2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

# Diagnosis Prediction based on Similarity of Patients Physiological Parameters

Carmela Comito
CNR-ICAR
Rende (CS), Italy
carmela.comito@icar.cnr.it

Deborah Falcone
CNR-ICAR
Rende (CS), Italy
deborah.falcone@icar.cnr.it

Agostino Forestiero
CNR-ICAR
Rende (CS), Italy
agostino.forestiero@icar.cnr.it

***Abstract*—Medical staff can be considerably supported in patient healthcare delivery thanks to the adoption of machine learning and deep learning methods by enhancing clinicians decisions and analysis with targeted clinical knowledge, patient information, and other health data. This paper proposes a learning methodology that, on the basis of the current patient health status, clinical history, diagnostic and laboratory results, provides insights for clinicians in the diagnosis and therapy decision processes. The approach relies on the concept that patients with similar vital signs patterns are, in all probability, affected by the same or very similar health problems. Thus, they can have the same or very similar diagnoses. Patients physiological signals are modeled as time series and the similarity among them is exploited. The method is formulated as a classification problem in which an ad-hoc multi-label k-nearest neighbor approach is combined with similarity concepts based on word embedding. Experimental results on real-world clinical data have shown that the proposed approach allows detecting diagnoses with a precision up to about 75%.**

***Index Terms*—Diagnosis Prediction, Patients Similarity, Word Embedding, Time series Analysis**

## I. Introduction

Clinicians knowledge can be significantly enriched by integrating it with suggestions and information from Clinical Decisions Support Systems (CDSS), so providing a great contribution to early detection, disease prevention, more effective treatment. Recently, artificial intelligence (AI) driven CDSS allow to leverage data and observations otherwise unavailable or uninterpretable by humans, by integrating patient data into useful reports for physicians, informing efficient and evidence-based diagnostic and triage decisions. Moreover, having medical information directly on hand, eliminates the need to access multiple databases or files, and healthcare personnel are prepared for each patient discussion or consultation with other specialists. The integration of Electronic health records (EHR), representing the digital version of a patient paper chart, provide several benefits to CDSS for clinical diagnosis. EHRs are characterized by high dimension, heterogeneous and incomplete data, often noisy and containing several missing values.

Data availability for digital clinical analysis has had an explosion thanks to the adoption of EHRs. Nevertheless, the possibility of using this enormous and complex data becomes real thanks to the advances in AI techniques. Using AI techniques allows to exploit many aspects of a patient's health state and generate predictions. Exploiting the ECG, the blood pressure and other vital signs of a patient can be very useful, but much better is to have the possibility to access multiple integrated information at once, e.g. patient's risk for stoke, coronary artery disease, lab test results, family history, socioeconomic status, and latest clinical trial data. This paper proposes a learning methodology based exploiting time series analysis of physiological signals (e.g. heart rate, respiratory rate, and oxygen saturation) and measurements (e.g. hemoglobin, glucose and bun) to improve healthcare delivery by enhancing medical decisions with targeted clinical knowledge, patient information, and other health information. The aim is to provide enriched and advanced information to clinicians so enabling accurate and better decisions about patient diagnoses.

A predictive model based on patients time series similarity allows us to design a *diagnosis prediction* method. In the proposed approach the charted data collected for a patient during its hospital stay is modeled as time series and a patient diagnosis can be predicted according to the evolution of the health status of other similar patients. Diagnosis prediction is framed as a classification problem, combining time series similarity and an ad-hoc multi-label k-nearest neighbor classification approach.

The proposed approach exploits the semantic similarity among diagnoses to model the classification classes to which the diagnoses will be associated. In particular, a sentence embedding technique is employed to catch patients similarity in terms of diagnosis. Data derived from the publicly available MIMIC-III critical care database [10], were exploited for the experimental evaluation of the proposed approach. The dataset contains information related to patients like vital signs, medications, laboratory measurements, observations and notes charted by care providers, procedure codes, diagnostic codes, imaging reports, hospital length of stay, survival data, and



*ASONAM '21*, November 8–11, 2021, Virtual Event, Netherlands

ACM ISBN 978-1-4503-9128-3/21/11...$15.00
http://dx.doi.org/10.1145/3487351.3490962

more. Experimental evaluations showed the effectiveness of the approach that can successfully detect diagnoses with a precision up to about 75%.

The rest of the paper is organized as follows. Section II presents related works. Section III outlines data modeling, including pre-processing and features selection and extraction. In Section IV the diagnosis prediction method is described, while Section V shows the results of the experimental evaluation of the approach. Finally, Section VI concludes the paper.

