Neural Networks 126 (2020) 170–177

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Prediction of admission in pediatric emergency department with deep neural networks and triage textual data

Bruno P. Roquette [a,1], Hitoshi Nagano [b,*], Ernesto C. Marujo [c], Alexandre C. Maiorano [a]

[a] Data Science Team, Itaú Unibanco, Praça Alfredo Egydio de Souza Aranha, 100, torre WMS, 10th floor, São Paulo - SP, 04344-902, Brazil
[b] Fundação Getúlio Vargas (FGV/EAESP), Rua Itapeva, 474/9th floor, São Paulo - SP, 01332-000, Brazil
[c] Instituto Tecnológico de Aeronáutica (ITA), Rua Mal. Eduardo Gomes, 50, Vila das Acacias, São José dos Campos - SP, 12228-900, Brazil

ARTICLE INFO

Article history:
Received 22 April 2019
Received in revised form 11 January 2020
Accepted 12 March 2020
Available online 18 March 2020

Keywords:
Deep neural networks
Emergency department admission
Gradient boosting
Prediction model
Triage
Unstructured data

ABSTRACT

Emergency department (ED) overcrowding is a global condition that severely worsens attention to patients, increases clinical risks and affects hospital cost management. A correct and early prediction of ED's admission is of high value and a motivation to adopt machine learning models. However, several of these studies do not consider data collected in textual form, which is a feature set that contains detailed information about patients and presents great potential for medical health care improvement. To this end, we propose and compare predictive models for admission that use both structured and unstructured data available at triage time. In total, our dataset comprised 499,853 pediatric ED's presentations (with an admission rate of 5.76%) of patients with age up to 18 years old observed over 3.5 years. Our best model consists of a 2-stage architecture with a deep neural network (DNN) to extract information from textual data followed by a gradient boosting classifier. This combined model achieved a value of 0.892 for the Area Under the Curve (AUC) in the test data. We highlight the importance of DNN-based text processing for better prediction, since the absence of text features resulted in AUC reduction of approximately two percentage points. Also, the feature importance of text was higher than that of the Manchester Triage System (MTS), which is a widely used risk classification protocol. These results suggest that activations from a trained DNN should be used in transfer learning setups in future studies.

1. Introduction

Emergency Departments present challenges for hospital managers due to increasing volume of patients, unpredictable nature of those arrivals and varying degrees of case urgency. A central part of this problem consists in correctly predicting admission among incoming patients. For the hospital, early detection of admissions often translates into better cost management. Several probabilistic models for predicting admissions were proposed in the literature aiming to optimize hospital resources, such as the preparation of beds and equipment necessary for those who will potentially need hospitalization. See, for example, Table 1 for a compilation of published material addressing the ED admission problem. The majority of these models use numerical and categorical variables related to the patient history and triage data. As for the patients and clinical body, the advantage of such system is the control of waiting times according to case acuity. This not only brings increasing levels of satisfaction, but also, most importantly, improves treatment results and reduces risks. Therefore, medical care should be provided at the earliest possible time, and attention should be prioritized for the most acute cases.

In this context, a widely accepted tool to achieve the desired goals of prioritization and efficiency is the classification protocol. This step is normally performed by a nurse who, after evaluating the patient, classifies the case on a priority scale according to the established protocol. The Canadian Triage and Acuity Scale (CTAS), the Emergency Severity Index (ESI) and the Manchester Triage System (MTS) are some of the widely used protocols in the world (Farrohknia et al., 2011) based on decision rules equivalent to those obtained in classification tree models, and can be applied at triage by the nurses (Gilboy, Tanabe, Travers, Rosenau, et al., 2012).

The work in van Veen and Moll (2009) provide an overview about the reliability and validity of these triage systems in children. In summary, the reliability of the MTS is good, the ESI is moderate to good and the reliability of CTAS is moderate. The authors indicate that MTS and CTAS both seem valid to triage

* Corresponding author.
E-mail addresses: bruno.roquette@itau-unibanco.com.br (B.P. Roquette), hitoshi.nagano@fgv.br (H. Nagano), marujo@ita.br (E.C. Marujo), alexandre.maiorano@itau-unibanco.com.br (A.C. Maiorano).

[1] Any opinions, findings, and conclusions expressed in this manuscript are those of the authors and do not necessarily reflect the views, official policy or position of Itaú Unibanco.

https://doi.org/10.1016/j.neunet.2020.03.012
0893-6080/

Table 1
Performance of binary classifiers for admissions prediction (sorted by year).

