UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Head Office: Università degli Studi di Padova

Department of Cardiac, Thoracic, Vascular Sciences and Public Health

Ph. D. COURSE IN: *Medicina Specialistica Traslazionale "G. B. Morgagni"*

CURRICULUM: *Biostatistica*

SERIES: *XXI*

SVILUPPO E APPLICAZIONE DI TECNICHE DI APPRENDIMENTO AUTOMATICO PER L'ANALISI E LA CLASSIFICAZIONE DEL TESTO IN AMBITO CLINICO

Development and Application of Machine Learning Techniques for Text Analyses and Classification in Clinical Research

Coordinator: Ch.ma Prof.ssa Annalisa Angelini

Supervisor: Ch.mo Prof. Dario Gregori

Ph. D. student: Corrado Lanera

ABSTRACT

The content of Electronic Health Records (EHRs) is hugely heterogeneous, depending on the overall health system structure. Possibly, the most present and underused unstructured type of data included in the EHRs is the free-text. Nowadays, with Machine Learning (ML), we can take advantage of automatic models to encode narratives showing performance comparable to the human ones. In this dissertation, the focus is on the investigation of ML Techniques (MLT) to get insights from free-text in clinical settings.

We considered two main groups of free-text involved in clinical research. The first is composed of extensive documents like research papers or study protocols. For this group, we considered 14 Systematic Reviews (SRs), including 7,494 studies from PubMed and a whole snapshot of 233,609 trials from ClinicalTrials.gov. Pediatric EHRs compose the second group, for which we considered two sources of data: one of 6,903,035 visits from the Italian Pedianet database, and the second of 2,723 Spanish discharging notes from pediatric Emergency Departments (EDs) of nine hospitals in Nicaragua.

The first contribution reported is an automatic system trained to replicate a search from specialized search engines to clinical registries. The model purposed showed very high classification performances (AUC from 93.4% to 99.9% among the 14 SRs), with the added value of a reduced amount of non-relevant studies extracted (mean of 472 and maximum of 2119 additional records compared to 572 and 2680 of the original manual extraction respectively). A comparative study to explore the effect of changing different MLT or methods to manage class imbalance is reported.

A whole investigation on pediatric ED visits collected from nine hospitals in Nicaragua was reported, showing a mean accuracy in the classification of discharge diagnoses of 78.31% showing promising performance of an ML for the automatic classification of ED free-text discharge diagnoses in the Spanish language.

A further contribution aimed to improve the accuracy of infectious disease detection at the population level. That is a crucial public health issue that can provide the background information necessary for the implementation of effective control strategies, such as advertising and monitoring the effectiveness of vaccination campaigns. Among the two studies reported of classify cases of Varicella-Zoster Virus and types of otitis, both the primary ML paradigms of shallow and deep models were explored. In both cases the results were highly promising; in the latter, reaching performances comparable to the human ones (Accuracy 96.59% compared with 95.91% achieved by human annotators, and balanced F1 score of 95.47% compared with 93.47%).

A further relevant side goal achieved rely on the languages investigated. The international research on the use of MLTs to classify EHRs is focused on English-based datasets mainly. Hence, results on non-English databases, like the Italian Pedianet or the Spanish of ED visits considered in the dissertation are essential to assess general applicability of MLTs at a general linguistic level.

Showing performances comparable to the human ones, the dissertation highlights the real possibility to start to incorporate ML systems on daily clinical practice to produce a concrete improvement in the health care processes when free-text comes into account.

SOMMARIO

Il contenuto delle cartelle cliniche elettroniche (EHR) è estremamente eterogeneo, dipendendo della struttura generale del sistema sanitario. Al loro interno, il testo libero è probabilmente la tipologia di dati non strutturato più presente e contemporaneamente sottoutilizzato. Al giorno d'oggi, grazie alle tecniche di Machine Learning (MLT), possiamo sfruttare modelli automatici per codificarne il contenuto testuale con prestazioni comparabili a quelle umane. In questa tesi, l'attenzione si concentra sull'investigazione delle MLT per l'ottenimento di informazioni utili non triviali dal testo libero in contesti clinici.

Abbiamo considerato due tipi principali di testo libero coinvolti nella ricerca clinica. Il primo è composto da documenti estesi come articoli scientifici o protocolli di studio. Per questo gruppo, abbiamo preso in considerazione 14 revisioni sistematiche (SR), tra cui 7.494 studi di PubMed e un'intera istantanea composta da 233.609 studi clinici da ClinicalTrials.gov. Le cartelle cliniche elettroniche pediatriche compongono il secondo gruppo, per il quale abbiamo considerato due fonti di dati: una di 6.903.035 visite dal database italiano Pedianet e la seconda da 2.723 note di dimissione ospedaliera scritte in spagnolo e provenienti dai dipartimenti di emergenza (DE) pediatrica di nove ospedali in Nicaragua.

Il primo contributo riportato è un sistema automatico addestrato per replicare una ricerca dai motori di ricerca specializzati ai registri clinici. Il modello proposto ha mostrato prestazioni di classificazione molto elevate (AUC dal 93,4% al 99,9% tra i 14 SR), con il valore aggiunto di una quantità ridotta di studi non rilevanti estratti (media di 472 e massimo di 2119 record aggiuntivi rispetto a 572 e 2680 dell'estrazione manuale originale rispettivamente). Viene riportato anche uno studio comparativo per esplorare l'effetto dell'utilizzo di differenti MLT e di metodi diversi per gestire gli effetti dello squilibro di numerosità nelle classi.

Nella tesi è riportata inoltre un'intera indagine sulle visite pediatriche presso i DE raccolte presso i nove ospedali del Nicaragua. In tale indagine emerge un'accuratezza media nella classificazione delle diagnosi di dimissione coi modelli proposti del 78,31%, mostrando promettenti prestazioni per un sistema ML per la classificazione automatica delle diagnosi di dimissione da testo libero in lingua spagnola.

Un ulteriore contributo riportato ha mirato a migliorare l'accuratezza del rilevamento delle malattie infettive a livello di popolazione. Questo è un problema cruciale per la salute pubblica che può fornire le informazioni di base necessarie per l'implementazione di strategie di controllo efficaci, come la notifica e il monitoraggio di efficacia di campagne di vaccinazione. Tra i due studi riportati, sono stati esplorati entrambi i paradigmi primari di ML classici e profondi. In entrambi i casi i risultati sono stati molto promettenti; nel secondo, raggiungendo prestazioni paragonabili a quelle umane (precisione del 96,59% rispetto al 95,91% raggiunta dagli annotatori umani e livello F1 bilanciato del 95,47% rispetto al 93,47%).

Un ulteriore obiettivo secondario ma rilevante raggiunto riguarda le lingue indagate. La ricerca internazionale sull'uso delle MLT per classificare gli EHR si concentra principalmente su set di dati testuali in lingua inglese. Pertanto, i risultati su database non inglesi, come il Pedianet italiano o quello spagnolo delle visite ED considerate nella tesi, risultano contributi chiave per valutare l'applicabilità generale delle MLT a livello linguistico generale.

Mostrando prestazioni paragonabili a quelle umane, la tesi evidenzia la reale possibilità di iniziare a incorporare i sistemi ML nella pratica clinica quotidiana per produrre un miglioramento concreto nei processi sanitari quando si tiene conto del testo libero.

AKNOWLEDGE

LIST OF PUBLICATIONS

WITHIN THE FIRST THREE AUTHORS' NAMES

1. G. Lorenzoni, S. S. Sabato, *C. Lanera*, D. Bottigliengo, C. Minto, H. Ocagli, P. De Paolis, D. Gregori, S. Iliceto, F. Pisanò, **Comparison of machine learning techniques for prediction of hospitalization in heart failure patients**, *J. of Clinical Medicine*, Vol. 8, Issue 9 (August 2019).
2. D. Bottigliengo, P. Berchialla, *C. Lanera*, D. Azzolina, G. Lorenzoni, M. Martinato, D. Giachino, I. Baldi, D. Gregori, **The role of genetic factors in characterizing extra-intestinal manifestations in Crohn's disease patients: are bayesian machine learning methods improving outcome predictions?** *J. of Clinical Medicine*, Vol. 8, Issue 6 (June 2019).
3. D. Gregori, D. Azzolina, *C. Lanera*, M. Ghidina, C. E. Gafare, G. Lorenzoni **Consumers' attitudes before and after the introduction of the Chilean regulation on food labelling** *Journal International Journal of Food Sciences and Nutrition*, Vol. 70 (June 2019).
4. G. Lorenzoni, S. Swain, *C. Lanera*, M. Florin, I. Baldi, S. Iliceto, D. Gregori, **High- and lowinpatients' serum magnesium levels are associated with in-hospital mortality in elderly patients: a neglected marker?** *Aging Clinical and Experimental Research* (May, 2019).
5. G. Lorenzoni, S. Bressan, *C. Lanera*, D. Azzolina, L. Da Dalt, D. Gregori, **Analysis of unstructured testbases data using machine learning techniques: the case of pediatric emergency department records in Nicaragua** *Medical Care Research and Review* (April 2019).
6. *C. Lanera*, C. Minto, A. Sharma, D. Gregori, P. Berchialla, I. Baldi, **Extending PubMed Searches to ClinicalTrials.gov Through a Machine Learning Approach for Systematic Reviews** *Journal of Clinical Epidemiology*, Vol. 103 (November 2018).
7. G. Lorenzoni, D. Azzolina, *C. Lanera*, G. Brianti, D. Gregori, D. Vanuzzo, I. Baldi, **Time trends in first hospitalization for heart failure in a community-based population** *International Journal of Cardiology*, Vol. 271 (November 2018).
8. I. Baldi, *C. Lanera*, P. Berchialla, D. Gregori, **Early termination of cardiovascular trials as a consequence of poor accrual: Analysis of ClinicalTrials.gov 2006-2015**, *BMJ Open*, Vol. 6, Issue 6, e013482 (June 2017).
9. [IZSTO:] G. Ru, M. I. Crescio, F. Ingravalle, C. Maurella; [UNIPD:] D. Gregori, *C. Lanera*, D. Azzolina, G. Lorenzoni, N. Soriani, S. Zec, [UNITO:] P. Berchialla, S. Mercadante, [Zetaresearch:] F. Zobec, M. Ghidina, S. Baldas, B. Bonifacio, A. Kinkopf, D. Kozina, L. Nicolandi, L. Rosati, **Machine Learning Techniques applied in risk assessment related to food safety**, *EFSA Supporting Publications*, 14.7 (May 2017).
10. D. Gregori, C. Minto, *C. Lanera*, G. Lorenzoni, **Feasibility and Reliability of Wearable Devices in Measuring Caloric Intake: Results from a Pilot Study**, *The FASEB Journal*, Vol. 31, Issue 1 Supplement, 302.6 (April 2017).
11. E. Menti, *C. Lanera*, G. Lorenzoni, D.F. Giachino, M. de Marchi, D. Gregori, P. Berchialla, **Bayesian Machine Learning Techniques for revealing complex interactions among genetic and clinical factors in association with extra-intestinal Manifestations in IBD patients**, *AMIA 2016 Annual Symposium proceedings*, 884-893 (February 2017). 46:101–121, 2014.

UNDER REVIEW (SUBMITTED AS FIRST AUTHOR)

1. *C. Lanera*, P. Berchialla, I. Baldi, G. Lorenzoni, L. Tramontan, A. Scamarcia, L. Cantarutti, C. Giaquinto, D. Gregori, **Use of Machine Learning techniques for case-detection of Varicella Zoster using routinely collected textual ambulatory records**, submitted to the Journal of Medical Internet Research (2019).
2. *C. Lanera*, P. Berchialla, A. Sharma, C. Minto, D. Gregori, I. Baldi **Screening PubMed Abstracts: is Class Imbalance Always a Challenge to Machine Learning?**, submitted to BMC Systematic Reviews (2019).

PREPRINT (AS FIRST AUTHOR)

1. *C. Lanera*, E. Barbieri, G. Piras, A. Maggie, D. Weissenbacher, D. Doná, A. Scamarcia, L. Cantarutti, G. Gonzalez, C. Giacquinto, D. Gregori, **Automatic identification and classification of different types of otitis from free-text pediatric medical notes in the Italian language: a deep-learning approach**, pre-print (2019)

PUBLISHED (AS CONTRIBUTOR)

1. G. Lorenzoni, D. Azzolina, S. Baldas, G. Messi, *C. Lanera*, M. A. French, L. Da Dalt, D. Gregori, **Increasing awareness of food-choking and nutrition in children through education of caregivers: the CHOP community intervention trial study protocol** *BMC Public Health* Vol. 19, Issue 1 (August 2019).
2. S. Poli, G. Boriani, M. Zecchin, D. Facchin, M. Gasparini, M. Landolina, R. P. Ricci, *C. Lanera*, D. Gregori, A. Proclemer, **Favorable Trend of Implantable Cardioverter-Defibrillator Service Life in a Large Single-Nation Population: Insights From 10-Year Analysis of the Italian Implantable Cardioverter-Defibrillator Registry**, *J. of the American heart Association*, Vol. 8, Issue 15 (July 2019).
3. S. Poli, D. Facchin, F. Rizzetto, L. Rebellato, E. Daleffe, M. Toniolo, A. Miconi, A. Altinier, *C. Lanera*, S. Indrigo, J. Comisso, A. Proclemer, **Prognostic role of non-sustained ventricular tachycardia detected with remote interrogation in a pacemaker population** *IJC Hearth and Vasculature*, Vol. 22 (March 2019).
4. M. Carrozzini, J. Bejko, A. Gambino, V. Tarzia, *C. Lanera*, D. Gregori, G. Gerosa, T. Bottio **Results of new-generation intrapericardial continuous flow left ventricular assist devices as a bridgeto-transplant**, *Journal of cardiovascular medicine*, Vol. 19, Issue 12 (December 2018).
5. E. Surkova, L.P. Badano, D. Genovese, G. Cavalli, *C. Lanera*, J. Bidviene, P. Aruta, C. Palermo, S. Iliceto, D. Muraru **Clinical and Prognostic Implications of Methods and Partition Values Used to Assess Left Atrial Volume by Two-Dimensional Echocardiography.** *J. of the American Society of Echocardiography*, Vol. 30, Issue 11, 1119-1129 (November 2017).
6. F. Folino, G. Buja, G. Zanotto, E. Marras, G. Allocca, D. Vaccari, G. Gasparini, E. Bertaglia, F. Zoppo, V. Calzolari, R.N. Suh, B. Ignatiuk, *C. Lanera*, A. Benassi, D. Gregori, S. Iliceto, **Association between air pollution and ventricular arrhythmias in high-risk patients (ARIA study): a multicentre longitudinal study,** *The Lancet Planetary Health*, Vol. 1, Issue 2, e58-e64 (May 2017).

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

***Statistics**: The study of the collection, analysis, interpretation, presentation, and organization of data*
--- The Oxford dictionary of statistical terms

***Machine Learning**: Field of study that gives computers the ability to learn without being explicitly programmed*
--- Arthur Samuel (1959)

***Well-posed Learning Problem**: A computer program is said to* learn *from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E*
--- Tom Mitchell (1998)

***Text Mining**: The discovery and the extraction of interesting, non-trivial knowledge from free, unstructured text*
--- Anne Kao and Steve Poteet (2007)

1.1 CLINICAL RESEARCH IN THE MACHINE LEARNING ERA

Clinical research, from the statistical and methodological point of view, is made up of observing, collecting, investigating, and analyzing data. Statistical methods and biostatistics have been widely exploited from the twentieth century on the different faces of research in clinical environments, like to draw reasonable inferences, to support decisions, to conduct agreement. In the Internet era, with the possibility to share and retrieve with near no effort the information, the amount of data available started to increase exponentially in time, and that led the development of new computational methods. The principal aim of those methods was to face the impossibility to manage a massive amount of data directly by humans. On the side of programs and languages born to assess specific low- or high-complexity tasks, there started to be necessary to have computational instruments to analyze data too. As a result, specialized programming languages explicitly oriented to statistical research, like S [1] appeared.

Nevertheless, clinical data were increasing more than humans could deal by themselves alone. Classical methods that permit to incorporate many human and expert decisions coded into algorithms started to be used along with other computational methods, exclusively designed to be data-driven. Those ecosystems of algorithms are well-known as machine learning (ML) techniques.

Thanks to classical programming and methods, a problem can be solved through a program defining its internal rules, like an exact mathematical formula or an expert human decision regarding possible interactions between the investigated factors. That kind of programs aims to answer a question by applying those rules to the data given them in input. They implement complex or simple formulas the program applies to data. If we have new data for which we do not know the related answer we are looking for, we can use that kind of programs to get an evaluation of that unknown answer. On the other hand, if we have new data for which we know the answer, then we can use that new pieces of information to enhancing our understanding of the problem, figuring out new ways to improve the rules, i.e., the formulas. That way, thanks to data, we can re-define the algorithm to have better evaluations on new data.

In contrast, the main feature of the ML approach is that the program is designed to be trained by the data, on the side of their already known related answers. In this case, the program aims to define a model, i.e., the formula which can represent at best the general rules linking data with their answers. Once a model is trained, i.e., our ML program has defined its optimal formulas to code the rule behind the data provided, it can be

applied to new data to infer the corresponding answers when those answers are unknown. On the other hand, if we know the answers related to some new data, then it is possible to add those data to the training ones, improving the model representation automatically and without changing anything in our ML algorithm (Figure 1). Develop an ML system is an act of coding an algorithm that from data and known answer given in input, conduct to formula as output. Hence, ML precisely tries to answer the considerable impact Big Data impose in daily life, that is too hard to be managed by humans thinking or with pre-defined expert rules alone.

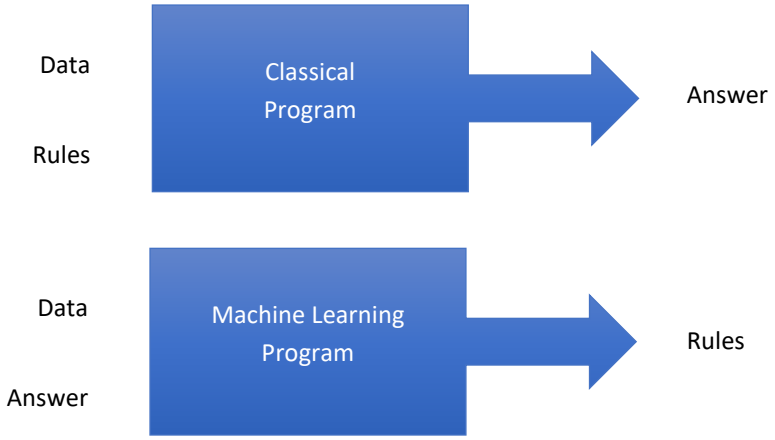

Figure 1 Classical and Machine learning paradigms to solve a given task

In the massive amount of digitally stored information we produce nowadays, clinical data which are stored electronically at the patient level comes with the name of Electronic Health Records (EHRs). The primary purpose of ML efficiently fits to analyze EHRs, which are a vast source of clinical information. Indeed, an emerging research paradigm lies in the extraction and in the processing of massive amounts of clinical data to gain clinical insights and ideally to complement the decision-making process at different levels, from the individual treatment to the definition of national public health policies. As acknowledged by others [2], the development and application of big data analysis methods on EHRs may help to create an effective, continually learning healthcare system [3].

On the other hand, the content inside EHRs is hugely heterogeneous and depends on the overall health system structure. It is worth noticing that the EHRs' content is generally characterized by structured data, i.e., information directly linked with numerical measurements that can be placed easily on a table. Those kinds of information can report continuous health or socio-demographic related measurements, like the Systolic Blood Pressure, the weights, or the age of a patient, as well as a discrete and coded information or factors, like the gender, the ethnicity, or an International Classification of Diseases (ICD) code. The main characteristic of that kind of information is that their usage as an input to a computer program is quite straightforward, i.e. the way we can communicate that information to software is similar to the way we can communicate them between us as humans. Systolic Blood Pressure, weight, age, codes are numbers we can directly pass to a computer, small discrete information like ethnicity or gender can be easily linked to numbers as a code to communicate them to a computer in a way that we still understand.

On the opposite, unstructured data are represented by a type of information that we can code into numbers to pass them to a computer program, but the code of a single object is not represented by a single number, nor it can be easily understood by a human, like images, sounds, or movements for which a single instance of them is represented by possibly multiple matrices of numbers. Possibly, the oldest and one of the most used unstructured type of data in clinical research (and not only in clinical research) is the free-text. We can easily encode letters, numbers, and symbols to single numbers to use them in a computer program with a one-to-

one map still understandable by humans — on the other hand, coding the meaning of a sentence still unlikely to be considered a structured type of information.

Natural Language Processing (NLP) represent the set of all the processes or techniques to automatically manage, annotate and, in general, process human languages. Using those tools, Text Mining (TM) is the field of computational science with the specific aim to extract non-trivial knowledge from natural text. Nowadays, with ML methods for NLP, we can take advantage of the knowledge of a vast number of expert physicians to train models for TM tasks, like the automatic encode of narratives. Often, the principal needs are collecting or retrieving enough textual data related to the environment of interest, along with their known and already encoded outcomes of interest. Especially in clinical environment, that is a task which is possibly costly or with legal limitations, but practically effortless nowadays.

In this dissertation, we will focus on the investigation of automatic techniques to get insights from free-text in clinical settings. Particular attention and space will be given to pediatric tasks of clinical interest which can be highly expensive in term of time and cost, or which can involve procedure to reduce the delay in the adaptation and the development of standardized data collection systems especially in Low- or Middle-Income Countries (LMIC).

For those reasons, our primary attention will be on text produced by researchers or physics, on the contrary of text produced by patients. We will consider two major free-text groups involved in clinical research: first, extensive electronic documents like research paper or study protocols; second, EHRs like discharging notes, medical reports, or diaries from family doctors or other professionals involved in the health care process.

To do that we will investigate several of the most effective ML techniques which have proved to be valid for text analyses, spanning from shallow classical techniques to deep learning (DL) architectures. On the same time, we will explore and compare different ways to manage textual sources, both considered at individual level and as a whole corpus of documents.

Moreover, to distinguish promising results to the ones that can be of real end practical utility for health care, we maintain a constant reference to the human performances on the same tasks investigated.

Free-text in clinical research is possibly both one of the most used resources by its professionals, and at the same time quite completely unused nor analyzed automatically by computer programs helping the same professionals which have produced all that text. This dual situation gives to clinical free-text and ML the potential to be of crucial help to improve the clinical research translationally among its branches in the very next future.

1.2 AUTOMATIC TEXT CLASSIFICATION FOR CLINICAL RESEARCH

Text analysis by a computer program is possible only after the establishment of a procedure to convert text, i.e., human-readable sequences of symbols, into numbers, i.e., computer-readable sequences of symbols. This process is called pre-processing, and it is the first [4] and probably the most critical step in TM and the stage in which NLP concentrate its efforts traditionally [5]. Generally, a TM strategy consists of i) text preprocessing, ii) training of the ML model and iii) estimation of its performance on new test data.

Classical text pre-processing steps aim to convert text into numbers in a way that humans can transfer their expert knowledge of the text directly inside to the coded representation. This process is also known as a *feature engineering*. Main classical pre-processing processes are the following ones:

- Conversion of the text to all lowercase; this is useful to have the same representation for words reported differently with uppercase o lowercase; this is useful to represent with the same token (i.e., a generalization of the concept of "words" to a piece of text representing a single digital entity) a word which is at the beginning of a sentence and the same word in another position in sentences.
- Removing non-words, like symbols, punctuations, and sometimes even numbers; this is useful to reduce the feature space, eliminating possible noise produced by non-words which do not contain information useful for the task.
- Stemming words, i.e., removing the ending part of a word to merge distinct tokens which contain information which is considered similar, e.g. English plurals "-s." This method can be useful in languages with strict rules about endings mainly.
- An approach with the same aim of the previous one but more agnostic regarding the language is the conversion of all the words to their *lemma*, i.e., the corresponding entry founded into the dictionary. In that way, tokens lose their meaning that language variations can incorporate in their variations like the time reference for a verb.
- Stripping white space is used mainly to avoid the possibility that two identical tokens were represented differently because of redundant spaces, and to save memory space too.
- Building n-grams, sequences of n adjacent words from the original text considered like a single token. The aim is to retain information stored directly in the exact sequences the words appear into the text, like negations, or questions (in English-language).

All the preprocessing procedures described still maintain text like it is: i.e., text, without converting it into numbers. A Document-Term Matrix (DTM) is how text can be classically represented as a matrix of numbers. It is made up of the coding of the collection of all the preprocessed tokens, i.e., the vocabulary. The DTM encode every document row-wise and every token column-wise like vectors. Initially, the DTM is made up by the so-known *one-hot* representation of the tokens. In this representation, every token is coded into a vector of the dimensionality of the entire vocabulary; only a single entry is filled with a 1 (the "one-hot") at the same position in the vector corresponding to the token's index in the vocabulary. All the other entries of the vector are filled with 0s. A document is then represented by the sum of its tokens' representation like a vector, creating a single vector which is all empty but in the entries related with the tokens included in that document. The resulting document-vector, in its non-zero's entries, is finally filled with the tokens' frequencies. All the document-vectors are then stacked together like rows of the DTM.

Often, to include information about the tokens' distribution in the corpus, the DTM is next weighted. One of the most used and useful weights are the TF-iDF ones: all the original Token Frequencies (TF) reported in the DTM are multiplied by the inverse of the Document Frequency (iDF) of every Token, i.e. the number of documents containing that token (often taken at the logarithmic scale). Hence, the iDF is a property (a number) of every distinct token, i.e., of every column of the DTM. This weighting schema is adequate to represent how much a token is relevant in distinguishing documents by each other. Indeed, the more time a token appears in a document, the more critical it is to define that document, but the more documents contain that tokens, the less useful it is to distinguish one document from another one in that corpus.

Like weighting strategies, other preprocessing techniques aim to prepare better the corpus to effectively analyzed by the ML program in a way to permit to encode information of the document collection like a whole and not like individuals only. An important one is related to the phenomenon of imbalance in the classes target of the classification. Often, in classification tasks, data are not separated in uniform classes respect to their frequencies. That means that one class can contain more documents than another one. Those differences can have an impact both in the training phase and in the evaluation of the performances. In the research conducted during my Ph.D. and reported in this dissertation, we have also explored the impact that different preprocessing approaches and strategies can affect the performances on classification tasks.

Adopting a completely different point of view, there is another effective way to represent the text tokens: the *dense* representation. In the dense representation, every token is represented by a vector in a D-dimensional space. Typically, D is much lower than the dimensionality of the entire vocabulary. The encoding of a token's vector is often the output of an ML model aimed to capture the syntactical and lexical information inside the text provided. That way the feature space tries to represent the whole complexity of the language, within all its possible variations. There, every token is represented by a vector, tokens with similar linguistic rules are supposed to be near, and also syntactical rules are supposed to be represented by the geometrical position of the vector in that space. E.g., one of the most famous examples reported by in the first work which purposed this representation is that subtracting from the vector related to the token "king" the vector related to the token "man", and next adding the vector related to the token "woman" to the previous result will lead to the vector related to the token "queen" [6].

In the last project at the end of this dissertation, we will explore a dense representation approach connected to deep ML architectures; hence, avoiding by design the needs of any other feature engineering.

Once the data and the information within them are encoded, the training phase can be conducted following one or more different strategies which are represented by the ML techniques considered. Choosing an ML strategy for a classification task means selecting an algorithm able to be trained with data labeled with their known classes to produce a model representation for the rules linking data with the classes.

Nowadays, we can distinguish two main categories of ML algorithms: shallow and deep ones. Shallow techniques take in input data, and after few or more simple or complex computations they provide the resulting model. The main shallow methods we will explore, describe and compare here are the following ones:

- Techniques that expands the family of Generalized Linear Model (GLM). In particular we will investigate Elastic-Net Regularized Generalized Linear Models (GLMNet), which are a regularized regression methods that linearly combines the L1 and the L2 penalties of the lasso and ridge methods applied in synergy with a link function and a variance function to overcome the linear model limitations; MaxEnt, which is an implementation of (multinomial) logistic regression aimed at minimizing the memory load on large datasets; and the LogitBoost, which use the boosting methods applied to single binary decision trees.
- k-Nearest Neighbor (k-NN), which is a geometrical approach that assigns data in classes considering their distances accordingly to a suitable norm.
- Support-Vector Machines (SVM), which is another geometrical approach that projects the vectors data in higher dimensional space respect to the original one with a dimension big enough (possibly infinite) to be all the data linearly separable by an hyper-plane accordingly to their classes.
- Random Forests (RF), which bring together the most useful properties of the decision trees, with boosting, while randomly sampling the feature to use in every tree and merging results with a voting system.

On the other side, we have deep models based on artificial neural networks methods. ANN is based on units also known as *neurons* which implements a linear transformation of the inputs followed by a non-linear differentiable function. Networks of neurons are stacked sequentially into layers. Each layer can report different number of neurons connected with different inputs from the previous layers to create from simple to highly complex architecture which is only bounded to be a Direct Acyclic Graph (DAG). Every neuron is identified by the parameters of its linear transformation. The DAG from the inputs to the outputs represent a high complex non-linear function which is differentiable. Hence, through a gradient descent process on the surface of a suitable loss-function, parameters of all the neurons are updated after every evaluation of the loss function on new known data.

In this dissertation, we will explore the applications of classical ML techniques, comparing their performances on the tasks investigated, as well as distinct deep neural network architectures for the last reported project.

The last step in an ML project regards the evaluation of the performances for the trained models. To evaluate a model there are necessary two main ingredients: a measure and data on which apply that measure. The selection of the data used to evaluate the measurements is crucial to have fair estimation of the model performance. Indeed, a testing set of data must be wholly hided from all the training process. If it does not be the case, the model trained could have learned structures from the test data too, producing biased and optimistic measure of its performance.

On the other hand, a ML model should be trained, or at least tuned, on data similar to the testing one, in a way that the distribution of the information stored in the training data on which the model-based its knowledge reflect the distribution of the information of the testing data on which we are interested in applying the trained model. That requires a strict and precise design of all the stages of the learning processes. The pre-processing must be designed to assure to be precisely the same for the training and the testing data, but at the same time it cannot incorporate information regarding the testing set into the pre-processed encoding of the training set. The training phase design is also crucial: there must be some data used to measure the performance during the training to permit the model to improve; on the other hand, those data must be completely disjoint from the testing without including any possible hidden information about the testing set, while they must be as similar as possible each other to be confident the model is learning to assess the right task in which we are interested.

One of the strengths of all the projects conducted during my Ph.D. and reported here in this dissertation relay in the rigorous design of the training and testing phases, which is, in my opinion, the most crucial aspect for a model definition and a fair performance evaluation and reporting.

1.3 MAIN CONTRIBUTIONS

Among the two groups of text considered in the projects reported in this dissertation, i.e. extensive literature of clinical research and EHRs, one of the crucial uses of the free-text in the first one relay on the retrieval of scientific documents to synthesize evidence at scale into Systematic Reviews (SR) and Meta-Analyses (MA). SRs and MAs summarize the results of controlled trials and provide the highest levels of evidence on the effectiveness of health care interventions [7]. The quality of the results depends on how much the identification is accurate and comprehensive of all the available knowledge on a specific topic. Also, the reliability of an SR is determined by the inclusion of up-to-date contents [8]. However, the increasing number of web repositories and the development of new scientific topics makes the SR process even more complex than it was used to be [9]. Moreover, time requirement and the need for involvement of different professionals make SRs very labor-intensive processes [10].

To retrieve published studies for SRs, researchers can quite easily use search engines like PubMed or Embase. Those platforms are organized in hierarchical branching structures (MeSH and EmTree respectively) facilitating paper's categorization and specific search. Anyway, SRs should be based on a broader set of literature dataset, and ML can be highly useful for the selection of scientific literature related to the investigated topic when the search is not conducted inside indexed search engines, e.g., when they are reported in specialized study registries like ClinicalTrials.gov. In those situations, often the platforms are not designed for hierarchical, or advanced searches and filtering based on text. At the contrary, they are often easy to be harvested retrieving all the textual information included in the documents they store.

The first contribution from my Ph.D. research reported in this dissertation is the development of an automatic system able to replicate a search conducted on specialized search engines to clinical registries. The model purposed showed performances at the same level of the human ones regarding the proportion of relevant study founded and extracted, but with reduced amount of non-relevant studies extracted respect to a manual search conducted by humans.

The second type of narratives investigated is the one reported into EHRs. Those narratives are used by doctors to store and find patients' clinical history, or they are explored by health professionals for public health purposes. Particularly in the latter case, the information extraction from free-text stored into the EHRs might be done through a manual, in-deep, review of individual medical records; however, such a strategy is costly and time-consuming [11]. Conversely, an automatic coding of free-text information reported into EHRs through ML trained models would be a promising opportunity [12], which is increasingly used also for the analysis of emergency department (ED) records, with encouraging results [13, 14].

Further contributions of my Ph.D. research reported here aim to improve the accuracy of infectious disease detection at the population level. That is a crucial public health issue that can provide the background information necessary for the implementation of effective control strategies, such as advertising and monitoring the effectiveness of vaccination campaigns [15]. Indeed, the need for fast, cost-effective, and accurate detection of infection rates has been widely investigated in recent literature [16]. Notably, the combination of increased EHR implementation in primary care, the growing availability of digital information within the EHR, and the development of data mining techniques offer great promise for accelerating pediatric infectious disease research [17]. Although EHRs data are collected prospectively in real-time at the point of health care delivery, observational studies intended to retrospectively assess the impact of clinical decisions are likely the most common type of EHR-enabled research [17].

Accessing daily data activities of pediatric general practitioners and family pediatricians is a unique resource, both for studying specific diseases, as well for pharmacoepidemiologic and pharmacoeconomic analysis [18–20]. For two of the studies reported in this dissertation, we had the opportunity to access to the Pedianet [21] database which focuses on children aged 0-14 years [22–25] and records reasons for accessing healthcare, diagnosis and clinical details. The sources of those data are primary care records written in the Italian language, and which are filled in by pediatricians with clinical details about diagnosis and prescriptions; they also contain details about the eventual hospitalization and specialist referrals.

A whole investigation on pediatric ED visits collected from nine hospitals in Nicaragua was made by one of the researches conducted during my Ph.D. and reported here. The availability of computerized and coded patients' information (e.g., signs, symptoms, admission diagnosis) is crucial for the successful monitoring of ED visits, with the purpose of epidemiological surveillance. Anyway, using information on ED visits for epidemiological research is still challenging [26]. The main barrier is represented by the employment of heterogeneous data collection systems, regarding methods of data collection, type of data collected, data structure, data format, lack of consistency and underuse of coding systems of diseases and injuries, and the widespread use of narrative free-text, especially in low- and mid-income countries.

The model developed shown results consistent with the consideration of Nicaragua like a pre-transitional country, characterized by a high prevalence of infectious diseases and adverse maternal and neonatal outcomes [27]. Moreover, it has highlighted the discrepancy in the performances according to the children's gender. Considering that the overall performances still high even considering variation inside the 95% confidence interval and that the model is trained on textual data only, this could suggest further investigation to clarify if the reason can rely on a lower accuracy in reporting the diagnoses for female children compared with males.

1.4 MAIN REMARKS

To develop models that can be of help in the clinical settings investigated, classical shallow ML strategies showed highly promising results. On the other hand, they all suffer from some issues.

Firstly, most of them cannot take full advantage of a massive amount of data [28, 29] reaching a plateau in the performances after a certain amount of data provided that cannot be improved providing more data [30]. That is not necessarily an issue if the performances showed are good enough concerning the level of interest, but should be a characteristic to take into account in developing an ML system.