## II. Related works

Clinical decision making is a complicated task in which the physician must attempt to bridge what has been referred to as an inferential gap between the information at hand in a given case and the clinical knowledge that is required to decide on the best treatment. EHR systems can narrow this gap by programmatically implementing clinical guidelines in CDSS systems that can process all of the EHR data that have been recorded about the patient [9]. Some approaches have focused on extraction of features from EMRs and their integration in predictive models [8] [7] [16].

These features include clinical time series, relate to multiple measurements, and handle sparseness of the data. Authors in [8] exploited features using logistic regression and k-nearest neighbor. Their main aim is to compare the two approaches, using the MIMIC-II dataset. This work has some similarities with ours in terms of adopted techniques, although it differs in terms of scope. In fact, their goal is to predict mortality among ICU patients by using their EMR data between ICU intake and discharge, while we focus on the prediction of patient's diagnosis. As in our work, the authors use the dynamic time warping for highly time varying features. They selected the heart rate, respiration, the nocturnal, systolic, and diastolic blood pressure, and the oxygen saturation as such features. Differently, we have selected a much broader pool of features. Again, similar to our approach, the authors adopt the k-nearest neighbors algorithm to predict the mortality risk for a specific patient.

In [16] and [7], authors present the benchmarking results for several clinical prediction tasks. In [16], the addressed tasks are mortality prediction, length of stay prediction, and ICD-9 code group prediction using deep learning models. The findings of this work show that deep learning models consistently perform better than the several existing machine learning models and severity scoring systems. In addition, they present benchmarking results on different feature sets including 'processed' and 'raw' clinical time series. The features pre-processing consists on drop outliers in the data according to medical knowledge and merge relevant features. The results show that deep learning models obtain better results on 'raw' features which indicates that rule-based pre-processing of clinical features is not necessary for deep learning models. Similarly, in [7], the covered range of clinical problems includes modeling risk of mortality, forecasting length of stay, detecting physiologic decline, and phenotype classification. They propose strong linear and neural baselines for all four tasks and evaluate the effect of deep supervision, multitask training and data-specific architectural modifications on the performance of neural models. Another work that use a deep learning approach to leverage EHR is [4]. The authors developed Doctor AI, a generic predictive model that covers observed medical conditions and medication uses. Doctor AI is a temporal model using recurrent neural networks (RNN) and was developed and applied to longitudinal time stamped EHR data. Encounter records (e.g. diagnosis codes, medication codes or procedure codes) were input to RNN to predict (all) the diagnosis and medication categories for a subsequent visit. Doctor AI assesses the history of patients to make multilabel predictions (one label for each diagnosis or medication category).

In our work, we model features as time series, and a key point is time series similarity measurement. When dealing with physiological data, signals are never equal although they might be similar. The degree of similarity may indicate if they are or not representative of the same health condition. If we can work with a similarity method capable of, besides small signals' variation, produce an effective method of finding the relationship among the time series, it will greatly increase precision of the analysis in time series databases, helping to improve accuracy in classification, prediction and clustering [11]. Application of similarity matching algorithm is commonly encountered in medical streaming data [1], arrhythmia detection [18] and several other sciences. Many studies have been conducted about the analysis of ECG series through the use of similarity comparison methods. In [17], authors present a novel method for the classification and identification of electrocardiograms of various heart rhythm disturbances. They use Dynamic time warping (DTW) as method to differentiate the ECGs of various arrhythmias. DTW is also used to analyze other biological signals, as to detect atypical functional connectivity in autism spectrum disorders [12], or to analyze the sleep states and the occurrence of some sleep-related problems [6].

## III. Data model

This section presents the real-world dataset used, the data model adopted together with the method to extract raw data from EHRs and model them into structured data to be used as input to the proposed learning methodology.

### A. MIMIC-III dataset

The data was extracted from the Medical Information Mart for Intensive Care III (MIMIC-III) database comprising information related to patients admitted to critical care units at the Beth Israel Deaconess Medical Centre (BIDMC), in Boston, USA. The dataset version used in this research is MIMIC-III v1.4, released on 2 September 2016 [10].

MIMIC-III contains data associated with 53,423 distinct hospital admissions for adult patients admitted to critical care units between 2001 and 2012. It contains data for 7870 neonates that we do not consider for the purposes of our study. In addition, to guarantee that each admission has some

relevant data, we do not consider those relating to patients discharged within 24 hours. For each patient admission, the dataset contains the following data: vital signs, medications, laboratory measurements from within the hospital (i.e. in-patient) and from clinics (i.e. out-patient), charted observations during a patient's stay in the intensive care unit, and de-identified notes regarding the patient's stay, including nursing notes, physician notes and discharge summaries. Note that one patient might have multiple admissions. In this work, for the sake of simplicity, we will consider every hospital stay as an independent sample.