Reference	AUC	Pediatric only	Year	Admission rate (%)
Leegon, Jones, Lanaghan, and Aronsky (2006)	0.897	Yes	2006	15
Sun, Heng, Tay, and Seow (2011)	0.849	No	2011	30
Peck, Benneyan, Nightingale, and Gaehde (2012)	0.887	No	2012	N/A
Peck et al. (2013)	0.846	No	2013	26 to 32
Cameron, Rodgers, Ireland, Jamdar, and McKay (2015)	0.877	No	2014	N/A
Dugas et al. (2016)	0.830	No	2016	14
Lucke et al. (2018) (<70 y.o.)	0.860	No	2017	24.1
Lucke et al. (2018) (≥70 y.o.)	0.770	No	2017	44.4
Levin et al. (2018)	0.840	No	2017	22.3 to 26
Zhang et al. (2017)	0.846	No	2017	13.4
Barak-Corren, Fine, and Reis (2017)	0.910	Yes	2017	20.3
Graham, Bond, Quinn, and Mulvenna (2018)	0.859	No	2018	24
Hong, Haimovich, and Taylor (2018)	0.920	No	2018	29.7
Parker et al. (2019)	0.825	No	2019	38.7
Raita et al. (2019)	0.820	No	2019	16.2
Goto, Camargo, Faridi, Freishtat, and Hasegawa (2019)	0.850	Yes	2019	4.5

children in pediatric emergency care. In our dataset, the hospital adopted MTS, so that is the only risk classification we could utilize as feature. By traversing this decision tree (also called flowchart in the references) the patient is assigned a level of risk between blue and red. Medical care should be delivered immediately for level 1 (color red — immediate), within 10 min for level 2 (color orange — very urgent), 60 min for level 3 (color yellow — urgent), 120 min for level 4 (color green — standard), and 240 min for level 5 (color blue — non-urgent). More details about the recent version of this triage system can be found in Mackway-Jones, Marsden, and Windle (2013).

While MTS is a valid mechanism for prioritization, its sole use as an admission predictor could lead to misclassification, particularly in young and elderly patients (Zachariasse et al., 2017). Thus, more information about patients is necessary to better predict admissions. This approach is taken by the works in Barak-Corren et al. (2017), Cameron et al. (2015), Goto et al. (2019), Graham et al. (2018) and Parker et al. (2019), where the risk classification is an input to the model.

An interesting type of information resource for clinical data is the electronic health records (EHR) and medical knowledge, which contain a considerable amount of information coming in free text (Meystre & Haug, 2005). As pointed by Hao, Chen, Li, and Yan (2018), this data could be analyzed using text mining and that has attracted much attention in the literature, with annual growth rate reaching 7.31%, on average over 2008–2017 period. However, only a few studies used textual information in the ED admission context (Lucini et al., 2017; Zhang et al., 2017) and, according to Zhang et al. (2017), further exploration of these techniques in improving the prediction of hospital admission is needed.

Therefore, we propose and evaluate some machine learning algorithms to predict admission in ED. Our best model consists of a new 2-stage architecture that considers deep learning to extract information about textual data followed by a boosting classifier. In practical terms, we treated and incorporated textual information about current use medications, chief compliant, triage notes and previous visit medical/laboratorial exam descriptions.

We present a cohort study of 499,853 presentations to a pediatric hospital located in São Paulo, Brazil, in the period between January/2015 and August/2018. The feature set contains a mix of structured (e.g., measurements) and unstructured data (e.g., texts). The response variable is a binary outcome, where positive class means that the patient visit turned into an admission, whereas the negative class represents the patient discharge. In our dataset, the admission rate, or the proportion of positive samples equals 5.76%. Due to the presence of class imbalance, the best model was selected using the Area Under the Curve (AUC), which is a practice in accordance with several works in the field.

Our results show that employment of triage textual data does improve the model classification beyond what could be achieved by structured data only. In other words, we found out that there are important signals in the texts that are present in the hospital's EHR. According to our experiments, a model without those text signals is limited to an AUC performance 1.9 percentage points lower than the one with texts. In an actual system deployment for a hospital production environment, use of structured-only data models could imply an increase in costs and case misjudgment.

This article is divided in five sections and is organized as follows: in Section 2 we describe related works dealing with this kind of problem. Section 3 contains a summary of the analyzed dataset, describing rules and steps used for pre-processing data and models utilized. The final model and the obtained performance are presented in Section 4 and, finally, in Section 5 we provide comments on goals achieved and points of further research.

2. Related work

In this section, we highlight some works that address the ED admission probability problem. Their modeling strategies helped us define basic variables and constituted a starting point for our own data collection. Also, the published AUC results provide baseline figures for comparison with our model. In Table 1, we summarize the AUC results, along with some other aspects of each paper.

Firstly, we note that studies pertaining only to pediatric ED are rare in the literature and their results are comparable to those of more generic models that do not restrict themselves to pediatric patients. In the general case, i.e., for both pediatric and non pediatric contexts, we find that the top performing models in Barak-Corren et al. (2017) and Hong et al. (2018) describe models with AUC performances equal to 0.92 and 0.91 respectively.