Secondly, all the models trained using those algorithms cannot be improved without a complete re-training from scratch of the full model. It is possible to expand self-ensembled models like RF, adding, or removing weak learners to the ensemble [31]; anyhow, update a trained model without retraining all the trees still a complex problem [32–34]. Moreover, to our knowledge, there are not options to factorize or stratify the "knowledge" they have learned in recyclable pieces suitable for other tasks. Their learning ability relay in a full single black-box from the input to the output because of their shallow intrinsic nature.

Third, excluding self-ensembled models like RF, the level of non-linear complexity of the boundary classification regions, as well as all the interaction and level of correlation between the features considered, must be hard defined by the human's knowledge, expertise or guess. GLM Net must explicitly state all the features involved in known or supposed interactions, as well as the level and nature of those [35, 36]. Moreover, the GLMNet link function, defining the non-linear hypothesis related to the classification task in the feature space, must be hard decided by a human too. SVM theoretically overcome the problem but projecting the feature space in a "higher enough" dimensional one in which the classes are linearly separable is a task often computationally intractable. That force to introduce the well-known kernel-trick to define the non-linear behavior of the problem [37], which makes SVM one of the fastest algorithms but, again, it introduces an hard-decision, made by a human, among a limited amount of reasonable options, for the non-linear hypotheses.

On the other hand, a deep-learning approach can take advantage of the amount of data that are orders of magnitude larger than classical shallow ML models [38]. Then, they can be improved over time, starting to learn from an already trained model and not from scratch only. Moreover, each layer provides a stratum for the model, which can be explored and even visualized to have a better understanding of the model learning process. That permit to consider reduced models, e.g., excluding the last decision layer, to achieve what is known as transfer-learning. Transfer-learning represents the possibility to use a substantial portion of the knowledge of a model, trained on a massive amount of data and possibly to achieve a different task, as a starting point to train a model on a similar, but likely different, task [39].

Among all of these advantages, the network learns non-linearities automatically [40–42]. The network design is the only limitation to the learning ability of non-linearities, and it still hard designed by a human. Anyway, as ML shifted the aim of a model definition from defining the structure of the problem to learning its rule, the deep-learning approach, as well as the RF ones, shifts the design of possible interaction and non-linearities from the exact human definition (e.g., as required for GLMNet or SVM) to the formulation of structural boundaries. Within those boundaries, by the training process, the network can explore all possible non-linearities, and interactions within the features or between them and the outcome(s) [43].

As a final consideration for the end of my research experience in clinical text classification during my Ph.D., I can retain that adopting a deep-learning approach leads to several specific advantages respect to the shallow counterparts. The first one is related to the pre-processing step. There are no more needs to hand-crafting features like n-grams, or even performing spelling corrections [44–46]. The ability of a network to find meaningful substructure and interaction inside the data has much more options and focus compared with a single,

multiple, and even expert, hard human decisions for feature engineering. Moreover, taking advantage of the transfer-learning ability of deep networks, scientific and professional communities already started to develop high performing models trained on a massive amount of language-specific textual data [47]. That gives access to model suitable to solve specific problems, different from the original one, even when only a small amount of data is present for the actual task. On the other hand, when enough amount of labeled data are present for the task, it still possible to train a tailored network reusing the architecture of some well-performed model already trained [48–50], as we did for the project aimed to classify otitis, reported in the study at the end of this dissertation.

A further relevant side goal achieved by the projects reported here, relay on the languages investigated. The international research on the use of Machine Learning Techniques (MLTs) to automatically extract or classify information from medical records is applied mainly to English-based datasets. That leads to reliable results for system English-based, but it left more uncertainty for systems developed for text on other languages. On that regard, it is well-known that different languages show different levels of linguistic, morphological, and syntactical complexities [51] (e.g., Spanish exhibits slightly higher levels of morphological complexity compared to English [52]). That inevitably influences how medical information is reported in EHRs which reflect the possible investigation of different ML approaches too. That highlights the need for testing MLTs algorithms on different languages other than the English ones too. Hence, results on non-English databases, like the Italian Pedianet or the Spanish collection of ED visits from the hospitals in Nicaragua are essential to assess general applicability of those techniques on topic which are possibly already explored on English-language environment.

As a closing comment, beside the extraction or the classification of information from free-text, a measure of comparison for the performance reached by the models purposed is crucial. Especially in a clinical environment, “good” or “promising” performance must not be enough, and economic or time-person motivations worth nothing if the performances are not comparable or possibly better than the human ones.

On both the tasks of detect literature of interest for SRs and to classify EHRs at the visit level with the aim of disease classification, the performances of the automatic system developed during my research Ph.D. program have shown levels comparable or higher respect to the human ones. This highlight the real possibility to start to incorporate ML systems on daily clinical practice to produce a concrete improvement when free-text come into account into health care processes.

1.5 DISSERTATION OUTLINE

The growing number of medical literature and textual data in online repositories led to an exponential increase in the workload of researchers involved in citation screening for SRs. Indexed search engines are unable to cover all relevant knowledge; hence, current literature recommends the inclusion of clinical trial registries in SR for MA. In Chapter 2, we provide an automated approach based on SVMs to extend a search on PubMed to ClinicalTrials.gov database, relying on a TM English free-text provided by search results. Fourteen SRs, covering a broad range of health conditions, are used as case studies for external validation.

One of the challenges dealing with SR is the high rate of imbalance between records “of interest” and “not of interest” among the retrieved ones. In Chapter 3, we combine four classical MLT with other four data preprocessing methods for the class imbalance to identify the outperforming strategy on the first stage of the task described in Chapter 2, i.e., screening articles in PubMed for inclusion in SRs. The same fourteen SRs were adopted as case studies. In those scenarios resampling techniques slightly improve the performance of the investigated MLT; on the other hand, their choice can have a considerable impact from the computational perspective.

Moving from providing support in what we can call *pure research*, to the application of MLTs on topics more focused on daily clinical interests, the detection of infectious diseases through the analysis of free-text on EHRs can be of interest and impact for health care. Indeed, it can provide timely and accurate background information for the implementation of preventative measures, such as advertising and monitoring the effectiveness of vaccination campaigns. In 4, we compare three of the main ML variations of the classical GLMs to the aim of detect diseases in pediatric medical records. We used the Italian Pedianet database as a data source for a real-world scenario on the identification of cases of Varicella-Zoster Virus (VZV) infections. We considered data from two different Italian regions' subset of Pedianet: the first including 7,631 patients from Padova and their 1,230,355 records, the second from the Sicilia region, with 2,347 patients and 569,926 records.

Free-text is not used on routinely collected data only: this type of (unstructured) type of data to store information is still widely used in EDs too. Considering such a framework, in Chapter 5, we tested the performance of a ML approach to a classification task on a dataset of pediatric Spanish-language ED's visits. For the analyses, we considered records from nine hospitals in Nicaragua and every free-text discharge diagnoses inside.

In the last chapter, we face a topic of high clinical interest because of one of the leading causes of antibiotic prescriptions to children: the detection and classification of otitis, one of the most common infections in pediatrics. Daily diaries are used by pediatricians to record an exhaustive status of their patients. However, using the very same diaries in a traditional manual human-driven analysis proved to be costly in terms both of person-time (years) and economic resources, and not feasible in practice. In Chapter 6, we purpose an automatic ML system trained to classify all the Pedianet records in six mutually excluding category: non-otitis, otitis, otitis media (OM), acute otitis media (AOM), AOM with tympanic membrane perforation or recurrent AOM. In this final work, all the 6,903,035 pediatric visits collected into Pedianet starting from 1st January 2004 to 23rd 2017 from 144 family pediatricians throughout Italy were used to learn the syntactical structure of all the text they contain. Next, we trained an ensemble model made upon five distinct Deep-Learning (DL) architectures. That model was able to reach 96.59% of accuracy with 95.47% of balanced F1 score through the classes, on a test set never involved in the training phases. It is worth to note that both those measures are comparable to the highest of their corresponding expert human-evaluator ones, i.e. 95.91% and 93.47% respectively.

2 EXTENDING PUBMED SEARCHES TO CLINICALTRIALS.GOV THROUGH A MACHINE LEARNING APPROACH FOR SYSTEMATIC REVIEWS

SUMMARY

Despite their essential role in collecting and organizing published medical literature, indexed search engines are unable to cover all relevant knowledge. Hence, current literature recommends the inclusion of clinical trial registries in Systematic Reviews. This study aims to provide an automated approach to extend a search on PubMed to ClinicalTrials.gov database, relying on Text Mining and Machine Learning Techniques. The procedure starts from a literature search on PubMed. Next, it considers the training of a classifier that can identify documents with a comparable word characterization in the ClinicalTrials.gov clinical trial repository. Fourteen Systematic Reviews, covering a broad range of health conditions, are used as case studies for external validation. Cross-Validated Support-Vector Machine was used as the classifier. The sensitivity was highest (100%) in all Systematic Reviews except one (87.5%) and the specificity ranged from 97.2 to 99.9%. The ability of the instrument to distinguish on-topic from off-topic articles ranged from AUC of 93.4 to 99.9%. The proposed Machine Learning instrument has been shown to have the potential to help researchers in identifying relevant studies along Systematic Reviews process by reducing workload, without losing sensitivity and at a small price in terms of specificity.

This chapter was published as:

Lanera, C., Minto, C., Sharma, A., Gregori, D., Berchialla, P., & baldi, I. (2018). **Extending PubMed Searches to ClinicalTrials.gov Through a Machine Learning Approach for Systematic Reviews.** *Journal of Clinical Epidemiology.* https://doi.org/10.1016/j.jclinepi.2018.06.015

2.1 INTRODUCTION

In medical practice and research, the highest level of evidence is represented by SRs [53]. An SR is the synthesis and evaluation of all relevant literature on a specific topic, aimed to make the available knowledge more accessible to physicians, care providers and decision makers [8]. Conducting an SR is not an easy task since it must follow specific guidelines and protocols, in order to ensure reproducibility of methodology. After the definition of review questions, researchers should accurately identify evidence from articles, studies, and any other relevant documentation. This selection process consists of active search through online and offline literature repositories and final identification of evidence among a large amount of irrelevant information [54]. In the search phase, researchers use keyword combinations to create queries which are able to filter documentations in large medical databases. This operational step is prone to potential bias related to the source of information, specificity, and completeness of search strings. After application of queries, researchers manually complete study selection process by a screening of titles, abstracts, and full-texts and eligibility assessment. Finally, they describe the process using the Preferred Reporting Items for SRs and MAs (PRISMA) flow diagram [55].

The increasing number of web repositories and the development of new scientific topics makes SR process even more complex [9]. Researchers can retrieve information using search engines, such as PubMed or Embase, that are organized in hierarchical branching structures (MeSH and EmTree) facilitating paper's categorization and specific search. This logical and hierarchical structure has important implication in literature search process. Firstly, it facilitates article retrieval by reducing or eliminating potential bias related to the difference in wording, language or brand names. Secondly, even if not exhaustive, MeSH or EmTree structures are useful for limiting the number of records to the relevant ones especially when the study topic is broad and non-specific.

Despite their essential role in collecting and organizing published medical literature, indexed search engines are often unable to cover all relevant knowledge. A meta-analysis based on this type of sources only may provide biased estimates due to the exclusion of relevant not-published information [56]. Furthermore, it has been proven how trial findings can influence the probability of publication and the presence of selective reporting outcomes [57]. World Health Organization stated how unreported studies could leave a misleading picture of the risks and benefits of a treatment, leading to the use and consumption of not-effective or harmful products[58]. For this reason, SRs should be based on a wide literature dataset, which is essential for clinicians and patients to have a reliable and complete picture of their condition and make informed decisions. Among alternative informative sources, current literature recommends the inclusion of clinical trial registries such as the ClinicalTrials.gov [59, 60]. ClinicalTrials.gov is an international web-based platform organized by US National Library of Medicine providing access to more than 263,373 clinical trials from 202 countries all over the world. Studies are registered and regularly updated by the principal investigator, and records are never removed from the site. On ClinicalTrials.gov, clinicians and patients can retrieve complete information about the disease, intervention, study design and phase, location and contacts, as well as the link to published papers. Some records also included results of the study, such as main characteristics of the population, incidence of adverse events and collected outcomes. Clinical trials registries are important literary sources contributing to an updated evidence-based medicine and may contain data that cannot be found in published papers [61]. It has been estimated that quite the 50% of results reported in ClinicalTrials.gov was not initially available elsewhere, while some other information on serious adverse events was not always reported in the corresponding publication [62, 63]. In a recent study, Baudard and colleagues confirmed how searching through registries does make a difference by identifying additional 122 trials for 41 SRs [60]. Despite their relevant role, clinical trial registries are seldom used as sources of studies for SRs, probably due to difficulties in records management. Main limitations are related to the absence of hierarchical order, poor interfaces, a limited number of synonyms and the impossible combination of different queries. In ClinicalTrials.gov the search strategy is based only on retrieval of one or more text words in the fields of Condition/Disease, Title, Brief Description,

Interventions, Locations, and Country. Text word variations include a limited number of synonyms, but no reference to any hierarchical order or subcategories. Recently the Clinical Trials Transformation Initiative (CTTI) proposed a solution to improve the usability of data included in ClinicalTrials.gov by creating a database for aggregate analysis (AACT) and categorization of clinical trials based on clinical specialty. However, this classification is limited to the definition of Disease/Condition and is not consistent with original MeSH classification that does not allow differentiation among clinical specialties.

This study aims to 1) provide an instrument based on TM and MLT able to perform an automated literature search on clinical trial registries; 2) evaluate usability and effectiveness of the proposed instrument. To reach our objectives, we present a case study based on results reported in a previous paper of Baudard et al [60].

2.2 METHODS

2.2.1 DATA SOURCES

To create and test the instrument for automated literature search, we used two different data sources. First, we used information reported in the article *Impact of searching clinical trials registries in Systematic Reviews of pharmaceutical treatments: methodological Systematic Review and reanalysis of meta-analyses* [60]. This study aimed to identify additional trials not included in original SRs, through a manual search on ICTRP (International Clinical Trials Registry Platform). Specifically, authors adapted and applied on ICTRP the search strings of fourteen SRs on effectiveness of pharmacological treatment for several health conditions (i.e., atrial fibrillation, psoriasis, colorectal cancer, gastric cancer, Alzheimer disease, Parkinson disease, diabetes, rheumatoid arthritis, hypertension). Then, they verified consistency of retrieved records with inclusion criteria listed in the original paper and included relevant trials in final estimation of treatment effectiveness. For our purpose, we used the same fourteen SRs listed in [60]. This information allowed us to recreate search strings for PubMed and compare results of automated search with those reported by the authors. Secondly, we used the full-database of ClinicalTrials.gov downloaded from the website of Clinical Trials Transformation Initiative. The database was organized in pipe-delimited files with data on each single study, such as identifier (NCT number), location, start date, sample size etc. Data could be reported as number, string (i.e., text), date or Boolean (i.e. true and false).

2.2.2 TRAINING DATASETS

We created a training dataset for each one of the fourteen SRs described above. Each training datasets included positive and negative records. Positive records were papers included in the original SRs, while negative records were a sample of papers off topic. Positive records were identified by running the original query in PubMed. When the search strategy did not allow to retrieve all relevant papers, missing citations were manually included in the training set. On the other hand, negative records were retrieved adding the Boolean operator NOT to the original query. In other words, we identified off-topic papers by subtracting records of original search strategy from the complete PubMed database. Negative records were filtered by “Text availability: abstract”, “Article types: Clinical trial”, “Species: Humans” and “Languages: English”. Since PubMed allows for downloading up to 200 citations at a time, “Sort by: Best Match” option was selected to avoid any potential bias in the selection of papers based on Entrez Date. Then, negative records were downloaded in group of 200 every time to achieve a ratio of at least twenty negative records for each positive one. The description of search strings and retrieved records is reported briefly in Table 1 (a more detailed description is reported as a supplementary material Table 2). Finally, first author, year, title and abstract from each positive and negative papers were collected and included in the training set.

2.2.3 TESTING DATASETS

A snapshot of the whole ClinicalTrials.gov was taken at January 5th, 2017. This was composed of a set of pipe-delimited files from which we extracted the following information:

- unique identifier (NCT number);
- brief title;
- official title;
- brief summary;
- detailed description;
- study type (nature of investigation, such as interventional or observational);
- overall recruitment status;
- month and year of study start (enrollment of first participant);
- month and year of primary completion (examination of final participant);
- allocation;
- number of arms;
- study phase;
- minimum age for participant eligibility;
- interventional study model (otherwise the strategy for assigning interventions to participants);
- inclusion of drug product subject to the US FDA (Federal Food, Drug and Cosmetic Act).

We used the brief title, official title and detailed description as textual information to perform our testing search. The other information was used to identify trials (NCT numbers) and to include filters similar to those applied in [60]. Specifically, Baudard and colleagues limited ICTRP results to clinical trials whose overall status were either completed or terminated. Moreover, we applied additional filters using fields consistently with inclusion and exclusion criteria described in the fourteen SRs replicating selection filters used in [60] and in original SRs (see *2.5 Procedure Workflow* for further details). Overall 233,609 trials were finally included in the testing dataset.

2.2.4 TEXT MINING

The TM strategy consists of (i) text preprocessing, (ii) training of the ML classifier, and (iii) estimation of the performance of the classifier on the testing dataset. We have also considered an option to handle the unbalanced data in training set. Text preprocessing steps converted the textual data into numbers. SVM, which is one of the most widely used classifiers for TM [64], was chosen as the classifier and trained using 5-Fold Cross-Validation (CV). In each of the training datasets, the ratio of positive to negative samples was at least 1:20 by construction. This type of data is known as unbalanced data. Hence, on the side of the straight application of the defined procedure we have also used the data handling strategy Random Undersampling (RUS), which randomly removes cases from the majority samples (in our case the negative samples) to make the classes more balanced [65]. We applied the RUS strategy up to obtain a final positive to negative ratio of the class samples of 35:65 according to [66]. This way we have an overall of 28 datasets, two for each SR, i.e., the original one and the one after the application of the RUS.

2.2.5 PROCEDURE WORKFLOW

For each one of the fourteen SRs, the title and the abstract of the retrieved records were merged together, and text preprocessing steps were applied in the following order: conversion to lowercase, removing non-words, stemming words, stripping white space, and building the sequences of every two adjacent words from the original text (bi-grams). Further, DTM was created with this collection of tokens (i.e., a unit of textual information) and the matrix was filled with Term Frequency (TF) weighting scheme. The sparsity of all 14 different DTMs was very high, ranging from 99-100 %. Top 4% of the features were selected according to TF-iDF rank as a tribute to (a double application of) Pareto's rule, i.e., the 80% of the effects comes from 20% of the causes.

These selected features were retained. The SVM was 5-fold cross-validated, and within the CV step, the balancing strategy that is RUS and ratio 35:65 (positive to negative samples), when applied, and reweighting with TF-iDF were applied.

Testing ClinicalTrials.gov dataset went through the same text preprocessing strategy, in the same order and then DTM was created with the TF weighing scheme initially. Further, it was adapted with same features retained from the training dataset and was reweighted with TF-iDF weighing scheme with the same retained iDF weights of the corresponding training DTM, which were retained when applied on the whole training dataset.

Each cross-validated SVM model was applied on the corresponding testing dataset for each SR. Procedure workflow is briefly described in Figure 1. Analyses were carried out in R version 3.4.2 [67] with the packages: *caret*, *tm*, *stringr* and *unbalanced* [68–71].

To compare consistency between manual search in [60] and automated search, we replicated selection filters used in [60] and in original SRs. Thus, positive citations identified by automated search, were limited adding all following filters: 1) recruitment status defined as completed or terminated; 2) interventional design; 3) start dated before search on ICTRP; 4) primary completion dated before the search on ICTRP as reported in [60]; 5) specific filters based on inclusion criteria reported in original SRs. Goodness and robustness of our results were evaluated by verifying the inclusion of the additional clinical trials previously identified by Baudard and colleagues.

2.3 RESULTS

Performance results of the most suitable filter are reported in

Table 3. The sensitivity was highest (100%) in all SRs except one (87.5%), and the specificity ranged from 97.2% to 99.9%. The AUC, which measures the ability of the instrument to distinguish relevant articles from off-topic articles, ranged from AUC 93.4% to 99.9%. Performance of the procedures in which an RUS strategy was implemented was similar (data are not shown). Table 4 reports the numbers of predicted positive citations before and after the application of a selection of filters. Comparison with results of Baudard and colleague's manual search's results on ICTRP is also reported. As shown in the table, filters progressively reduced number of citations (predicted positives), without excluding additional clinical trials identified in **[60]** (true positives).

The only false negative (1 out of 8 positives) pertained to an SR on the role of biological therapy in metastatic colorectal cancer [72] and referred to the study with ClinicalTrials.gov identifier: NCT00079066.

Noteworthy, the total number of records from automated search (predicted positives) was lower than the number of records from manual search in half cases, with a mean of 472 and 2119 maximum records compared with a mean of 572 and 2680 maximum ones retrieved in [60].

2.4 DISCUSSION

Time requirement and the need for involvement of different professionals makes SR a very labor-intensive process [10]. Quality of results depends on how much identification is accurate and comprehensive of all available knowledge on a specific topic. Also, the reliability of an SR is determined by the inclusion of up-to-date contents [8].

Our study proposes a classifier that can extend PubMed searches to clinical trials registries, by reducing efforts and time expenditure without losing accuracy and sensitivity.

Previous researchers highlighted how ML could make the standard SR process more efficient [73]. They focused on living SR, considering starting point as the existence of an initial SR provided by humans. Accordingly, we have provided an instrument which is also usable for the "living" step of updating an SR dataset, but it is specially tailored for the more complex and tricky step of contributing to the base dataset definition/extraction for new sources of data (work left to humans in [73]). Our procedure reached high performance in detecting true positive citations of interest in completely different sources of data from the original one regarding the way of storing meta-data, access, and structure of information, and leaving out only 1 of 133 human-detected positive citations through fourteen independent SRs. From this starting point we have also highlighted how, with simple and quick filtering, the number of false positives can be easily and drastically reduced without affecting the sensitivity of the procedure. This way the work left to humans can be reduced and quite limited on the first run of the living update of the SR, i.e. the part of dataset definition which was completely based on human research until now.

Other studies have shown how an ML approach for the classification of information based on clinical text could be very effective [74] also when tested on databases different (and not subsampled) from the original one [75]. On the other hand, to our knowledge, no other study was conducted on this wide range of differentiated dataset test with hundred-thousands of entries.

Anyway, the strict procedure followed, maintaining all the test sets blinded both from the training ones at every stage and from all the training procedure, make us confident in the quality of the results themselves. In an SR, both very specific positive and very specific negative sets can be selected to create a high-quality training set. This characteristic together with the ability of the SVMs to distinguish the well-separated type of data and the high disproportion of few positive records against a huge number of negative ones have led to the quite perfect results in sensitivity, which is the main characteristic of interest in this contest.

Our study demonstrates the usefulness of ML when scientific literature is not reported in indexed search engines. This is the case of clinical trial registries such as ClinicalTrials.gov, whose interfaces are usually not

sophisticated. Limited functionality has an important impact on process workload and often requires the application of long search strings, multiple searches and the screening of a high number of not-specific records. Moreover, when a researcher wants to use the same query on different search engines and registries, he must adapt each singular term and string according to the specific requirement of each platform. In the case of registries, an adaptation from common search engines (PubMed or Embase) is even more complex due to frequent absence of text functionalities such as truncation or brackets. The use of ML could allow a more accurate and easier translation of queries by reducing the number of not-relevant records.

The main strength of the study is the robustness of the training and testing procedure which was designed to be stable and unbiased. Furthermore, an R-package and a companion Graphical User Interface (GUI) are under development (preliminary version publicly available at https://github.com/UBESP-DCTV/costumer). They are intended as a user friendly tool for healthcare researchers, who will only have only to provide: a) the set of citations finally retained, b) a personalized set of negative citations or the search string used on PubMed (to automatically identify a suitable set of random negative citations), c) an optional set of false positives already known from a previous run or directly the set of filters to be applied on non-textual meta-data. The first part of the feature c) highlights also the usability of the package for a very quick update of the SR, e.g., after the first run (for which the false positives must be manually identified).

2.4.1 LIMITATIONS

Our study has some limitations. Firstly, the adoption of a defined ML algorithm and the use of only one strategy for managing the unbalanced data. We acknowledge that other techniques such as Convolutional Neural Networks (CNNs) are effective in achieving slightly better F-scores [76] over more traditional approaches to biomedical text classification, such as SVM, especially, when there is significant label imbalance. Nevertheless, CNNs typically take at least an order of magnitude more time than traditional classifiers, especially when compared with SVM [77]. Hence, we decided to start our investigation by considering SVM only. We are already working on testing both a wider range of ML techniques and more methods for unbalanced datasets. Anyway, the performance with the choice adopted in terms of number of positives, number of true positives and negatives as well as in terms of computational speed, is already good and we do not expect more improvement. Though small relative increases in specificity can still have a big impact on absolute numbers of false positives. Moreover, filters were manually applied after automatic search and were not yet included in ML instrument. The reason for this choice is related to the fact that inclusion/exclusion criteria are rarely reported in title, abstract or description. Thus it was not possible to make a more accurate automatic selection of trials. That said, similar studies were able to reach a very high level of sensitivity at the cost of a discrete specificity [78]. Explanations for our results lie in the choice of the training set and of the task itself. I.e., positives and negatives are highly-separated by design: the set of positives is the (human filtered) output of a PubMed query string and not of a “rule-free” human selection on the whole PubMed as it was made, e.g., for the classification task performed in [78]. The same applies to the negatives, which are sampled from the output of the negative search of PubMed query string used for positives. As a result, positives share a similar word characterization which is easily identified by SVM and can lead to a near perfect sensitivity and also an excellent specificity.

2.5 CONCLUSION

Following the recommended paradigm for model validation [79, 80], this predictive tool underwent internal validation through CV and external validation on an independent data source. This aspect, in conjunction with the broad range of health conditions analyzed, strongly argues in favor of credibility of the proposed instrument.

3 SCREENING PUBMED ABSTRACTS: IS CLASS IMBALANCE ALWAYS A CHALLENGE TO MACHINE LEARNING?

SUMMARY

The growing number of medical literature and textual data in online repositories led to an exponential increase in the workload of researchers involved in citation screening for systematic reviews . This work aims to combine machine learning techniques and data preprocessing for class imbalance to identify the outperforming strategy to screen articles in PubMed for inclusion in systematic reviews. We trained four binary text classifiers (support-vector machines, k-nearest neighbour, random forest, and elastic-net regularised generalised linear models) in combination with four techniques for class imbalance: random under-sampling, and over-sampling with 50:50 and 35:65 positive to negative class ratios, and none as a benchmark. We used textual data of fourteen systematic reviews as case studies. Difference between cross-validated area under the receiver operating characteristic curve (AUC-ROC) for machine learning techniques with and without preprocessing (delta-AUC) was estimated within each systematic review, separately for each classifier. Meta-analytic fixed-effects models were used to pool delta-AUCs separately by classifier and strategy. Cross-validated AUC-ROC for machine learning techniques (excluding k-nearest neighbour) without preprocessing were prevalently above 90 per cent. Except for k-nearest neighbour, machine learning techniques achieved the best improvement in conjunction with random over-sampling 50:50, and random under-sampling 35:65. Resampling techniques slightly improved the performance of the investigated machine learning techniques. From a computational perspective, random under-sampling 35:65 may be preferred.

This chapter was submitted and currently under revision as:

Lanera, C., Berchialla, P., Sharma, A., Minto, C., Gregori, D., & Baldi, I. (2019). **Screening PubMed Abstracts: is Class Imbalance Always a Challenge to Machine Learning?** *BMC Systematic Reviews.*

3.1 BACKGROUND

The growing number of medical literature and textual data in online repositories led to an exponential increase in the workload of researchers involved in citation screening for SRs. The use of TM tools and MLT to aid citation screening is becoming an increasingly popular approach to reduce human burden and increase efficiency to complete SRs.[10, 73, 78, 81–83]

Thanks to its 28 million citations, PubMed is the most prominent free online source for biomedical literature, continuously updated and organised in a hierarchical structure that facilitates article identification.[84] When searching through PubMed by using keyword queries, researchers usually retrieve a minimal number of papers relevant to the review question and a higher number of irrelevant papers. In such a situation of imbalance, most common ML classifiers, used to differentiate relevant and irrelevant texts without human assistance, are biased towards the majority class and perform poorly on the minority one.[85, 86]. Mainly, three sets of different approaches can be applied to deal with imbalance.[86] The first is the pre-processing data approach. With this approach either majority class samples are removed (i.e., undersampling techniques), or minority class samples are added (i.e., oversampling techniques), to make the data more balanced before the application of an MLT.[65, 85] The second type of approaches is represented by the set of algorithmic ones, which foresee cost-sensitive classification, i.e., they put a penalty to cases misclassified in the minority class, this with the aim to balance the weight of false positive and false negative errors on the overall accuracy.[87] Third approaches are represented by the set of ensemble methods, which apply to boosting and bagging classifiers both resampling techniques and penalties for misclassification of cases in the minority class.[88, 89].

This study examines to which extent class imbalance challenges the performance of four traditional MLTs for automatic binary text classification (i.e., relevant vs irrelevant to a review question) of PubMed abstracts. Moreover, the study investigates whether the considered balancing techniques may be recommended to increase MLTs accuracy in the presence of class imbalance.

3.2 METHODS

3.2.1 DATA USED

We considered the fourteen SRs used and described in [90]. The training datasets contain the positive and negative citations retrieved from the PubMed database, where positives were the relevant papers finally included in each SR. To retrieve positive citations, for each SR, we ran the original search strings using identical keywords and filters. We selected negative citations from those resulting by the addition of the Boolean operator NOT to the original search string (see Figure 3). The whole set of these negative citations was then sampled up to retain a minimum ratio of 1:20 (positives to negatives).

Further details on search strings and records retrieved in PubMed can be found in the supplementary material in [90]. The search date was the 18th of July 2017. For each document (n = 7,494), information about the first author, year, title and abstract were collected and included in the final dataset.

3.2.2 TEXT PRE-PROCESSING

We applied the following text pre-processing procedures to the title and abstract of each retrieved citation: each word was converted to lowercase, non-words were removed, stemming was applied, whitespaces were stripped away, bi-grams were built and considered as a single token like a single word. The whole collection of tokens was finally used to get fourteen document-term matrices (DTMs), one for each SR. The DTMs were initially filled by the TF weights, i.e., the simple counting number of each token in each document. The sparsity

(i.e., the proportion of zero-entries in the matrix) of the DTM was always about 99 per cent (see Table 5). TF-iDF [91] weights were used both for reducing the dimensionality of the DTMs by retaining the tokens ranked in the top 4 per cent and as features used by the classifiers. The TF-IDF weights where applied to DTMs during each CV step, accordingly to the same process described in [90].

3.2.3 CHOSEN LEARNERS

We selected four commonly used classifiers in TM: SVMs [92], k-NN [93], RFs [26], and GLMNet [28]. SVM and k-NN are among the most widely used MLTs in the text classification with low computational complexity [64]. Although computationally slower, RFs have also proved effective in textual data classification [64]. We selected GLMNets as benchmark linear model classifiers.[94]

3.2.4 DEALING WITH CLASS IMBALANCE

Random over-sampling (ROS) and random under-sampling (RUS) techniques were implemented to tackle the issue of class imbalance [65]. RUS removes the majority samples randomly from the training dataset to the desired ratio of the minority to majority classes. Since it reduces the dimensionality of the training dataset, it reduces the overall computational time as well, but there is no control over the information being removed from the dataset [65]. ROS adds the positive samples, i.e., the ones in the minority class, randomly in the dataset with replacement up to the desired minority to majority class ratio in the resulting dataset.

We included two different ratios for the balancing techniques: 50:50 and 35:65 (the minority to the majority). The standard ratio considered is the 50:50. On the other hand, we also examined the 35:65 ratio as suggested in [66].

3.2.5 ANALYSIS

The 20 modelling strategies resulting from any combination of MLTs (SVM, k-NN, RF, GLMNet), balancing techniques (RUS, ROS) and balancing ratios (50:50, 35:65) plus the ones resulting from the application of MLTs without any balancing technique were applied to the SRs reported in [90].

Five-fold CV was performed to train the classifier. The Area Under Receiver Operating Characteristic Curve (AUC-ROC) was calculated for each of the ten random combinations of the tunable parameters of the MLTs. The considered parameters were the number of variables randomly sampled as candidates for the trees to be used at each split for RF, the cost (C) of constraints violation for SVM, the regularization parameter (lambda) and the mixing parameter (alpha) for GLMNet, and the neighborhood size (k) for k-NN. The parameters with the best cross-validated AUC-ROC were finally selected.

RUS and ROS techniques were applied to the training dataset. However, the validation data set was held out before using the text preprocessing and balancing techniques to avoid possible bias in the validation.[95] The whole process is represented in Figure 4.

To compare the results, separately for each MLT, we computed the within SR difference between the cross-validated AUC-ROC values resulting from the application of some balancing technique and the AUC-ROC resulting from the crude application of the MLT (i.e., by the “none” strategy to managing the unbalanced data). For all those delta-AUCs we computed 95% confidence intervals, estimated by the observed CV standard deviations and sample sizes. Next, we pooled the results by MLT using meta-analytic fixed-effects models. To evaluate the results, sixteen forest plots were gridded together with MLTs by rows and balancing techniques by columns, in Figure 5.

3.3 RESULTS

Table 6 reports cross-validated AUC-ROC values for each strategy, stratified by SR. In general, all the strategies achieved a very high cross-validated performance. Regarding the methods to handle class imbalance, ROS-50:50 and RUS-35:65 reported the best results. The application of no balancing technique resulted in a high performance only for the k-NN classifiers. Notably, for k-NN, the application of any method for class imbalance dramatically hampers its performance. A gain is observed for GLMnet and RF when coupled with a balancing technique. Conversely, no gain is observed for SVM.

Meta-analytic analyses (see Figure 5) show a significant improvement of the GLMNet classifier while using any strategy to manage the imbalance (minimum delta-AUC of +0.4 with [+0.2, +0.6] 95% CI, reached using ROS-35:65). Regarding the application of strategies in combination with k-NN, all of them drastically and significantly hamper the performance of the classifier in comparison with the use of the k-NN alone (maximum delta-AUC of -0.38 with [-0.39, -0.36] 95% CI reached using RUS-50:50). About the RF classifier, the worst performance was reached using ROS-50:50 which is the only case the RF did not show a significant improvement (delta-AUC +0.01 with [-0.01, +0.03] 95% CI), in all the other cases the improvements were significant. Last, the use of an SVM in combination with strategies to manage the imbalance shows no clear pattern in the performance: i.e., using RUS-50:50, the performance decreases significantly (delta-AUC -0.13 with [-0.15, -0.11] 95% CI); ROS-35:65 does not seem to have any effect (delta-AUC 0.00 with [-0.02, +0.02] 95% CI); for both ROS-50:50 and RUS-35:56 the performance improves in the same way (delta-AUC 0.01 with [-0.01, +0.03] 95% CI), though not significantly.

3.4 DISCUSSION

Application of MLTs in TM has proven to be a potential model to automatize the literature search from online databases.[10, 73, 78, 81, 82] Although it is difficult to establish any overall conclusions about best approaches, it is clear that efficiencies and reductions in workload are potentially achievable.[83]

This study compares different combinations of MLTs and pre-processing approaches to deal with the imbalance in text classification as part of the screening stage of an SR. The application of such a methodology would allow researchers to make comprehensive SRs, by extending existing literature searches from PubMed to other repositories such as ClinicalTrials.gov where documents with a comparable word characterisation could be accurately identified by the classifier trained on PubMed, as illustrated in [90].

Regardless of the balancing techniques applied, all the MLTs considered in the present work have shown the potential to be used for the literature search from the online databases with AUC-ROCs across the MLTs (excluding k-NN) ranging prevalently above 90 per cent.