### B. Patients physiological parameters extraction

A set of parameters are identified in the MIMIC-III dataset to catch the key features characterizing the health state of patients. Specifically, for each admission are examined the charted observations and the laboratory measurements. A mean of 4579 charted observations and 380 laboratory measurements are available for each hospital admission. During hospital stay, the primary repository of a patient's health state is the electronic chart that displays patients' routine vital signs and any additional information relevant to their care: ventilator settings, laboratory values, code status, mental status, and so on. Table I shows the selected features. We selected as physiological parameters only features having continuous and ordinal measurements with an occurrence in the admissions greater than 65%. For each feature are reported a set of information like the label, the percentage of occurrences in admissions, and the sampling rate. The final column of this table represents the sampling rate of the feature and is better explained in the next subsection. The physiological parameters selected are relative to several categories as cardiovascular, respiratory, blood pressure, hematology, and chemistries measurements.

The physiological parameters are modeled as time series. An important issue in the time series modeling process was handling parameters with different temporal sampling rate as many of them were charted hourly (H), while others were charted daily (D). Even more, during emergency episodes, the sampling frequency often increases, but in other circumstances the frequency decreases determining several missing values in the time series. To overcome this problem a resampling of the parameters is performed: features charted hourly are sampled into non-overlapping one-hour windows, while those charted daily are sampled into non-overlapping one-day windows. In particular, when several data points were available within a window we replaced them by the mean, instead, to account for missing data we utilized linear interpolation as in [13]. After time series extraction and resampling, we normalized them by applying the Frobenius norm.

TABLE I
PHYSIOLOGICAL PARAMETERS EXTRACTED FROM MIMIC-III DATASET.

| Parameter | Label | Admissions Occurrence(%) | Sampling rate |
|---|---|---|---|
| Heart Rate | heart rate | 0.99 | H |
| Respiratory Rate | respiratory rate | 0.99 | H |
| Hemoglobin | hemoglobin | 0.98 | D |
| Capillary refill rate | calprevflg | 0.92 | H |
| Oxygen Saturation | spo2 | 0.92 | H |
| Glasgow Coma Scale | gcs total | 0.92 | D |
| | eye opening | 0.92 | D |
| | verbal response | 0.92 | D |
| | motor response | 0.92 | D |
| Potassium | potassium (3.5-5.3) | 0.92 | D |
| | potassium | 0.65 | D |
| Hematocrit | hematocrit | 0.91 | D |
| Glucose | glucose (70-105) | 0.91 | D |
| | glucose | 0.65 | D |
| Bun | bun (6-20) | 0.91 | D |
| | bun | 0.73 | D |
| Creatinine | creatinine (0-1.3) | 0.91 | D |
| | creatinine | 0.73 | D |
| Sodium | sodium (135-148) | 0.91 | D |
| Braden Score | braden score | 0.91 | D |
| Platelets | platelets | 0.91 | D |
| Carbon Dioxide | carbon dioxide | 0.91 | D |
| Chloride | chloride | 0.91 | D |
| White Blood Cell | wbc (4-11.000) | 0.90 | D |
| | wbc (4-11.000) | 0.85 | D |
| | wbc | 0.73 | D |
| Red Blood Cell | rbc | 0.90 | D |
| Temperature | temperature c (calc) | 0.90 | D |
| | temperature f | 0.90 | D |
| Noninvasive Blood Pressure Systolic | nbp [systolic] | 0.89 | H |
| Noninvasive Blood Pressure Diastolic | nbp [diastolic] | 0.89 | H |
| Noninvasive Blood Pressure Mean | nbp mean | 0.89 | H |
| Magnesium | magnesium (1.6-2.6) | 0.88 | D |
| | magnesium | 0.72 | D |
| Phosphorous | phosphorous(2.7-4.5) | 0.83 | D |
| | phosphorous | 0.69 | D |
| Calcium | calcium (8.4-10.2) | 0.82 | D |
| Tnternational Normalized Ratio | inr (2-4 ref. range) | 0.81 | D |
| | inr | 0.66 | D |
| Partial Thromboplastin Time | ptt(22-35) | 0.81 | D |
| | ptt | 0.66 | D |
| Prothrombin Time | pt(11-13.5) | 0.81 | D |
| Oxygen Flow Rate | o2 flow (lpm) | 0.79 | D |

## IV. THE DIAGNOSIS PREDICTION METHOD

In this section is described the diagnosis prediction model and the main components of the methodology.

The approach proposed tries to predict a diagnosis for a patient's admission exploiting a set of historical hospitalization data about the patient itself but also other patients, like vital signs, laboratory and chemistry measurements and physicians prescriptions and diagnoses. Specifically, the prediction model relies on the similarity of the physiological signals (e.g. heart rate, respiratory rate, and oxygensaturation), laboratory measurements (e.g. hemoglobin, glucose andbun) and diagnoses of patients.