Compared to Hong et al. (2018), our pediatric patients tend to have less past data on important admission predictors, such as a long history of chronic condition therapeutic medication or advanced age. In their work, out of the ten most important features, six correspond to binned outpatient medication subgroups (e.g., cardiovascular, gastrointestinal), while two are age related (age, employment status = retired). With Barak-Corren et al. (2017), the notable difference is the use of arrival mode (e.g., emergency vehicles, walk in). The addition of that feature could have improved the our model performance beyond what could be attained by clinical and text data only. The works in Hong et al. (2018) and Sun et al. (2011) also use arrival mode as a feature. Unfortunately, our dataset did not contain such data.

In other works related to pediatric data, such as in Leegon et al. (2006) authors make use of features indicating presence of laboratory tests, images and electrocardiogram. In our case, these exams were not available at triage, but only after a medical consultation. Basically, our model focuses on early prediction, where only limited clinical data are available.

In Goto et al. (2019), the authors predict clinical outcomes and disposition in pediatric ED and compare the performance of four machine learning–based models such as lasso regression, random forest, gradient-boosted decision trees and deep neural network. In that paper, the high quantity of ED visits by children in the United States (of 137 million annual emergency department visits, 30 million visits are made by children) is emphasized. Also, we note that their admission rate (equal to 4.5%) is lower than ours.

Features extracted from free texts in medical records were considered in recent works by Lucini et al. (2017) and Zhang et al. (2017). Lucini et al. (2017) provide the first study that utilizes NLP (Natural Language Processing) techniques for hospital admission and uses bag-of-words encoding followed by univariate feature selection and term-frequency inverse document frequency (tf–idf) transformation. Zhang et al. (2017) also uses univariate feature selection but relies on PCA (principal component analysis) to reduce dimensionality of bag-of-words vectors. We build on those models by devising other ways to capture signal from texts.

Particularly, our model attempts to extract a predictor from unstructured text data manually typed by the nurse during triage, using a deep neural network (DNN). We chose to use a DNN in a supervised setup, where the response variable is the admission event, to give our model a single purpose one that suits our medical oriented objectives. There are several frameworks to accomplish this objective, including, for example, word embedding in Mikolov, Sutskever, Chen, Corrado, and Dean (2013) that produces a compact and informative representation for text, which is independent from context.

Recently, a comprehensive literature review of the use of clinical text in medical applications was published in Mujtaba et al. (2019). According to that article, only 3 out of the 72 studies have used deep learning and we note that none of them address the admission problem specifically. Furthermore, that publication highlights the use of more deep learning frameworks as a future research direction.

3. Methodology

In this section, we first cover structured and unstructured data in Sections 3.1 and 3.2. Then, 3.3 describes the models, and 3.4 presents tuning and validation strategies. We utilized R version 3.5 and Python version 3.6 languages executed in GNU-Linux, kernel: 4.15 operational system. For the deep learning part, we utilized Keras version 2.2.4 with Tensorflow backend version 1.10.0 throughout all the steps involved in this work.

3.1. Data collection and pre-processing for structured data

Each entry register in the dataset represents a patient presentation characterized by 62 dimensions of various types, namely, numerical, binary, categorical and texts. In this subsection, we discuss treatments to all variables except for text, which comes in Section 3.2. Among the numerical variables, we have for example: temperature, heart rate, age and weight, while for binary we can mention city (whether or not the patient is local), gender and indicators of numerical variables absence (missing data). An overview of the study population according to the considered variables is presented in Tables 2 and 3.

Before the imputation of values, some variables were transformed as follows:

Table 2
Univariate statistics of patient arrivals. Total quantity: 499,853 arrivals.

Variable	Statistics
Gender	264,636 (52.9%) Men / 235,217 (47.1%) Women
Age [years]	1Q = 1.4 / Median = 2.9 / 3Q = 5.7
Health plan	492,032 (98.4%) Yes / 7821 (1.6%) No
Pain score (0,2)	463,422 (92.7%) 0 / 33,621 (6.7%) 1 / 2810 (0.6%) 2
Weight [kg]	1Q = 10.5 / Median = 14.5 / 3Q = 21.6
Heart rate [bpm]	1Q = 111 / Median = 128 / 3Q = 143
Oxygen saturation [%]	1Q = 95 / Median = 96 / 3Q = 97
Temperature [°C]	1Q = 36.1 / Median = 36.6 / 3Q = 37.3

Table 3
Univariate statistics of patient population. Total quantity: 135,516 patients.

Variable	Statistics
Gender	70,778 (52.2%) Men / 64,738 (47.8%) Women
State	134,621 (99.4%) Local state / 895 (0.6%) Other state
City	100,286 (74.0%) Local city / 35,230 (26%) Other city

- Time interval between arrival in hospital and triage: values lower than 0 or greater than 199th 200-quantile were converted to NA. 2311 records were changed.
- Time interval between current and last visit to ED: values lower than 0 were converted to NA. 843 records were changed.
- Weight: if age was lower than 11 and weight was higher than 100 kg, weight was converted to NA. 123 records were changed.