Among study findings, the resampling pre-processing approach showed a slight improvement in the performance of the MLTs. ROS-50:50 and RUS-35:65 techniques showed the best results in general. Consistent with the literature, the use of k-NN does not seem to require any approach for imbalance.[96] On the other hand, for straightforward computational reasons directly related to the decrease in the sample size of the original dataset, the use of RUS 35:65 may be preferred. Moreover, k-NN showed unstable results when data had been balanced using whatever technique. It is also worth noting that k-NN-based algorithms returned an error, with no results, three times out of the seventy applications, while no other combination of MLT and pre-processing method encountered any errors. The problem occurred only in the SR of Kourbeti [97] which is the one with the highest number of records (75 positives and 1600 negatives), and only in combination with one of the two ROS techniques or when no technique was applied to handle unbalanced data, i.e., when the dimensionality does not decrease. The issue is known (see for instance the discussion in

https://github.com/topepo/caret/issues/582) when using the caret R interface to MLT algorithms, and manual tuning of the neighbourhood size could be a remedy [68].

According to the literature, the performance of various MLTs was found sensitive to the application of approaches for imbalanced data.[87, 98] For example, SVM with different kernels (linear, radial, polynomial and sigmoid kernels) was analysed on a genomics biomedical text corpus using resampling techniques and reported that normalised linear and sigmoid kernels and the RUS technique outperformed the other approaches tested.[99] SVM and k-NN were also found sensitive to the class imbalance in the supervised sentiment classification.[98] Addition of cost-sensitive learning and threshold control has been reported to intensify the training process for models such as SVM and Artificial Neural-Network, and it might provide some gains for validation performances, not confirmed in the test results.[100]

However, the high performance of MLTs in general and when no balancing techniques were applied are not in contrast with the literature. The main reason could be that each classifier is already showing good performance without the application of methods to handle unbalanced data, and there is no much scope left for the improvement. A possible explanation for such a good performance lies in the type of the training set and features, where positives and negatives are well-separated by design, and based on search strings performing word comparison into the metadata of the documents.[90] Nevertheless, the observed small relative gain in performance (around 1%) may translate into a significant absolute improvement depending on the intended use of the classifier (i.e., an application on textual repositories with millions of entries).

Study findings suggest that there is not an outperforming strategy to recommend as a convenient standard. However the combination of SVM and RUS-35:65 may be suggested when the preference is for a fast algorithm with stable results and low computational complexity related to the sample size reduction.

3.4.1 LIMITATIONS

Other approaches to handle unbalanced data could also be investigated, such as the algorithmic or the ensemble ones. Also, we decided to embrace the data-driven philosophy of ML and compare the different methods without any *a-priori* choice and manual tuning of the specific hyper-parameter for each technique. This with the final aim of obtaining reliable and not analyst-dependent results.

3.5 CONCLUSIONS

Resampling techniques slightly improved the performance of the investigated MLT. From a computational perspective, random under-sampling 35:65 may be preferred.

4 USE OF MACHINE LEARNING TECHNIQUES FOR CASE-DETECTION OF VARICELLA ZOSTER USING ROUTINELY COLLECTED TEXTUAL AMBULATORY RECORDS

SUMMARY

The detection of infectious diseases through the analysis of free text on electronic health reports (EHRs) can provide prompt and accurate background information for the implementation of preventative measures, such as advertising and monitoring the effectiveness of vaccination campaigns. Purpose of this paper is to compare Machine Learning Techniques with application to EHR analysis for disease detection. The PEDIANET database was used as a data source for a real-world scenario on the identification of cases of varicella. The models' training and test sets were based on two different Italian regions' dataset of 7,631 patients and 1,230,355 records, and 2,347 patients and 569,926 records, respectively, for whom a gold standard of varicella diagnosis was available. GLMNet, Maximum Entropy (MAXENT) and LogitBoost (Boosting) algorithms were implemented in a supervised environment and 5-fold cross-validated. The DTM generated by the training set involves a dictionary of 1,871,532 tokens. The analysis was conducted on a subset of 29,096 tokens, corresponding to a matrix with no more than 99% of sparsity ratio. The highest test accuracy was reached by Boosting (96.0% and 95% CI (93.8%, 98.1%)). GLMNet delivered superior predictive accuracy compared to MAXENT (86.6% vs 66.0%). MAXENT and GLMNet predictions weakly agree with each other (AC1 = 0.60, 95% CI of (0.58, 0.62)), as well as with LogitBoost ((AC1 = 0.64, 95% CI of (0.63, 0.66) and AC1 = 0.53, 95% CI of (0.51, 0.55) respectively)). Boosting has demonstrated promising performance in large-scale EHR-based infectious disease identification.

This chapter was submitted and currently under revision as:

Lanera, C., Berchialla, P., Baldi, I., Lorenzoni, G., Tramontan, L., Scamarcia, A., Cantarutti, L., Giaquinto, C., & Gregori, D. (2019). **Use of Machine Learning techniques for case-detection of Varicella Zoster using routinely collected textual ambulatory records.** *Journal of Medical Internet Research.*

4.1 INTRODUCTION

Improving the accuracy of infectious disease detection at the population level is an important public health issue that can provide the background information necessary for the implementation of effective control strategies, such as advertising and monitoring the effectiveness of vaccination campaigns [15].

The need for fast, cost-effective and accurate detection of infection rates has been widely investigated in recent literature [16]. Particularly, the combination of increased Electronic Health Reports (EHR) implementation in primary care, the growing availability of digital information within the EHR, and the development of data mining techniques offer great promise for accelerating pediatric infectious disease research [17].

Although EHRs data are collected prospectively in real time at the point of healthcare delivery, observational studies intended to retrospectively assess the impact of clinical decisions are likely the most common type of EHR-enabled research [17].

Among the high-impact diseases, the prompt identification of Varicella Zoster viral infections is of key interest due to the debate around the need and cost-benefit dynamics of a mass-vaccination program in young children [101, 102].

Challenges in this context arise from both the unique epidemiological characteristics of VZV with respect to information extraction, such as age-specific consultation rates, seasonality, force of infection, hospitalization rates and inpatient days [103], and from the way that medical records are organized, often in free-format and un-coded fields [104]. A critical step is to transform this large amount of healthcare data into knowledge.

Data extraction from free text for disease detection at the individual level can be based on manual, in-depth examinations of individual medical records or, to contain costs and ensure time-tightening and control, by automatic coding. MLTs are the most commonly used approaches [105], which show good overall performance [106, 107]. Nevertheless, few indications are currently available on the most appropriate technique to be used, and comparative evidence is still lacking on the performances of each available technique [108] in the field of pediatric infectious disease research.

In recent years, GLM based techniques have been largely used for TM of EHRs, both as a technique of choice [12] and as a benchmark [94]. The performance of GLMs, specially multinomial or in the simplest case logistic regression, has been indicated as unsatisfactory [109] because they are prone to overfitting and are sensitive to outliers. Enhancements to GLMs have been proposed recently in the form of the lasso, and elastic-net regularized GLM [110] (GLMNet), multinomial logistic regression (MAXENT) and the boosting approach implemented in LogitBoost algorithm [111] to overcome the limitations of naïve GLMs. Nevertheless, to the best of our knowledge, no comparisons have been made among these techniques to determine to what extent improvements are needed.

The purpose of this study is to make comparisons among enhanced GLM techniques in the setting of automatic disease detection [112]. Particularly, these methods will be assessed on their ability of identifying cases of VZV from a large set of EHRs.

4.2 METHODS

4.2.1 ELECTRONIC MEDICAL RECORD DATABASE

The Italian PEDIANET database [21] collects anonymized clinical data from more than 300 pediatricians throughout the country. This database focuses on children aged 0-14 years [22–25] and records reasons for

accessing healthcare, diagnosis and clinical details. The sources of those data are primary care records written in Italian, which are filled in by pediatricians with clinical details about diagnosis and prescriptions; they also contain details about the eventual hospitalization, and specialist referrals.

For the purpose of the study, we were allowed to access only two subsets of the PEDIANET database, corresponding to the data collected between 2004 and 2014 in the Italian regions of Veneto (Northern Italy) and Sicilia (South Italy). Since the Veneto region dataset was larger, it was considered for carrying out the training of the model. The dataset of the Sicilia region provided an independent dataset for testing the model. The main characteristics of the two datasets are reported in Table 7. It is worth to note that the proportion of positive cases of VZV is different in the two databases. Interpreting differences in prevalence between Regions is beyond the purpose of this study; nevertheless, given the smaller prevalence, it is expected a lower Positive Predictive Value (PPV) and a higher in Negative Predictive Value (NPV) on the test set.

The PEDIANET source data includes five different tables. In Table 8, we report a short description of them:

All the tables can be linked at the individual level; i.e., each row of all the tables contains the fields reporting information on dates, the assisting pediatrician's anonymous identifier, and the patients' anonymous identifier, which constitutes the linking key.

4.2.2 *CASE* DEFINITION

The case definition comes directly from the gold standard provided, and the training set for ML was created using those dichotomous labels, i.e., '0' = non-case, i.e., the case was not a VZV case, and '1' = case, i.e., the case was a VZV case.

Training and test sets for ML

Linking by Patient ID, Pediatrician ID, and reporting date, we merged the five tables into a single table consisting of several entries, each of which represents a visit/evaluation of a patient carried out by a pediatrician on a specific day. At this step, the information, other than Patient ID, Pediatrician ID and reporting date, are contained in fifteen columns containing free text mixed with coded text, considered by us as free text as well. Finally, all remaining columns of the table were merged into a single corpus, i.e., a body of text. This process was applied to training the models on 1,230,355 entries (database of the Veneto region) and to test them on 569,926 entries (database of the Sicily region) separately.

4.2.3 *PRE-PROCESSING*

Text analysis by a computer program is possible only after establishing a way to convert text, i.e., human readable, into numbers, i.e., computer-readable. This process is called pre-processing, and it is the first [4] and probably the most important step in data mining [5]. To process the corpus of PEDIANET EHRs included in the training set, we used the following strategy: first, we converted all fields in a text type, lowered the content and cleared it of symbols, punctuation, numbers and extra white spaces. Next, we stemmed the words, i.e., reducing them to their basic form, or "root", which is recognized as one of the most important procedures to perform [113] and constructed 2-gram tokens, which has been shown to be the optimal rank for gram tokenization [114]. After that, we removed all the (stemmed) stopwords, i.e., very common and non-meaningful words, such as articles or conjunctions, from the set of tokens as well as all bigrams containing any of them. We decided for this strategy after exploring different approaches described in [115]. Finally, we created the DTM as a patient-token matrix. To take into the account both the importance of the tokens within a patient, i.e., a row of the DTM, and its discrimination power between patients' records, i.e., the rows of the DTM, we computed the TF-iDF (Term Frequencies – inverse Document Frequencies) weights. TF-iDF weights help to adjust for the presence of words that are more frequent but less meaning [116]. TF-iDF-ij entry is equal to the product of the frequency of the j-th token in the i-th document by the logarithm of the inverse of the number

of documents that contain that token; i.e., the more frequent a word appears in a document the more its weight rises for that document, while the more documents contain the j-th token the more the weight shrink across all the documents [117]. In the initial DTM there were 1,871,532 tokens that appear at least once, with non-sparse/sparse entries ratio of (18,951,304/14,262,709,388). We decided to reduce it up to achieve 99% of overall sparsity. Filtering out the tokens that do not appear in at least 1% of the documents lead to a final sparsity of 94%, i.e. 29,096 tokens that appear at least once for a non-sparse/sparse entries ratio of 13,140,370/208,891,206. The choice of 99% level of sparsity was a tradeoff between the need to retain as many tokens as possible and the computational effort.

The corpus of PEDIANET EHRs comprised in the test set went through the same text preprocessing strategy, in the same order and then DTM was created with the TF weighing scheme initially. Further, it was adapted with the same tokens retained in the training phase, (i.e., adding the missing tokens, weighting them like zero, and removing the ones not included in the training DTM), and was finally reweighted with TF-iDF weighing scheme with the same retained iDF weights of the corresponding training DTM, which were retained when applied on the whole training dataset. Those are necessary step to guarantee that the two feature spaces are the same and that the models trained can be evaluated on the test set.

4.2.4 MACHINE LEARNING TECHNIQUES

Enhancements of GLMs for carrying out TM on EHR have been proposed in the form of the lasso, and elastic-net regularized GLM [111] (GLMNet), multinomial logistic regression (MAXENT) and boosting approach (LogitBoost) [111].

GLMNet is a regularized regression method that linearly combines the L1 and L2 penalties of the lasso and ridge methods applied in synergy with a link function and a variance function to overcome linear model limitations (such as the constant variability among the mean and the normality of the data). The link function selected was the binomial, i.e., the model fit a regularized logistic regression model for the log-odds; while the amount of regularization was automatically selected by the algorithm trough an exploration of one hundred values between the minimum value that drop down all the coefficients to zero and its 0.01 fraction.

MAXENT is an implementation of (multinomial) logistic regression aimed at minimizing the memory load on large datasets in R and is primarily designed to work with the sparse DTM provided by the R package tm [118]. It is proven to provide results mathematically equivalent to a GLM with a Poisson link function [119].

Boosting is a general approach for improving the accuracy of any given learning algorithm. We used the adaptations of Tuszynski [120] to the original algorithm, i.e., LogitBoost [121, 122], which is aimed at making the entire process more efficient while applying it on large datasets. The standard boosting technique [122] is applied to the sequential use of a decision stump classification algorithm as a weak learner, i.e., a single binary decision tree. The number of stumps considered is the same of the columns provided in the training set.

Those techniques are chosen among computationally treatable algorithms for use with large datasets [118]. GLMNet and MAXENT, respectively, represent classical benchmark approaches to linear and logistic classification in a manner that differs from LogitBoost, which is a modern boosted tree-based ML approach [123, 124]. Moreover, LogitBoost generalizes the classical logistic models by fitting a logistic model at each node [125] and shows an alternative point of view with regards to models, such as the GLMs, for which the structure of the learner must be chosen *a priori* [126].

4.2.5 TRAINING AND TESTING

We addressed the issue of internal validation by performing CV on the training set comprising records from the Veneto region. We dealt with external validation by accessing a truly external sample of Pedianet EHRs from another Italian region, Sicily. This accomplishes two tasks: 1) to preserve precision in the training phase; 2) to

complement study findings with external validation results using data that were not available when the predictive tool was developed.

We used a 5-fold CV approach to validate each of the three MLTs on the DTM with the corresponding (by row) "case/non-case" attached labels. All MLTs were simultaneously fitted on the same set of folds to ensure a proper comparison between techniques. Values of k = 10 or k = 5 (especially for large dataset) has been shown empirically to yield acceptable (in term of bias-variance trade-off) error rate [95, 127]. Thus, the choice of 5-folds was driven by the computational complexity, the fewer fold, the fewer complexity.

As measures of performance, we calculated point estimates and 95% Confidence Intervals (95% CI) of:

- PPV or Precision: $\frac{True(T)Positives(Po)}{Predicted(Pr)Po}$ i.e., the fraction of positively identified cases that are true positives;

- NPV: $\frac{TNegatives(N)}{PrN}$, i.e., the fraction of positively identified non-cases that are true negative;

- Sensitivity or Recall : $\frac{TPo}{Real(R)Po}$, i.e., the true positive rate;

- Specificity: $\frac{TN}{RN}$, i.e., the true negative rate;

- Accuracy: $\frac{TP+TN}{Samples(S)}$ i.e., the fraction of cases correctly classified;

- F score: $\frac{2}{\frac{1}{Precision}+\frac{1}{Recall}} = 2\frac{Precision*Recall}{Precision+Recall}$, i.e., the harmonic mean of the PPV (or Precision) and Sensitivity (or Recall).

The Gwet's Agreement Coefficient 1 (AC1) statistics of agreement [128] between the techniques are computed and reported, along with their corresponding 95% Cis. Given A = the number of times both models classify a record as non-case, D = the number of times both models classify a record as a case, and N = the total sample size, then $AC1 = \frac{p-e^{\gamma}}{1-e^{\gamma}}$, where $p = \frac{(A+D)}{N}$, and e^{γ} is the agreement probability by chance and is equal to $2q(1-q)$, where $q = \frac{A1+B1}{2}N$, A1 is the number of records classified as non-case by model 1, and B1 is the number of records classified as non-case by model 2 , AC1 has been used given its propensity to be weakly affected by marginal probability, and therefore it was chosen to manage unbalanced data [129].

All the analyses were implemented in the R System [67] with the computing facilities of the Unit of Biostatistics, Epidemiology and Public Health. The R packages used were: *SnowballC* to stem the words and *RWeka* (to create n-grams) for the pre-processing step; *Matrix* and *SparseM* to manage sparse matrices; *GLMNet*, *MAXENT* and *caTools* for the GLMNet, MAXENT and LogitBoost MLT implementation; *caret* to create and evaluate the CV folds; and *ROCR* to estimate the performance and the *tidyverse* bundle of packages for data management, functional programming and plots.

4.3 RESULTS

The flow chart, from data acquisition to pre-processing, is shown in Figure 6. In the training set, 29,096 initial terms out of 1,871,532 were retained by the sparsity reduction step.

Boosting significantly outperforms all other MLTs on the training set, with a predictive accuracy of 94.8% with 95% CI (94.0%, 95.5%), and a positive predictive value of 95.8 with 95% CI (93.2%, 98.5%). The GLMNet predictor delivered superior predictive accuracy compared to MAXENT, with 85.02% versus 79.7% accuracy, and 73.2% versus 66% positive predictive value, respectively (Table 9).

The same considerations hold on the test set where Boosting, GLMNet and MAXENT reach a predictive accuracy of 96% (93.8%, 98.1%), 86.6% (84.6%, 88.7%), and 66% (56.4%, 75.5%), and a positive predictive value of 63.1% (42.7% – 83.5%), 24.5% (21.0% – 28.0%), and 11.0% (9.5% – 12.5%) respectively (Table 10).

Agreement between MLT predictions on the training set was good as measured by AC1 statistics (Table 11).

MAXENT and GLMNet achieved an AC1 of 0.68 with a 95% CI of (0.67, 0.70) between each other, and an AC1 of 0.66, 95% CI of (0.65, 0.68) and 0.74, 95% CI of (0.72, 0.75) with LogitBoost.

With the aim to analyze the most relevant errors, we explored if any records were wrongly classified by all the techniques. It turned out there were three records: one wrongly classified as positive and two wrongly classified negatives by all the MLTs.

4.4 DISCUSSION

The application of MLTs to EHRs constitutes the analytical component of an emerging research paradigm which rests on the capture and pre-processing of massive amounts of clinical data to gain clinical insights and ideally to complement the decision-making process at different levels, from individual treatment to definition of national public health policies. As acknowledged by others [2], the development and application of big data analysis methods on EHRs may help create a continually learning healthcare system [3].

This study trains and compares three different ML approaches towards infectious disease detection at the population level based on clinical data collected in primary care EHRs. In line with the recommended paradigm for model validation [95], MLTs performance underwent internal validation through CV and external validation on an independent set of EHRs.

The predictive capabilities of the developed MLTs are promising even if quite different from each other, e.g., validation accuracy ranges from 80% to 94% and test accuracy from 66% to 96%. Findings on the higher level of accuracy reached by LogitBoost are in line with recent evidence that shows an improvement in general classification problems moving from maximum entropy algorithms to LogitBoost-based ones [130]. LogitBoost is thus confirmed to be a useful technique for solving health-related classification problems [122].

Only three records were wrongly classified by all the models. The first one was wrongly classified as positive probably because the text entry was “vaccini:varicella e mpr” and after the pre-processing the bigram “vaccin varicell” was removed, because the TF-iDF weight was very low. Thus the relationship between VZV and vaccine was lost and remained only the token “varicell”.

The other two records were wrongly classified as negative. For one of them, the misclassification was probably due to an issue in the tokenization. In fact, an anomalus sequence of dashes (“-”) and blanks leaded to the token “- varicella”, which was removed from the feature space, leaving no reference to the disease. The second negative misclassified record referred to a child who was vaccinated for measles, mumps, rubella, and VZV (quadrivalent vaccine). The pediatrician wrote “vaccinazione morbillo parotite rosolia varicella”. The bigram “rosol varicell” (e.g. “rubell varicell”) was weighted 0.361 and hence retained in the feature space and was considered by all the MLTs a pattern of non-infection.

The strength of tree-based models such as LogitBoost also lies in their high scalability. In fact, their computational complexity, i.e., the asymptotical time needed for a complete run, grows linearly with the sample size and quadratically with the number of features used, i.e., the number of tokens considered [125]. Assuming that the richness of the pediatric EHRs vocabulary, i.e., the number of tokens reaches a plateau as data

accumulates over time, the further increase in computational time will only depend linearly on the number of patients.

Any attempt to use EHRs to identify patients with a specific disease would depend on the algorithm, the database, the language and the true prevalence of the disease. As to the generalization of these models to other contexts, we hypothesize that they could be successfully applied also in public health systems with EHR charting in other languages (https://apha.confex.com/apha/2017/meetingapp.cgi/Paper/387228).

We acknowledge that one metric, i.e., sensitivity, specificity, PPV, or NPV may be more important than another, depending on the intended use of the classification algorithm. Thus, LogitBoost model is adequate for ascertaining VZV cases, with a preference for case identification with good sensitivity and excellent specificity.

Finally, if the aim of using ML techniques is to help creating gold standard database, limited agreement between the ML techniques reported in Table 11 suggests that these classification algorithms are not reliable as set of annotators.

4.4.1 *LIMITATIONS*

Some limitations must be acknowledged. First, it is acknowledged that text pre-processing is a crucial step. The way to convert free text into numbers and numbers into features is an essential step of the process and one of the most impactful on the results [5]. For the same reason as before, we decided to follow a standard pre-processing procedure without searching for the best one to obtain results that are at most independent of human tuning.

Furthermore, we set the number of boosting iterations at the same number of features considered. This is suboptimal in computational time because the same performance can be reached with fewer iterations [125]. Nevertheless, we aimed to reach an upper-bound value for the performance estimated in an optimal situation.

4.5 CONCLUSIONS

Given their promising performance in identifying VZV cases, LogitBoost, and MLTs, in general, could be effectively used for large-scale surveillance, minimizing time and cost in a scalable and reproducible manner.

5 ANALYSIS OF UNSTRUCTURED TEXT-BASED DATA USING MACHINE LEARNING TECHNIQUES: THE CASE OF PEDIATRIC EMERGENCY DEPARTMENT RECORDS IN NICARAGUA

SUMMARY

Free text information is still widely used Emergency Department (ED) records. Machine Learning Techniques (MLT) are useful for analyzing narratives, but they have been used mostly for English-language datasets. Considering such a framework, it was tested the performance of an ML classification task of a Spanish-language ED visits database. ED visits collected in the EDs of nine hospitals in Nicaragua were analyzed. Spanish-language, free-text discharge diagnoses were considered in the analysis. Five-hundred RFs were trained on a set of bootstrap samples of the whole dataset (1789 ED visits) to perform the classification task. For each one, after having identified optimal parameter value, the final validated model was trained on the whole bootstrapped dataset and tested. The classification accuracies had a median of 0.783 (95% C.I. 0.779-0.796). MLTs seemed to be a promising opportunity for the exploitation of unstructured information reported in ED records in low- and middle-income Spanish-speaking countries.

This chapter was published as:

Lorenzoni, G., Bressan, S., Lanera, C., Azzolina, D., Da Dalt, L., & Gregori, D. (2019). **Analysis of Unstructured Text-Based Data Using Machine Learning Techniques: The Case of Pediatric Emergency Department Records in Nicaragua**. *Medical Care Research and Review*. https://doi.org/10.1177/1077558719844123

5.1 INTRODUCTION

Monitoring ED visits represents a powerful tool for public health surveillance [26]. It allows for the analysis of frequency (e.g., time trends, seasonality) and distribution of diseases and injuries referred to ED, the early detection of outbreaks (through syndromic surveillance [131, 132] which is currently employed in a growing number of application fields other than the ones for which it has been initially developed, i.e., the early detection of bioterrorism attack [133]), the quality assessment of health services, and, not least, the evaluation of the effectiveness of intervention programs.

The availability of computerized and coded patients' information (e.g., signs, symptoms, admission diagnosis) is crucial for the successful monitoring of ED visits with the purpose of epidemiological surveillance. In view of making ED information readily accessible, since the beginning of the 2000s, several signs of progress have been made in the computerization and coding of ED health records, especially in high-income countries (e.g., in the USA [134]). However, using information on ED visits for epidemiological research is still challenging [26]. The main barrier is represented by the employment of heterogeneous data collection systems, regarding methods of data collection, type of data collected, data structure, data format, lack of consistency and underuse of coding systems of diseases and injuries, and the widespread use of narrative free-text. Particularly, the documentation of ED visits using unstructured free-text is still widely used, since several coding systems are available and are continually being developed, but their use is not straightforward [11].

Such barriers in the analysis of ED datasets for epidemiological research are even more relevant for low- and middle-income countries (LMICs), where the care of acute conditions is not as well established as in high-income countries [135]. Fortunately, in recent years, several initiatives have been put forward to improve the performance of EDs in LMICs, and especially in Latin American ones [136, 137]. However, the wide use of free-text information instead of coded and computerized data collection systems makes the analysis of ED visits epidemiology difficult. These data are useful to monitor ED performance and to target ad hoc interventions to develop emergency care systems in such countries further [138].

5.1.1 CONCEPTUAL *FRAMEWORK*

Given such a framework, besides a progressive development of a standardized data collection system for ED visits, in both high- and LMICs, it is crucial to adopt approaches of analysis allowing for the exploitation of unstructured, text-based, ED medical records currently available. Data extraction from free-text ED health records might be done through a manual, in-deep, review of individual medical records; however, such a strategy is extremely expensive and time-consuming [11]. Conversely, the automatic coding of free-text information reported in ED health records through appropriate MLTs would be a promising opportunity [12], which is increasingly used also for the analysis of ED records, with encouraging results [13, 14]. However, the research on the use of MLTs to automatically extract information from medical records is still at an early stage, and it is applied mainly to the English-based datasets. Only a few examples are available in the literature about the application of MLT to the Spanish language [139–142], which is one of the most widespread languages worldwide. In addition to that, it is well-known that different languages show different levels of linguistic, morphological, and syntactical complexities [51] (e.g., Spanish exhibits slightly higher levels of morphological complexity compared to English [52]). This inevitably influences how medical information is reported in ED health records and, consequently, the accuracy of automatic classification algorithms. This highlights the need for testing MLTs algorithms on different languages other than the English ones.

5.1.2 NEW CONTRIBUTION

Considering the usefulness of ED data for monitoring population's health care needs, but the wide heterogeneity of data collection systems employed in the EDs and, not least, the wide use of free text information instead of coded ones, it is crucial to develop analysis approaches able to exploit the ED data available for deriving

useful information to monitor population's health. MLTs would be a promising approach of analysis of free-text medical information, but their use is still limited, and most of the studies have been done on English language datasets. Considering such a framework, it was tested the performance of a ML classification task of Spanish free-text discharge diagnoses reported in an ED visits database from Nicaragua.

5.2 METHODS

5.2.1 *ITALY*-NICARAGUA *COOPERATION PROJECT*

Data were derived from an international cooperation project between Italian and Nicaraguan pediatricians aimed at setting up a pediatric emergency clinical network in Nicaragua. The project started in 2011 and was carried out thanks to the partnership between the Regione Lombardia; the IRCCS Fondazione Ca' Grande – Policlinico Milano, the Department of Women's and Children's Health– University of Padova, the Nicaraguan government and La Mascota Hospital in Managua.

Nine Nicaraguan hospitals were included in the project: one referral center, La Mascota Hospital located in Managua, the capital city of Nicaragua, and eight referring hospitals located in the towns of Chinandega, Granada, Juigalpa, Jinotega, Matagalpa, Masaya, Bluefields, and Puerto Cabeza. Clinical resources and pediatrician coverage greatly varied between hospitals making pediatric emergency care of acutely ill or injured patients challenging.

5.2.2 *DATA SOURCE*

An electronic data collection system was developed, using *FileMaker Pro 11.0v3* (Santa Clara, CA, USA), as part of the international cooperation project to monitor the clinical outcomes of patients presenting to the ED with urgent or emergent clinical conditions based on the inclusion criteria available as Supplementary Material (Table S1). All the ED visits entered in the data collection system, according to the inclusion criteria, were used in the analysis. Such a system, initially developed with the goal to use it as a base for telemedicine communication with the referral hospital, worked within an intranet system between the referring hospitals and the referral center.

Data available in the system were represented by children's demographic characteristics (age and gender) and clinical history, vital signs (body temperature, blood pressure, heart and breathing rates, and oxygen saturation), results of laboratory tests, diagnostic and therapeutic interventions (if performed), discharge diagnosis, outcomes of the ED visit (hospitalization, transfer to another hospital, death, discharge from ED). Most information was reported in Spanish narrative free-text.

For the study, we focused on ED visits reported in the data collection system in 2012 for which discharge diagnosis was available. The full dataset (ED visits collected in 2012) was represented by 2723 ED visits, and those for which discharge diagnosis was available were 1789 (66%).

5.2.3 *DISCHARGE DIAGNOSIS CLASSIFICATION: THE GOLD STANDARD*

The free-text discharge diagnoses were manually revised and classified by an independent peer-review group of expert pediatricians. The classification comprised ten different classes, including diseases of the cardiovascular, gastrointestinal, metabolic, neurological, respiratory systems, tropical diseases, injuries, poisonings, burns, and others. Such classification was considered as the gold standard. Table 12 reports the variables available in the dataset after the manual classification. The variable reporting the final discharge diagnosis (i.e., discharge diagnosis) was the basis to create the set of tokens used as predictors. The variable reporting the manual classification (i.e., manual classification, which represents the gold standard) was used as the target variable in the classification procedure.

5.2.4 DATA IMPORT, PRE-PROCESSING, AND MANAGEMENT

Original data were available in Excel file format. For the analysis using MLT, they were converted in CSV using the UTF-8 character's encoding. Data pre-processing [5] consisted in the transformation of all characters in lower-case letters, in the removal of all non-alphabetical characters and extra white spaces, and the transformation of each word to its corresponding *lemmata* (i.e., term reported in the dictionary). Every single word and every consecutive sequence of two words (bigrams) were considered as *tokens*.

A DTM was then built up. Each column in a DTM corresponds to a *token* and each row to a discharge diagnosis. It was reported the TF-iDF [143] in each cell of the DTM. The TF-iDF consists in the product between the TF (number of times that a token was reported in a free text diagnosis record), and the inverse of the logarithm of DF (number of free text diagnosis records in which a token appeared), thus providing information on the frequency a token appeared in the diagnoses. The most important tokens (including bigrams) are reported in Table S2 of the Supplementary Material.

5.2.5 DATA ANALYSIS AND MLT TRAINING

To obtain a fair estimation of the performance ranges, the strategy adopted for the analyses was to repeat the whole training procedure on five hundred bootstrap resamples of the dataset. Each training procedure involved the fitting of a set of RFs MLT [144, 145]. The classification task was to classify the manual-identified diagnoses' classes (i.e., the gold standard) using only the text of discharge diagnoses. Each RF was trained considering a forest with 500 trees. The number was set large enough to reach the stability of the votes in the classification model (Figure 7). For each RF, the optimal number of variables (*tokens*) to be sampled and selected for the training procedure, namely *mtry* parameter, was established independently for each one. The *mtry* selection strategy was to perform five repetitions of a 10-fold CV procedure [146]. This was the optimal *mtry* selected to guarantee the optimal trade-off between bias and variance of the models estimated. As a set of options for the *mtry* search, the procedure considered a pseudo-exponential sequence of possible values (i.e., 3, 10, 30, 100, 300, 1000 up to the maximum number of variables -*tokens*- available).

Once the optimal *mtry was* chosen (through the five repetitions of the 10-fold CV procedure), a final validated model (i.e., a brand new RF made up of new 500 trees), was trained on the whole bootstrapped dataset (1789 bootstrapped ED visits), and tested on its Out-Of-Bag (OOB) set, i.e., the observation initially excluded by the bootstrap selection and hence never seen by the whole training procedure. The strategy is reported in Figure 8.

5.2.6 STATISTICAL ANALYSIS AND ESTIMATION OF MLT PERFORMANCE

Descriptive statistics were reported as median (I and III quartiles) for continuous variables, and percentages (absolute numbers) for categorical variables. Thanks to the bootstrap procedure adopted, the classification task could have been evaluated by the Out-Of-Bag (OOB) classification performance of the final trained RF [147] for each one of the 500 bootstrapped RFs, i.e., the performance of every one of this final set of forests were assessed on the set of observations not included in the bootstrapped dataset used to train the trees in the forest. The quality of the classification task was assessed by computing the accuracy (rate of discharge diagnosis correctly classified, according to the gold standard, by the algorithm) overall and stratified by each class of discharge diagnosis. The set of accuracies of the 500 bootstrapped RFs was computed and reported with their median and the corresponding 95% confidence interval.

5.2.7 SOFTWARE

R software (ver. 3.4.2) [67] was used for the analyses, within the packages rms [148] for the statistical analyses, tidyverse [149] for the data management, lubridate [150] for the date-time data management. Packages stringr [70] and glue [151] were used for the text management, while tm [69], randomForest [145] and caret

[68] were employed for text analyses and ML interface. All the analyses run on a Windows 10 Enterprise desktop computer powered by an Intel(R) quad Core (TM) i7-6700 CPU @ 3.4GHz with x64-based operating system and processor, equipped with 40 GB of RAM. The scripts were implemented to train the trees of the RFs in parallel on 3 (i.e., n-1) cores.

5.3 RESULTS

One thousand seven hundred eighty-nine pediatric ED records reported in 2012 in the data collection system set-up in the context of the Italy-Nicaragua Cooperation Project were considered in the analysis. Most of the children admitted to ED were young children (median age of two years) of male gender (56%). According to the gold standard (manual classification), the discharge diagnoses' class most represented was that about the respiratory system (mainly pneumonia), followed by that of the gastrointestinal tract (diarrhea) (

Table 13). The male gender was the most prevalent in all the discharge diagnoses classes except for the metabolic and the poisoning ones. Children admitted to ED with diagnoses about the metabolic system and affected by tropical diseases were the oldest (median age of 13 and 9 years, respectively).

5.3.1 MACHINE LEARNING CLASSIFICATION TASK PERFORMANCE

Overall three thousand eight hundred ninety-one distinct tokens were considered in the analyses, in particular, they range from two hundred fifty-six distinct tokens for Hospital Juigalpa to one thousand five hundred fifty-two distinct tokens for Hospital La Mascota, and a median of 461 tokens. The overall CPU time (on Intel(R) quad Core (TM) i7-6700 CPU @ 3.4GHz with x64-based operating system and processor, equipped with 40 GB of RAM) to train all the models was of 3968.68 seconds, ranging from 35.33 seconds for Hospital Puerto Cabezas to 3090.29 seconds for Hospital La Mascota, and a median CPU time of 95.56 seconds.

Looking at the classification task, it showed an accuracy of 0.7831 (95% C.I. 0.7792-0.7965) on the dataset overall (Table 14). The analysis of the accuracy of the RF according to discharge diagnoses' classes generally showed good performance. Figure 7 shows the trend of the OOB error from 1 to 500 trees considered for each of the validated bootstrap RF models, showing very good performance of the ML algorithm, with a very low and stable error rate at 500 trees.

The analysis of the RF performance according to the sample characteristics (age and gender) showed a good performance for age (Figure 9). Conversely, the accuracy of the models was better (p-value <0.001) for male gender (0.788 95% C.I. 0.783-0.785) compared to the female ones (0.777 95% C.I. 0.772-0.773).