The prediction method is formulated as a multi-label classification problem based on the k-nearest neighbors (ML-KNN) algorithm. It predicts one or more classes (diagnoses) for each admission. Fundamental aspects, that we dealt with, are the choice of the similarity measure, the identification of diagnoses classes, and the definition of an ad-hoc ML-KNN algorithm tailored for the specific application domain. Each of such aspects is detailed in the following of the section.

### A. Similarity measure

The physiological parameters are modeled by using the time series representation. Accordingly, the patients similarity problem is reformulated as similarity of the time series of their physiological parameters.

In this context, identifying the proper similarity measure is a key issue. Due to the non-stationary of the time series of the physiological data, the Euclidean distance is not suitable as similarity measure. For instance, heartbeat variation can produce nonlinear time fluctuation of time series. Since the Euclidean distance is very sensitive to small distortions in the time axis, it may fail to produce an intuitively correct measure of similarity between two series. In order to eliminate this fluctuation, time normalization is frequently employed; in particular, Dynamic Time Warping (DTW), is a pattern matching algorithm with a nonlinear time-normalization effect [17] that can be effective in the faced scenario. It is a technique employed to align two sequences in order to obtain a dissimilarity measure using non-linear temporal alignment.

By using the DTW measure, time series of different length can be compared, because the measure replaces the one-to-one point comparison, used in Euclidean distance, with a many-to-one (and viceversa) comparison. The main feature of this measure is that it allows to recognize similar shapes, even if they present signal transformations, such as shifting and/or scaling. This is also the case of of patients admission measurements that cover differently sized time windows, or shifted data.

Formally, let us consider two time series $X = (x_0, \ldots, x_{n-1})$ and $Y = (y_0, \ldots, y_{m-1})$ of respective lengths $n$ and $m$. The DTW similarity between $X$ and $Y$ is formulated as the following optimization problem:

$$DTW(X,Y) = \overset{min}{\pi} \sqrt{\sum_{(i,j)\in\pi} d(x_i, y_j)^2} \tag{1}$$

where $d(x_i, y_j)$ corresponds to the distance of i-th point of X and j-th point of Y, with $1 \leq i \leq n$ and $1 \leq j \leq m$, and $\pi = [\pi_0, \ldots, \pi_z]$ is the warping path. That path can be seen as a temporal alignment of time series such that the distance between aligned time series is minimal. In many cases, this method can bring to undesired effects. An example is when a large number of points of a time series is mapped to a single point of another time series. A common way to overcome this problem is enforcing the recursion to stop at a certain depth $\delta$. This constraint, besides limiting extreme or degenerate mappings, allows to speed-up DTW distance calculation. Introducing the above constraint, reduces the computational complexity from $O(nm)$ to $O((n+m)\delta)$ [3]. The above proposed constraint is known as Sakoe-Chiba band, and it is classified as global constraint. Sakoe-Chiba band is parametrized by a radius $r$ (also called warping window size). After several tests, we set the radius $r = 2$.

Accordingly, given two admissions $HADM_i$ and $HADM_j$, we define a similarity score $\rho$ between them as the sum of the similarity of the respective time series. Note that, for some admissions, certain features might be completely missing and we performed null imputation for these cases. Formally, considering the set $F$ of the time series of all the features, and $X_{f,i}$ the time series of a given feature $f$ for an admission $HADM_i$, the similarity score $\rho$ between admissions $HADM_i$ and $HADM_j$ is defined according to equation 2.

$$\rho = \begin{cases} \sum_{f\in F} 1 - DTW(X_{f,i}, X_{f,j}) \\ \qquad if\ |X_{f,i}| > 0\ and\ |X_{f,j}| > 0 \\ 0 \qquad otherwise \end{cases} \tag{2}$$

In order to predict the diagnosis for a new patient admission, the aim is to identify the historical admissions that maximize equation 2.

### *B. Diagnosis classes definition*

In this section is described the methodology adopted to identify the classes with which label the diagnoses through the classification algorithm that will be introduced in the following section. To this purpose, we designed an algorithm that clusters known diagnoses based on their semantic similarity. Each cluster represents a possible class (diagnosis) for an admission. The approach is based on the simple observation that very similar diagnoses describe the same disorder. So, instead of considering each distinct diagnosis as a class, we consider as class a set of diagnoses that are very similar in terms of semantic meaning. In other words, we perform a semantic-based clustering. This approach allows us to restrict the number of possible diagnoses (classes) to consider to correctly label an admission.

To compute the semantic similarity between diagnosis, we relay on sentence embedding, a natural language processing technique where sentences from a vocabulary are mapped to vectors of real numbers in a low-dimensional space. The similarities between the vectors correlate with the sentences semantic similarity. For the purpose, we use the method, named *sent2vec*, proposed in [15].

