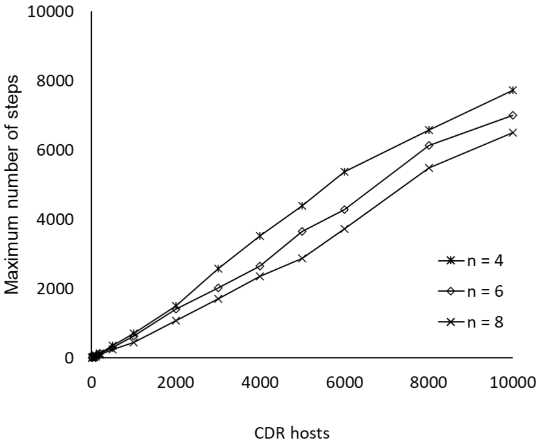


FIGURE 8. The "worst case", that is the maximum value of possible steps necessary to obtain the logical sorting.

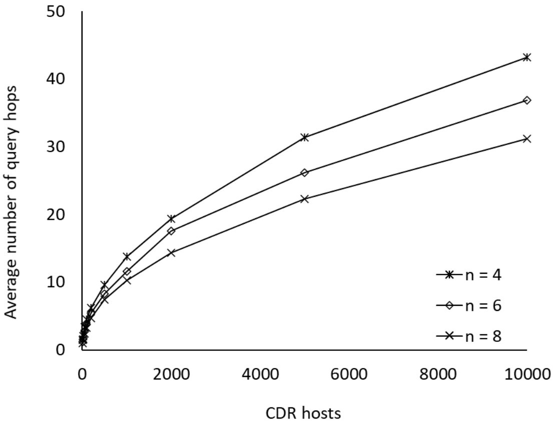


FIGURE 9. Mean number of query hops to locate the objective.

disordered. An intuitive informed search mechanism can be designed, simply by forwarding (at each step) the query towards the CDR host with the highest similarity value with the target. The procedure assures that, at each step, the current resource is that with the highest similarity with the target vector. The search procedure finishes when none of the linked servers improve the similarity value with respect to the current CDR host. Figure 9 reports the mean number of query steps needed to locate the best resource for different number of linked CDR hosts. Notice that the number of steps is always limited also for high numbers of CDR hosts involved.

VII. CONCLUSION

In this paper, an artificial intelligence driven Clinical Decision Support system is proposed. The system is able to integrate heterogeneous health data from different sources, and implements a set of intelligent services exploiting innovative machine learning and deep learning approaches to support physicians in disease diagnosing and treating. In particular, the paper presented a neural network model for predicting patients' future health information. The model is based on patients similarity in terms of symptoms and diseases. The approach employs word embedding to model the semantic relations of symptoms and diagnoses, and it introduces a mechanism to measure the semantic relationship of different diagnoses in terms of symptoms similarity for the prediction. Experimental results, performed on a real-world EHR dataset, shown that the proposed approach is effective and accurate and provides clinically meaningful interpretations.

REFERENCES

[1] A. D. Black, J. Car, C. Pagliari, C. Anandan, K. Cresswell, T. Bokun, B. McKinstry, R. Procter, A. Majeed, and A. Sheikh, "The impact of eHealth on the quality and safety of health care: A systematic overview," *PLoS Med.*, vol. 8, no. 1, Jan. 2011, Art. no. e1000387.

[2] C. Comito, A. Forestiero, and G. Papuzzo, "Exploiting social media to enhance clinical decision support," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. (Companion)*, New York, NY, USA, Oct. 2019, pp. 244–249.

[3] A. E. W. Johnson, T. Pollard, L. Shen, L.-W. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Celi, and R. Mark, "MIMIC-III, a freely accessible critical care database," *Sci. Data*, vol. 3, p. 160035, May 2016.

[4] E. Choi, N. Du, R. Chen, L. Song, and J. Sun, "Constructing disease network and temporal progression model via context-sensitive hawkes process," in *Proc. IEEE Int. Conf. Data Mining*, Nov. 2015, pp. 721–726.

[5] X. Wang, D. Sontag, and F. Wang, "Unsupervised learning of disease progression models," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, Aug. 2014, pp. 85–94, doi: 10.1145/2623330.2623754.