Table 4 presents information about missing values (NA) in the analyzed dataset. A zero quantity of NA's means that no imputation was required, even though imputation procedures were created for all columns. Per-column imputed values are described below.

- City = São Paulo. This is the most frequent value. Other cities had very small proportions (less than 1%).
- State = São Paulo. Most frequent value. Same idea.
- Geographical distance to patient city of birth = 0. Same idea.
- Weight and Heart rate. We used the result of a linear regression model trained with age and sex variables. The p-values of both variables and constant coefficient were lower than 2.2×10^{-16}.
- Temperature = 36.5 (absence of fever and equal to the median of training data). Assumption that the lack of recorded temperature is due to the fact that there is no fever.
- Pain score. If symptom is equal to *pain* or the word *pain* was present in the chief complaint or in nurse observations, this variable was converted to 1 and for the others cases, 0. Same assumption of previous item.
- Oxygen saturation = 96 (normal value for healthy patient and equal to median of training data). Same assumption of previous item.
- Capillary blood glucose = 92 (normal value for healthy patient and equal to median of training data). Same assumption of previous item.
- Time interval between arrival in hospital and triage = 281. Median value of training data.
- Correct MTS classification in the last visit = True. Required in the patient's first visit.
- Was admitted in the last visit = False. This imputation value could be due to: (a) first visit; (b) no data, which may be caused by the frequency of admittance being considerably lower than the frequency of discharge.
- Blood pressure = 105 for systolic and 71 for diastolic pressures. Median of training data.

Table 4
List of variables.

Type	Description	NA
binary	was MTS classification correct in the last visit?	135,516
binary	is São Paulo city?	11
numeric	MTS classification (Red = 1, . . . , Blue = 5)	2,249
numeric	geographical distance to patient city of birth	11
numeric	pain score	16,167
text	drugs	449,415
text	triage notes	79,841
text	past visit triage notes	197,526
text	chief complaint	2,252
text	past visit chief complaint	137,375
categorical	symptom from a pre-defined list	2,248
numeric	time interval between current and past visits	136,354
numeric	time interval between arrival and triage	4,738
categorical	season of the year	0
binary	is São Paulo state?	11
numeric	past image exam/visit ratio	0
numeric	past admission/visit ratio	0
numeric	past laboratory exam/visit ratio	0
numeric	heart rate	26,478
numeric	capillary blood glucose	499,311
categorical	arrival time of day	0
numeric	age	0
text	last visit medical image exams requested	381,096
binary	was admitted in last visit?	135,516
binary	has health insurance?	0
binary	doctor acknowledge last triage MTS classification	0
binary	is city empty?	0
binary	is MTS classification empty?	0
binary	is geographical distance empty	0
binary	is pain empty?	0
binary	is drugs empty?	0
binary	is triage notes empty?	0
binary	is chief complaint empty?	0
binary	is symptom empty?	0
binary	is first visit?	0
binary	is time interval between arrival and triage empty?	0
binary	is state empty?	0
binary	is heart rate empty?	0
binary	is respiratory rate empty?	0
binary	is capillary blood glucose empty?	0
binary	is past image exam empty?	0
binary	is past admitted empty?	0
binary	is past laboratory exam empty?	0
binary	is paste visit triage notes empty?	0
binary	is weight empty?	0
binary	is diastolic pressure empty?	0
binary	is systolic pressure empty?	0
binary	is past chief complaint empty?	0
binary	is oxygen saturation empty?	0
binary	is temperature empty?	0
text	last visit laboratory tests requested	418,144
categorical	month of visit	0
numeric	weight	5,073
numeric	diastolic pressure	498,884
numeric	systolic pressure	499,225
numeric	number of past image exam requests	0
numeric	number of past ED admissions	0
numeric	number of past laboratory test requests	0
numeric	number of past ED visits	0
numeric	oxygen saturation	24,726
binary	sex	0
numeric	temperature	10,229

- Symptom = create new category named ‘empty’.
- MTS classification = 4. Median of training data.
- Time interval between current and last visit to ED = 2 years. This is above the 95th percentile and was chosen to reduce the effect of high correlation between this variable and readmission. Absence of this value is primarily due to first time visits.

Finally, all numerical variables were normalized. All these imputation rules were discussed with medical doctors and experienced health care professionals and they considered them all reasonable assumptions.

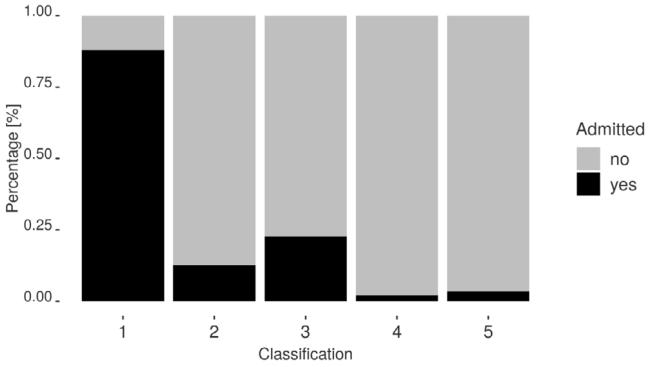

Fig. 1. Admissions percentage by MTS classification.