The method uses a simple but efficient unsupervised objective function to train distributed representations of sentences. It can be regarded as an unsupervised version of FastText [2], and an extension of word2vec [14] to sentences. As first stage, we train a model using a large training text file consisting of known diagnoses. We preprocess the input data performing tokenization and lowercasing. Then we applied sent2vec to compute the 700-dimensional sentence embeddings. On the basis of the generated model we extracted the vectors for each diagnosis.

To extract patients final diagnosis, the reference table in MIMIC-III is the DRGCODES table, that contains diagnosis related groups (DRG) codes for the hospital admissions. More precisely, a final diagnosis can consist of several terms based on the number of DRG codes associated with an admission. A DRG code represents the diagnosis 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 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 types and the descriptions 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, more specific, code which could be used in conjunction with the HCFA codes. Accordingly, each term $t_i$ is a pair $t_i = DRG_{TYPE} : DRG_{DESCRIPTION}$, and each diagnosis is a set of terms $d = t_1, t_2, \ldots, t_n$.

Given two diagnosis $d_i$ and $d_j$, after extracting their semantic vectors through the sentence embedding method, we apply an algorithm to calculate the semantic similarity of the diagnoses. The proposed technique uses the *cosine* similarity measure and compare all terms in $d_i$ with all terms in $d_j$. For each term in $d_i$, it selects the $d_j$ term with which has the maximum similarity. The two diagnoses are considered to belong to the same cluster when their similarity is higher than a threshold $\gamma$. The class threshold can assume values between 0 and 1. A similarity of 100% between the diagnoses is obtained by setting $\gamma$ to 1. A representative example of this case is the following. Consider three diagnosis, each related to a different admission:

- ms:septicemia or severe sepsis w/o mv 96+ hours w mcc, apr:septicemia & disseminated infections
- apr:septicemia & disseminated infections, hcfa:septicemia w mechanical ventilator w/0 96+ hours age>17
- hcfa:septicemia w mechanical ventilator 96+ hours age>17, apr:septicemia & disseminated infections

Those diagnoses describe the same disease even though they are not entirely identical. Our algorithm will create a cluster that contains them all. This cluster is a possible class for a future admission of a patient with septicemia.

### C. The classification model

We formulate the diagnosis prediction as a multi-label classification problem, by combining the time series similarity measure based on the DTW metrics and the sentence embedding. More precisely, the multi-label classification algorithm applies the k-nearest neighbors (KNN) principle in order to classify a new admission, assigning it multiple possible diagnoses. We refer to the algorithm as Time Series k-Nearest Neighbors (TSNN) classification algorithm.

The training examples are the historical admissions, each with a class label. In our model the classes are group of diagnoses representing the same disease, and they are identified using the semantic-based clustering as described earlier.

In the classification phase, the first step is to compare an unlabeled admission with the training samples. Comparison that is performed through the similarity function implemented using the DTW metrics. The $k$ admissions which maximize the similarity function are selected as the nearest neighbors of the unlabeled admission. The classification is carried out through a weighted plurality vote($wpv$) of the $k$ neighbors, namely weights are assigned to the contributions of the neighbors, so that the nearer neighbors contribute more than the more distant ones. The weights take into account the similarity of the test admission to each of its $k$ nearest neighbors.

Formally, given a test admission $HADM_i$, for each class label $c$ of the nearest admissions $HADM_j$, with $0 \leq j \leq k$, the $wpv_c$ is computed based on the following equation:

$$wpv_c = \sum_j \rho(HADM_i, HADM_j) \qquad \forall\ j \mid d_j \in c \quad (3)$$

where $d_j$ is the diagnosis of the training admission $HADM_j$ belonging to the class $c$.

This approach allows to overcome a drawback of the basic *majority voting* classification algorithm, occurring when the class distribution is skewed. That is, examples of a more frequent class tend to dominate the prediction of the new example, because they tend to be common among the $k$ nearest neighbors due to their large number [5]. As classification output, the algorithm provide the top-N most likely diagnoses for a new admission.

## V. Experimental Evaluation

The approach was validated by using a subset of the MIMIC-III V1.4 dataset, obtained as described in Section III. A test environment was implemented in Python, while PostgreSQL was used as the database management. Two well-known performance metrics, Precision and Recall measures, as reported in the following, were exploited.

- *Precision* $= TP/(TP + FP)$
  It is the fraction of the number of successfully detected ground-truth diagnosis (TP : true positive), out of the total number of diagnosis detected TP + FP. Where FP (false positive) is the number of mistakenly detected ground-truth diagnosis. A ground-truth diagnosis is considered successfully detected if exists a predicted diagnosis that matches it.
- *Recall* $= TP/(TP + TN)$
  It is the percentage of ground truth events successfully detected by a method, where TP +TN (TN :True Negative) is the total number of ground-truth diagnosis.