[6] H. Xiao, L. Vu, and D. Turaga, "Learning temporal state of diabetes patients via combining behavioral and demographic data," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2017, pp. 2081–2089.

[7] J. Zhou, L. Yuan, J. Liu, and J. Ye, "A multi-task learning formulation for predicting disease progression," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, New York, NY, USA, 2011, pp. 814–822, doi: 10.1145/2020408.2020549.

[8] E. Choi, M. T. Bahadori, E. Searles, C. Coffey, M. Thompson, J. Bost, J. Tejedor-Sojo, and J. Sun, "Multi-layer representation learning for medical concepts," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, Aug. 2016, pp. 1495–1504.

[9] E. Choi, M. Bahadori, J. Kulas, A. Schuetz, W. Stewart, and J. Sun, "Retain: An interpretable predictive model for healthcare using reverse time attention mechanism," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2016, pp. 3512–3520.

[10] Q. Suo, F. Ma, G. Canino, J. Gao, A. Zhang, P. Veltri, and G. Agostino, "A multi-task framework for monitoring health conditions via attention-based recurrent neural networks," in *Proc. AMIA Annu. Symp.*, Jan. 2017, pp. 1665–1674.

[11] J. Zhou, J. Sun, Y. Liu, J. Hu, and J. Ye, "Patient risk prediction model via top-K stability selection," in *Proc. SIAM Int. Conf. Data Mining*, May 2013, pp. 55–63.

[12] Z. Che, D. Kale, W. Li, M. T. Bahadori, and Y. Liu, *Deep Computational Phenotyping*. New York, NY, USA: Association for Computing Machinery, 2015, pp. 507–516.

[13] P. Jensen, L. Jensen, and S. Brunak, "Mining electronic health records: Towards better research applications and clinical care," *Nature Rev. Genetics*, vol. 13, no. 6, pp. 395–405, 2012.

[14] C. Liu, F. Wang, J. Hu, and H. Xiong, *Temporal Phenotyping From Longitudinal Electronic Health Records: A Graph Based Framework*. New York, NY, USA: Association for Computing Machinery, 2015, pp. 705–714, doi: 10.1145/2783258.2783352.

[15] F. Ma, C. Meng, H. Xiao, Q. Li, J. Gao, L. Su, and A. Zhang, "Unsupervised discovery of drug side-effects from heterogeneous data sources," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, Aug. 2017, pp. 967–976, doi: 10.1145/3097983.3098129.

[16] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. NIPS*. Red Hook, NY, USA: Curran Associates, 2013, pp. 3111–3119.

[17] Y. Cheng, F. Wang, P. Zhang, and J. Hu, "Risk prediction with electronic health records: A deep learning approach," in *Proc. SIAM Int. Conf. Data Mining*, Jun. 2016, pp. 432–440.

[18] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Sci. Rep.*, vol. 8, no. 1, p. 6085, Dec. 2018.

[19] Z. C. Lipton, D. C. Kale, and R. Wetzel, "Modeling missing data in clinical time series with RNNs," 2016, *arXiv:1606.04130*.

[20] E. Choi, M. T. Bahadori, L. Song, W. Stewart, and J. Sun, "GRAM: Graph-based attention model for healthcare representation learning," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2017, pp. 787–795.

[21] F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun, and J. Gao, "Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2017, pp. 1903–1911.

[22] R. Cohen, I. Aviram, M. Elhadad, and N. Elhadad, "Redundancy-aware topic modeling for patient record notes," *PLoS ONE*, vol. 9, no. 2, Feb. 2014, Art. no. e87555.

[23] B. J. Marafino, J. M. Davies, N. S. Bardach, M. L. Dean, and R. A. Dudley, "N-gram support vector machines for scalable procedure and diagnosis classification, with applications to clinical free text data from the intensive care unit," *J. Amer. Med. Inform. Assoc.*, vol. 21, no. 5, pp. 871–875, Sep. 2014.

[24] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. Workshop (ICLR)*, 2013.

[25] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. 31st Int. Conf. Mach. Learn.*, vol. 32, 2014, pp. 1188–1196.

[26] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2014, pp. 655–665.