Among all variables explored, the triage classification is especially meaningful. Fig. 1 depicts the admitted patient proportion by this variable. A hypothesis raised a priori would be that higher degrees of risk in the triage protocol correlate to higher probabilities of admission. However, the observations do not match that expectation.

We must note that this risk assessment used in isolation by hospital managers could lead to, for example, overestimation for class 2 and underestimation for class 3.

3.2. Data collection and pre-processing for unstructured data

In order to exemplify the contents of the text, raw samples are presented in Table 5. Each one of the five text fields is described as follows:

- drugs: current use medications
- chief complaint
- triage notes: textual triage notes
- past image exam: medical image exams requested in the last visit
- past laboratory exam: laboratory tests requested in the last visit

We also add the previous visit texts for chief complaint and triage notes, when available, i.e., when it is not the first visit. Thus, we have seven text fields for feature extraction.

Each text field pre-processing followed the steps: 1 — conversion of all characters to lowercase, 2 — replacement of special characters, numbers, single characters and multiple spaces for single space, 3 — character conversion with accents to characters without accent, for example, the replacement of *ç* by *c*, 4 — cropping of the lengths to 20 tokens, a process known as pad sequence, 5 — conversion of the token to one-hot-encoded sparse vectors that were used to feed to DNN described in 3.3. No imputation was performed on these variables and this procedure resulted in a total of 32 129 unique tokens.

In Table 5 we also present in separate columns the corrected version of the original text, along with an English translation of the corrected text. Note that the token *ccça*, which actually refers to the term *criança* (*child*), is an acknowledged token among the clinical body. Our pre-processing did not substitute the correct word in Portuguese language. Instead, the DNN “finds” the similarities between these two words. The same occurs with several medical terms that are abbreviated.

3.3. Description of the models

We built a total of six models: (1) SVM (Support Vector Machine); (2) ElasticNet; (3) DNN; (4) Catboost with structured data

Table 5
Text samples in open text fields.

Original text	Corrected text	Translation of corrected text
Field#1: drugs		
• montelar flixotide	• montelair flixotide	• montelair flixotide
• medicada triagem dipirna gts	• medicada na triagem dipirona gotas	• medicated in triage dipyrone drops
• medicado casa ibuprofeno gts	• medicado em casa ibuprofeno gotas	• medicated at home ibuprofen drops
Field#2: chief complaint		
• febre, dor de garganta e dor de cabeça	• febre, dor de garganta e dor de cabeça	• fever, sore throat and headache
• dor na nuca, dificuldade para engolir e respirar	• dor na nuca, dificuldade para engolir e respirar	• neck pain, difficulty of swallowing and breathing
• febre, nausea, e hiperemia pelo corpo e edema de mmssii	• febre, nausea, e hiperemia pelo corpo e edema de membros superiores e inferiores	• fever, nausea and hyperemia throughout body and edema of upper and lower limbs
Field#3: triage notes		
• 22:30 medicado com alivium pelo pai	• 22:30 medicado com alivium pelo pai	• 10:30PM medicated by father with alivium
• cça agitada e chorosa, não foi possivel verificar fc e sat	• criança agitada e chorosa, não foi possivel verificar frequencia cardıiaca e saturação de oxigênio	• agitated and crying child, not possible to check heart rate and oxygen saturation
• em uso de loratadina por conta própria	• em uso de loratadina por conta própria	• in use of loratadine on their own
Field#4: last visit medical image exams requested		
• rx seios da face − fn + mn + lat, rx torax − frente e perfil	• raio-x seios da face − fronto-naso + mento-naso + lateral, raio-x thorax − frente e perfil	• X-ray facial sinus − occipitofrontal + occipitomental + lateral, X-ray thorax − frontal and lateral
• usg cervical	• ultrassonografia cervical	• cervical ultrasonography
Field#5: last visit laboratory tests requested		
• streptococcus - a, teste rapido, hemograma	• streptococcus - a, teste rapido, hemograma	• streptococcus - a, rapid test, hemogram
• hemograma, transaminase piruvica, glicose, urina tipo i, transaminase oxalacetica	• hemograma, transaminase piruvica, glicose, urina tipo I, transaminase oxalacetica	• hemogram, pyruvate transaminase, glucose, type I urine, oxalacetic transaminase
• sodio no sangue, calcio, proteina c reativa, magnesio	• sodio no sangue, calcio, proteina c reativa, magnesio	• sodium in blood, calcium, c-reactive protein, magnesium

only; (5) XGBoost; and (6) Catboost. Except for (4), models used all available data, that is, structured and unstructured data.