In the experiments, the value of first $N$ items was considered belonging to the range $\{5, 10, 15, 20, 25, 30\}$. In particularly, we take the top-N model predictions with higher probability. If one of them matches with the ground-truth diagnosis, it classifies the prediction as correct. Each set of experiments is performed considering 5-Fold Cross Validation.

First, we show the results that led us to the selection of the features listed in Table I. During the experimentation we set the number of nearest neighbors $k$ to 400 and the class threshold $\gamma$ to 1. In this regard we considered different sets of features, based on their percentage of occurrence $Occur\ x\%$ in admissions. More precisely, $Occur\ x\%$ includes features that are measured in more than x% of hospital stays. Note that the percentage of occurrence is inversely proportional to the cardinality of the set of features. The results, reported in Figure 1 and Figure 2, allow us to make two main observations. The first is that a small set of vital parameters, see $Occur\ 95\%$, describes less the similarity of admissions, leading to a worse

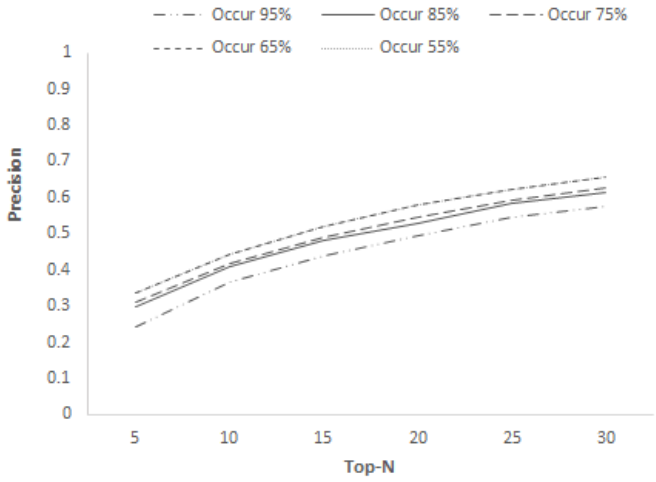


Fig. 1. Precision of TSNN w.r.t. the features percentage occurrence in the hospital admissions.

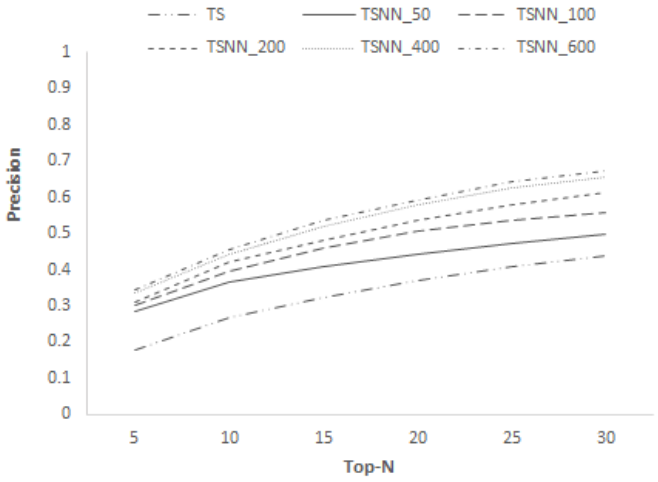


Fig. 3. Precision comparison between TS and TSNN, considering for TSNN different numbers of nearest neighbors $k$.

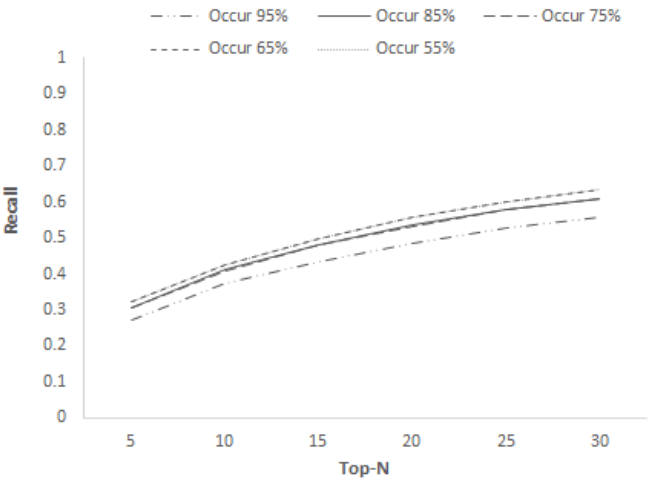


Fig. 2. Recall of TSNN w.r.t. the features percentage occurrence in the hospital admissions.