5.4 DISCUSSION

The present study aimed at assessing the performance of RF-based classification strategy in the automatic classification of free-text discharge diagnoses reported in pediatric ED records from the country of Nicaragua.

Nicaragua is one of the poorest countries in the Western world. In recent years, several efforts have been put forward to try to improve the Nicaraguan healthcare system, although hampered by a lack of resources. From the epidemiological point of view, Nicaragua is still considered a pre-transitional country, characterized by a high prevalence of infectious diseases and adverse maternal and neonatal outcomes [27]. This is consistent with the present analysis since most of the children were admitted to the ED with respiratory and gastrointestinal diseases (mainly respiratory infections and diarrhea).

The analysis of RFs accuracy according to sample characteristics showed that the performance of the classification algorithm was stable over children's age, even though the age group most represented was that of young children. Conversely, the RFs performance varied according to gender. The accuracy of the classification task was better for boys compared to girls. One potential explanation of such finding could be represented by the fact that the algorithm was unsuitable to classify discharge diagnoses in female children. However, this seems very unlikely, given the good performance of the classification algorithm for the overall sample. The lower accuracy in reporting the diagnoses for female children compared with males is more likely to explain our finding. However, there are no available data to support either hypothesis.

Overall, the algorithm's performance was found to be very good, providing new insights about the application of such techniques to ED data. MLTs have been increasingly used in the field of emergency medicine, as it has been shown by a recent literature review [152]. It is worth pointing out that the ED visits included in the analyses were the most severe ones corresponding to 1-2% of all the ED visits. This is even more relevant from the public health perspective since the most severe ED visits are those that require the most careful monitor and the most complex clinical management since they are related to higher morbidity and mortality compared to

the less severe ones. For this reason, an accurate classification of such ED visits is essential to allow for careful planning of the ED activities and resources, especially in LMIC where the care of acute conditions is not as well established as in high-income countries.

The main applications of such techniques to emergency medicine data are the development of predictive risk models, the patients' monitoring, and the integration of such techniques with EDs activities (e.g., in the triage) [152]. Present findings further improve our knowledge about the potentials of the application of MLTs to emergency medicine data. Such an algorithm would be a promising tool to automatically classify information from ED health records for the Nicaraguan government since the only requirement for MLTs use is that the ED records are extractable. This means that the application of the algorithm to free text information might improve *(i)* the epidemiological surveillance of ED visits (e.g., seasonality, identification of infectious diseases outbreaks) to allow for a better plan of ED activities and resources' allocation, *(ii)* the identification of pediatric population healthcare needs, *(iii)* the monitor of the performance of the EDs, and *(iv)* the evaluation of the effectiveness of public health interventions.

5.4.1 *LIMITATIONS*

The main limitations were represented by the fact that the MLTs was applied to a small (1789 ED records) dataset in the Spanish language, which has been only rarely analyzed using MLTs. The fact that the dataset was small represents the main reason why the actual discharge diagnosis categories were broader than those identified by the manual classification (gold standard) and, as a consequence, some discharge diagnosis categories were underrepresented. However, the performance of the ML algorithm in classifying the discharge diagnoses was very good, both overall and by discharge diagnoses' groups. This in line with the very few studies available from international literature about the application of MLTs to the Spanish language, suggesting a good performance of MLT also in this linguistic context [139–142]. Looking specifically at the studies on the analysis of free-text ED records using MLT, our results are in line with those of previous studies, showing good performance of RF [14] and the usefulness of analyzing free-text information to enhance information from medical records [153].

5.5 CONCLUSIONS

Results of the present study showed a good performance of a ML approach for the automatic classification of ED free-text discharge diagnoses in the Spanish language, providing insights for the use of MLT for the exploitation of unstructured information reported in ED records for epidemiologic surveillance in LMICs Spanish-speaking countries and communities. Clearly, further work should be done in testing the algorithm on wider pediatric ED datasets allowing for a more detailed classification, through a strict collaboration between physicians, epidemiologists, and big data specialists.

6 AUTOMATIC IDENTIFICATION AND CLASSIFICATION OF DIFFERENT TYPES OF OTITIS FROM FREE-TEXT PEDIATRIC MEDICAL NOTES IN THE ITALIAN LANGUAGE: A DEEP-LEARNING APPROACH

SUMMARY

tabThere is a high clinical interest in the detection and classification of otitis being one of the most common infections in pediatrics and the main cause of antibiotic prescriptions. Daily diaries are useful for pediatricians to record a more exhaustive status of their patients. However, using the very same diaries in a traditional manual human-driven analysis proved to be costly in terms both of person-time (years) and economic resources. The present work aims to develop an automatic machine learning system trained to classify all the Pedianet records in six mutually-exclusive categories: non-otitis, otitis, otitis media, acute otitis media (AOM), AOM with tympanic membrane perforation or recurrent AOM. Data used comes from the Pedianet database containing 6,903,035 pediatric visits starting from 1st January 2004 to 23rd August 2017 from 144 family pediatricians throughout Italy. A gold standard composed by 4,928 records for training, 723 records for validation and 880 records for tests was developed with a high rate of agreement between two expert evaluators (0.89 weighted Cohen's kappa). A pediatrician specialized in infectious diseases validated the gold standard, allowing us to estimate the expert evaluators' performances too. Six deep-learning architectures were explored and tuned on the validation set, an ensemble model was constructed based on them. The ensemble model reached 96.59% of accuracy with 95.47% of balanced F1 score through the classes. Our ensemble obtained performance higher than our expert evaluators (max accuracy: 95.91%, max balanced F1: 93.47%). Our analysis confirmed that deep learning models could indeed have a practical application in the differential diagnosis of otitis from free text.

This chapter is a preprint for :

Lanera, C., Barbieri, E., Piras, G., Maggie, A., Weissenbacher, D., Doná, D., Scamarcia, A., Cantarutti, L, Gonzalez, G., Giacquinto, C., & Gregori, D. (2019). **Automatic identification and classification of different types of otitis from free-text pediatric medical notes: a deep-learning approach**

6.1 INTRODUCTION

Data from the daily consults of pediatric general practitioners and family pediatricians is an important resource, both for studying specific diseases and for pharmacoepidemiologic and pharmacoeconomic analysis [18–20]. In Italy, Pedianet is an example of an efficient pediatric outpatient network which collects specific data from electronic clinical files filled out by pediatricians during their daily professional activities [21]. With more than 300 Italian pediatricians enrolled throughout the country, this network has been shown its value for conducting epidemiological studies on major pediatric diseases or pharmacovigilance [19, 154–156].

There is a high clinical interest in the differential diagnosis of otitis for public health, as it is one of the most common infections in pediatrics and the main reason for antibiotic prescriptions [157].

Queries on diagnoses in healthcare databases form the basis for clinical research and are usually based on ICD9-CM and ICD10-CM codes. The ability (or desire) to use free text fields is rare, despite increasing evidence of the value of mining the free text portions of the Electronic Health Record (EHRs) [158–161]. Anyway, databases don't always include structured or closed fields for the diagnoses' codes, and often codes are reported as free-text. For instance, Nunes et al. found that amongst patients with Type II diabetes, reports of episodes of hypoglycemia found using ICD codes showed 12.4% of the cohort reported at least one, whilst using the free text fields with automatic language processing methods showed 25.1%, and a combination of the two yielded 32.2% [160]. In order to include not-coded diagnosis, free text fields could be searched using a search string strategy sensitive for the specific diagnosis. However, considering the non-specificity of symptoms, to clinically classify an otitis case could be quite challenging, especially if the recommended diagnostic instruments are not used (e.g. pneumatic otoscope is used in only 3.7% of the single acute otitis media episodes instead of the static otoscope) [162]. Those uncertainties are reflected in the healthcare report, thus in order to include all possible cases, the string for the diagnosis identification could become quite complex and lose specificity. Moreover, the potential episodes should then be manually evaluated and validated in order to exclude any false positive cases.

An automatic machine learning (ML) approach to the problem could effectively use all the textual information and record-time at disposal to detect the diagnosis and classify it efficiently based on severity or other specifications has two primary advantages. Long term, it may reduce the time that humans are directly involved on the task, i.e. humans need to conduct the classification for gold standard preparation, its possible improvements or update, and software implementation only, not requiring involvement with the full dataset. Second, a ML approach to differential diagnosis may reduce the time-to-diagnosis: once implemented and trained, the system would classify new records in real-time to inform a physician's diagnosis. However, this would only be true if the ML system reaches performances comparable to human-levels [163, 164].

The aim of the present work is to develop an automatic deep learning ML system trained to classify otitis from outpatient clinical records and to classify them into six mutually exclusive categories: non-otitis, otitis (not media nor acute), OM (not acute), AOM, AOM with perforation or recurrent AOM (when explicitly stated by the pediatrician into the corresponding EHR). We opted for a deep-learning approach, instead of classical shallow MLTs, for several reason: it can take effectively advantage amounts of data that are orders of magnitude larger than classical shallow ML models [38]. Then, a deep learning model can improve over time, starting to learn from an already trained model and not from scratch only. Moreover, a deep network can learn non-linearities and possible interactions among the features automatically [40–42].

6.2 MATERIALS AND METHODS

6.2.1 DATA *SOURCES*

Data used for the present work comes from a snapshot (DB0) of the Pedianet database containing 6,903,035 visits of 216,976 children collected by 144 family pediatricians starting from 1st January 2004 to 23rd August 2017. The Internal Scientific Committee approved the study and access to the database. Pedianet [21] is an Italian pediatric general practice research database. It contains fields stating the reason for the visit, health status (according to the Guidelines of Health Supervision of the American Academy of Pediatrics), personal details, growth parameters, diagnosis and clinical detail (free text or the 9th International Statistical Classification of Diseases and Related Health Problems system (ICD-9) code), prescriptions (pharmaceutical prescriptions identified by the Anatomical-Therapeutical-Chemical code, specialist appointments, diagnostic procedures, hospital admissions) and outcome data of the children habitually seen by more than 300 family pediatricians distributed throughout Italy. The pediatricians are filling these fields using a standard software (JuniorBit) during routine patient care, then data are anonymized and sent monthly to a centralized database in Padua for validation.

6.2.2 GOLD STANDARD

From DB0, records relevant to the classification were selected (DB1) through a search string similar to the one used by Barbieri et al. [19] but looking at all the free-text fields. The string was built to include a wide range of potential typographical errors and abbreviations, assuming QWERTY standard Italian layout keyboard usage (Regular expression

Table 24, Supplementary WEB materials). Variations included in the developed search string aim to include most of the possible reasonable misspelling or abbreviation which can still be reasonably accurate for decision in the classification.

We sampled the records included for annotation from DB1 in three main sets: training set (DBtrain), validation set (DBvali) and test set (DBtest). For all those three set, we considered the following rules for the sampling strategy: i) same proportion of patients each pediatricians like in DB1, ii) retrieve at least one record from every pediatrician, to grasp at least some of all the possible style variability in DB1, and iii) collecting at least 500 records each set following the suggestion in [165]. DBtrain was sampled from historical records from 2004 to 2007 including near ten times the threshold mentioned, i.e., 4,928 records. The other two datasets, i.e., DBvali and DBtest, were created from recent records from 2008 to 2017, and include 723 and 880 records respectively (**Error! Reference source not found.**Figure 10). Table 15 reports the primary metrics of the database.

To build the DBtrain, DBvali and DBtest gold standard classes, two independent evaluators, expert in pediatric EHR, labelled all the records independently accordingly to six, mutually exclusive classes: 0 - non-otitis; 1 - otitis (not acute); 2 - acute otitis (not media); 3 - AOM (without tympanic membrane perforation nor recurrent); 4 - AOM with tympanic membrane perforation ; 5 - recurrent AOM. Next, a pediatrician specialized in infectious diseases classified the records showing disagreement between the first two evaluators. Class distribution from the DBtrain, DBvali, and DBtest are reported in Figure 11.

With regards to the class related to recurrent AOM cases (i.e., label “5”), the definition used was the one coined by Goycoolea “the condition in a child is defined as having at least three episodes of acute otitis media (AOM) in a period of 6 months, or four or more episodes in 12 months” [166], or with an explicit statement of the pediatricians which mark the case as recurrent. The classification model was asked to consider the latter definition only, i.e., the model does not need to consider dates or to count the number of records previously classified like OMA cases.

Considering that the gold standard has been built from scratch and it is not passed through along period of publicly revisions that can reasonably guarantee it does not contains errors, we considered the agreement between the annotators like an initial measure for the quality of the gold standard itself. Moreover, we set as

human-level performance to the task the better performances among the two annotators measured after the revision of the specialist on the records classified differently by the expert annotators. That provided the base of comparison for the model performances.

Given that the classes are theoretically in order, the weighted Cohen's Kappa ($\boldsymbol{k^w}$) index of agreement between the evaluators is used to compute raters'agreement [167]. Given $\boldsymbol{N}$ is the number of possible classes (in our case $\boldsymbol{N = 6}$) , to compute the weighted Cohen's Kappa a symmetric $\boldsymbol{N \times N}$ matrix of weights ($\boldsymbol{w}$) is built with zeros in the main diagonal and positive values elsewhere in a way that the farther apart the judgments are, the higher the weights assigned. Next, two additional $\boldsymbol{N \times N}$ matrices are needed: the confusion matrix ($\boldsymbol{p}$) given by the evaluators' decisions and the one reporting the values as if they were assigned by chance ($\boldsymbol{e}$), i.e., the cell $\boldsymbol{p_{ij}}$ reports the number of observation that the evaluator A assigned to the class $\boldsymbol{i}$ and the evaluator B assigned to the class $\boldsymbol{j}$, during the cell $\boldsymbol{e_{ij}}$ reports the product between the proportion of records the evaluator A assigned to the class $\boldsymbol{i}$ and the proportion of records that the evaluator B assigned to the class $\boldsymbol{j}$, i.e., the probability that a record would be in $\boldsymbol{p_{ij}}$only by chance. Finally, $\boldsymbol{k^w = 1 - \frac{\sum_{ij=1}^{N} w_{ij} p_{ij}}{\sum_{ij=1}^{N} w_{ij} e_{ij}}}$.

On the other hand, to assess the level of human performances to the task to set a reference to rate the performance of the model, the balanced precision (average of precisions for each class, i.e., $\boldsymbol{\frac{\sum_{i=1}^{N} \frac{correctly\ labelled\ like\ i}{labelled\ like\ i}}{N}}$), the balanced recall (average of recalls for each class, i.e., $\boldsymbol{\frac{\sum_{i=1}^{N} \frac{correctly\ labelled\ like\ i}{records\ in\ the\ class\ i}}{N}}$, also known as balanced sensitivity), the balanced F1 score (harmonic mean of balanced precision and balanced recall, i.e., $\boldsymbol{2 * \frac{balanced\ precision * balanced\ recall}{balanced\ precision + balanced\ recall}}$), and the overall accuracy (i.e., the gross proportion of correctly classified records) were computed comparing the classifications provided by each one of the two expert evaluators with the final gold standard approved by the specialist, i.e., the pediatrician.

6.2.3 PRE-PROCESSING

To consider word similarities, like synonyms or spell-errors, we decided to use a dense representation for words in our models [168]. To create the dense representation model, i.e., the one which converts words into the corresponding dense vectors, we applied the fastText algorithm to the full DB0, i.e., 6,903,035 records, choosing the skip-gram architecture and a feature space of 300 dimensions [169]. Our final look-up dictionaries linking each word to the corresponding 300-dimensional dense vector count 122,591 entries. Nowadays, it is possible to find pretrained embedding models for general language trained from a large amount of text, e.g., Wikipedia. On the other hand, when a context-specific corpus of text is accessible and huge enough, it is possible to achieve better performances by training a context-specific representation [XXX]. With near seven million records of free-text, including diaries, diagnoses, prescription, and specialistic visits, we considered the (unlabeled) DB0 huge enough to train the word embeddings representation like a context-specific one.

With regards to the datasets, we merged the fields from each visit to a single stream of text, i.e. all the text in diagnosis, signs-and-symptoms, diary, prescription, visit description, and visit result fields provided by Pedianet for every visit were considered as a single field. We used the token "__SEP__" to mark the separation between one field and the following one. Furthermore, to avoid a considerable amount of different tokens given by every different possible number appearing in some record, and also considering that our task is not focused on, nor related to, the detection of measurements, dosages or other kinds of numbers, we substitute all the numbers with the token "__NUM__." In DBtrain and DBvali there were 11,544 distinct words/tokens (including "__SEP__" and "__NUM__") we used to define the embedding representation from DB0.

We considered the full embedding dictionary in the models; hence, every word in the sets considered must be present and has a representation. On the other hand, it could be possible than future records will include words which are not present in the current dictionary. For that kind of possible records, the model would

throw an errors when asked to classify them because it does not have a representation for the new unknown tokens. To avoid the problem, we added an additional token "__OOV__" representing any Out-Of-Vocabulary ones, on which the embedding will map every token not present in the vocabulary. We set the vector representation for "__OOV__" to a small random vector with all the components included between -0.1 and 0.1; that is to represent a general word, non-polarized to any preferred direction. Taking into account that every token in the actual sets considered is included in the vocabulary used, this will not affect the current estimation in any way, given that it will never be used. The aim of this vector is only to provide a representation for possible future unknown words, the choice to use a small random vector was driven by the consideration of using a vector which is central on the feature representation space, hence without polarized meaning, but different from the zero vector of the embedding space, which is reserved to another one: the __PAD__.

The input to each network trained was a $(. \times \boldsymbol{d})$ 2-dimensional tensor where "." represent the batch size, i.e., the number of records the network receive in input at every training step, and $\boldsymbol{d}$ represent the number of words to considers from each record. We set $\boldsymbol{d} = \mathbf{1000}$ as a standard value already adopted on similar analyses [170]. Possible records longer than 1000 tokens were truncated to the first 1000 ones, while records smaller to 1000 tokens were padded with the tokens "__PAD__" to the right up to reach a length of 1000 total tokens as well. So, the first hidden layer of each network was the embedding layer which converts the $(. \times \mathbf{1000})$ input tensor to the $(. \times \mathbf{1000} \times \mathbf{300})$ one, accordingly to the embedding lookup dictionary used of size 122,593 (the number of tokens in the dictionary plus the "__OOV__," and "__PAD__").

6.2.4 LEARNING AND TUNING STRATEGIES

Our DBtrain and DBtest come from different distributions, i.e., the first collects record from 2004 to 2007, the second from 2008 to 2017. Extracted from the same distribution of DBtest, the DBvali is used to tune the networks trained, i.e., to evaluate the performance of them with different values of the hyperparameters to select the best set [171–173].

On the other hand, having training set records that match the distribution of the test set can be useful to train better models. That is because our validation (and test) sets, comes from a distinct distribution respect to the training set one. Hence, while validating the models on a set with the same distribution of the test set is crucial, train a model on records from a different distribution respect to the test set only could be suboptimal. To supply data from the distribution of the test set for the learning process, while maintaining data from that distribution but disjoint from the test set on which validate the learning progresses, we retained 300 randomly sampled records from DBvali to evaluate the performance of the models during the learning phase, training the models adding (the same) 418 remaining records from DBvali to the full DBtrain. To asses the concrete impact of this strategy, we will retrain and evaluate the final models selected on the DBtrain alone too.

The networks set up with the weights used in the earliest epoch in which the model reached the best accuracy in the first phase. The aim of this second phase is to allow the embedding weights, which were frozen in the first phase, to be tuned by the learning process. That can have the potential to achieve a further small improvement [174]. Considering this as an experimental approach, we decided that we would consider the fine-tuned models if improvements were shown only.

The best model of every architecture considered was retrained on the whole union set of DBtrain and DBvali sets. The final set of best models, as well as their ensemble, were asked to classify all the records in the DBtest, providing the final measures of performance (Figure 10).

To evaluate the relation between the final performances obtained and the amount of data considered, the final ensemble model were also trained considering firstly the DBtrain only, and next the DBtrain +20%, +40%, +60%, and +80% of the records in the DBval, on the side of the training on the full union of the DBtrain and DBval already mentioned.

6.2.5 ARCHITECTURES EXPLORED

All the networks were trained by the Adam optimizer [175] to minimize the average training cross-entropy loss function among the batches, i.e. $\frac{1}{m}\sum_{m=1}^{M}\sum_{i_m=1}^{I}\sum_{c=0}^{5}\boldsymbol{\delta}_{i_m}^{c}\boldsymbol{log}(\boldsymbol{p}_{i_m}^{c})$, where M is the number of batches in which the training set is divided, $\boldsymbol{I}$ is the size of each butch, i.e., $\boldsymbol{M}\cdot\boldsymbol{I}=|\boldsymbol{training\ set}|$, c is the index for the classes, $\boldsymbol{\delta}_{i_m}^{c}$ is the Kronecker's delta for the class of the i-th record of the m-th batch as compared with c, while $\boldsymbol{p}_{i_m}^{c}$ is the probability assigned by the network for the i-th record of the m-th batch to inherit to the class c.

We explored several different architectures. Common parts of all of them are the input provided, the first hidden layer, i.e., the embedding, and the output layer. Like the output layer, we considered a layer with six neurons to represent all the possible classes. It was activated by the logit function and processed by the softmax function, i.e. $\boldsymbol{p}_{i_m}^{s}=\boldsymbol{softmax}(\boldsymbol{z}_{i_m})^{s}=\frac{e^{z_{i_m}^{s}}}{\sum_{c=0}^{5}e^{z_{i_m}^{c}}}$, where $\boldsymbol{z}_{i_m}=\{\boldsymbol{z}_{i_m}^{0},\cdots,\boldsymbol{z}_{i_m}^{5}\}$ is the vector of logits of the output layer. We use the relu function to activate all the hidden layers, i.e., $\boldsymbol{relu}(\boldsymbol{x})=\mathbf{max}\{\boldsymbol{x},\mathbf{0}\}$.

To maintain under control both the exploding or vanishing gradient events, we apply a batch normalization after each hidden layer [176]. To take under control the overfitting, we considered a drop-out layer, i.e., a layer which randomly ignores a random set of neurons given a rate, after each hidden layer, once batch-normalized. For the embedding layer, we considered a drop-out ratio of 0.2, while for the others, we explored two ratios, i.e., 0.5 and 0.7 [177]. With regards to the batch size, for each network, we explored two options, i.e., $\boldsymbol{M}=\mathbf{8}$ or $\boldsymbol{M}=\mathbf{16}$ [178].

The architectures explored are reported in the section Networks of the Supplementary WEB materials. The final ensemble model was made up on the four networks described (simple embedding excluded) considering the mean of all their probability prediction for each class estimated by their output layer before the application of their softmax activation function. The same softmax was applied next to decide the class assigned to the record by the final ensemble.

6.3 RESULTS

A high weighted Cohen's Kappa level of agreement between the two expert annotators was achieved (0.89). Based on the specialist decisions made on the records they classified differently, the evaluations of their performances on the test set have shown a mean accuracy of 95.86% and a mean balanced F1 score of 91.80. Individual performances are reported in Table 16.

In the first training phase aimed to select the best set of parameters for every architecture, the base architecture considered like reference was the simple embedding which reached 88% of accuracy on the validation set in both the configuration explored. All the other architecture's best models were able to reach 98% accuracy on the validation set, without fine-tuning of the embeddings. On the opposite, the simple embedding architecture was the only one showing improvement after the fine-tuning stage. Indeed, fine-tuning hampers the performances of all the best models of the other architectures. Table 17, Table 18, Table 19, and Table 20reports the accuracies on the validation set for all the models trained. Computational time is also reported in the tables. Given that, fine-tuning was not applied to the final models included in the ensemble one.

Performances on the test set for the best model selected and their ensemble are reported in Table 22. Composed by all the models but the simple embedding one, the ensemble reached 96.59% of accuracy with 95.47% of balanced F1 score. Both those measures are over the highest of their corresponding expert annotator ones, i.e., 95.91% and 93.47%. The classification matrix of the predicted classes vs. the ones reported in the gold standard is reported in Table 23.

Moreover, within the 30 records misclassified by the ensemble model, a subsequent check made by the pediatrician highlighted five errors in the gold standard. Among the other, real, errors the most frequents are the class 0 predicted as 1 (7 times in which there were presence of negations of otitis in 4 times, and in the others the doctor mentioned other doctor's diagnoses, predisposition, and a doubt case), and the class 4 predicted as 3 (6 times, in all cases the doctors reported the perforation with uncommon expressions).

The ensemble model showed slow and almost flat improvement while data were added to DBtrain from DBvali (Figure 12) after the first threshold in which DBtrain was considered alone. That could be a marker of the high amount of information the architectures explored grasped from the data provided from one side, and to the necessity to include in the training set records coming from the distribution of the test set.

6.4 DISCUSSION

In this work, we considered a deep-learning approach for a multiclass classification problem. In particular, we used the Pedianet database as a source of information to classify children visits as visit reporting: i) other than an otitis case, ii) a not media otitis, iii) an OM non acute, iv) an AOM, v) an AOM with tympanic membrane perforation or vi) a recurrent OMA; we trained models using five different deep-learning architectures.

The final ensemble model developed was comparable on accuracy the best human performances observed among two expert evaluators (model: 96.59% vs. best-human: 96.33%), balanced precision (model: 97.03% vs. best-human: 95.91%) and balanced F1 score (model: 95.47% vs. best-human: 93.47%). Regarding the recall, the two evaluators performed quite differently, showing the first one recall of 95.30% and the second one recall of 84.66%. Our model reached a slightly lower lever (model: 93.97%) compared to the best of the two human performances, but considerably higher than the other one.

The achieved performances are highly promising to classify diagnoses based on free-text in pediatric EHRs. To compare them with other similar studies, in one of our recent projects on binary decision for a set of possible discharge diagnoses from the notes of pediatric emergency departments [179] we adopted a bootstrap Random Forest approach [180], reaching a median accuracy of 78.3%. In another one on the same Pedianet database aimed to binary decide if children have been affected or not by Varicella-Zoster Virus (unpublished) we explored main ML generalizations of classical generalized linear models, and we were able to achieve an accuracy of 96% but with an F1 score of 68.5% caused by the low precision of the trained model (63.1%). Results are also consistent in comparison with a recent study on diagnoses evaluation for pediatric diseases from EHRs by Huiying Liang et al. [181] on more than one hundred million of EHRs of over a million of children, which achieve a mean F1 score of 96.33% while showing a mean exact match of 94.10% across the category of clinical data they considered applying a deep recurrent architecture on free-text.

Indeed, a deep-learning model can take advantage of datasets that are orders of magnitude larger than classical shallow machine learning models [38]. Considering the rate of grown of the learning curves of our model, we can be confident that it has already saturated the amount of information it can effectively use from the data. That from one side is an indirect proof that classical shallow model would perform worst on the same task with the same data [38], while on the other side it stimulates the investigation of ways to improve further.

Another advantage of the deep-learning approaches is that they can be improved over time, starting from an already trained model and not every time from scratch, like traditional shallow models. That permit also to consider reduced models, e.g., excluding some of the last layers, to reuse a substantial portion of the knowledge of a model, which has already achieved high performance on a task, as a starting point to train a new model on a different task which can take advantage of similar knowledge [39]. That means, on one side, that our training model could possibly be useful as a starting basis to train other deep-learning models to classify different infections; on the other side, that we can improve our model without restarting the training from scratch.

Moreover, considering that deep-learning models can be merged to combine their knowledge, our results and methodology could also be of interest to improve other deep learning models. Rashidian S. et al. [170] conducted an analysis aimed to identify diabetes, chronic kidney disease and acute renal failure diagnosis considering structured data. They reached an accuracy of 87.12%, 90.91% and 89.06% (F1 score of 80.04%, 75.77%, 66.86%) respectively. Including free text could have had a significant impact on such analyses, especially regarding the understanding of patients' clinical history.

For text analysis, adopting a deep-learning approach leads to further specific advantages, especially when joined with the use of embedding layers for the token representation. The first one is related to the pre-processing step. There are no more needs to hand-crafting features like n-grams, stem the words or taking their lemmata, removing the stopwords, or even performing spelling corrections, as well as for deciding weighting strategies to represent the tokens [44–46]. The ability of a network to find meaningful substructure and interaction inside the data has much more options and focus compared with even expert human hard-decisions to hand-craft new features or select a particular set of weights.

6.4.1 LIMITATIONS

From a clinical point of view, given the difficulties in the diagnosis of otitis, it was not always easy to understand the clinical situation described by the pediatricians. Moreover, in order to fasten the daily clinical practice, some family pediatricians have a pre-compiled sheet for reporting general parameters, which, if amended and not corrected later, sometimes were in contrast with the diagnosis. In this regard, it was useful to have information about other symptoms in order to understand the veracity of the principal diagnosis.

Our gold standard was developed with a high agreement between the two evaluators, which reached expected high performance when compared with the external professional reviewer. On the other hand, a future option of developing a gold standard from specialist expert directly, evaluated by consensus from a team of them, would provide a more accurate estimation of the upper limit of the human performances in classifying that kind, and quantity, of data on the investigated task.

A second possible improvement could be connected to the decision to collapse all the fields provided by Pedianet in one, separating them by a specific token. With more computational effort, we will investigate the usage of a set of initial parallel layer to evaluate each field individually, before to merge them deeply in the network. On the other hand, that would lead to a more complex network requiring different and more powerful computational resources.

Indeed, a limitation of our approach is the computational power: powerful and modern deep-learning architectures like BERT [182] or XLNet [183], which reached the highest performance in text mining challenges, were not computationally feasible on our systems.

6.5 CONCLUSIONS

Our analysis confirmed the potential of deep learning models in identifying and classifying diagnosis from free text. These methodologies could be adopted in other health care databases and can improve healthcare research limiting human errors and time speeding databases interrogations.

6.6 ACKNOWLEDGMENTS

The authors thank all the family pediatricians collaborating in Pedianet.

Alongi Angelo ,Angelini Roberta ,Avarello Giovanni ,Azzoni Lucia ,Balliana Franco ,Barbazza Maria Carolina ,Barberi Frandanisa Maria ,Barbieri Patrizia ,Belluzzi Gabriele ,Benetti Eleonora ,Bezzi Roberto ,Bit Junior ,Boe Franca ,Bollettini Stefano ,Bonfigli Emanuela ,Bruna Andrea ,Brusaterra Ivana ,Budassi Roberto ,Caccini

Massimo ,Cantalupi Laura ,Cantarutti Luigi ,Caprio Luigia ,Castaldo Massimo ,Castelli Stefano ,Castronuovo Serenella ,Cavedagni Monica ,Censini Stefania ,Cera Giuseppe Egidio ,Ciscato Carla ,Clerici Schoeller Mariangela ,Collacciani Giuseppe ,Comaita Fabrizio ,Conte Ugo Alfredo ,Costanzo Nicola ,Cozzani Sandra ,Cuboni Giancarlo ,Curti Valentino ,D'Amanti Vito Francesco ,De Angelis Rita ,De Clara Roberto ,De Marchi Annamaria ,De Nicolò Emanuele ,Del Bono Gian Piero ,Del Ponte Gigliola ,Dell'Antonia Fabio ,Di Giampietro Tiziana ,Di Mauro Giuseppe ,Di Renzo Anna Paola ,Di Santo Giuseppe ,Dolci Marco ,Doria Mattia ,Drago Stefano ,Falco Pietro ,Fama Mario ,Faraci Marco ,Favilli Tania ,Federico Mariagrazia ,Felice Michele ,Ferrara Enrico ,Ferrarese Marta ,Ferretti Michele ,Forcina Paolo ,Frattini Claudio Paolo ,Frison Ezio ,Fusco Fabrizio ,Gallo Giovanni ,Galvagno Andrea ,Gentili Alberta ,Gentilucci Pierfrancesco ,Giampaolo Giuliana ,Giancola Giuseppe ,Giaretta Letizia ,Girotto Silvia ,Gobbi Costantino ,Grelloni Mauro ,Grugnetti Mirco ,Lagrasta Urania Elisabetta ,Landi Massimo ,Lasalvia Paola ,Letta Maria Rosaria ,Lietti Giuseppe ,Lista Cinzia ,Lucantonio Ricciardo ,Luise Francesco ,Luotti Diego ,Macropodio Nadia ,Marine Francesca ,Mariniello Lorenzo ,Marostica Gabriella ,Masotti Sergio ,Meneghetti Stefano ,Milani Massimo ,Milone Stella Vittoria ,Monteleone Angela Maria ,Mussinu Pierangela ,Muzzolini Carmen ,Nicoloso Flavia ,Olimpi Laura Maria ,Palma Maria Maddalena ,Pandolfini Vittorio ,Pasinato Angela ,Passarella Andrea ,Pazzola Pasquale ,Perri Danilo ,Pescosolido Silvana Rosa ,Petrazzuoli Giovanni ,Petrotto Giuseppe ,Picco Patrizia ,Pirola Ambrogina ,Pisanello Lorena ,Pittarello Daniele ,Porro Elena ,Profumo Elisabetta ,Puma Antonino ,Ragazzon Ferdinando ,Rosas Paolo ,Rosignoli Rino ,Rossitto Mariella ,Ruffato Bruno ,Ruggieri Lucia ,Ruscitti Annamaria ,Russo Annarita ,Salamone Pietro ,Sambugaro Daniela ,Saretta Luigi ,Sarno Vittoria ,Sciolla Nico Maria ,Semenzato Flavio ,Senesi Paolo ,Silvan Carla ,Spanevello Valter ,Speciale Sergio Maria ,Speranza Francesco ,Sticco Maura ,Storelli Francesco ,Tamassia Gianni ,Tambaro Paolo ,Toffol Giacomo ,Tonelli Gabriele ,Tummarello Angelo ,Ulliana Antonella ,Venditti Sergio ,Volpe Concetta ,Volpe Francescopaolo ,Vozzi Aldo.