In models (1), (2) and (3) structured and unstructured feature vectors are concatenated to form a single vector that is used as input to algorithms. Model (3) consists of using a classical multi-layer deep neural network for all structured and unstructured data combined. In models (5) and (6) we attempted a hybrid approach where initially neural networks are used only for text to produce a single scalar, which we call *text2num*. Then, this number is concatenated to the feature vector of structured variables to produce the input to the boosting algorithm. The two boosting algorithms, namely, XGBoost and Catboost are described in Chen and Guestrin (2016) and Prokhorenkova, Gusev, Vorobev, Dorogush, and Gulin (2018), respectively.

The model architecture to obtain *text2num* is depicted in Fig. 2 and is explained as follows. The inputs are the sparse vectors from one-hot-encoding step followed up by an embedding step producing a vector of 16 dimensions. These vectors are then input to Long Short-Term Memory (LSTM, Hochreiter and Schmidhuber (1997)) layers, one layer per text field. Then, the output of these seven LSTM layers are concatenated and conveyed through a single dense layer in a feed-forward fashion until a single final sigmoid unit. Note that this represents a stand-alone supervised classifier trained with the admission event ground truth. The outcome of the sigmoid unit is the feature used along with structured part for the downstream boosting algorithm. We also ran model (4), which only takes structured variables,i.e., without *text2num*. This was done to estimate the impact extent of the natural language processing techniques on performance.

On the specific aspect of the imbalance nature of our dataset, we tuned *class weights* (Catboost, 2020; Xgboost, 2020), a hyper-parameter available in the boosting algorithms that acts on the way that loss is computed between minority and majority classes. This unbalancing mitigation strategy is an alternative to data resampling, but we note that it was not applied to the DNN part of the model.

3.4. Model tuning and validation

We subdivided the dataset into two parts, according to the following dates:

- train: from Jan/2015 until Apr/2018 (467,571 observations)
- test: from May/2018 until Aug/2018 (32,282 observations)

For the text processing DNN, we used a train-validation split where the last four months of the train set, namely, Jan/2018 through Apr/2018 were utilized as validation set. This single split was chosen because the neural network training poses high processing demands. A time-series *n*-split, as described in Scikit-learn.org (2019), would have required more computational power, since successive folds with increasing training data sizes would have to be built. We note that the small train size in the first folds would entail a reduced number of tokens and sentences, thus impairing the validity of the cross validation.

Now, for the downstream boosting models, we conducted a time-series 3-split cross validation. These tree-based algorithms are very sensitive to hyper-parameter settings, which justified a more comprehensive strategy.

Lastly, for the test data, a bootstrapping of 100 rounds was considered and the average and confidence intervals of AUC were collected.

4. Results

The results obtained in the test set are found in Table 6. We can see that the use of unstructured data improved the model's performance, increasing the AUC by 1.9 percentage points, which is equivalent to a performance gain of 2.2% approximately in relation to the model without *text2num*. This gain means reducing classification errors by approximately 141 patients per month on average, which can directly impact hospital costs. The thresholds used in this comparison are given by Youden's Indices,

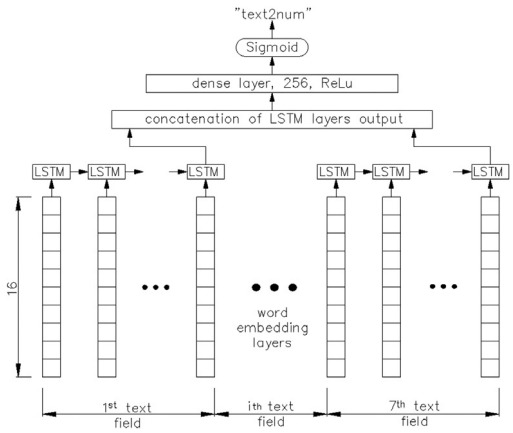

Fig. 2. DNN architecture with Long Short Term Memory (LSTM) layers for text variables conversion.

Table 6
Model choices in the 2nd stage and their performances.

Model	AUC
SVM	0.687
ElasticNet	0.840
CatBoost without text features	0.872
DNN	0.877
XGBoost	0.890
CatBoost with text features	0.891

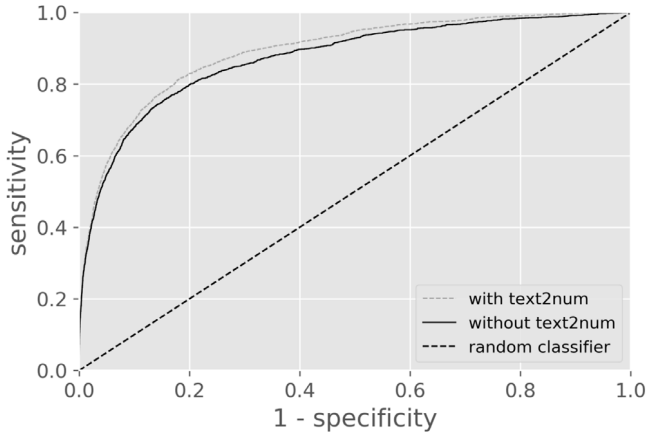

Fig. 3. ROC curves.

a common practice when equal weight is given to sensitivity and specificity, (Schisterman, Perkins, Liu, & Bondell, 2005).