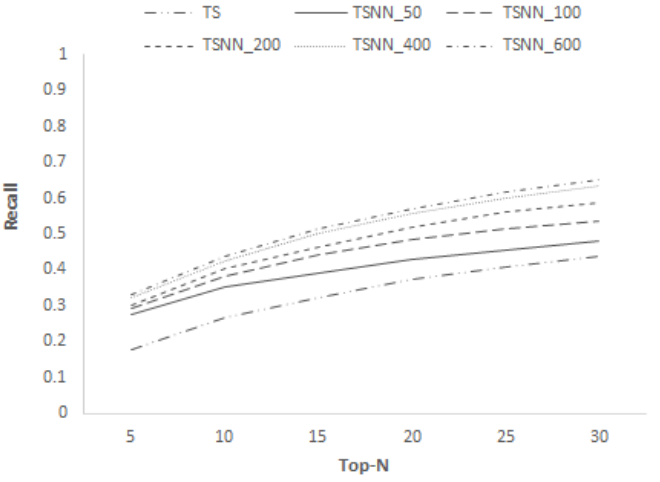


Fig. 4. Recall comparison between TS and TSNN, considering for TSNN different numbers of nearest neighbors $k$.

trend of precision and recall. The second is that a too large set, see $Occur$ $55\%$, containing less common features among the various admissions, does not lead to significant performance improvements. Consequently, for our analysis, we chose the features that occur in at least 65% of admissions.

In a second set of experiments we evaluated precision (Figure 3) and recall (Figure 4) of the proposed TSNN algorithm considering different values of the $k$ nearest neighbors. For class identification we set $\gamma = 1$.

In those experiments we compared the performance of TSNN with a baseline algorithm to which we refer to as TS. The baseline algorithm simply predicts diagnosis by choosing the training admissions that have a greater similarity in terms of time series. Even this algorithm calculates the similarity between two admissions, according to equation 2, as the sum of time series similarities of the considered physiological parameters.

The graphs of Figure 3 and Figure 4 show that, as expected, precision and recall have an increasing trend as the number of top predicted diagnoses increases. It is also evident that the greater the number of $k$ neighbors, the greater the performance of the algorithm, both in terms of precision and recall. However, for values of $k > 400$ the performances present a very slight improvement. Therefore the best configuration is $k = 400$. The comparison with the baseline shows the effectiveness of our approach which, in addition to the similarity of time series, is enriched by the semantic similarity of the diagnoses and the principle of multi-label classification.

In the last set of experiments we examined the performance of the approach by varying the class threshold $\gamma$. The aim is to show what happens when a lower semantic similarity is considered between diagnoses belonging to the same class. We

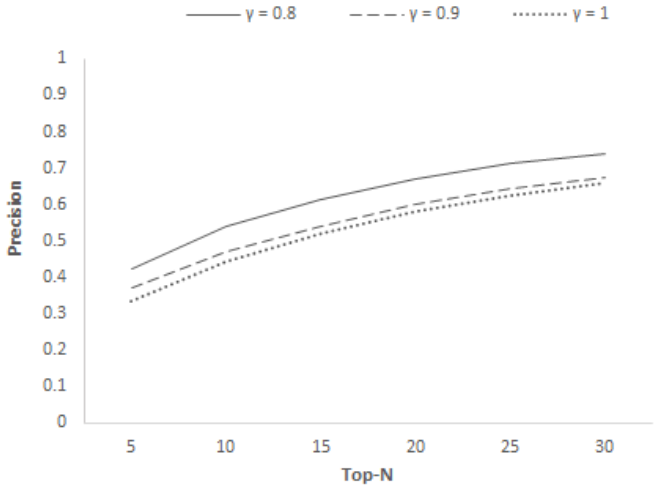


Fig. 5. Precision of TSNN w.r.t. the class threshold $\gamma$.

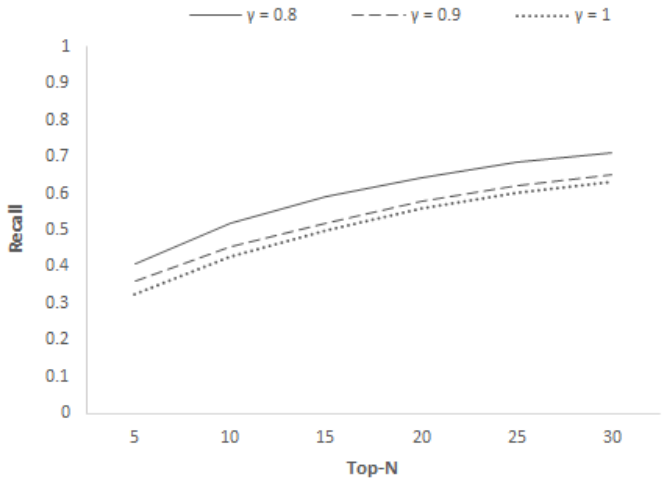


Fig. 6. Recall of TSNN w.r.t. the class threshold $\gamma$.

set $k = 400$. The results show that the fraction of correctly detected ground truth diagnosis increases by decreasing the threshold, in fact both precision (Figure 5) and recall (Figure 6) present an increasing trend. However, the value of this threshold cannot be decreased too much because the classes must represent the same diagnosis.