A LIST OF ABBREVIATIONS

AC: Agreement Coefficient

AOM: Acute Otitis Media

AUC-ROC: Area Under Receiver Operating Characteristic Curve

CI: Confidence Intervals

CNN: Convolutional Neural Network

CSV: Comma-Separated Values

CV: Cross-Validation

DAG: Directed Acyclic Graph

DL: Deep Learning

DTM: Document-Term Matrix

ED: Emergency Department

EHR: Electronic Health Record

GLM: Generalized Linear Model

GLMNet: Elastic-Net Regularized Generalized Linear Models

GUI: Graphical user Interface

ICD: International Classification of Diseases

ICTRP: International Clinical Trials Registry Platform

iDF: inverse Document Frequency

k-NN: k-Nearest Neighbors

LMIC: Low- Middle-Income Country

LSTM: Long- Short-Term Memory

MA: Meta Analysis

MaxEnt: Maximum Entropy

ML: Machine Learning

MLT: Machine Learning Technique

NLP: Natural Language Processing

NPV: Negative Predictive Value

OM: Otitis Media

OOB: Out-Of-Bag

PPV: Positive Predictive Value (or, Precision)

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

RF: Random Forest

SOAP: Symptoms, Objectivity, Diagnosis or Prescriptions

SR: Systematic Review

RNN: Recurrent Neural Network

ROS: Random Over-Sampling

RUS: Random Under-Sampling

SR: Systematic Review

SVM: Support-Vector Machine

TF: Term Frequency

TF-iDF: Term Frequencies – inverse Document Frequencies

TM: Text Mining

UTF: Unicode Transformation Format

VZV: Varicella-Zoster Virus

A EXTENDING PUBMED SEARCHES TO CLINICALTRIALS.GOV THROUGH A MACHINE LEARNING APPROACH FOR SYSTEMATIC REVIEWS

A.1 FIGURE

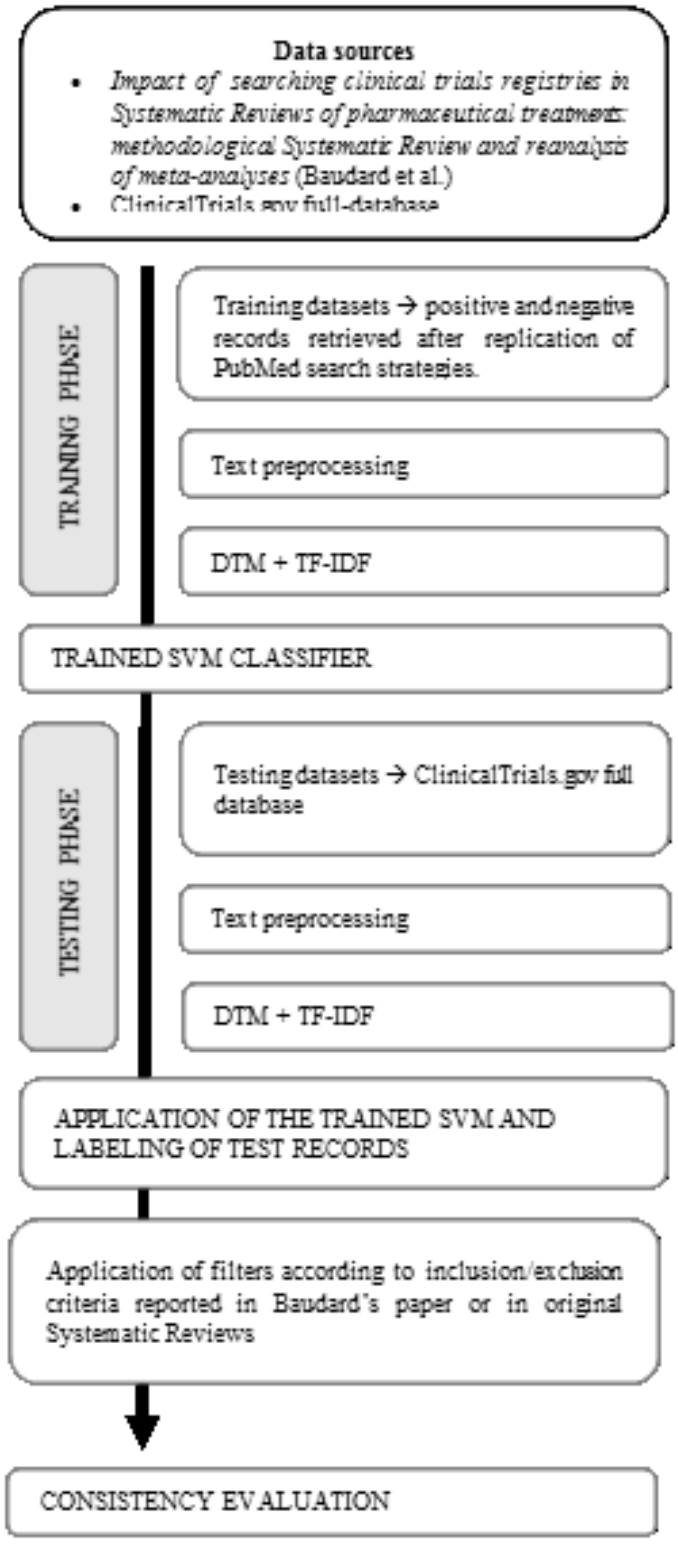

Figure 2 General procedure workflow.

A.2 TABLES

Table 1 Results of PubMed search strategies for the fourteen Systematic Reviews included in [60]. Final training datasets included the sum of positive and negative citations.

Systematic Review	Health condition	Positive records	Negative records
Yang et al. 2014 [184]	Atrial fibrillation	18	400
Meng et al 2014 [185]	Psoriasis	9	200
Segelov et al. 2014 [72]	Colorectal cancer	13	400
Li et al. 2014 [186]	Gastric cancer	6	200
Lv et al. 2014 [187]	Colorectal cancer	12	400
Wang et al. 2015 [188]	Alzheimer's disease	32	800
Zhou at al. 2014 [189]	Parkinson's disease	9	200
Liu et al. 2014 [190]	Type 2 diabetes mellitus	23	600
Douxfils et al. 2014 [191]	Venous thromboembolic events	13	400
Kourbeti et al. 2014 [97]	Rheumatoid arthritis	75	1600
Li et al. 2014 [192]	Primary hypertension	9	200
Cavender et al. 2014 [193]	Venous thromboembolic events	14	400
Chatterjee et al. 2014 [194]	Venous thromboembolic events	18	400
Funakoshi et al 2014 [195]	Solid cancers	43	1000

Table 2 Replication of PubMed search strategies for the fourteen Systematic Reviews included in [60]. Final training datasets included the sum of positive and negative citations reported in bold characters.

Study	Positive search strategy *Sorted by* Most recent	Negative search strategy *Filter for* Abstract & Clinical trial & Humans & English *Sorted by* Best match	Positive records	Negative records
Yang et al. 2014 [184]	(atorvastatin) AND atrial fibrillation AND ("0001/01/01"[PDat]:"2014/04/30"[PDat])	(("0001/01/01"[Date - Publication] : "2014/04/30"[Date - Publication])) NOT ((atorvastatin) AND atrial fibrillation)	n=76 total records n=5 manually added	n=563037 total records
		Citations finally included in the main database	**n=18 positives**	**n=400 negatives**
Meng et al 2014 [185]	(((((((ustekinumab) OR CNTO-1275) OR interleukin 12 23 monoclonal antibody) OR stelara)) AND ((psoriasis) OR (pustulosis of palms and soles))) AND randomized) AND ("0001/01/01"[PDat] : "2013/08/01"[PDat])	((("0001/01/01"[Date - Publication] : "2013/08/01"[Date - Publication]) AND ("0001/01/01"[PDat] : "2013/08/01"[PDat]))) NOT ((((((((ustekinumab) OR CNTO-1275) OR interleukin 12 23 monoclonal antibody) OR stelara)) AND ((psoriasis) OR (pustulosis of palms and soles))) AND randomized))	n=91 total records n=0 manually added	n=538257 total records
		Citations finally included in the main database	**n=9 positives**	**n=200 negatives**
Segelov et al. 2014 [72]	(((((((("Antibodies, Monoclonal"[Mesh]) OR "Antineoplastic Combined Chemotherapy Protocols"[Mesh]) OR "Antineoplastic Agents"[Mesh])) AND ((((("Bevacizumab"[Mesh]) OR "Camptothecin"[Mesh]) OR "Fluorouracil"[Mesh]) OR "Leucovorin"[Mesh])) AND ((((("Colorectal Neoplasms"[Mesh]) OR "Adenocarcinoma"[Mesh])) AND (advanced OR metastatic OR metastases OR metastasis))) AND "Humans"[Mesh]) AND "Randomized Controlled Trial" [Publication Type]) AND ("0001/01/01"[PDat] : "2012/05/31"[PDat])	((("0001/01/01"[Date - Publication] : "2012/05/31"[Date - Publication])) NOT ((((((((("Antibodies, Monoclonal"[Mesh]) OR "Antineoplastic Combined Chemotherapy Protocols"[Mesh]) OR "Antineoplastic Agents"[Mesh])) AND ((((("Bevacizumab"[Mesh]) OR "Camptothecin"[Mesh]) OR "Fluorouracil"[Mesh]) OR "Leucovorin"[Mesh])) AND ((((("Colorectal Neoplasms"[Mesh]) OR "Adenocarcinoma"[Mesh])) AND (advanced OR metastatic OR metastases OR metastasis))) AND "Randomized Controlled Trial" [Publication Type]))))	n=913 total records n=10 manually added	n=499438 total records
		Citations finally included in the main database	**n=13 positives**	**n=400 negatives**
Li et al. 2014 [186]	(((((stomach cancer) OR gastric cancer)) AND S-1) AND fluorouracil) AND ("0001/01/01"[PDat] : "2014/02/20"[PDat])	(("0001/01/01"[Date - Publication] : "2014/02/20"[Date - Publication])) NOT (((((stomach cancer) OR gastric cancer)) AND S-1) AND fluorouracil)	n=1248 total records n=2 manually added	n=557386 total records
		Citations finally included in the main database	**n=6 positives**	**n=200 negatives**
Lv et al. 2014 [187]	((((((("Colorectal Neoplasms"[Mesh]) OR ((colorectal) AND neoplasms)) OR colorectal neoplasms)) AND ((("Cetuximab"[Mesh]) OR cetuximab)) AND ((("Clinical Trial"[Publication Type]) AND "Humans"[Mesh])) AND ("0001/01/01"[PDat] : "2014/02/16"[PDat])	((("0001/01/01"[Date - Publication] : "2014/02/16"[Date - Publication])) NOT ((((("Colorectal Neoplasms"[Mesh]) OR ((colorectal) AND neoplasms)) OR colorectal neoplasms)) AND ((("Cetuximab"[Mesh]) OR cetuximab)))	n=201 total records n=0 manually added	n=557094 total records
		Citations finally included in the main database	**n=12 positives**	**n=400 negatives**
Wang et al. 2015 [188]	(((((((alzheimer's disease[Title/Abstract]) OR (alzheimer[Title/Abstract) OR (AD[Title/Abstract])) AND ((cholinesterase inhibitors[Title/Abstract]) OR (donepezil[Title/Abstract]) OR (galantamine[Title/Abstract]) OR (rivastigmine[Title/Abstract]) OR (metrifonate[Title/Abstract]) OR (tacrine[Title/Abstract]) OR (antipsychotics[Title/Abstract]) OR (haloperidol[Title/Abstract]) OR (thioridazine[Title/Abstract]) OR (thiothixene[Title/Abstract]) OR (chlorpromazine[Title/Abstract]) OR (acetophenazine[Title/Abstract]) OR (clozapine[Title/Abstract]) OR (olanzapine[Title/Abstract]) OR (risperidone[Title/Abstract]) OR (quetiapine[Title/Abstract]) OR (aripiprazole[Title/Abstract]) OR (antidepressants[Title/Abstract]) OR (setraline[Title/Abstract]) OR (fluoxetine[Title/Abstract]) OR (citalopram[Title/Abstract]) OR (trazodone[Title/Abstract]) OR (mood stabilizers[Title/Abstract]) OR (valproate[Title/Abstract]) OR (carbamazepine[Title/Abstract]) OR (lithium[Title/Abstract]) OR (anticonvulsants[Title/Abstract]) OR (benzodiazepines[Title/Abstract]) OR (memantine[Title/Abstract]) OR (psychotropic drugs[Title/Abstract])) AND ((behavioral and psychological symptoms of dementia) OR (BPSD) OR (neuropsychiatric symptoms) OR (behavior)))) AND (	((("0001/01/01"[Date - Publication] : "2013/11/30"[Date - Publication])) NOT ((((((alzheimer's disease[Title/Abstract]) OR (alzheimer[Title/Abstract) OR (AD[Title/Abstract])) AND ((cholinesterase inhibitors[Title/Abstract]) OR (donepezil[Title/Abstract]) OR (galantamine[Title/Abstract]) OR (rivastigmine[Title/Abstract]) OR (metrifonate[Title/Abstract]) OR (tacrine[Title/Abstract]) OR (antipsychotics[Title/Abstract]) OR (haloperidol[Title/Abstract]) OR (thioridazine[Title/Abstract]) OR (thiothixene[Title/Abstract]) OR (chlorpromazine[Title/Abstract]) OR (acetophenazine[Title/Abstract]) OR (clozapine[Title/Abstract]) OR (olanzapine[Title/Abstract]) OR (risperidone[Title/Abstract]) OR (quetiapine[Title/Abstract]) OR (aripiprazole[Title/Abstract]) OR (antidepressants[Title/Abstract]) OR (setraline[Title/Abstract]) OR (fluoxetine[Title/Abstract]) OR (citalopram[Title/Abstract]) OR (trazodone[Title/Abstract]) OR (mood stabilizers[Title/Abstract]) OR (valproate[Title/Abstract]) OR (carbamazepine[Title/Abstract]) OR (lithium[Title/Abstract]) OR (anticonvulsants[Title/Abstract]) OR (benzodiazepines[Title/Abstract]) OR (memantine[Title/Abstract]) OR (psychotropic drugs[Title/Abstract])) AND ((behavioral and psychological symptoms of dementia) OR (BPSD) OR (neuropsychiatric symptoms) OR (behavior))))))	n=1091 total records n=5 manually added	n=547357 total records

	"0001/01/01"[PDat] : "2013/11/30"[PDat]))) AND English[lang])			
		Citations finally included in the main database	**n=32 positives**	**n=800 negatives**
Zhou at al. 2014 [189]	((((((pramipexole extended release) OR ropinirole prolonged release) OR rotigotine transdermal patch)) AND (((parkinson's disease) OR parkinson's) OR PD))) AND ("0001/01/01"[PDat] : "2013/02/10"[PDat])	(("0001/01/01"[Date - Publication] : "2013/02/10"[Date - Publication])) NOT ((((((pramipexole extended release) OR ropinirole prolonged release) OR rotigotine transdermal patch)) AND (((parkinson's disease) OR parkinson's) OR PD)))	n=107 total records n=2 manually added	n=523622 total records
		Citations finally included in the main database	**n=9 positives**	**n=200 negatives**
Liu et al. 2014 [190]	((("Diabetes Mellitus, Type 2"[Mesh]) AND (((((((((dpp-iv inhibitors) OR vildagliptin) OR sitagliptin) OR saxagliptin) OR alogliptin) OR linagliptin) OR dutogliptin) OR metformin) OR sulfonylureas))) AND Randomized Controlled Trial[ptyp] AND ("0001/01/01"[PDat] : "2013/01/31"[PDat]) AND Humans[Mesh] AND English[lang]	((("0001/01/01"[Date - Publication] : "2013/01/31"[Date - Publication])) NOT ((("Diabetes Mellitus, Type 2"[Mesh]) AND (((((((((dpp-iv inhibitors) OR vildagliptin) OR sitagliptin) OR saxagliptin) OR alogliptin) OR linagliptin) OR dutogliptin) OR metformin) OR sulfonylureas))) AND Randomized Controlled Trial[ptyp]))	n=1427 total records n=0 manually added	n=521120 total records
		Citations finally included in the main database	**n=23 positives**	**n=600 negatives**
Douxfils et al. 2014 [191]	(((((dabigatran) OR dabigatran etexilate) OR BIBR 1048)) AND ((((((randomized controlled trial) OR randomized clinical trial) OR randomized trial) OR randomised controlled trial) OR randomised clinical trial) OR randomised trial)) AND ("0001/01/01"[PDat] : "2013/12/08"[PDat]) AND English[lang]	(("0001/01/01"[Date - Publication] : "2013/12/08"[Date - Publication])) NOT (((((dabigatran) OR dabigatran etexilate) OR BIBR 1048)) AND ((((((randomized controlled trial) OR randomized clinical trial) OR randomized trial) OR randomised controlled trial) OR randomised clinical trial) OR randomised trial))	n=276 total records n=3 manually added	n=548647 total records
		Citations finally included in the main database	**n=13 positives**	**n=400 negatives**
Kourbeti et al. 2014 [97]	(((((((rheumatoid) AND arthritis)) AND randomized) AND (((((((((infliximab) OR etanercept) OR adalimumab) OR certolizumab) OR golimumab) OR anakinra) OR abatacept) OR tocilizumab) OR rituximab)) AND ("0001/01/01"[PDat] : "2013/06/24"[PDat]) AND English[lang])) AND ("0001/01/01"[PDat] : "2013/06/24"[PDat])	(("0001/01/01"[Date - Publication] : "2013/06/24"[Date - Publication])) NOT (((((rheumatoid) AND arthritis)) AND randomized) AND (((((((((infliximab) OR etanercept) OR adalimumab) OR certolizumab) OR golimumab) OR anakinra) OR abatacept) OR tocilizumab) OR rituximab))	n=827 total records n=4 manually added	n=534353 total records
		Citations finally included in the main database	**n=75 positives**	**n=1600 negatives**
Li et al. 2014 [192]	(((((((("Angiotensin Receptor Antagonists"[Mesh]) OR ((((((((((((((((abitesartan) OR azilsartan) OR candesartan) OR elisartan) OR embusartan) OR eprosartan) OR forasartan) OR irbesartan) OR losartan) OR milfasartan) OR olmesartan) OR saprisartan) OR tasosartan) OR telmisartan) OR valsartan) OR zolasartan))) AND ((((("Angiotensin-Converting Enzyme Inhibitors"[Mesh]) OR angiotensin converting enzyme inhibit*)) OR ((acei) OR alacepril) OR altiopril) OR ancovenin) OR benazepril*) OR captopril) OR ceranapril) OR ceronapril) OR cilazapril*) OR deacetylalacepril) OR delapril) OR enalapril*) OR epicaptopril) OR fasidotril*) OR foroxymithine) OR fosinopril*) OR gemopatrilat) OR idapril) OR imidapril*) OR indolapril) OR libenzapril) OR lisinopril) OR moexipril*) OR moveltipril) OR omapatrilat) OR pentopril*) OR perindopril*) OR pivopril) OR quinapril*) OR ramipril*) OR rentiapril) OR saralasin) OR s nitrosocaptopril) OR spirapril*) OR temocapril*) OR teprotide) OR trandolapril*) OR utibapril*) OR zabicipril*) OR zofenopril*))) AND (((hypertension) OR hypertens*) OR "Blood Pressure"[Mesh])) AND (((((((("Randomized Controlled Trial" [Publication Type]) OR "Controlled Clinical Trial" [Publication Type]) OR randomi*[Title/Abstract]) OR placebo[Title/Abstract]) OR "Clinical Trials as Topic"[Mesh]) OR randomly[Title/Abstract]) OR trial[Title])) AND "Humans"[Mesh]) AND ("0001/01/01"[PDat] : "2014/02/15"[PDat])	((((("0001/01/01"[Date - Publication] : "2014/02/15"[Date - Publication])) NOT ((((((("Angiotensin Receptor Antagonists"[Mesh]) OR ((((((((((((((((abitesartan) OR azilsartan) OR candesartan) OR elisartan) OR embusartan) OR eprosartan) OR forasartan) OR irbesartan) OR losartan) OR milfasartan) OR olmesartan) OR saprisartan) OR tasosartan) OR telmisartan) OR valsartan) OR zolasartan))) AND ((((("Angiotensin-Converting Enzyme Inhibitors"[Mesh]) OR angiotensin converting enzyme inhibit*)) OR ((acei) OR alacepril) OR altiopril) OR ancovenin) OR benazepril*) OR captopril) OR ceranapril) OR ceronapril) OR cilazapril*) OR deacetylalacepril) OR delapril) OR enalapril*) OR epicaptopril) OR fasidotril*) OR foroxymithine) OR fosinopril*) OR gemopatrilat) OR idapril) OR imidapril*) OR indolapril) OR libenzapril) OR lisinopril) OR moexipril*) OR moveltipril) OR omapatrilat) OR pentopril*) OR perindopril*) OR pivopril) OR quinapril*) OR ramipril*) OR rentiapril) OR saralasin) OR s nitrosocaptopril) OR spirapril*) OR temocapril*) OR teprotide) OR trandolapril*) OR utibapril*) OR zabicipril*) OR zofenopril*))) AND (((hypertension) OR hypertens*) OR "Blood Pressure"[Mesh])) AND (((((((("Randomized Controlled Trial" [Publication Type]) OR "Controlled Clinical Trial" [Publication Type]) OR randomi*[Title/Abstract]) OR placebo[Title/Abstract]) OR randomly[Title/Abstract]) OR trial[Title]))))	n=1441 total records n=0 manually added	n=556668 total records
		Citations finally included in the main database	**n=9 positives**	**n=200 negatives**
Cavender et al. 2014 [193]	(((((bivalirudin) OR Angiomax) OR Hirulog)) AND ((((((stent) OR percutaneous coronary intervention) OR acute coronary syndromes) OR st-elevation myocardial infarction) OR non-ST-elevation myocardial infarction) OR unstable angina)) AND ("0001/01/01"[PDat] : "2014/04/09"[PDat])	(("0001/01/01"[Date - Publication] : "2014/04/09"[Date - Publication])) NOT ((((((((bivalirudin) OR Angiomax) OR Hirulog)) AND ((((((stent) OR percutaneous coronary intervention) OR acute coronary syndromes) OR st-elevation myocardial infarction) OR non-ST-elevation myocardial infarction) OR unstable angina)))))	n=745 total records n=1 manually added	n=561617 total records
		Citations finally included in the main database	**n=14 positives**	**n=400 negatives**

Chatterjee et al. 2014 [194]	((((((("Rivaroxaban"[Mesh]) OR dabigatran) OR "apixaban" [Supplementary Concept]) OR "new oral anticoagulants") OR "oral thrombin inhibitors") OR "oral factor Xa inhibitors") AND ("2001/01/01"[PDat] : "2013/09/15"[PDat]) AND English[lang]	(("0001/01/01"[Date - Publication] : "2001/01/01"[Date - Publication])) NOT (((((((("Rivaroxaban"[Mesh]) OR dabigatran) OR "apixaban" [Supplementary Concept]) OR "new oral anticoagulants") OR "oral thrombin inhibitors") OR "oral factor Xa inhibitors"))	n=2034 total records n=0 manually added	n=223263 total records
		Citations finally included in the main database	**n=18 positives**	**n=400 negatives**
Funakoshi et al 2014 [195]	((((((((((axitinib) OR cabozantinib) OR erlotinib) OR gefitinib) OR lapatinib) OR pazopanib) OR regorafenib) OR sorafenib) OR sunitinib) OR vandetanib)) AND "Randomized Controlled Trial" [Publication Type]) AND ("1966/01/01"[PDat] : "2013/03/31"[PDat]) AND English[lang]	(("0001/01/01"[Date - Publication] : "2013/03/31"[Date - Publication])) NOT ((((((((((((axitinib) OR cabozantinib) OR erlotinib) OR gefitinib) OR lapatinib) OR pazopanib) OR regorafenib) OR sorafenib) OR sunitinib) OR vandetanib)) AND "Randomized Controlled Trial" [Publication Type]))	n=418 total records n=5 manually added	n=527068 total records
		Citations finally included in the main database	**n=43 positives**	**n=1000 negatives**

Table 3 Number of training (PubMed) and testing (ClinicalTrial.gov) positive and negatives records on the side of the number of predicted positives and the relevant statistics for each Systematic Reviews considered (AUC = Area Under the receiver operator Curve; PREV = prevalence of positive in ClinicalTrial.gov; PPV = Positive Predictive Value; SENS = sensitivity; SPEC = specificity; LR+ = positive likelihood ratio LR- = negative likelihood ratio).

Systematic Review	Training positives	Training negatives	Testing positives	Testing negatives	Predicted positives	AUC	PPV	SENS	SPEC	LR+	LR-
Yang et al. 2014 [184]	18	400	5	233604	1718	0.9963	0.0029	1	0.9927	136.9863	0
Meng et al 2014 [185]	9	200	4	233605	462	0.9990	0.0087	1	0.9980	500.0000	0
Segelov et al. 2014 [72]	13	400	8	233601	1595	0.9341	0.0044	0.875	0.9932	128.6765	0.1259
Li et al. 2014 [186]	6	200	3	233606	1635	0.9965	0.0018	1	0.9930	142.8571	0
Lv et al. 2014 [187]	12	400	3	233606	1429	0.9969	0.0021	1	0.9939	163.9344	0
Wang et al. 2015 [188]	32	800	5	233604	1901	0.9959	0.0026	1	0.9919	123.4568	0
Zhou at al. 2014 [189]	9	200	3	233606	1011	0.9978	0.0030	1	0.9957	232.5581	0
Liu et al. 2014 [190]	23	600	30	233579	2178	0.9954	0.0138	1	0.9908	108.6957	0
Douxfils et al. 2014 [191]	13	400	10	233599	378	0.9992	0.0265	1	0.9984	625.0000	0
Kourbeti et al. 2014 [97]	75	1600	25	233584	1843	0.9961	0.0136	1	0.9922	128.2051	0
Li et al. 2014 [192]	9	200	2	233607	6558	0.9860	0.0003	1	0.9719	35.5872	0
Cavender et al. 2014 [193]	14	400	7	233602	149	0.9997	0.0470	1	0.9994	1666.666	0
Chatterjee et al. 2014 [194]	18	400	17	233592	771	0.9984	0.0220	1	0.9968	312.5000	0
Funakoshi et al 2014 [195]	43	1000	11	233598	3851	0.9918	0.0029	1	0.9836	60.9756	0

Table 4 The number of predicted positives and true positives in manual and automated searches after filter application. Records of the manual search are those retrieved on ICTRP by Baudard and colleagues [60]. Records of the automated search are those retrieved on ClinicalTrials.gov using our ML instrument. Predicted positives are a pool of citations resulting from manual search strings or from automated search. True positives are clinical trials added by Baudard and colleagues in each Systematic Review. Description of filters reports data element entries and number of retrieved records. Filters are applied sequentially from Filter 0 to Filter 5.

	Manual search		**Automated search**													
			Filter 0		**Filter 1**		**Filter 2**		**Filter 3**		**Filter 4**		**Filter 5**		**All**	
Systematic Review	*Predicted positives*	*True positives*	*none*	*Predicted positives*	*study_type*	*Predicted positives*	*overall_status*	*Predicted positives*	*start_before*	*Predicted positives*	*primary_completion before/within*	*Predicted positives*	*specific filters*	*Predicted positives*	*True positives*
Yang et al. 2014 [184]	12	1	-	1718	interventional	1341	completed OR terminated	759	April 2014	705	April 2014	588	allocation = randomized numer_of_arms ≠ 1	457	1
Meng et al 2014 [185]	26	1	-	462	interventional	399	completed OR terminated	282	August 2013	243	August 2013	202	allocation = randomized numer_of_arms ≠ 1	144	1
Segelov et al. 2014 [72]	684	2	-	1595	interventional	1432	completed OR terminated	836	May 2012	770	May 2012	588	allocation = randomized numer_of_arms ≠ 1	274	2
Li et al. 2014 [186]	201	1	-	1635	interventional	1545	completed OR terminated	376	February 2014	837	February 2014	695	allocation = randomized numer_of_arms ≠ 1 phase = 2 OR 3	289	1
Lv et al. 2014 [187]	665	1	-	1429	interventional	1294	completed OR terminated	727	February 2014	716	February 2014	583	allocation = randomized numer_of_arms ≠ 1 minimum_age ≥ 18 years	243	1
Wang et al. 2015 [188]	227	1	-	1901	interventional	1690	completed OR terminated	1191	December 2013	1118	December 2013	972	allocation = randomized numer_of_arms ≠ 1 intervention_model = parallel OR crossover	729	1
Zhou at al. 2014 [189]	3	1	-	1011	interventional	793	completed OR terminated	534	February 2013	468	February 2013	372	allocation = randomized numer_of_arms ≠ 1	263	1
Liu et al. 2014 [190]	1661	21	-	2178	interventional	2028	completed OR terminated	1587	January 2013	1317	January 2013	1112	allocation = randomized numer_of_arms ≠ 1 minimum_age ≥ 18 years phase ≥ 3	622	21
Douxfils et al. 2014 [191]	76	1	-	378	interventional	270	completed OR terminated	190	December 2013	174	December 2013	150	allocation = randomized numer_of_arms ≠ 1	116	1
Kourbeti et al. 2014 [97]	581	4	-	1843	interventional	1449	completed OR terminated	1023	June 2013	941	June 2013	756	allocation = randomized numer_of_arms ≠ 1 is_fda_regulated = true	409	4
Li et al. 2014 [192]	909	2	-	6558	interventional	5483	completed OR terminated	3629	January 2014	3412	January 2014	2771	allocation = randomized numer_of_arms ≠ 1	2119	2
Cavender et al. 2014 [193]	71	1	-	149	interventional	130	completed OR terminated	87	April 2014	84	April 2014	74	allocation = randomized numer_of_arms ≠ 1	60	1
Chatterjee et al. 2014 [194]	217	1	-	771	interventional	509	completed OR terminated	279	March 2014	263	January 2001 - March 2014	207	allocation = randomized numer_of_arms ≠ 1	169	1
Funakoshi et al 2014 [195]	2680	2	-	3851	interventional	3762	completed OR terminated	2147	February 2014	2111	January 2004 - February 2014	1699	allocation = randomized numer_of_arms ≠ 1 phase = 2 OR 3	711	2

B SCREENING PUBMED ABSTRACTS: IS CLASS IMBALANCE ALWAYS A CHALLENGE TO MACHINE LEARNING?

B.1 FIGURES

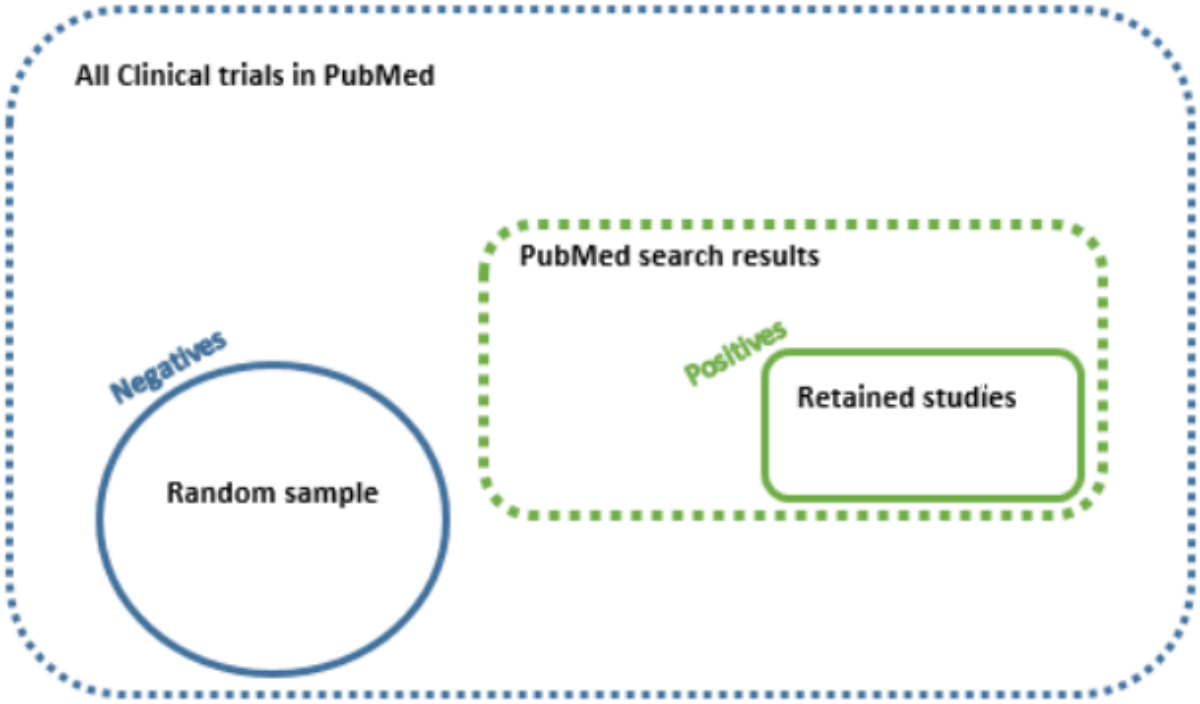

Figure 3 *Building process of the training dataset*. The positive citations are papers included in a systematic review. The negative citations are papers randomly selected from those completely off-topic. To identify positive citations, we recreate the input string in the PubMed database, using keywords and filters proposed in the original systematic review. Among retrieved records (dashed green line delimited region), we retain only papers finally included in the original systematic review (solid green line delimited region). On the other side, we randomly selected the negative citations (solid blue line delimited region) from those completely off-topic, by adding the Boolean operator NOT to the input string (region between green and blue dashed lines).

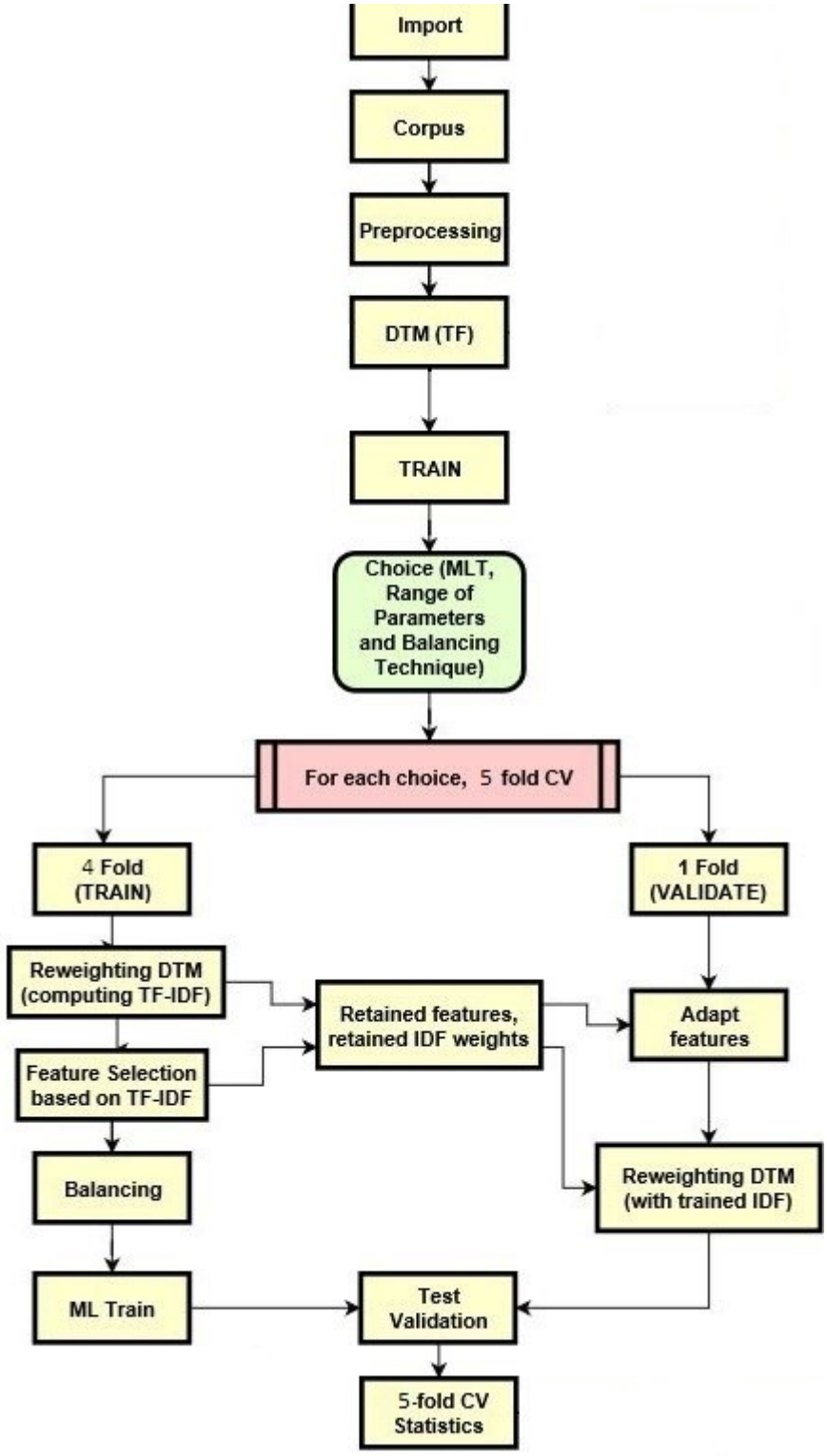

Figure 4 Computational plan. The set of documents for each systematic review considered was imported and converted into a corpus, preprocessed, and the corresponding Document-Term Matrix (DTM) was created for the training. Next, for each combination of machine learning technique (MLT), each one of the corresponding ten randomly selected tuning parameters, and balancing technique adopted, the training was divided in 5 -fold for the Cross-Validation (CV) process. In each step of the CV, the DTM was rescaled to the Term Frequencies-Inverse Document Frequencies (TF-IDF) weights (which are retained to rescale all the samples in the corresponding, i.e., the out-fold, test set). Next, the imbalance was treated with the selected algorithm, and the classifier was trained. Once the features in the test set were adapted to the training set, i.e., additional features were removed, missing ones were added with zero-weight, and all of them were reordered accordingly, the trained model was applied to the test set to provide the statistics of interest.