In order to reduce the dimensionality of the problem and anticipate a practical implementation scenario, we conducted a recursive feature elimination procedure as follows. A random variable with uniform distribution was added to the dataset to be used as cut-off point after calculation of the features importance. Initially, we run the model with all features, evaluate features importance and drop those with importance equal to or less than the random feature. In a recursive way, with decreasing set sizes, we repeat this operation until the random variable becomes the least important variable. When this is achieved the random variable is discarded. This procedure was executed 30 times with different random seeds. Each execution produced a different feature set and its intersection was considered to be the final feature set selection, composed of 18 features.

With this reduced feature set, the cross-validation strategy was done again and the result obtained in the test data after 100 rounds through bootstrap was 0.892 for the AUC average with a 95% confidence interval between 0.885 and 0.900. The ROC of this classifier can be verify in Fig. 3. We also plot the Catboost curve without text in the same figure.

Table 7
Confusion matrices: Catboost without text features for threshold = 0.21; Catboost with text features for threshold = 0.22. Youden's Index in both cases.

Catboost without *text2num* (A)		Predict	
		Negative	Positive
True Class	Negative	24,193	6066
	Positive	406	1617
Catboost with *text2num* (B)		**Predict**	
		Negative	Positive
True Class	Negative	24,704	5555
	Positive	370	1653

Fig. 4 illustrates the importance of the 18 variables that stood out. *text2num* comes in first, an outcome that corroborates our findings that text fields do hold interesting and predictive signals. The runner up is the assigned MTS classification. A possible interpretation here is that *text2num* is helping to adjust the MTS classification in the right direction. In third and fourth places, we have the time interval between the current and the last visit and the oxygen saturation. The remaining ones are vital signs, demographics and temporal (visit hour and month) features. We observe that none of highly imputed variables made the final list and distance is the only one present out of the geographical set variables. Here, feature importance with *Prediction Values Change* concept was utilized and its details can be found in Catboost (2019). Basically, the importance is high when the changes in the feature value causes a high average change to prediction.

The performance of this reduced feature set model is comparable with that of the full feature set, but it runs approximately twice as fast. In addition, we can expect that in a production environment, a less complex model makes the system more robust to lack of data completeness.

Table 7 compares the confusion matrices for the two scenarios: (A) without text; (B) with texts. These figures correspond to the test set, equivalent to a complete 4 months interval. Overall, there is an improvement of 547 correct classifications. Then, there is a reduction of 36 false negatives, and this improvement comes with no penalty to false positives. In fact, this metric also improves the results by 511 cases. In addition, if we consider that bed preparation is carried out for predicted positive cases,

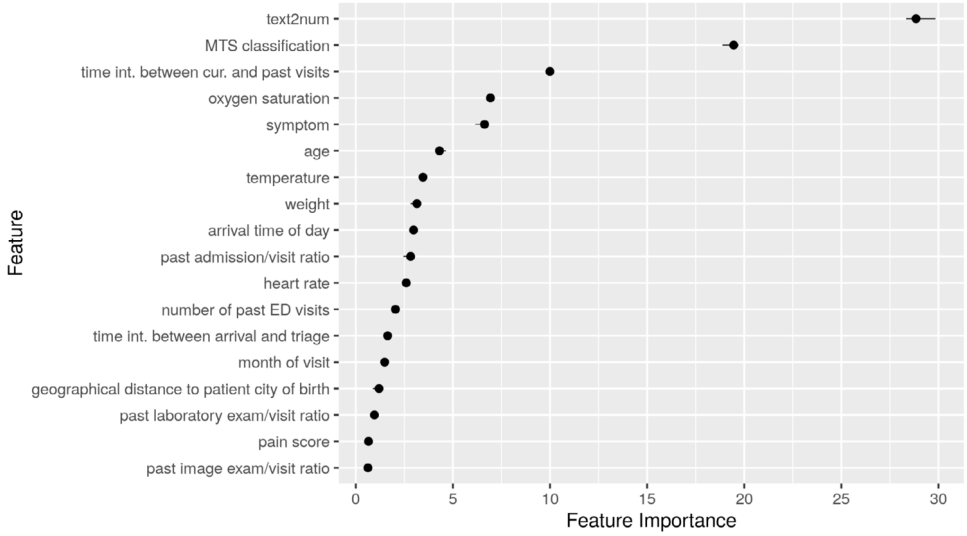

Fig. 4. Variables importance after a recursive feature elimination. Average value as dots; minimum, maximum interval in horizontal lines.