## VI. Conclusion

The paper presented a supervised diagnosis prediction model, leveraging historical data, to support physicians in the clinical decision process. The predictive approach relies on patients similarity and is formulated as a multi-label classification algorithm that combines time series similarity of patients clinical parameters with the semantic similarity of the diagnoses. Specifically, time series are extracted for vital signs, laboratory and chemistries measurements. Their similarity is measured using the dynamic time warping technique. The diagnoses are classified according to a set of labels obtained through an algorithm that clusters them on the basis of the semantic similarity. Each cluster represents a possible class for an admission. The semantic-aware clustering algorithm is based on sentence embedding to compute the semantic similarity of diagnoses.

A thorough experimental evaluation has been performed using the real-world MIMIC-III dataset. Results shown that the proposed diagnosis prediction approach is effective, reaching considerable performance metrics in terms of precision and recall rates in different experimental settings. Overall, the achieved outcomes indicate that physiological parameters of patients present high predictive ability for detecting disease diagnoses.

## References

[1] Jolita Bernataviciene, Gintautas Dzemyda, Gediminas Bazilevicius, Viktor Medvedev, Virginijus Marcinkevicius, and Povilas Treigys. Method for visual detection of similarities in medical streaming data. *International Journal of Computers Communications & Control*, 10(1):8–21, 2014.

[2] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[3] Carmelo Cassisi, Placido Montalto, Marco Aliotta, Andrea Cannata, and Alfredo Pulvirenti. Similarity measures and dimensionality reduction techniques for time series data mining. *Advances in data mining knowledge discovery and applications'(InTech, Rijeka, Croatia, 2012,*, pages 71–96, 2012.

[4] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In Finale Doshi-Velez, Jim Fackler, David Kale, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 1st Machine Learning for Healthcare Conference*, volume 56 of *Proceedings of Machine Learning Research*, pages 301–318, Northeastern University, Boston, MA, USA, 18–19 Aug 2016. PMLR.

[5] D. Coomans and D.L. Massart. Alternative k-nearest neighbour rules in supervised pattern recognition: Part 1. k-nearest neighbour classification by using alternative voting rules. *Analytica Chimica Acta*, 136:15–27, 1982.

[6] Chunxiao Fu, Pengle Zhang, Jiang Jiang, Kewei Yang, and Zhihan Lv. A bayesian approach for sleep and wake classification based on dynamic time warping method. *Multimedia Tools and Applications*, 76(17):17765–17784, 2017.

[7] Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):1–18, 2019.

[8] Mark Hoogendoorn, Ali El Hassouni, Kwongyen Mok, Marzyeh Ghassemi, and Peter Szolovits. Prediction using patient comparison vs. modeling: A case study for mortality prediction. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2464–2467. IEEE, 2016.

[9] Peter B Jensen, Lars J Jensen, and Søren Brunak. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395–405, 2012.

[10] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[11] A Kianimajd, MG Ruano, P Carvalho, J Henriques, T Rocha, S Paredes, and AE Ruano. Comparison of different methods of measuring similarity in physiologic time series. *IFAC-PapersOnLine*, 50(1):11005–11010, 2017.

[12] AC Linke, LE Mash, CH Fong, MK Kinnear, JS Kohli, M Wilkinson, R Tung, RJ Jao Keehn, RA Carper, I Fishman, et al. Dynamic time warping outperforms pearson correlation in detecting atypical functional connectivity in autism spectrum disorders. *NeuroImage*, 223:117383, 2020.

[13] Daniel Lopez-Martinez, Patrick Eschenfeldt, Sassan Ostvar, Myles Ingram, Chin Hur, and Rosalind Picard. Deep reinforcement learning for optimal critical care pain management with morphine using dueling double-deep q networks. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3960–3963. IEEE, 2019.

[14] Tomas Mikolov, Kai Chen, G.s Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*, 2013, 01 2013.

[15] Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. Unsupervised learning of sentence embeddings using compositional n-gram features. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 528–540, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[16] Sanjay Purushotham, Chuizheng Meng, Zhengping Che, and Yan Liu. Benchmarking deep learning models on large healthcare datasets. *Journal of biomedical informatics*, 83:112–134, 2018.

[17] V. Tuzcu and S. Nas. Dynamic time warping as a novel tool in pattern recognition of ecg changes in heart rhythm disturbances. In *2005 IEEE International Conference on Systems, Man and Cybernetics*, volume 1, pages 182–186 Vol. 1, 2005.

[18] Yun-Chi Yeh. An analysis of ecg beats by using the mahalanobis distance method. In *2009 Fourth International Conference on Innovative Computing, Information and Control (ICICIC)*, pages 1460–1463, 2009.