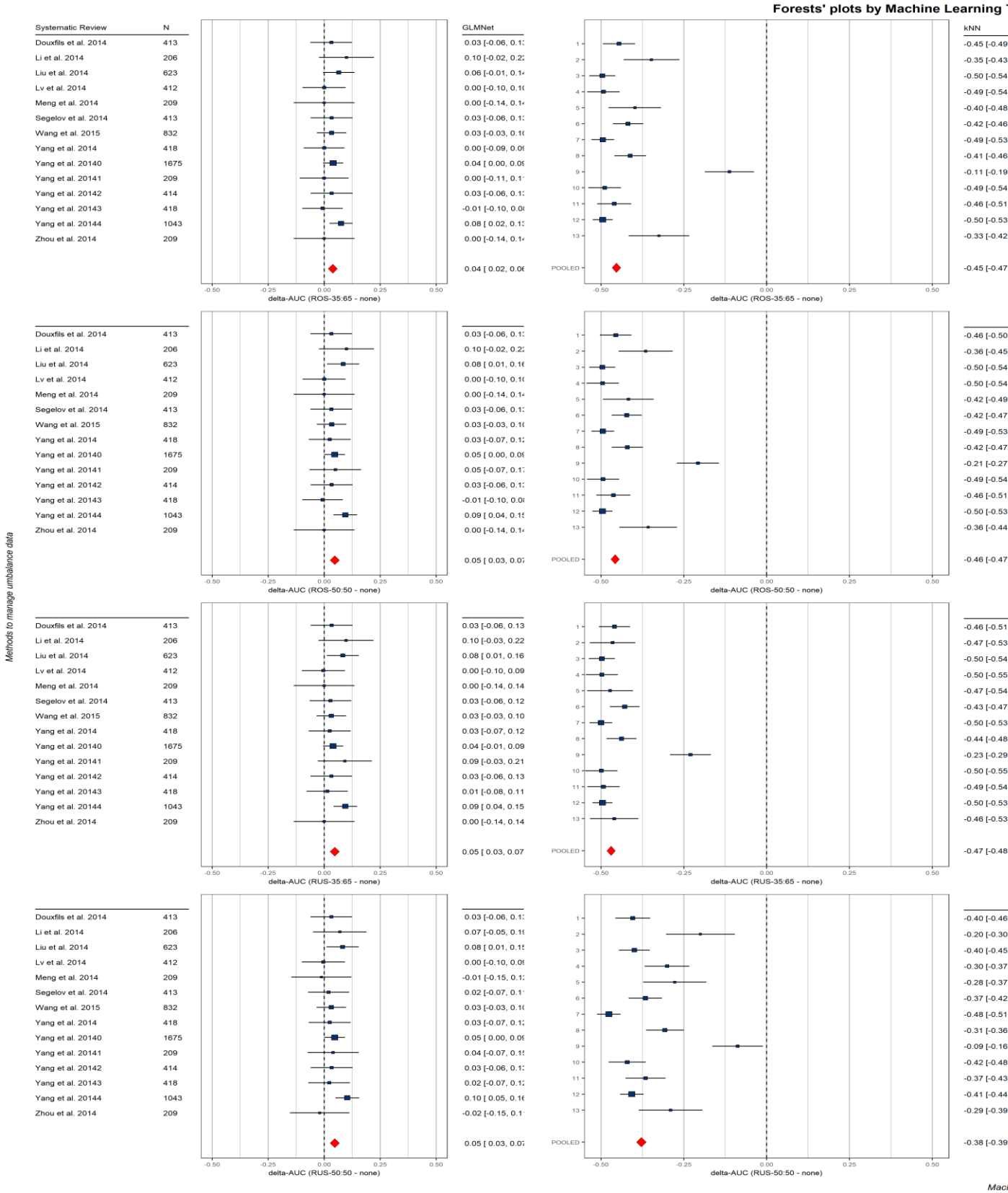

Figure 5 Forest plots of Delta-AUCs by balancing and machine learning techniques (MLTs). Forest plots that show differences in AUC (delta-AUCs) between the AUCs obtained with each balancing technique (i.e. RUS-50:50, RUS-35:65, ROS-50:50, and ROS-35:65) and the AUC obtained without the application of any of them for each combination of MLT and systematic reviews. Red diamonds report to pooled results obtained with a by-MLT meta-analytic fixed-effect model. The first author and year of systematic review corresponding to each row of the forest plots are reported in the first column only, the MLTs are reported in the first row only, and the balancing techniques are reported in each forest plot's x-axis label.

B.2 TABLES

Table 5 Characteristics of the Document-Term Matrices (DTMs). For each DTM are reported the number of documents included (number of rows), the number of tokens included/computed within those documents (number of columns), the number of cells of the matrix which are filled with a 0 (zero), or a positive weight; the ratio of these last two numbers (i.e., the sparsity) is also reported.

Systematic Reviews	Documents	Tokens	*Zero* entries	*Non-zero* entries	Sparsity
Yang et al. 2014 [184]	418	61208	147445	25437499	0.99
Meng et al 2014 [185]	209	35821	73977	7412612	0.99
Segelov et al. 2014 [72]	413	58351	125027	23963936	0.99
Li et al. 2014 [186]	206	33851	68826	6904480	0.99
Lv et al. 2014 [187]	412	57485	138846	23544974	0.99
Wang et al. 2015 [188]	832	101418	288432	84091344	1.00
Zhou at al. 2014 [189]	209	33389	69854	6908447	0.99
Liu et al. 2014 [190]	623	88108	219258	54672026	1.00
Douxfils et al. 2014 [191]	413	58133	141721	23869208	0.99
Kourbeti et al. 2014 [97]	1675	187947	603479	314207746	1.00
Li et al. 2014 [192]	209	33653	69130	6964347	0.99
Cavender et al. 2014 [193]	414	59572	141105	24521703	0.99
Chatterjee et al. 2014 [194]	418	54458	130782	22632662	0.99
Funakoshi et al 2014 [195]	1043	131172	370385	136442011	1.00

Table 6 AUC-ROC values by combination of MLTs, balancing techniques and balancing ratios across 14 systematic reviews. AUC-ROC: Area Under the Receiver Operator Characteristic Curve; ROS: Random Oversampling; RUS: Random Under-Sampling; RF: Random Forest; k-NN: k-Nearest Neighbours; SVM: Support-Vector Machines; GLMNet: elastic-net regularised generalised linear model. In boldface the best value(s) by row.

MLT	Systematic review	Method for imbalance				
		None	ROS-35:65	ROS-50:50	RUS-35:65	RUS-50:50
GLMNet	Cavender et al. 2014 [193]	0.9667	1	1	0.9988	1
	Chatterjee et al. 2014 [194]	0.9738	0.9667	0.9667	0.9875	0.9963
	Douxfils et al. 2014 [191]	413	58133	141721	23869208	0.99
	Funakoshi et al 2014 [195]	0.8851	0.9602	0.9799	0.9794	0.9885
	Kourbeti et al. 2014 [97]	0.9518	0.9921	0.9991	0.9918	0.9991
	Li et al. 2014 [186]	0.9	1	1	0.9975	0.97
	Li et al. 2014 [192]	0.8975	0.8975	0.9475	0.99	0.9375
	Liu et al. 2014 [190]	0.915	0.98	1	0.9983	0.9975
	Lv et al. 2014 [187]	1	1	1	0.9963	0.9963
	Meng et al 2014 [185]	1	1	1	1	0.9875
	Segelov et al. 2014 [72]	0.9667	1	0.9988	0.995	0.9863
	Wang et al. 2015 [188]	0.9667	1	1	0.9988	0.9988
	Yang et al. 2014 [184]	0.975	0.975	1	1	1
	Zhou at al. 2014 [189]	1	1	1	1	0.98
k-Nearest Neighbors	Cavender et al. 2014 [193]	1	0.5113	0.5063	0.5013	0.5792
	Chatterjee et al. 2014 [194]	0.9988	0.5388	0.5363	0.5063	0.6333
	Douxfils et al. 2014 [191]	0.9667	0.5213	0.5113	0.5075	0.5625
	Funakoshi et al 2014 [195]	0.9955	0.5005	0.5	0.5	0.5885
	Kourbeti et al. 2014 [97]	NA	NA	NA	0.5	0.5661
	Li et al. 2014 [186]	0.9775	0.63	0.6125	0.5125	0.7775
	Li et al. 2014 [192]	0.7975	0.685	0.59	0.5675	0.71
	Liu et al. 2014 [190]	0.9975	0.5017	0.5017	0.5	0.5983
	Lv et al. 2014 [187]	1	0.5075	0.505	0.5025	0.6996
	Meng et al 2014 [185]	0.9875	0.59	0.57	0.515	0.71
	Segelov et al. 2014 [72]	0.9283	0.51	0.5063	0.5	0.5625
	Wang et al. 2015 [188]	1	0.5056	0.5056	0.5	0.5237
	Yang et al. 2014 [184]	0.9404	0.5288	0.52	0.5025	0.6333
	Zhou at al. 2014 [189]	1	0.675	0.6425	0.54	0.71
	Cavender et al. 2014 [193]	1	1	1	1	1

Random Forest	Chatterjee et al. 2014 [194]	0.9167	0.975	0.975	0.9963	1
	Douxfils et al. 2014 [191]	1	1	1	1	1
	Funakoshi et al 2014 [195]	0.9184	0.9517	0.9299	0.9895	0.9895
	Kourbeti et al. 2014 [97]	0.9918	0.9854	0.9854	0.9988	0.9984
	Li et al. 2014 [186]	0.95	1	1	1	1
	Li et al. 2014 [192]	0.8	0.9	0.9	0.9	0.9475
	Liu et al. 2014 [190]	0.98	0.9992	0.9783	0.9992	0.9992
	Lv et al. 2014 [187]	1	1	1	0.9988	0.9988
	Meng et al 2014 [185]	0.95	0.95	0.95	1	1
	Segelov et al. 2014 [72]	0.9988	0.9988	0.9988	0.9975	0.9963
	Wang et al. 2015 [188]	0.9815	0.9821	0.9827	0.9994	0.9975
	Yang et al. 2014 [184]	0.95	0.975	0.95	0.9083	0.9046
	Zhou at al. 2014 [189]	1	1	1	1	0.995
Support Vector Machines	Cavender et al. 2014 [193]	1	1	1	1	0.825
	Chatterjee et al. 2014 [194]	1	1	0.9988	1	0.9263
	Douxfils et al. 2014 [191]	1	1	1	0.9963	0.8338
	Funakoshi et al 2014 [195]	0.999	0.999	0.9985	0.9945	0.975
	Kourbeti et al. 2014 [97]	0.9927	0.9927	0.9991	0.9988	0.9875
	Li et al. 2014 [186]	1	0.9975	0.9975	0.9325	0.5625
	Li et al. 2014 [192]	0.85	0.9	0.9925	0.98	0.6775
	Liu et al. 2014 [190]	1	1	1	0.9992	0.96
	Lv et al. 2014 [187]	1	1	1	0.9988	0.785
	Meng et al 2014 [185]	1	1	1	0.99	0.62
	Segelov et al. 2014 [72]	0.9333	0.9333	1	0.995	0.8013
	Wang et al. 2015 [188]	1	0.9857	1	0.9988	0.9681
	Yang et al. 2014 [184]	0.975	0.9417	0.9654	0.995	0.8825
	Zhou at al. 2014 [189]	1	1	1	1	0.7425

C USE OF MACHINE LEARNING TECHNIQUES FOR CASE-DETECTION OF VARICELLA ZOSTER USING ROUTINELY COLLECTED TEXTUAL AMBULATORY RECORDS

C.1 FIGURE

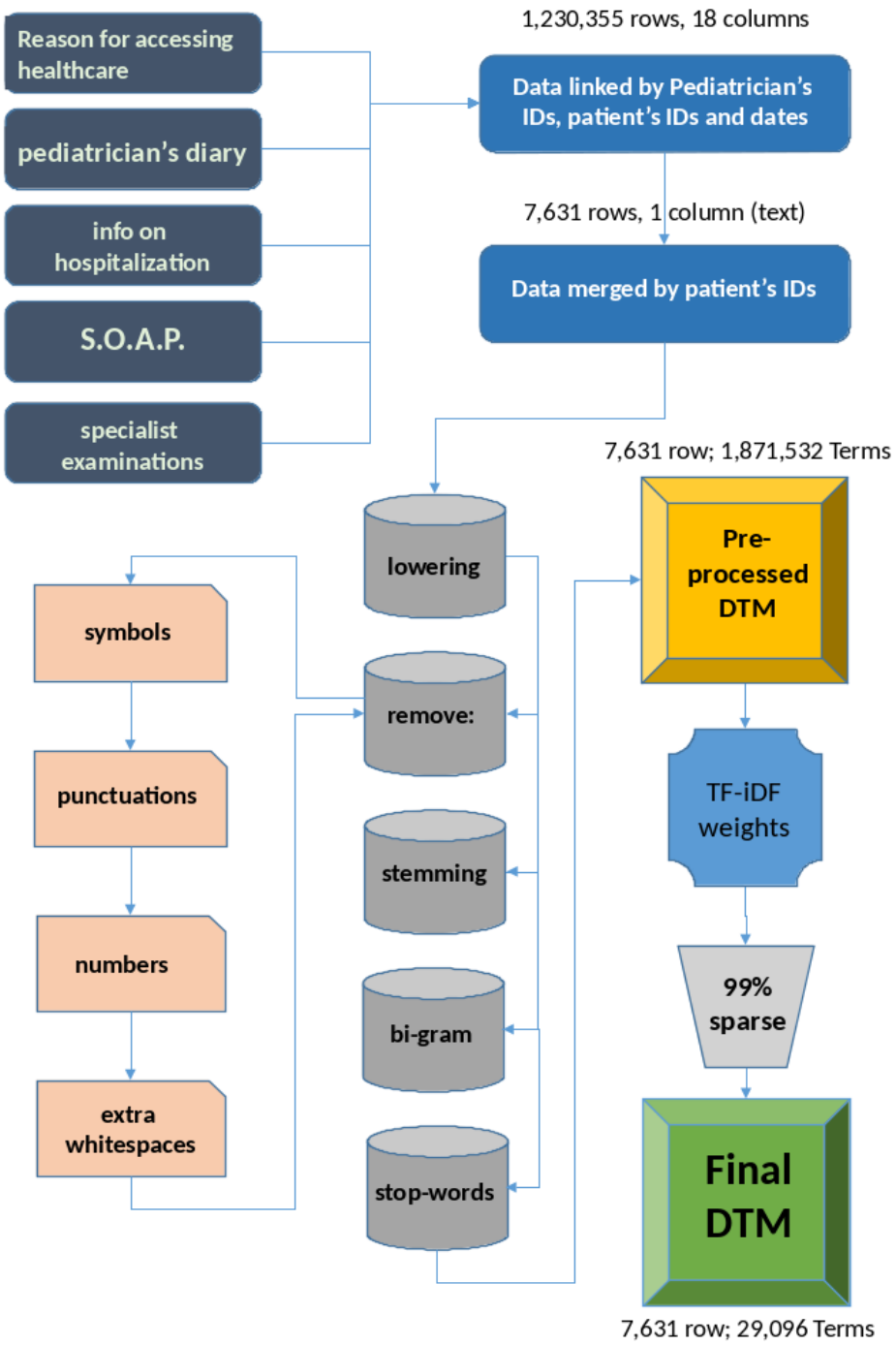

Figure 6 Flowchart from the acquisition of the five tables containing the Electronic Health Records (EHRs) (dark gray) in the training set that were merged into a single table (dark blue), preprocessed (gray) with the specification of what was removed (pink) prior to the creation of the Document-Term Matrix (DTM) (yellow), the computation of the weights (light blue), the dimensionality reduction, i.e., the reduction of the terms used (light gray), and the final DTM used (green).

C.2 TABLES

Table 7 ***Main characteristics for the train (Veneto) and test (Sicilia) datasets used.***

	Train	Test
Database	PEDIANET	PEDIANET
Language	Italian	Italian
Italian Region	Veneto	Sicilia
Date span	2004/01/02 - 2014/12/31	2004/01/07 - 2014/12/30
Records (No.)	1,230,355	569,926
Children (No.)	7,631	2,347
Pediatricians (No.)	46	13
Positive cases (No. [%])	3,481 [45.6%]	128 [5.4%]

Table 8 Tables used from the PediaNET database, including topic, content, type of data and examples.

Table topic	Content	Type of data	Example
Accessing	Reasons for accessing the pediatrician and diagnoses	Free-text codes	Ritardo di crescita <783.4>
Diaries	Pediatrician's free-text diaris	Free-text	DIBASE OS GTT 10ML 10000UI/ML n° conf. 2\r\n per Visita di controllo e di follow up\r\n\r\n
Hospitalizations	Details on hospital admissions, diagnoses and length of stays	Free-text	Divisione di pediatria Tosse, difficolta' respiratoria e di alimentazione
SOAP	Symptoms, Objectivity, Diagnosis or Prescriptions	Free-text codes	<SOAP> "P", <SOAP_code> "77469", <SOAP_text> "visita otorinolaringoiatrica<89.7>"
Specialistic visits	visit type and its diagnosis	Free-text codes	<codice_visitaSP> "89.01", <visita> "ecografia anche sec. Graaf per screening", <diagnosi> "problemi della vista <V41.0>"

Table 9 Performance on the training set of the three Machine Learning Techniques under consideration using a 5-fold cross-validation method (e.g., GLMNet, MAXENT, and Boosting). The values represent the mean across the folds of the point estimates, with the 95% Confidence Intervals between the parentheses, of Sensitivity, Positive Predictive Value (PPV), Negative Predictive Value (NPV), Specificity, Accuracy and F score (F).

Technique	Sensitivity	PPV	NPV	Specificity	Accuracy	F
GLMNet	80.2 (77.7-82.7)	73.2 (70.9-75.6)	90.9 (89.6-92.2)	87.1 (85.6-88.7)	85.0 (84.2-85.8)	76.5 (75.6-77.5)
MAXENT	68.8 (66.8-70.7)	66.0 (62.5-69.5)	86.1 (85.2-86.9)	84.5 (82.7-86.3)	79.7 (78.1-81.3)	67.4 (64.7-70.0)
Boosting	**86.6 (82.1-91.1)**	**95.8 (93.2-98.5)**	**94.4 (92.4-96.3)**	**98.3 (97.0-99.6)**	**94.8 (94.0-95.5)**	**90.9 (89.7-92.1)**

Table 10 Performance on the test set of the three Machine Learning Techniques under consideration. The values represent the mean across the folds of the point estimates, with the 95% Confidence Intervals between the parentheses, of Sensitivity, Positive Predictive Value (PPV), Negative Predictive Value (NPV), Specificity, Accuracy and F score (F).

Technique	Sensitivity	PPV	NPV	Specificity	Accuracy	F
GLMNet	72.3 (66.4–78.1)	24.5 (21.0–28.0)	98.3 (97.9–98.6)	87.4 (85.4–89.5)	86.6 (84.6–88.7)	36.5 (32.2–40.8)
MAXENT	74.8 (62.2–87.5)	11.0 (9.5–12.5)	98.0 (97.3–98.6)	65.5 (54.7–76.2)	66.0 (56.4–75.5)	19.1 (17.2–20.9)
Boosting	**79.2 (69.7–88.7)**	**63.1 (42.7–83.5)**	**98.8 (98.3–99.3)**	**96.9 (94.2–99.6)**	**96.0 (93.8–98.1)**	**68.5 (59.3–77.7)**

Table 11 Agreement between GLMNet, MAXENT, and Boosting using 5-fold cross-validation. The "Wrongly Agree" column refers to the number of records misclassified by both techniques. The "Correctly Agree" column states the number of records correctly classified by both techniques. The "Disagree" column lists the number of records for which the techniques disagree in the classification. Gwet's AC1 represents the index of agreement between the identified techniques along with the 95% Confidence Interval. Legend for AC1 is: -1 ≤ AC1 < 0 = disagreement; 0 ≤ AC1 ≤ 0.4 = poor; 0.4 < AC1 ≤ 0.6 = discrete; 0.6 < AC1 ≤ 0.8 = good; 0.8 < AC1 ≤ 1 = optimal.

Techniques	Wrongly Agree	Correctly Agree	Disagree	Gwet's AC1
GLMNet vs. MAXENT	669	5,609	1,353	0.68 (0.67-0.70)
GLMNet vs. Boosting	195	6,269	1,146	0.74 (0.72-0.75)
MAXENT vs. Boosting	224	5,895	1,491	0.66 (0.65-0.68)

D ANALYSIS OF UNSTRUCTURED TEXT-BASED DATA USING MACHINE LEARNING TECHNIQUES: THE CASE OF PEDIATRIC EMERGENCY DEPARTMENT RECORDS IN NICARAGUA

D.1 FIGURES

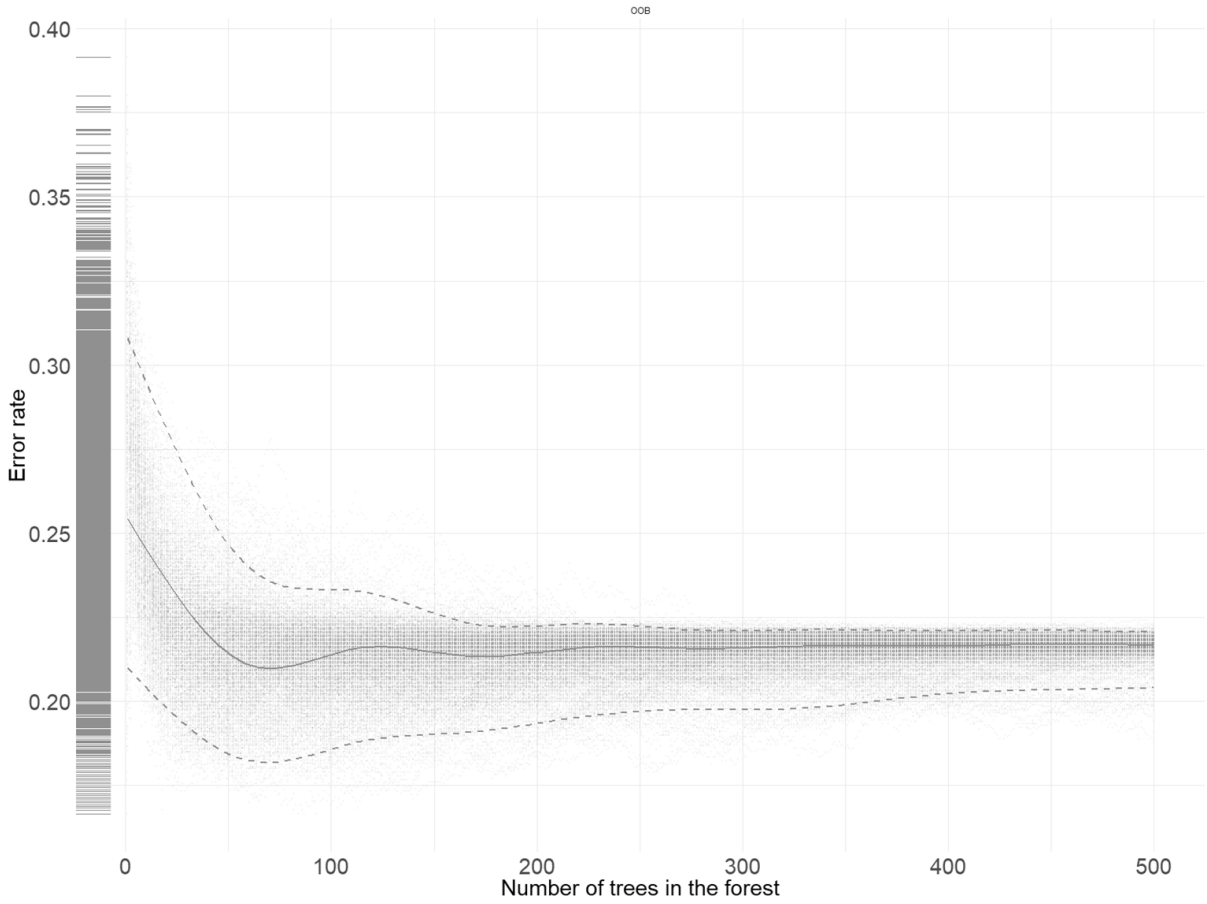

Figure 7 Out-of-bag error of the final validated models (calculated considering the Out-Of-Bag performance of 500 bootstrap repetitions of evaluation a 500-trees RF classifier by 5 repetition of 10-fold CV procedure). Dashed lines represent the performance corresponding to the 95% Confidence Interval borders for the bootstrapped classifiers, the solid line represents the median one, and each semi-transparent dot corresponds to the performance of a single RF into the pool created by the bootstrapped procedure

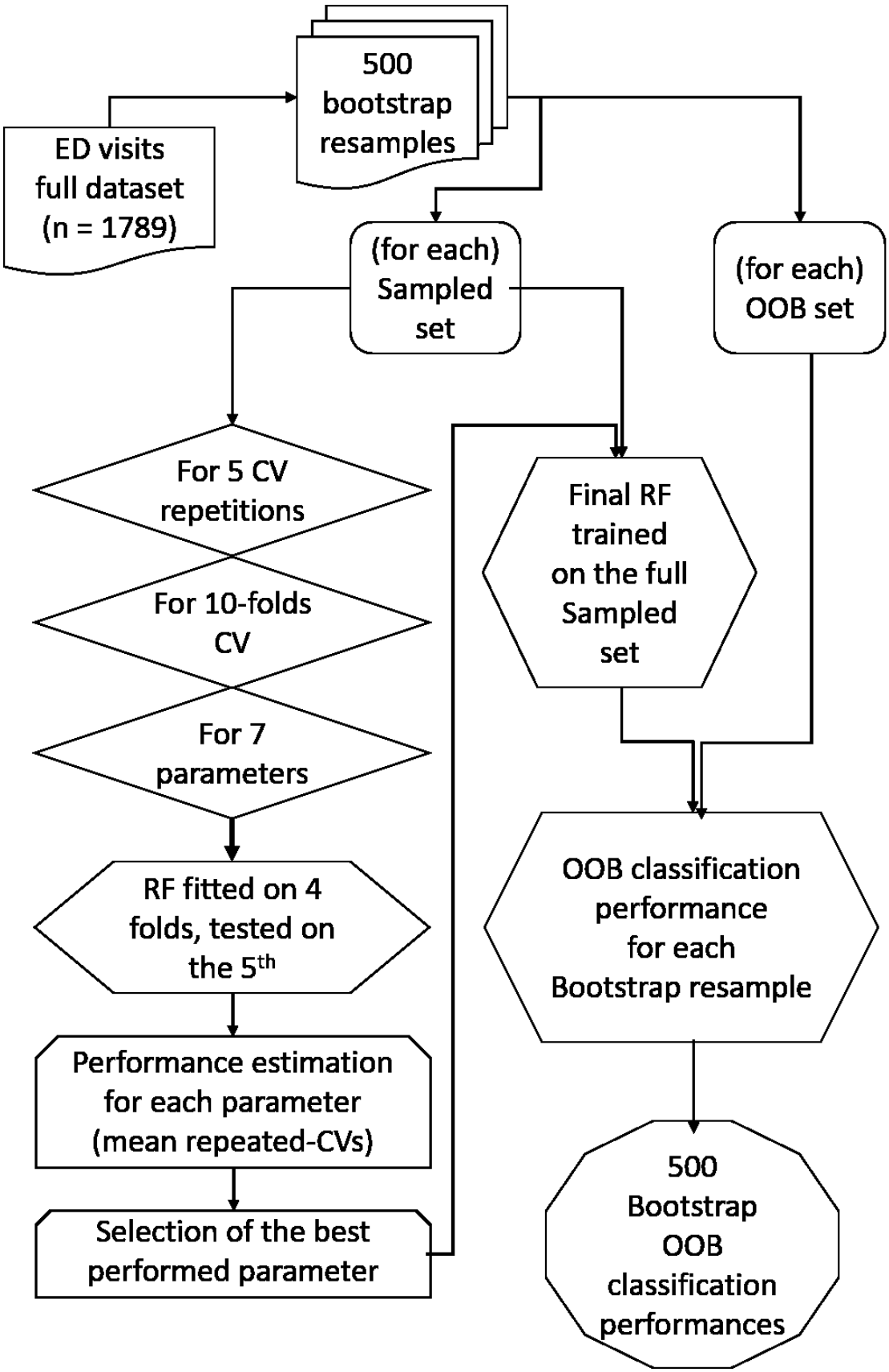

Figure 8 Training Procedure: (ED: Emergency Department; CV: Cross-Validation; RF: Random Forest; OOB: Out-Of-Bag) For each of the 500 bootstrap resampled dataset the performance estimation was calculated on its OOB set, which was never seen by the training procedure and different for every sample. For the final model trained on each bootstrap sample, the optimal parameter was selected by 5 repetition of 10-fold CV estimation.

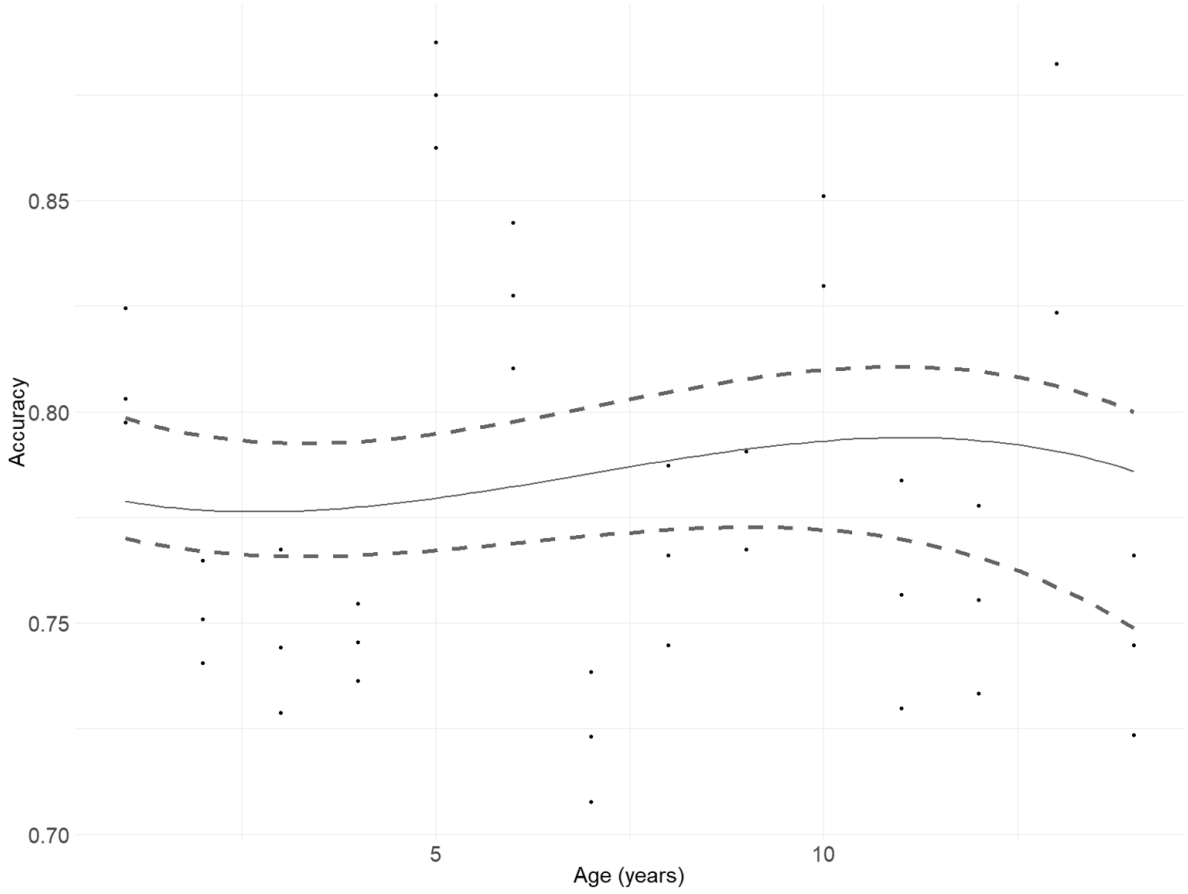

Figure 9 Accuracy according to children's age. Dashed lines represent 95% C.I. (calculated considering 500 bootstrap repetitions), solid line represents the median.

D.2 TABLES

Table 12 Variables included in the dataset with the corresponding type and examples

Variables	Data type	Examples
Age (in years)	Numerical	6; 38; 13
Gender	Categorical	Male; Female
Vital signs (i.e., body temperature, blood pressure, heart rate, and breathing rate, oxygen saturation)	Numerical	Body temperature -°C (37.6; 39; 36.7) Blood pressure (BP)-mmHg (Systolic BP: 91; 79; 107) (Diastolic BP: 65; 79; 50) Heart rate -bpm (145; 138; 149) Breathing rate -bpm (22; 42; 70) Oxygen saturation -% (97; 100; 78)
Laboratory tests (i.e., White blood cells count; Creatinite; Glucose; Natremia; Urea)	Numerical	
Diagnostic and therapeutic interventions (i.e., radiological examinations; respiratory support; medications administered; vascular access)	Free-text	Radiological examinations: "rx torax: infiltrado basal derecha"; "eco cardiograma: hap severa, falla cardiaca aguda, derrame pericardio moderado."; "rx de abdomen. radiopoacidad en fid."
Outcome of ED visit	Categorical	Ingresado; Fallecido
Discharge diagnosis	Free-text	"dengue hemorrágico" "crisis convulsiva febril"
Manual classification of the discharge diagnosis (gold standard)	Categorical	Gastrointestinal; Cardiovascular; Neurological

Table 13 Children's characteristics according to diagnosis category. Data are expressed as medians [I; III quartile] for continuous data and percentages (absolute number) for categorical ones

	N, *gold standard*	Age, *years*	Gender, *male*
Burn	1% (20)	4.0 [4.0; 7.5]	70% (14)
Cardiovascular	5% (98)	1.5 [1.0; 9.0]	57% (55)
Gastrointestinal	12% (208)	2.0 [1.0; 7.0]	52% (107)
Injury	4% (80)	8.0 [5.0; 11.0]	68% (54)
Metabolic	2% (29)	13.0 [10.0; 15.0]	31% (9)
Neurological	8% (141)	4.0 [2.0; 9.0]	59% (82)
Poisoning	1% (13)	4.0 [3.0; 7.0]	46% (6)
Respiratory	56% (1003)	1.0 [1.0; 3.0]	56% (560)
Tropical disease	6% (104)	9.0 [6.0; 11.0]	51% (53)
Other	5% (93)	4.0 [2.0; 9.0]	57% (52)
Overall	100% (1789)	2.0 [1.0; 6.0]	56% (992)

Table 14 Median accuracy (rate of diagnosis correctly classified by the final validated model) of the ML algorithms together with 95% Confidence Interval (C.I.) (calculated considering the Out-Of-Bag performance of 500 bootstrap repetitions of evaluation a 500-trees RF classifier by 5 repetition of 10-fold CV procedure). The C.I. was not estimated for Burn, Metabolic and discharge diagnosis' classes because of the small size of the sample of children in these classes.