[27] P. N. Srinivasu, J. G. Sivasai, M. F. Ijaz, A. K. Bhoi, W. Kim, and J. J. Kang, "Classification of skin disease using deep learning neural networks with MobileNet V2 and LSTM," *Sensors*, vol. 21, no. 8, p. 2852, Apr. 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/8/2852

[28] M. F. Ijaz, M. Attique, and Y. Son, "Data-driven cervical cancer prediction model with outlier detection and over-sampling methods," *Sensors*, vol. 20, no. 10, p. 2809, May 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/10/2809

[29] G. Alfian, M. Syafrudin, M. Ijaz, M. Syaekhoni, N. Fitriyani, and J. Rhee, "A personalized healthcare monitoring system for diabetic patients by utilizing BLE-based sensors and real-time data processing," *Sensors*, vol. 18, no. 7, p. 2183, Jul. 2018.

[30] A. El Saddik, "Digital twins: The convergence of multimedia technologies," *IEEE Multimedia Mag.*, vol. 25, no. 2, pp. 87–92, Apr. 2018.

[31] K. Bruynseels, F. S. di Sio, and J. van den Hoven, "Digital twins in health care: Ethical implications of an emerging engineering paradigm," *Frontiers Genet.*, vol. 9, p. 31, Feb. 2018. [Online]. Available: https://www.frontiersin.org/article/10.3389/fgene.2018.00031

[32] M. Pagliardini, P. Gupta, and M. Jaggi, "Unsupervised learning of sentence embeddings using compositional N-gram features," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, vol. 1, 2018, pp. 528–540.

[33] J. Wieting, M. Bansal, K. Gimpel, and K. Livescu, "Towards universal paraphrastic sentence embeddings," in *4th Int. Conf. Learn. Represent. (ICLR)*, 2016, pp. 1–17.

[34] S. Arora, Y. Liang, and T. Ma, "A simple but tough-to-beat baseline for sentence embeddings," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 1–14.

[35] M. Pagliardini, P. Gupta, and M. Jaggi, "Unsupervised learning of sentence embeddings using compositional N-gram features," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, vol. 1, 2018, pp. 528–540, doi: 10.18653/v1/N18-1049.

CARMELA COMITO received the master's degree in computer engineering and the Ph.D. degree in systems and computer engineering from the University of Calabria, Italy. In 2006, she was a Visiting Researcher with the School of Computer Science, The University of Manchester, U.K. In 2017, she was a Visiting Researcher with LIRMM, University of Montpellier, France. She is currently a Researcher with the Institute of High Performance Computing and Networking of the Italian National Research Council (ICAR-CNR), Italy. She is also an Adjunct Professor with the University of Calabria. She coauthored over 80 articles in international journals, conference proceedings, and edited volumes. Her research interests include big data analysis and mining, mobility mining, social network data analysis and mining, and health informatics. She served as the chair, a program committee member and a reviewer of several international conferences.

DEBORAH FALCONE received the master's degree in computer engineering and the Ph.D. degree in systems and computer engineering from the University of Calabria, Italy. In 2013, she was a Visiting Researcher with the Computer Laboratory, University of Cambridge, U.K. She worked as a Research Fellow and a Teaching Assistant in concurrent programming and object oriented programming, with experience in designing, developing and testing distributed applications, mobile applications and a good data science background and experience in social media data analysis. She is currently a Research Fellow with the Institute of High Performance Computing and Networking of the Italian National Research Council (ICAR-CNR), Italy. Her research interests include artificial intelligence, big data analysis and mining, social network data analysis and mining, and health informatics.

AGOSTINO FORESTIERO received the Laurea degree in computer engineering and the Ph.D. degree in computer engineering from the University of Calabria, Cosenza, Italy, in 2002 and 2006, respectively. He is currently a Researcher with the Institute for High Performance Computing and Networking of the CNR, Rende, Italy. He published more than 90 scientific papers on international conferences and journals among which IEEE/ACM Transactions on Networking (TON), IEEE Transactions on Evolutionary Computation (TEVC), IEEE Transactions on Green Communications and Networking (TGCN), IEEE Internet of Things Journal, *Information Sciences*, *FGCS*, and *ACM TAAS*. His research interests include the Internet of Things, cyber-physical systems, pervasive computing, cloud, fog, and edge computing, social mining, artificial intelligence, and cyber security. He serves as a PC member of several conferences and journal.