	Accuracy	(95% C.I.)
Burn	0.900	-
Cardiovascular	0.683	(0.663; 0.704)
Gastrointestinal	0.759	(0.745; 0.769)
Injury	0.837	(0.825; 0.850)
Metabolic	0.758	-
Neurological	0.602	(0.588; 0.624)
Poisoning	0.692	-
Respiratory	0.801	(0.797; 0.826)
Tropical disease	0.971	(0.971; 0.980)
Other	0.752	(0.731; 0.763)
Overall	0.783	(0.779; 0.796)

D.3 SUPPLEMENTARY MATERIALS

Table S 1 Criteria for inclusion in the study registry of urgent-emergent paediatrics visits to Paediatric Emergency Departments in Nicaragua

Neurologic	Persistent altered mental status (GCS < 15)
	Signs of raised intracranial pressure
	Signs of severe neuroinfection
	Active seizures on arrival
	Acute focal neurological signs
Respiratory	Signs of airway obstruction
	Severe respiratory distress (based on PALS 2015)
	Bradipnoea/apnoea
Cardiovascular	Cardiac arrest
	Signs of shock (based on PALS 2015)
	Tachycardia/bradycardia
	Signs of cardiac failure
	Suspected sepsis
	Hypoxic spells
Gastrointestinal	Acute gastroenteritis (vomiting and/or diarrhea) with severe dehydration (based on clinician judgment)
	Gastrointestinal bleeding
	Acute abdomen
Metabolic	Diabetic ketoacidosis
Injury	Potentially severe isolated (single site) or multiple trauma
	Burns > 20% BSA
	Venomous snake bite
	Poisoning (high risk- based on respiratory, neurologic, cardiovascular, and/or gastrointestinal signs or symptoms
Suspected Dengue with warning signs	

BSA= Body Surface Area; GCS= Glasgow Coma Scale;

Table S 2 Common tokens (including bigrams) among the 500 bootstrapped final validated model appearing in the top 100 of each model (according to the TF-iDF weight)

Tokens (including bigrams)
cetoacidosis
dengue
diabetes
grado
intoxicacion
“intoxicacio por”
quemadura
sepsis
shock
sustancia
trauma

E AUTOMATIC IDENTIFICATION AND CLASSIFICATION OF DIFFERENT TYPES OF OTITIS FROM FREE-TEXT PEDIATRIC MEDICAL NOTES IN THE ITALIAN LANGUAGE: A DEEP-LEARNING APPROACH

E.1 FIGURES

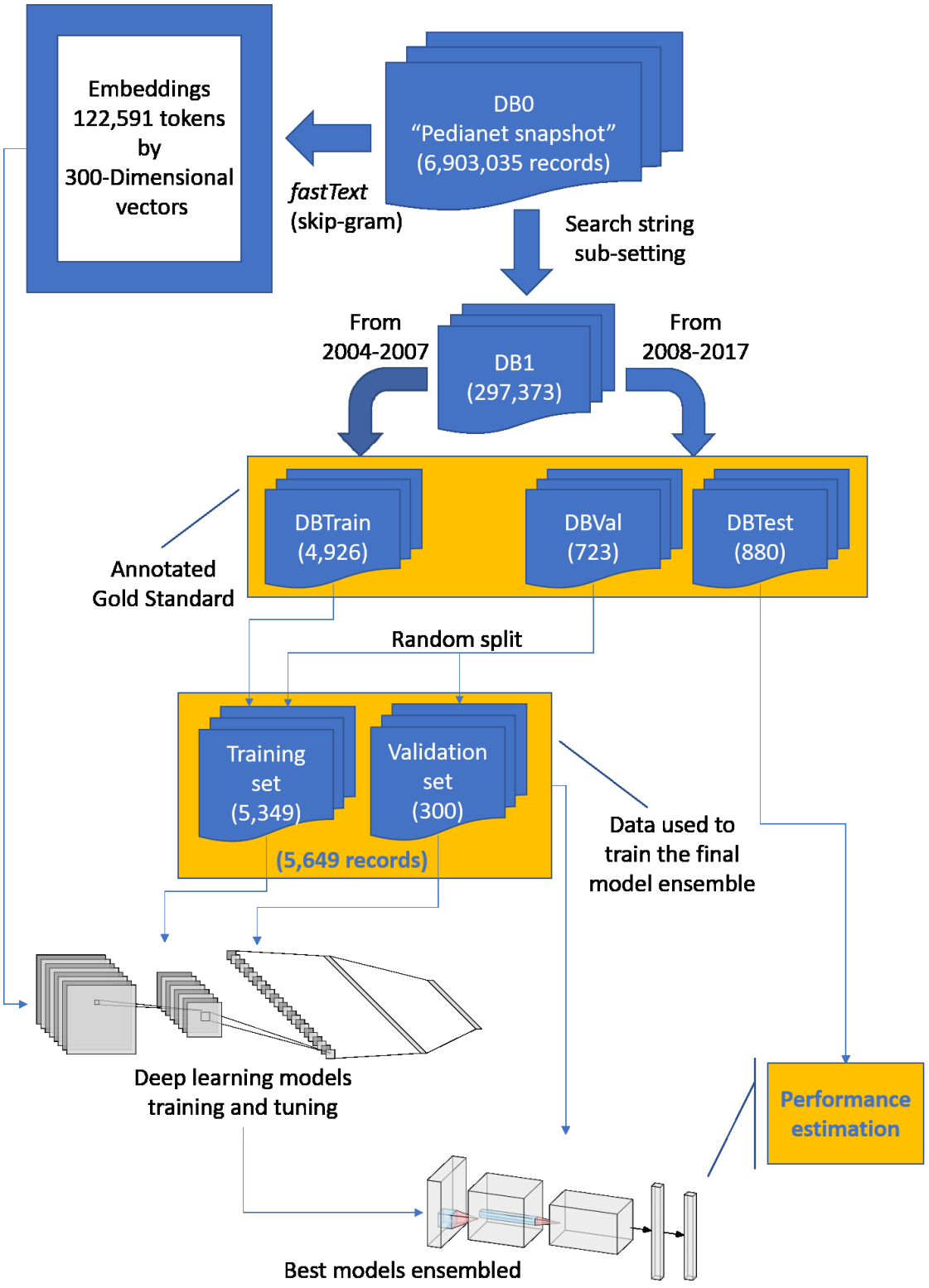

Figure 10 Flowchart for the project.

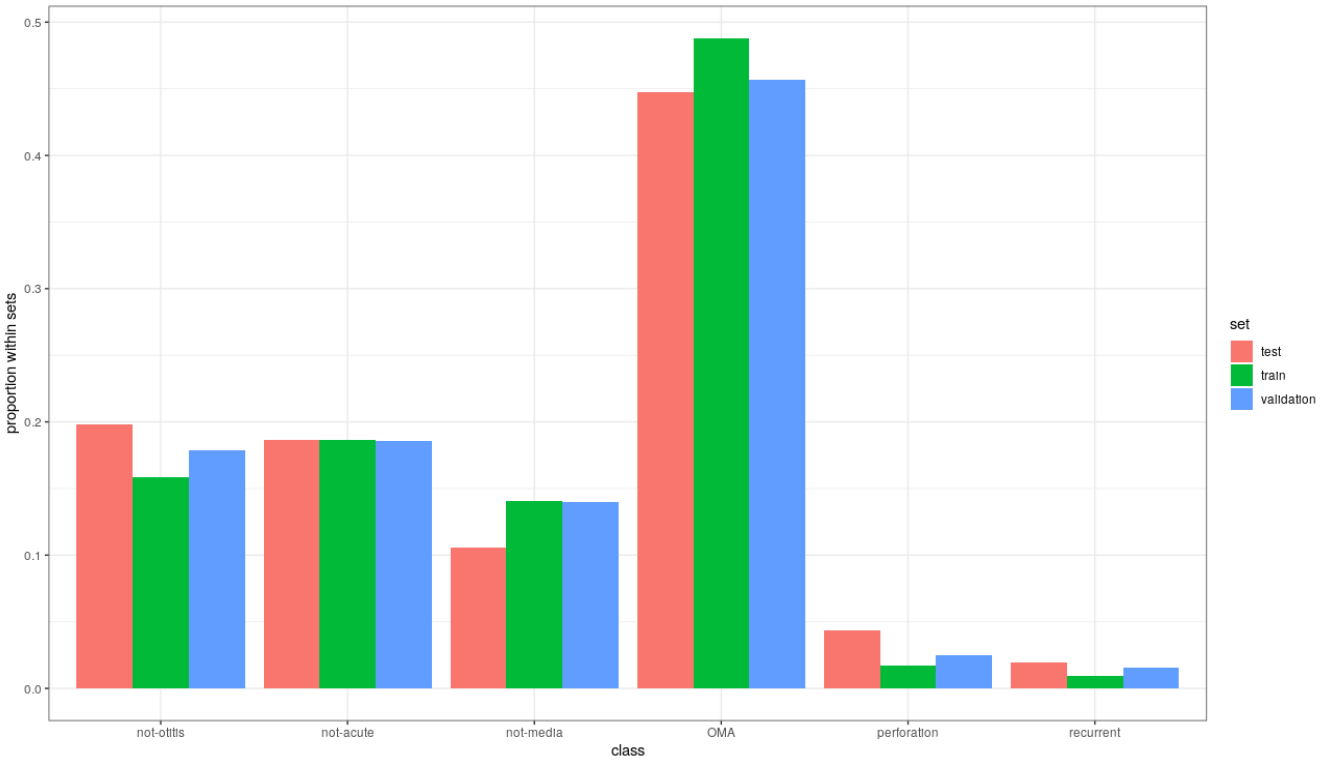

Figure 11 The proportion of classes for the train, validation, and test set.

.

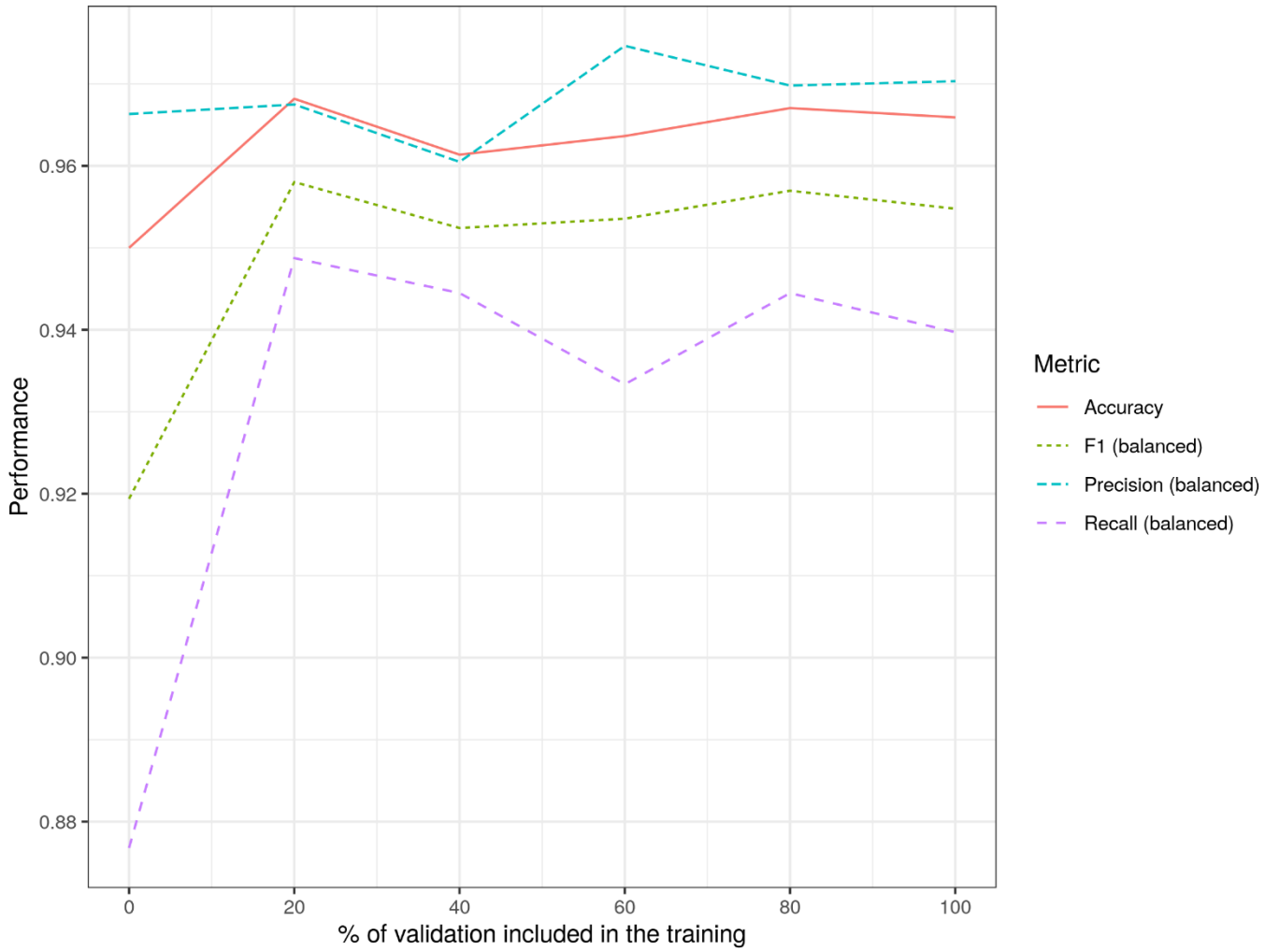

Figure 12 Accuracy and balanced precision, recall, and F1 performances for the ensemble model when the base models ensembled are trained using only a subset of the validation set added to the training one. On the x-axis is reported the proportion of DBvali added to DBtrain for the training.

E.2 TABLES

Table 15 The number of visits, children, pediatricians (MDs), male, female, and years for the databases considered for the study. I.e., DB0 is the Pedianet snapshot considered, DB1 contains data from DB0 positives to the search string. DBtrain, DBvalidation, and DBtest are the gold standard sets of data containing visits from DB1.

Dataset	Visits	Children	MDs	Male [%]	Female [%]	Years (range)
DB0	6,903,035	216,976	143	112,413 [51.8%]	104,563 [48.2%]	2004-1017
DB1	297,373	99,896	142	53,159 [52.2%]	47,737 [47.8%]	2004-2017
DBtrain	4,926	4,475	138	2,349 [52.5%]	2,078 [46.4%]	2004-2007
DBval	723	718	142	377 [52.5%]	341 [47.5%]	2008-2017
DBtest	880	873	142	463 [53.0%]	410 [47.0%]	2008-2017

Table 16 Agreement between evaluators A and B (weighted Cohen's Kappa), and measurements to assess the human-level performances to the task, i.e., balance precision (BalPrec), balanced recall (BalRec), balanced F1 score (BalF1), and overall accuracy (Acc) for both A and B using as reference the final gold standard approved by the third specialist.

Database	Weighted Cohen's Kappa	BalPrec (A)	BalRec (A)	BalF1 (A)	Acc (A)	BalPrec (B)	BalRec (B)	BalF1 (B)	Acc (B)
DBtest	0.89	91.70	95.30	93.47	95.91	96.33	84.66	90.12	95.80

Table 17 Performances for the simple embedding architecture.

Simple embedding	Batch [size]	Accuracy (phase 1)	Accuracy (fine-tuned)	Running time [min]
b8	8	88.00 [75]	94.00 [296]	147.79
b16	16	88.00 [64]	93.00 [280]	88.47

Table 18 Performances for the single kernel CNN architectures.

Single kernel	Kernel [size]	Batch [size]	Filters [#]	Drop-out [%]	Accuracy (frozen) [epochs]	Accuracy (fine-tuned) [epochs]	Running time [min]
b8-k2-f128-do05	2	8	128	50	98.00 [203]	96.67 [48]	154.80
b8-k2-f128-do07	2	8	128	70	97.00 [114]	96.67 [3]	108.57
b8-k2-f256-do05	2	8	256	50	97.33 [73]	97.00 [1]	119.82
b8-k2-f256-do07	2	8	256	70	97.00 [48]	96.67 [1]	110.68
b16-k2-f128-do05	2	16	128	50	96.33 [23]	96.67 [121]	112.60
b16-k2-f128-do07	2	16	128	70	96.33 [92]	95.67 [1]	87.15
b16-k2-f256-do05	2	16	256	50	97.00 [39]	97.33 [24]	104.40
b16-k2-f256-do07	2	16	256	70	97.33 [69]	96.67 [1]	103.79
b8-k3-f128-do05	3	8	128	50	97.00 [37]	96.67 [1]	109.31
b8-k3-f128-do07	3	8	128	70	96.33 [140]	95.67 [1]	149.97
b8-k3-f256-do05	3	8	256	50	97.33 [53]	96.33 [1]	141.60
b8-k3-f256-do07	3	8	256	70	97.33 [111]	97.33 [1]	174.00
b16-k3-f128-do05	3	16	128	50	97.67 [128]	97.67 [1]	125.11
b16-k3-f128-do07	3	16	128	70	96.33 [83]	95.33 [1]	108.10
b16-k3-f256-do05	3	16	256	50	97.00 [30]	96.67 [69]	151.93
b16-k3-f256-do07	3	16	256	70	96.67 [47]	95.67 [20]	131.40

Table 19 Performances for the sequential kernels CNN architectures.

Sequential kernels	Filters [#-#]	Batch [size]	Drop-out [%]	Accuracy (frozen) [epochs]	Accuracy (fine-tuned) [epochs]	Running time [min]
b8-2x128-3x256-do05	128-256	8	50	97.67 [57]	98.00 [1]	119.67
b8-2x128-3x256-do07	128-256	8	70	97.33 [68]	96.67 [1]	131.40
b16-2x128-3x256-do05	128-256	16	50	97.33 [107]	97.00 [1]	119.09
b16-2x128-3x256-do07	128-256	16	70	97.00 [80]	97.00 [1]	110.03
b8-2x256-3x512-do05	256-512	8	50	97.67 [50]	97.67 [1]	172.20
b8-2x256-3x512-do07	256-512	8	70	97.67 [65]	98.33 [2]	183.00
b16-2x256-3x512-do05	256-512	16	50	98.00 [19]	97.67 [1]	140.40
b16-2x256-3x512-do07	256-512	16	70	97.33 [95]	97.33 [1]	184.80

Table 20 Performances for the parallel kernels CNN architectures.

Parallel kernels	Filters [#]	Batch [size]	Drop-out [%]	Accuracy (frozen) [epochs]	Accuracy (fine-tuned) [epochs]	Running time [min]
b8-(emb+2+3)x128-do05	128	8	50	97.33 [71]	97.33 [1]	178.80
b8-(emb+2+3)x128-do07	128	8	70	97.33 [47]	97.33 [1]	163.20
b16-(emb+2+3)x128-do05	128	16	50	97.33 [46]	97.00 [1]	174.00
b16-(emb+2+3)x128-do07	128	16	70	97.33 [107]	97.33 [1]	212.40
b8-(emb+2+3)x256-do05	256	8	50	97.00 [7]	97.33 [18]	190.80
b8-(emb+2+3)x256-do07	256	8	70	98.00 [82]	97.33 [2]	232.80
b16-(emb+2+3)x256-do05	256	16	50	97.33 [48]	97.33 [1]	243.60
b16-(emb+2+3)x256-do07	256	16	70	97.67 [65]	97.67 [1]	257.40

Table 21 Performances for the deep-parallel kernels CNN architectures.

Deep-parallel kernels	Filters [#-#]	Batch [size]	Dropout [%]	Accuracy (frozen) [epochs]	Accuracy (fine-tuned) [epochs]	Running time [min]
b8-(emb+2+3)x128-(2+3)x256-do05	128-256	8	50	98 [30]	98 [11]	308.56
b8-(emb+2+3)x128-(2+3)x256-do07	128-256	8	70	97.33 [92]	97.67 [1]	325.84
b16-(emb+2+3)x128-(2+3)x256-do05	128-256	16	50	97.67 [101]	98.00 [17]	379.12
b16-(emb+2+3)x128-(2+3)x256-do07	128-256	16	70	97.00 [70]	96.67 [2]	311.08
b8-(emb+2+3)x256-(2+3)x512-do05	256-512	8	50	98.00 [133]	97.67 [8]	786.6
b8-(emb+2+3)x256-(2+3)x512-do07	256-512	8	70	97.67 [98]	97.33 [1]	580.6
b16-(emb+2+3)x256-(2+3)x512-do05	256-512	16	50	96.67 [12]	97.00 [2]	423.04
b16-(emb+2+3)x256-(2+3)x512-do07	256-512	16	70	97.33 [45]	96.67 [1]	529.8

Table 22 Performances on the test set evaluated on the best model of each architecture, re-trained on the whole training data available (DBTrain + DBval), and their ensemble model. Each model was re-trained using the best hyper-parameters set from the architectures explored, for the same number of epochs selected in the validation stage.

Selected network	Balanced precision	Balanced recall	Accuracy	Balanced F1
Simple embedding: b8	84.51	68.63	81.70	75.75
single kernel: b8-k2-f128-do05	92.60	91.87	94.66	92.23
sequential CNN: b16-2x256-3x512-do05	95.94	81.26	93.64	87.99
parallel CNN: b8-(emb+2+3)x256-do07	96.95	94.78	96.59	95.86
deep CNN: b8-(emb+2+3)x128-(2+3)x256-do05	96.38	93.36	96.25	94.85
Ensemble (w/o simple embeddings)	97.03	93.97	96.59	95.47

Table 23 Confusion matrix for the classes predicted on the DBtest set by the ensemble model (by row) and reported on the gold standard (by columns)

Predicted\Gold	0	1	2	3	4	5	Sum
0	155	0	2	0	0	0	157
1	7	168	7	0	0	0	182
2	1	0	101	1	1	0	104
3	2	1	1	389	6	0	399
4	0	0	0	1	28	0	29
5	0	0	0	0	0	9	9
Sum	165	169	111	391	35	9	880

E.3 SUPPLEMENTARY MATERIALS

6.6.1 REGULAR EXPRESSION

Table 24 Regular expressions used to filter possible cases of otitis from Pedianet databases. The final regular expression applied was the disjunction of the reported ones (i.e., linked with the "OR" boolean operator)

Regular expression
"o{1,2}tit|ot{1,2}it|oti{1,2}t|otit{1,2}|toite|oitte|otiet|^om|^o\\.m|\\som|\\so\\.m\\.|\\Wom|\\Wo\\.m\\.|\\dom|\\do\\.m\\.|\\nom|\\no\\.m\\."
"\\so\\stit|ot\\sit|oti\\st|\\so\\m\\s|\\so\\.\\sm\\.\\s"
"\\sitite|\\s9tite|\\s0tite|\\sptite|\\sltite|\\sktite"
"\\sorite|\\so5ite|\\so6ite|\\soyite|\\sogite|\\sofite"
"\\sotute|\\sot8te|\\sot9te|\\sotote|\\sotkte|\\sotjte"
"\\sotire|\\soti5e|\\soti6e|\\sotiye|\\sotige|\\sotife]"

6.6.2 NETWORKS

were the following benchmark one (i.e., a simple embedding: "0") plus four others of increasing complexity:

0. Simple embedding: the only hidden layers were the embedding layer, followed by the batch normalization end the drop-out ones (Figure 13).
1. Single kernel Convolutional Neural Network (CNN): after the embedding layer we attached a convolutional layer considering the *same* padding with a stride of 1, i.e., a sufficient quantity of zeros were added to each side of the input tensor to produce an output tensor with the same dimensions as the one received in input after applying the convolution kernel window sliding by one neuron at time. We explored the kernel sizes of 2, and 3, both with 128 or 256 filters. After the batch normalization and the drop-out layer, we flatted the tensor by the application of a global-max pooling layer before to pass it to the fully connected output layer (Figure 14).
2. Sequential single kernel CNN: we considered like the first hidden layer after the embedding group (i.e., embedding + batch normalization + drop-out) a CNN group of layer made up by a CNN layer with kernel size equal to 2 and 128 filters, followed by batch normalization, drop-out and a max-pooling layer with both the window kernel and the stride equal to 2. After that, we attached a similar CNN group sequentially, but with a kernel size equal to 3, 256 filters, and a global max-pooling layer at the end, before the output layer (Figure 15).
3. Multiple parallel kernel CNN: we consider for the first hidden layer after the embedding group, a layer composed by the concatenation of the simple embedding itself, with the two output of the application of the embedding to two CCNs layers with *same* padding, kernel = 2, and 3, and 128 or 256 filters each one. To its output, we applied a batch normalization, a drop-out, and a global max-pooling layers before the application of the fully connected output layer (Figure 16).
4. Deep multiple parallel kernel CNN: we stack sequentially two layers like the one designed in the case of multiple parallel kernel CNN, considering 128 or 256 filters for the first one which considers a kernel size equal to 2 for its max-pooling layer, while 256 or 512 filters for the CNNs in the second group which use a global max-pooling layer at the end. The embedding layer was concatenated only to the first of those hidden layers, and not in the second (Figure 17).

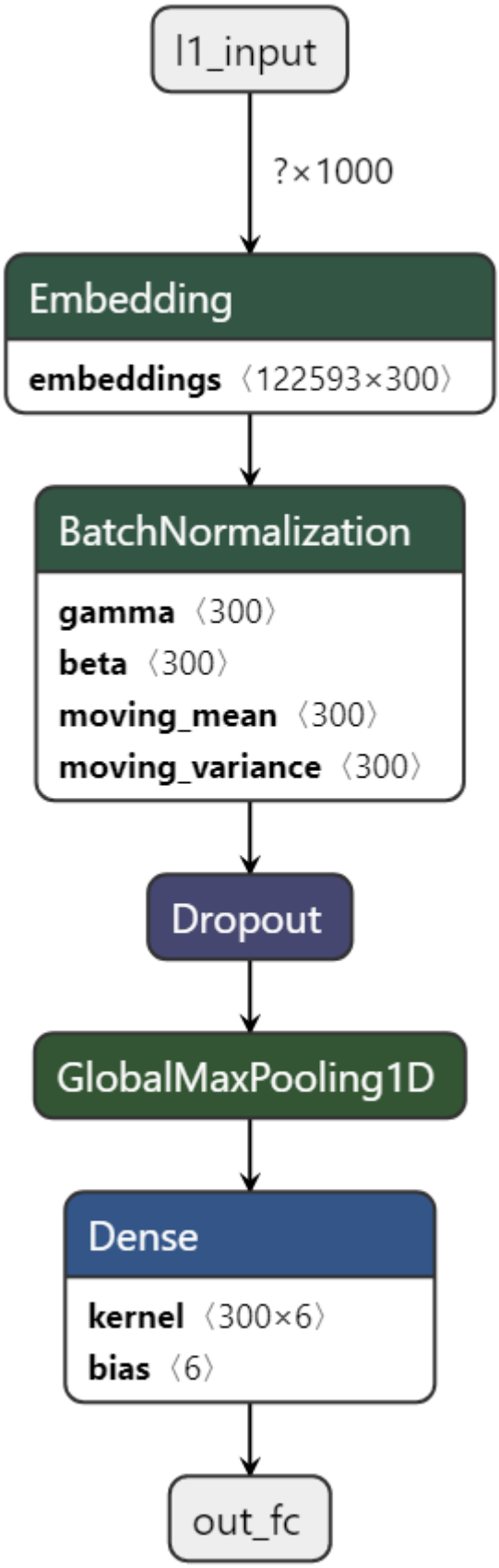

Figure 13 Diagram for the simple-embedding architecture. The dropout rate was 20%.

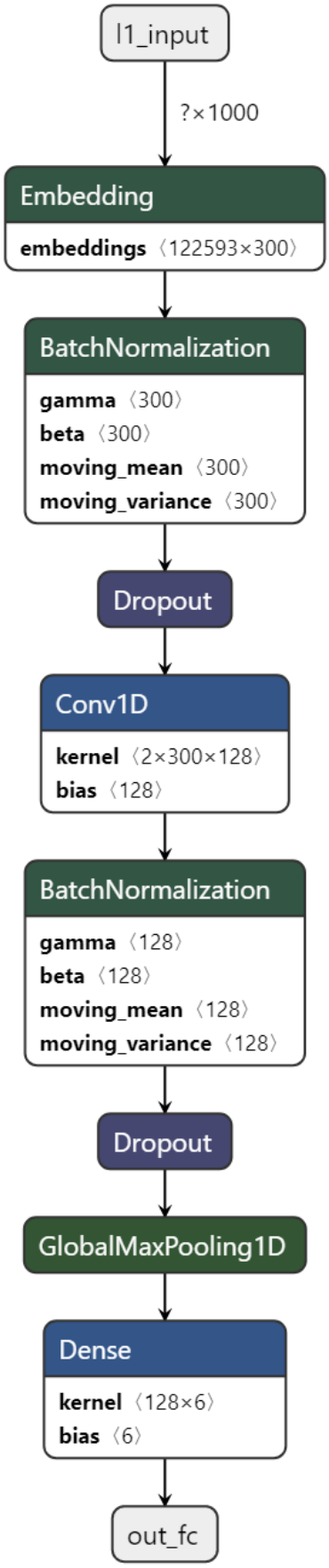

Figure 14 Diagram for the single kernel architecture. The dropout rate was 20% after the embeddings, 50% after the convolution

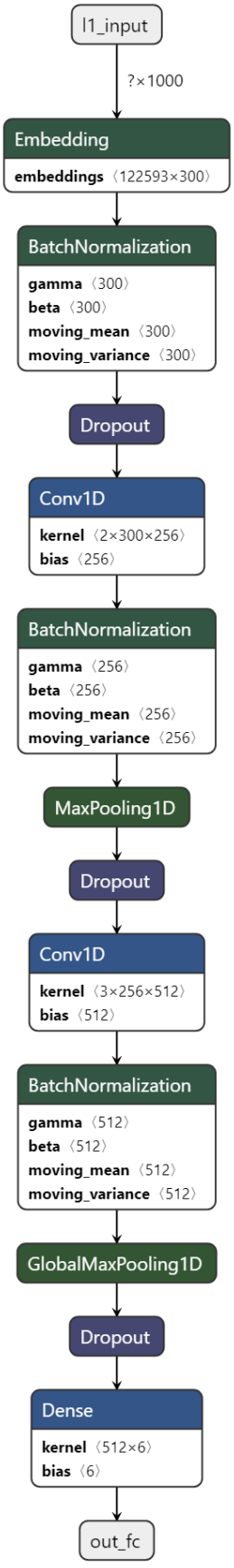

Figure 15 Diagram for the sequential kernel architecture. The dropout rate was 20% after the embeddings, 50% after the convolution. MaxPooling after the first convolution layer has a window of size and stride both equals to 2, with valid padding.

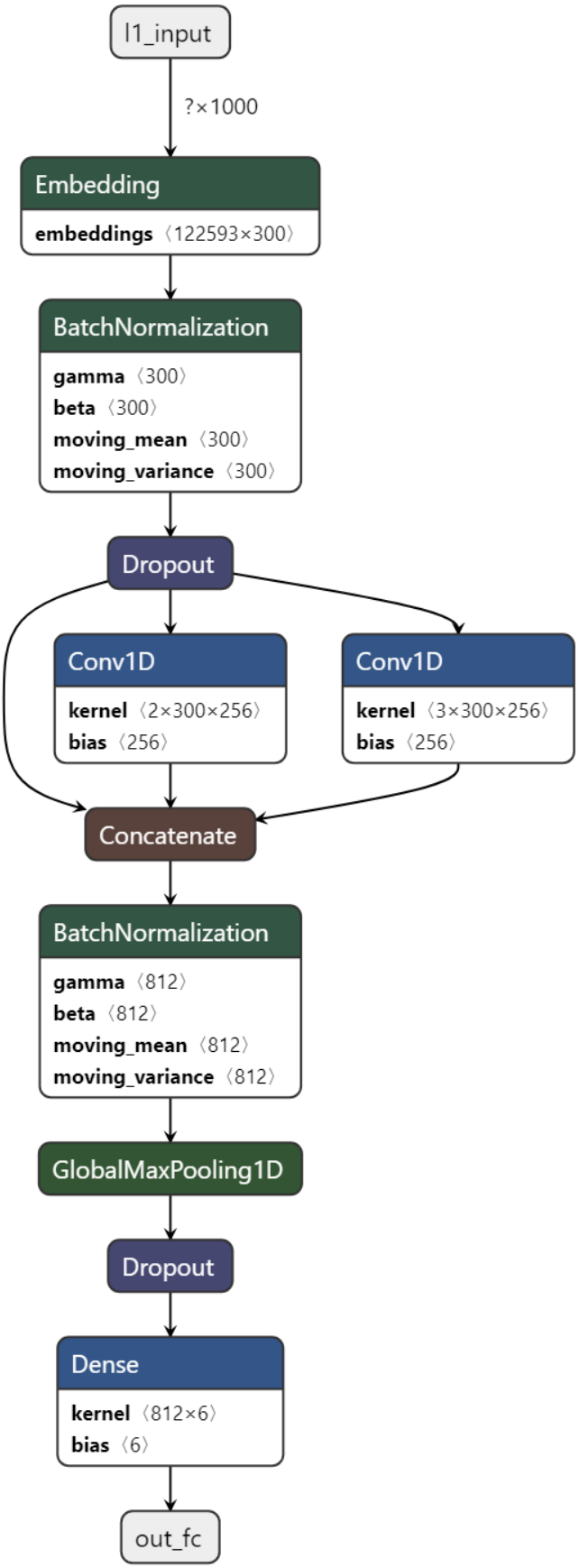

Figure 16 Diagram for the parallel kernel architecture. The dropout rate was 20% after the embeddings, 50% after the convolutions.

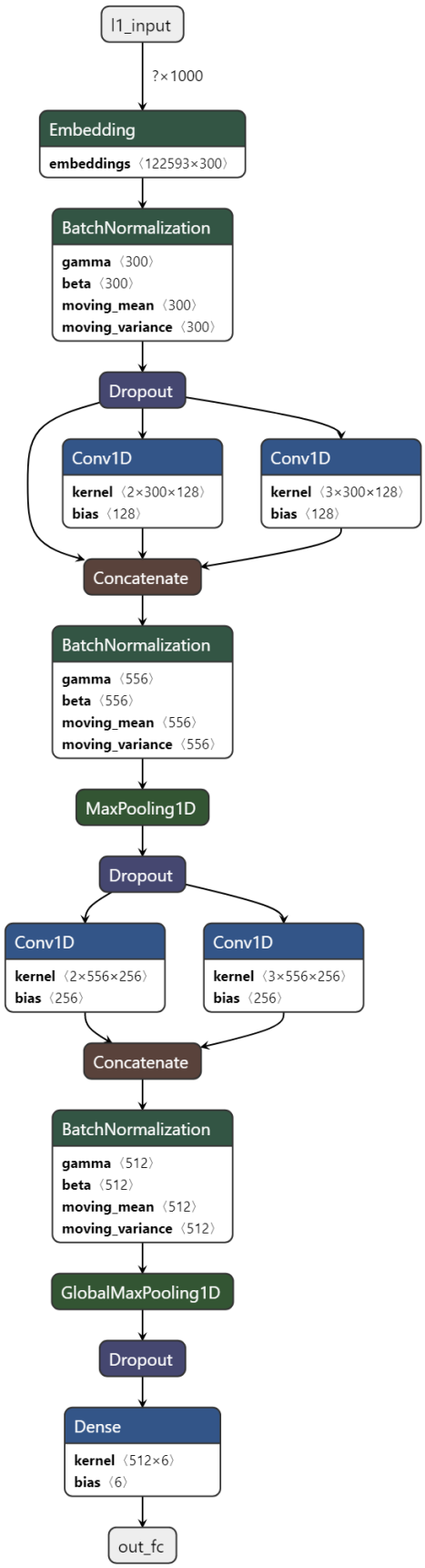

Figure 17 Diagram for the deep-sequential kernel architecture. The dropout rate was 20% after the embeddings, 50% after the convolution. MaxPooling after the first convolution layer has a window of size and stride both equals to 2, with valid padding.

6.6.3 SYSTEM

We run all the computations on an Ubuntu 18.04.3 LTS GNU/Linux 4.15.0-58-generic x86_64 virtualized server of the Unit of Biostatistics Epidemiology and Public Health of the University of Padova, equipped with 16 cores from Intel® Xeon® CPUs E5-2640 v4 @ 2.40GHz, and 96 GiB-RAM. We implement all the networks and codes for the analyses in *R* (v3.6.1) powered by the *Keras* (v2.2.4.1.9001) *R* interface to the *TensorFlow* (v1.14) backend, built from source enabling the usage of Intel® AVX set of instruction extensions. To learn word representation was used *fastText* (v0.9.1). Diagram for the networks were automatically drowned from the *Keras* models trained using *netron* (v3.3.5). All the development and code was tracked on a GitHub repository publicly available at www.github.com/UBESP-DCTV/limpido.

REFERENCES

[1] Becker RA, Chambers JM. *S : an interactive environment for data analysis and graphics*. Belmont, Calif. : Wadsworth Advanced Book Program, http://archive.org/details/sinteractiveenvi00beck_0 (1984, accessed 30 September 2019).

[2] Ross MK, Wei W, Ohno-Machado L. 'Big data' and the electronic health record. *Yearb Med Inform* 2014; 9: 97–104.

[3] Wiens J, Shenoy ES. Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology. *Clin Infect Dis* 2018; 66: 149–153.

[4] Sebastiani F. Machine learning in automated text categorization. *Acm Comput Surv* 2002; 34: 1–47.

[5] Denny MJ, Spirling A. Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It. *Polit Anal* 2018; 26: 168–189.

[6] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space. *arXiv:13013781 [cs]*, http://arxiv.org/abs/1301.3781 (2013, accessed 3 September 2019).

[7] Cochrane Handbook for Systematic Reviews of Interventions. *Wiley Online Library*. DOI: 10.1002/9780470712184.

[8] CRD's guidance for undertaking reviews in health care, https://www.york.ac.uk/media/crd/Systematic_Reviews.pdf (2009).

[9] Balan PF, Gerits A, Vanduffel W. A practical application of text mining to literature on cognitive rehabilitation and enhancement through neurostimulation. *Front Syst Neurosci* 2014; 8: 182.

[10] Khabsa M, Elmagarmid A, Ilyas I, et al. Learning to identify relevant studies for systematic reviews using random forest and external information. *Mach Learn* 2016; 102: 465–482.

[11] Biese KJ, Forbach CR, Medlin RP, et al. Computer-facilitated Review of Electronic Medical Records Reliably Identifies Emergency Department Interventions in Older Adults. *Academic Emergency Medicine* 2013; 20: 621–628.

[12] Ford E, Carroll JA, Smith HE, et al. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *J Am Med Inform Assoc* 2016; 23: 1007–15.

[13] Gerbier S, Yarovaya O, Gicquel Q, et al. Evaluation of natural language processing from emergency department computerized medical records for intra-hospital syndromic surveillance. *BMC medical informatics and decision making* 2011; 11: 50.

[14] Metzger M-H, Tvardik N, Gicquel Q, et al. Use of emergency department electronic medical records for automated epidemiological surveillance of suicide attempts: a French pilot study. *International journal of methods in psychiatric research*; 26.

[15] Magill SS, Dumyati G, Ray SM, et al. Evaluating Epidemiology and Improving Surveillance of Infections Associated with Health Care, United States. *Emerg Infect Dis* 2015; 21: 1537–42.

[16] Lloyd-Smith JO, Funk S, McLean AR, et al. Nine challenges in modelling the emergence of novel pathogens. *Epidemics* 2015; 10: 35–9.

[17] Sutherland SM, Kaelber DC, Downing NL, et al. Electronic Health Record-Enabled Research in Children Using the Electronic Health Record for Clinical Discovery. *Pediatr Clin North Am* 2016; 63: 251–68.

[18] Corrao G, Cantarutti A. Building reliable evidence from real-world data: Needs, methods, cautiousness and recommendations. *Pulmonary Pharmacology & Therapeutics* 2018; 53: 61–67.

[19] Barbieri E, Donà D, Cantarutti A, et al. Antibiotic prescriptions in acute otitis media and pharyngitis in Italian pediatric outpatients. *Ital J Pediatr* 2019; 45: 103.

[20] Trifirò G, Gini R, Barone-Adesi F, et al. The Role of European Healthcare Databases for Post-Marketing Drug Effectiveness, Safety and Value Evaluation: Where Does Italy Stand? *Drug Saf* 2019; 42: 347–363.

[21] Pedianet, http://pedianet.it/en (accessed 30 August 2019).

[22] Giaquinto C, Sturkenboom M, Mannino S, et al. [Epidemiology and outcomes of varicella in Italy: results of a prospective study of children (0-14 years old) followed up by pediatricians (Pedianet study)]. *Ann Ig* 2002; 14: 21–7.

[23] Nicolosi A, Sturkenboom M, Mannino S, et al. The incidence of varicella: correction of a common error. *Epidemiology* 2003; 14: 99–102.

[24] Cantarutti A, Dona D, Visentin F, et al. Epidemiology of Frequently Occurring Skin Diseases in Italian Children from 2006 to 2012: A Retrospective, Population-Based Study. *Pediatr Dermatol* 2015; 32: 668–78.

[25] Dona D, Mozzo E, Scamarcia A, et al. Community-Acquired Rotavirus Gastroenteritis Compared with Adenovirus and Norovirus Gastroenteritis in Italian Children: A Pedianet Study. *Int J Pediatr* 2016; 2016: 5236243.

[26] Hirshon JM, Warner M, Irvin CB, et al. Research using emergency department–related data sets: current status and future directions. *Academic emergency medicine* 2009; 16: 1103–1109.

[27] Sequeira M, Espinoza H, Amador JJ, et al. The Nicaraguan health system. *Seattle, WA: PATH*; 201.

[28] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, pp. 785–794.

[29] Sheridan RP, Wang WM, Liaw A, et al. Extreme Gradient Boosting as a Method for Quantitative Structure–Activity Relationships. *J Chem Inf Model* 2016; 56: 2353–2360.

[30] Kasai J, Qian K, Gurajada S, et al. Low-resource Deep Entity Resolution with Transfer and Active Learning. *arXiv:190608042 [cs]*, http://arxiv.org/abs/1906.08042 (2019, accessed 8 September 2019).

[31] Lakshminarayanan B, Roy DM, Teh YW. Mondrian Forests: Efficient Online Random Forests. In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. Cambridge, MA, USA: MIT Press, pp. 3140–3148.

[32] Domingos P, Hulten G. Mining High-speed Data Streams. In: *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, pp. 71–80.

[33] Shaoqing Ren, Cao X, Yichen Wei, et al. Global refinement of random forest. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 723–730.

[34] Krawczyk B, Minku LL, Gama J, et al. Ensemble learning for data stream analysis: A survey. *Information Fusion* 2017; 37: 132–156.

[35] Czado C, Raftery AE. Choosing the link function and accounting for link uncertainty in generalized linear models using Bayes factors. *Statistical Papers* 2006; 47: 419–442.

[36] Friedman JH, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* 2010; 33: 1–22.

[37] Pattern Recognition - 4th Edition, https://www.elsevier.com/books/pattern-recognition/theodoridis/978-1-59749-272-0 (accessed 30 August 2019).

[38] Qian K, Burdick D, Gurajada S, et al. Learning Explainable Entity Resolution Algorithms for Small Business Data Using SystemER. In: *Proceedings of the 5th Workshop on Data Science for Macro-modeling with Financial and Economic Datasets*. New York, NY, USA: ACM, pp. 8:1–8:6.

[39] Dahl GE, Yu D, Deng L, et al. Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition. *IEEE Transactions on Audio, Speech, and Language Processing* 2012; 20: 30–42.

[40] Zhong G, Ling X, Wang L-N. From shallow feature learning to deep learning: Benefits from the width and depth of deep architectures. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. Epub ahead of print 1 January 2019. DOI: 10.1002/widm.1255.

[41] Bengio Y. Learning Deep Architectures for AI. *Found Trends Mach Learn* 2009; 2: 1–127.

[42] Ranzato MA, Boureau Y-L, LeCun Y. Sparse Feature Learning for Deep Belief Networks. In: *Proceedings of the 20th International Conference on Neural Information Processing Systems*. USA: Curran Associates Inc., pp. 1185–1192.

[43] Mhaskar H, Liao Q, Poggio T. When and Why Are Deep Networks Better Than Shallow Ones? In: *Thirty-First AAAI Conference on Artificial Intelligence*, https://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14849 (2017, accessed 30 August 2019).

[44] Wu Y, Schuster M, Chen Z, et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv:160908144 [cs]*, http://arxiv.org/abs/1609.08144 (2016, accessed 30 August 2019).

[45] Chorowski J, Jaitly N. Towards better decoding and language model integration in sequence to sequence models. *arXiv:161202695 [cs, stat]*, http://arxiv.org/abs/1612.02695 (2016, accessed 30 August 2019).

[46] Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need. *arXiv:170603762 [cs]*, http://arxiv.org/abs/1706.03762 (2017, accessed 30 August 2019).

[47] Erhan D, Bengio Y, Courville A, et al. Why Does Unsupervised Pre-training Help Deep Learning? *Journal of Machine Learning Research* 2010; 11: 625–660.

[48] Huang Z, Pan Z, Lei B. Transfer Learning with Deep Convolutional Neural Network for SAR Target Classification with Limited Labeled Data. *Remote Sensing* 2017; 9: 907.

[49] Caruana R. Multitask Learning. *Machine Learning* 1997; 28: 41–75.

[50] Press TM. Advances in Neural Information Processing Systems. *The MIT Press*, https://mitpress.mit.edu/books/advances-neural-information-processing-systems (accessed 30 August 2019).

[51] Ehret K, Szmrecsanyi B. An information-theoretic approach to assess linguistic complexity. *Complexity and isolation Berlin: de Gruyter*.

[52] Bentz C, Ruzsics T, Koplenig A, et al. A comparison between morphological complexity measures: typological data vs. language corpora. In: *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*. 2016, pp. 142–153.

[53] Howick J, Chalmers I, Glasziou P, et al. Explanation of the 2011 Oxford Centre for Evidence-Based Medicine (OCEBM) Levels of Evidence (Background Document), https://www.cebm.net/2016/05/ocebm-levels-of-evidence/ (2016, accessed 13 January 2018).

[54] Hirschman L, Burns GAPC, Krallinger M, et al. Text mining for the biocuration workflow. *Database (Oxford)* 2012; 2012: bas020.

[55] Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration. *PLOS Medicine* 2009; 6: e1000100.

[56] Guyatt GH, Oxman AD, Montori V, et al. GRADE guidelines: 5. Rating the quality of evidence--publication bias. *J Clin Epidemiol* 2011; 64: 1277–1282.

[57] Hopewell S, Loudon K, Clarke MJ, et al. Publication bias in clinical trials due to statistical significance or direction of trial results. *Cochrane Database Syst Rev* 2009; MR000006.

[58] WHO. Major research funders and international NGOs to implement WHO standards on reporting clinical trial results, http://www.who.int/mediacentre/news/releases/2017/clinical-trial-results/en/ (2017, accessed 13 January 2018).

[59] Hughes S, Cohen D, Jaggi R. Differences in reporting serious adverse events in industry sponsored clinical trial registries and journal articles on antidepressant and antipsychotic drugs: a cross-sectional study. *BMJ Open* 2014; 4: e005535.

[60] Baudard M, Yavchitz A, Ravaud P, et al. Impact of searching clinical trial registries in systematic reviews of pharmaceutical treatments: methodological systematic review and reanalysis of meta-analyses. *BMJ* 2017; 356: j448.

[61] Halfpenny NJ, Thompson JC, Quigley JM, et al. Clinical Trials Registries For Systematic Reviews – An Alternative Source For Unpublished Data. *Value in Health* 2015; 18: A12.

[62] Zarin DA, Tse T, Williams RJ, et al. The ClinicalTrials.gov Results Database — Update and Key Issues. *N Engl J Med* 2011; 364: 852–860.

[63] Tang E, Ravaud P, Riveros C, et al. Comparison of serious adverse events posted at ClinicalTrials.gov and published in corresponding journal articles. *BMC Med* 2015; 13: 189.

[64] Jindal R, Malhotra R, Jain A. Techniques for text classification: Literature review and current trends. *Webology* 2015; 12: 1.

[65] Liu AC. The effect of oversampling and undersampling on classifying imbalanced text datasets. *The University of Texas at Austin*, https://pdfs.semanticscholar.org/cade/435c88610820f073a0fb61b73dff8f006760.pdf (2004, accessed 11 October 2017).

[66] Khoshgoftaar TM, Seiffert C, Van Hulse J, et al. Learning with limited minority class data. In: *Machine Learning and Applications, 2007. ICMLA 2007. Sixth International Conference on*. IEEE, pp. 348–353.

[67] R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, https://www.R-project.org/ (2017).

[68] Wing MKC from J, Weston S, Williams A, et al. *caret: Classification and Regression Training*, https://CRAN.R-project.org/package=caret (2017).

[69] Feinerer I, Hornik K. *tm: Text Mining Package*, https://CRAN.R-project.org/package=tm (2017).

[70] Wickham H. *stringr: Simple, Consistent Wrappers for Common String Operations*, https://CRAN.R-project.org/package=stringr (2017).

[71] Pozzolo AD, Caelen O, Bontempi G. *unbalanced: Racing for Unbalanced Methods Selection*, https://CRAN.R-project.org/package=unbalanced (2015).

[72] Segelov E, Chan D, Shapiro J, et al. The role of biological therapy in metastatic colorectal cancer after first-line treatment: a meta-analysis of randomised trials. *Br J Cancer* 2014; 111: 1122–1131.

[73] Thomas J, Noel-Storr A, Marshall I, et al. Living systematic reviews: 2. Combining human and machine effort. *J Clin Epidemiol* 2017; 91: 31–37.

[74] Rochefort CM, Verma AD, Eguale T, et al. A novel method of adverse event detection can accurately identify venous thromboembolisms (VTEs) from narrative electronic health record data. *J Am Med Inform Assoc* 2015; 22: 155–165.

[75] Connolly B, Matykiewicz P, Bretonnel Cohen K, et al. Assessing the similarity of surface linguistic features related to epilepsy across pediatric hospitals. *J Am Med Inform Assoc* 2014; 21: 866–870.

[76] Rios A, Kavuluru R. Convolutional Neural Networks for Biomedical Text Classification: Application in Indexing Biomedical Articles. *ACM BCB* 2015; 2015: 258–267.

[77] Majumder S, Balaji N, Brey K, et al. 500+ Times Faster Than Deep Learning (A Case Study Exploring Faster Methods for Text Mining StackOverflow). *arXiv:180205319 [cs, stat]*, http://arxiv.org/abs/1802.05319 (2018, accessed 12 May 2018).

[78] Marshall IJ, Noel-Storr A, Kuiper J, et al. Machine learning for identifying Randomized Controlled Trials: An evaluation and practitioner's guide. *Research Synthesis Methods*; 0. Epub ahead of print January 2018. DOI: 10.1002/jrsm.1287.

[79] Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol* 2016; 69: 245–247.

[80] Altman DG, Royston P. What do we mean by validating a prognostic model? *Statistics in Medicine* 2000; 19: 453–473.

[81] Wallace BC, Noel-Storr A, Marshall IJ, et al. Identifying reports of randomized controlled trials (RCTs) via a hybrid machine learning and crowdsourcing approach. *Journal of the American Medical Informatics Association* 2017; 24: 1165–1168.

[82] Miwa M, Thomas J, O'Mara-Eves A, et al. Reducing systematic review workload through certainty-based screening. *Journal of Biomedical Informatics* 2014; 51: 242–253.

[83] O'Mara-Eves A, Thomas J, McNaught J, et al. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic Reviews* 2015; 4: 5.

[84] Kritz M, Gschwandtner M, Stefanov V, et al. Utilization and perceived problems of online medical resources and search tools among different groups of European

physicians. *Journal of medical Internet research*; 15, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3713956/ (2013, accessed 22 September 2017).

[85] Wallace BC, Trikalinos TA, Lau J, et al. Semi-automated screening of biomedical citations for systematic reviews. *BMC Bioinformatics* 2010; 11: 55.

[86] Longadge R, Dongre S. Class imbalance problem in data mining review. *arXiv preprint arXiv:13051707*, https://arxiv.org/abs/1305.1707 (2013).

[87] Laza R, Pavón R, Reboiro-Jato M, et al. Evaluating the effect of unbalanced data in biomedical document classification. *Journal of Integrative Bioinformatics (JIB)* 2011; 8: 105–117.

[88] Chawla NV, Bowyer KW, Hall LO, et al. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 2002; 16: 321–357.

[89] Wang S, Yao X. Diversity analysis on imbalanced data sets by using ensemble models. IEEE, pp. 324–331.

[90] Lanera C, Minto C, Sharma A, et al. Extending PubMed searches to ClinicalTrials.gov through a machine learning approach for systematic reviews. *Journal of Clinical Epidemiology* 2018; 103: 22–30.

[91] Naderalvojoud B, Bozkir AS, Sezer EA. Investigation of term weighting schemes in classification of imbalanced texts. In: *Proceedings of European Conference on Data Mining (ECDM), Lisbon*, pp. 15–17.

[92] Lessmann S. Solving Imbalanced Classification Problems with Support Vector Machines. In: *IC-AI*, pp. 214–220.

[93] Tan S. Neighbor-weighted k-nearest neighbor for unbalanced text corpus. *Expert Systems with Applications* 2005; 28: 667–671.

[94] Zheng T, Xie W, Xu L, et al. A machine learning-based framework to identify type 2 diabetes through electronic health records. *International journal of medical informatics* 2017; 97: 120–127.

[95] Friedman J, Hastie T, Tibshirani R. *The elements of statistical learning*. Springer series in statistics New York, http://statweb.stanford.edu/~tibs/book/preface.ps (2001, accessed 30 August 2017).

[96] KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction | BibSonomy, https://www.bibsonomy.org/bibtex/2cf4d2ac8bdac874b3d4841b4645a5a90/diana (accessed 4 September 2018).

[97] Kourbeti IS, Ziakas PD, Mylonakis E. Biologic therapies in rheumatoid arthritis and the risk of opportunistic infections: a meta-analysis. *Clin Infect Dis* 2014; 58: 1649–1657.

[98] Mountassir A, Benbrahim H, Berrada I. An empirical study to address the problem of unbalanced data sets in sentiment classification. In: *Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on*. IEEE, pp. 3298–3303.

[99] González RR, Iglesias EL, Diz LB. Applying balancing techniques to classify biomedical documents: An Empirical study. *International Journal of Artificial Intelligence™* 2012; 8: 186–201.

[100] Liu S, Forss T. Text Classification Models for Web Content Filtering and Online Safety. In: *Data Mining Workshop (ICDMW), 2015 IEEE International Conference on*. IEEE, pp. 961–968.

[101] Baracco GJ, Eisert S, Saavedra S, et al. Clinical and economic impact of various strategies for varicella immunity screening and vaccination of health care personnel. *Am J Infect Control* 2015; 43: 1053–60.

[102] Damm O, Ultsch B, Horn J, et al. Systematic review of models assessing the economic value of routine varicella and herpes zoster vaccination in high-income countries. *BMC Public Health* 2015; 15: 533.

[103] Kawai K, Gebremeskel BG, Acosta CJ. Systematic review of incidence and complications of herpes zoster: towards a global perspective. *BMJ Open* 2014; 4: e004833.

[104] Pierik JG, Gumbs PD, Fortanier SA, et al. Epidemiological characteristics and societal burden of varicella zoster virus in the Netherlands. *BMC Infect Dis* 2012; 12: 110.

[105] Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet* 2012; 13: 395–405.

[106] Afzal Z, Schuemie MJ, van Blijderveen JC, et al. Improving sensitivity of machine learning methods for automated case identification from free-text electronic medical records. *BMC Med Inform Decis Mak* 2013; 13: 30.

[107] Wang Z, Shah AD, Tate AR, et al. Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning. *PLoS One* 2012; 7: e30412.

[108] Kavuluru R, Rios A, Lu Y. An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. *Artif Intell Med* 2015; 65: 155–66.

[109] Wu PY, Cheng CW, Kaddi CD, et al. -Omic and Electronic Health Record Big Data Analytics for Precision Medicine. *IEEE Trans Biomed Eng* 2017; 64: 263–273.

[110] Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* 2010; 33: 1–22.

[111] Friedman J, Hastie T, Tibshirani R. Additive logistic regression: A statistical view of boosting - Rejoinder. *Ann Stat* 2000; 28: 400–407.

[112] Mani S, Chen Y, Arlinghaus LR, et al. Early prediction of the response of breast tumors to neoadjuvant chemotherapy using quantitative MRI and machine learning. *AMIA Annu Symp Proc* 2011; 2011: 868–77.

[113] Liu M, Hu Y, Tang B. Role of text mining in early identification of potential drug safety issues. *Methods Mol Biol* 2014; 1159: 227–51.

[114] Marafino B, Davies JM, Bardach NS, et al. N-gram support vector machines for scalable procedure and diagnosis classification, with applications to clinical free text data from the intensive care unit. *J Am Med Inform Assn* 2014; 21: 871–875.

[115] Lanera C, Paola B, Baldi I, et al. Maximizing Text Mining Performance: the Impact of preprocessing. In: Association AS (ed). 2016, pp. 3265–3270.

[116] Wu HC, Luk RWP, Wong KF, et al. Interpreting TF-IDF term weights as making relevance decisions. *Acm T Inform Syst*; 26. Epub ahead of print 2008. DOI: 10.1145/1361684.1361686.

[117] Goodall CR. Data mining of massive datasets in healthcare. *J Comput Graph Stat* 1999; 8: 620–634.

[118] Jurka TP. maxent: An R Package for Low-memory Multinomial Logistic Regression with Support for Semi-automated Text Classification. *R J* 2012; 4: 56–59.

[119] Renner IW, Warton DI. Equivalence of MAXENT and Poisson Point Process Models for Species Distribution Modeling in Ecology. *Biometrics* 2013; 69: 274–281.

[120] Jarek T. *caTools: Tools: moving window statistics, GIF, Base64, ROC AUC, etc.*, https://CRAN.R-project.org/package=caTools (2019).

[121] Dettling M, Buhlmann P. Boosting for tumor classification with gene expression data. *Bioinformatics* 2003; 19: 1061–9.

[122] Freund Y SR. Experiments with a new boosting algorithm. Morgan Kaufmann, 1996, pp. 148–156.

[123] Boughorbel S, Al-Ali R, Elkum N. Model Comparison for Breast Cancer Prognosis Based on Clinical Data. *PLoS One* 2016; 11: e0146413.

[124] Andrews PJ, Sleeman DH, Statham PF, et al. Predicting recovery in patients suffering from traumatic brain injury by using admission variables and physiological data: a comparison between decision tree analysis and logistic regression. *J Neurosurg* 2002; 97: 326–36.

[125] Landwehr N, Hall M, Frank E. Logistic model trees. *Mach Learn* 2005; 59: 161–205.

[126] Shane A. *Comparisons of boosted regression tree, GLM and GAM performance in the standardization of yellowfin tuna catch-rate data from the Gulf of Mexico lonline [sic] fishery*. Louisiana State University and Agricultural and Mechanical College, https://digitalcommons.lsu.edu/gradschool_theses/2880 (2009).

[127] Borra S, Di Ciaccio A. Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods. *Comput Stat Data An* 2010; 54: 2976–2989.

[128] Lilem L. G. *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters*. Advanced Analytics, LLC, 2014.

[129] Zec S, Soriani N, Comoretto R, et al. High Agreement and High Prevalence: The Paradox of Cohen's Kappa. *Open Nurs J* 2017; 11: 211–218.

[130] Xing C, Geng X, Xue H. Logistic Boosting Regression for Label Distribution Learning. *Proc Cvpr Ieee* 2016; 4489–4497.

[131] Heffernan R. Syndromic surveillance in public health practice, New York City.

[132] Henning KJ. What is syndromic surveillance? *Morbidity and Mortality Weekly Report* 2004; 7–11.

[133] Lall R, Abdelnabi J, Ngai S, et al. Advancing the use of emergency department syndromic surveillance data, New York City, 2012-2016. *Public Health Reports* 2017; 132: 23S-30S.

[134] Geisler BP, Schuur JD, Pallin DJ. Estimates of electronic medical records in US emergency departments. *PLoS One* 2010; 5: e9274.

[135] Obermeyer Z, Abujaber S, Makar M, et al. Emergency care in 59 low-and middle-income countries: a systematic review. *Bulletin of the World Health Organization* 2015; 93: 577–586.

[136] Taira BR, Orue A, Stapleton E, et al. Impact of a novel, resource appropriate resuscitation curriculum on Nicaraguan resident physician's management of cardiac arrest. *Journal of educational evaluation for health professions*; 13.

[137] Crouse HL, Torres F, Vaides H, et al. Impact of an Emergency Triage Assessment and Treatment (ETAT)-based triage process in the paediatric emergency department of a Guatemalan public hospital. *Paediatrics and international child health* 2016; 36: 219–224.

[138] Johnson T, Gaus D, Herrera D. Emergency Department of a Rural Hospital in Ecuador. *Western Journal of Emergency Medicine* 2016; 17: 66.

[139] Pérez A, Weegar R, Casillas A, et al. Semi-supervised medical entity recognition: A study on Spanish and Swedish clinical corpora. *Journal of Biomedical Informatics* 2017; 71: 16–30.

[140] Cotik V, Filippo D, Castaño J. An Approach for Automatic Classification of Radiology Reports in Spanish. *Studies in health technology and informatics* 2014; 216: 634–638.

[141] Castillo JJ. A machine learning approach for recognizing textual entailment in Spanish. In: *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on*

Computational Approaches to Languages of the Americas. Association for Computational Linguistics, 2010, pp. 62–67.

[142] Tanev H, Zavarella V, Linge J, et al. Exploiting machine learning techniques to build an event extraction system for portuguese and spanish. *Linguamática* 2009; 1: 55–66.

[143] Salton G, Buckley C. Term-weighting Approaches in Automatic Text Retrieval. *Inf Process Manage* 1988; 24: 513–523.

[144] Breiman L. Random Forests. *Machine Learning* 2001; 45: 5–32.

[145] Liaw A, Wiener M. Classification and regression by randomForest. *R news* 2002; 2: 18–22.

[146] Kim J-H. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis* 2009; 53: 3735–3745.

[147] *An Introduction to Statistical Learning - with Applications in R | Gareth James | Springer*, //www.springer.com/it/book/9781461471370 (accessed 22 October 2017).

[148] Harrell FEJ. rms: Regression Modeling Strategies. R package version 4.1-3, http://CRAN.R-project.org/package=rms (2014).

[149] Wickham H. *tidyverse: Easily Install and Load the 'Tidyverse'*, https://CRAN.R-project.org/package=tidyverse (2017).

[150] Grolemund G, Wickham H. Dates and Times Made Easy with lubridate. *Journal of Statistical Software* 2011; 40: 1–25.

[151] Hester J. *glue: Interpreted String Literals*, https://CRAN.R-project.org/package=glue (2017).

[152] Liu N, Zhang Z, Wah Ho AF, et al. Artificial intelligence in emergency medicine. *Journal of Emergency and Critical Care Medicine*; 2.

[153] Worster A, Bledsoe RD, Cleve P, et al. Reassessing the methods of medical record review studies in emergency medicine research. *Annals of emergency medicine* 2005; 45: 448–451.

[154] Donà D, Mozzo E, Scamarcia A, et al. Community-Acquired Rotavirus Gastroenteritis Compared with Adenovirus and Norovirus Gastroenteritis in Italian Children: A Pedianet Study. *Int J Pediatr* 2016; 2016: 5236243.

[155] Cantarutti A, Donà D, Visentin F, et al. Epidemiology of Frequently Occurring Skin Diseases in Italian Children from 2006 to 2012: A Retrospective, Population-Based Study. *Pediatr Dermatol* 2015; 32: 668–678.

[156] Gabutti G, Rota MC, Guido M, et al. The epidemiology of Varicella Zoster Virus infection in Italy. *BMC Public Health* 2008; 8: 372.

[157] Patterns in acute otitis media drug prescriptions: a survey of Italian pediatricians and otolaryngologists: Expert Review of Anti-infective Therapy: Vol 12, No 9, https://www.tandfonline.com/doi/abs/10.1586/14787210.2014.944503?journalCode=ierz20 (accessed 17 September 2019).

[158] Wang Y, Chen ES, Pakhomov S, et al. Investigating Longitudinal Tobacco Use Information from Social History and Clinical Notes in the Electronic Health Record. *AMIA Annu Symp Proc* 2017; 2016: 1209–1218.

[159] Karystianis G, Nevado AJ, Kim C, et al. Automatic mining of symptom severity from psychiatric evaluation notes. *Int J Methods Psychiatr Res*; 27. Epub ahead of print March 2018. DOI: 10.1002/mpr.1602.

[160] Nunes AP, Yang J, Radican L, et al. Assessing occurrence of hypoglycemia and its severity from electronic health records of patients with type 2 diabetes mellitus. *Diabetes Res Clin Pract* 2016; 121: 192–203.

[161] Rumshisky A, Ghassemi M, Naumann T, et al. Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. *Transl Psychiatry* 2016; 6: e921–e921.

[162] Marchisio P, Cantarutti L, Sturkenboom M, et al. Burden of acute otitis media in primary care pediatrics in Italy: a secondary data analysis from the Pedianet database. *BMC Pediatr* 2012; 12: 185.

[163] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019; 25: 44–56.

[164] Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. *NPJ Digit Med*; 1. Epub ahead of print 28 August 2018. DOI: 10.1038/s41746-018-0048-y.

[165] Juckett D. A method for determining the number of documents needed for a gold standard corpus. *Journal of Biomedical Informatics* 2012; 45: 460–470.

[166] Goycoolea MV, Hueb MM, Ruah C. Otitis media: the pathogenesis approach. Definitions and terminology. *Otolaryngol Clin North Am* 1991; 24: 757–761.

[167] Hallgren KA. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutor Quant Methods Psychol* 2012; 8: 23–34.

[168] On the dimensionality of word embedding, https://dl.acm.org/citation.cfm?id=3327026 (accessed 2 September 2019).

[169] Bojanowski P, Grave E, Joulin A, et al. Enriching Word Vectors with Subword Information. *arXiv:160704606 [cs]*, http://arxiv.org/abs/1607.04606 (2016, accessed 30 August 2019).

[170] Deep Learning on Electronic Health Records to Improve Disease Coding Accuracy. - PubMed - NCBI, https://www.ncbi.nlm.nih.gov/pubmed/31259017 (accessed 20 September 2019).

[171] Masters T. *Practical Neural Network Recipes in C++*. San Diego, CA, USA: Academic Press Professional, Inc., 1993.

[172] James G, Witten D, Hastie T, et al. *An Introduction to Statistical Learning: with Applications in R*. New York: Springer-Verlag, https://www.springer.com/gp/book/9781461471370 (2013, accessed 5 September 2019).

[173] Artificial Intelligence: A Modern Approach, 3rd Edition, http://www.mypearsonstore.com/bookstore/artificial-intelligence-a-modern-approach-9780136042594?xid=PSED (accessed 5 September 2019).

[174] Deep Learning with R. *Manning Publications*, https://www.manning.com/books/deep-learning-with-r (accessed 21 September 2018).

[175] Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. *arXiv:14126980 [cs]*, http://arxiv.org/abs/1412.6980 (2014, accessed 5 September 2019).

[176] Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv:150203167 [cs]*, http://arxiv.org/abs/1502.03167 (2015, accessed 5 September 2019).

[177] Hinton GE, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv:12070580 [cs]*, http://arxiv.org/abs/1207.0580 (2012, accessed 5 September 2019).

[178] Keskar NS, Mudigere D, Nocedal J, et al. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. *arXiv:160904836 [cs, math]*, http://arxiv.org/abs/1609.04836 (2016, accessed 5 September 2019).

[179] Lorenzoni G, Bressan S, Lanera C, et al. Analysis of Unstructured Text-Based Data Using Machine Learning Techniques: The Case of Pediatric Emergency Department Records in Nicaragua. *Med Care Res Rev* 2019; 1077558719844123.

[180] Breiman L. Random forests. *Machine Learning* 2001; 45: 5–32.

[181] Liang H, Tsui BY, Ni H, et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nat Med* 2019; 25: 433–438.

[182] Devlin J, Chang M-W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:181004805 [cs]*, http://arxiv.org/abs/1810.04805 (2018, accessed 20 September 2019).

[183] Yang Z, Dai Z, Yang Y, et al. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv:190608237 [cs]*, http://arxiv.org/abs/1906.08237 (2019, accessed 20 September 2019).

[184] Yang Q, Qi X, Li Y. The preventive effect of atorvastatin on atrial fibrillation: a meta-analysis of randomized controlled trials. *BMC Cardiovasc Disord* 2014; 14: 99.

[185] Meng Y, Dongmei L, Yanbin P, et al. Systematic review and meta-analysis of ustekinumab for moderate to severe psoriasis. *Clin Exp Dermatol* 2014; 39: 696–707.

[186] Li D-H, Pan Z-K, Ye F, et al. S-1-based versus 5-FU-based chemotherapy as first-line treatment in advanced gastric cancer: a meta-analysis of randomized controlled trials. *Tumour Biol* 2014; 35: 8201–8208.

[187] Lv Z-C, Ning J-Y, Chen H-B. Efficacy and toxicity of adding cetuximab to chemotherapy in the treatment of metastatic colorectal cancer: a meta-analysis from 12 randomized controlled trials. *Tumour Biol* 2014; 35: 11741–11750.

[188] Wang J, Yu J-T, Wang H-F, et al. Pharmacological treatment of neuropsychiatric symptoms in Alzheimer's disease: a systematic review and meta-analysis. *J Neurol Neurosurg Psychiatr* 2015; 86: 101–109.

[189] Zhou C-Q, Zhang J-W, Wang M, et al. Meta-analysis of the efficacy and safety of long-acting non-ergot dopamine agonists in Parkinson's disease. *J Clin Neurosci* 2014; 21: 1094–1101.

[190] Liu X, Xiao Q, Zhang L, et al. The long-term efficacy and safety of DPP-IV inhibitors monotherapy and in combination with metformin in 18 980 patients with type-2 diabetes mellitus—a meta-analysis. *Pharmacoepidemiology and drug safety* 2014; 23: 687–698.

[191] Douxfils J, Buckinx F, Mullier F, et al. Dabigatran etexilate and risk of myocardial infarction, other cardiovascular events, major bleeding, and all-cause mortality: a systematic review and meta-analysis of randomized controlled trials. *J Am Heart Assoc* 2014; 3: e000515.

[192] Li ECK, Heran BS, Wright JM. Angiotensin converting enzyme (ACE) inhibitors versus angiotensin receptor blockers for primary hypertension. *Cochrane Database Syst Rev* 2014; CD009096.

[193] Cavender MA, Sabatine MS. Bivalirudin versus heparin in patients planned for percutaneous coronary intervention: a meta-analysis of randomised controlled trials. *Lancet* 2014; 384: 599–606.

[194] Chatterjee S, Sardar P, Giri JS, et al. Treatment discontinuations with new oral agents for long-term anticoagulation: insights from a meta-analysis of 18 randomized trials including 101,801 patients. *Mayo Clin Proc* 2014; 89: 896–907.

[195] Funakoshi T, Latif A, Galsky MD. Safety and efficacy of addition of VEGFR and EGFR-family oral small-molecule tyrosine kinase inhibitors to cytotoxic chemotherapy in solid cancers: a systematic review and meta-analysis of randomized controlled trials. *Cancer Treat Rev* 2014; 40: 636–647.