Table 8
Summary of statistical measures using different thresholds. The threshold value of 0.22 corresponds to the maximum Youden's Index (J-index); the value of 0.85 corresponds with the maximum accuracy.

Threshold	J-Index	Sensitivity	Specificity	PPV	NPV	Accuracy
0.10	0.550	0.907	0.643	0.145	0.990	0.659
0.20	0.628	0.829	0.799	0.216	0.986	0.801
0.22	0.633	0.817	0.816	0.229	0.985	0.816
0.30	0.624	0.759	0.865	0.273	0.982	0.858
0.40	0.593	0.688	0.905	0.326	0.977	0.891
0.50	0.560	0.627	0.933	0.383	0.974	0.913
0.60	0.511	0.557	0.954	0.450	0.970	0.930
0.70	0.444	0.473	0.971	0.522	0.965	0.940
0.80	0.333	0.347	0.986	0.628	0.958	0.946
0.85	0.255	0.261	0.994	0.733	0.953	0.948
0.90	0.180	0.183	0.997	0.800	0.948	0.946

we can find that 475 less beds would be needed, indicating an opportunity for better cost management.

An interesting form to present the results is shown in Table 8. In that table, we can easily analyze the effect of using various probability thresholds for classification. Those metrics, commonly found in medical literature, are derived from the confusion matrices. PPV and NPV stands for Positive Predicted Value and Negative Predicted Value, respectively. The value of J-index equal to 0.633 that maximizes the sum of sensitivity and specificity, corresponds to the threshold of 0.22.

5. Conclusions and future steps

In this paper, we focused on early prediction, so we utilized only triage data and past visits history. The modeling strategies of data collection, pre-processing, architecture and validation were described along with the obtained results. We also built simpler models for comparison and to support our finding about text data effect on the performance increment. In a pure AUC baseline comparison, our classifier performed nearly as well as the best models found in the literature. However, it is not straightforward to compare models trained and tested in different datasets due to different data distribution.

The use of this model could provide significant gains in terms of efficient use of ED resources since it would be possible to better anticipate needed equipment, staff and other resources for the admitted patient. In particular, we note that current EHR systems commonly store several fields of unstructured data. However, we think that very few, if any, make use of such data for concrete prediction purposes in the Portuguese language. We could not find any published material on such application.

Some of the variables shown to be of major importance in other studies were not available to us in this work. For example: race and form of arrival at the hospital. The use of these in future studies may potentially improve the performance of our current models.

In addition, we intend to test methodological enhancements to perform orthographic corrections of text variables, possibly using metrics such as Levenshtein distance. In an architectural perspective, we would like to focus on new DNN architectures such as: Bidirectional Encoder Representations from Transformers (BERT) language representation (Devlin, Chang, Lee, & Toutanova, 2019) for text variables and Wide & Deep Learning (Cheng et al., 2016) for the entire dataset.

Another interesting candidate for further research resides on other prediction problems that can use the history of admissions and symptoms as features. We note that the DNN gives us *text2num* and intermediate activations that can be used in other downstream classifiers. This is an instance of transfer learning and could give a multi-purpose status to our proposed architecture. These triage-text-embeddings could be features in predictions of, e.g., mortality in Intensive Care Unit (ICU), chronic diseases, high cost treatments, among others.

Finally, results on test data indicate that the proposed framework has a good chance of being successfully implemented in pediatric ED in a production setup. We expect that a real test can be carried out to find beneficial impacts to health care improvement, risk decrease, improved patient satisfaction and higher managerial efficiency.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We want to express our gratitude to the doctors, nurses and staff of Sabará Children's Hospital. They were always supportive and not only made considerable efforts to make the appropriate data available for the research but also always had a word of encouragement for the pursuit of better medical services with the use of data science methodology.

References

Barak-Corren, Y., Fine, A. M., & Reis, B. Y. (2017). Early prediction model of patient hospitalization from the pediatric emergency department. *Pediatrics*, [ISSN: 0031-4005] *139*(5).

Cameron, A., Rodgers, K., Ireland, A., Jamdar, R., & McKay, G. A. (2015). A simple tool to predict admission at the time of triage. *Emergency Medical Journal*, *32*(3), 174–179.

Catboost (2019). Feature importance - predictionvalueschange. Available at: https://catboost.ai/docs/concepts/fstr.html. Accessed on 2019-12-26.

Catboost (2020). Python package training parameters. Available at: https://catboost.ai/docs/concepts/python-reference_parameters-list.html. Accessed on 2020-01-05.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *KDD '16, Proceedings of the 22Nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785–794). New York, NY, USA: ACM, ISBN: 978-1-4503-4232-2.

Cheng, H.-T., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., et al. (2016). Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems* (pp. 7–10). ACM.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics.

Dugas, A. F., Kirsch, T. D., Toerper, M., Korley, F., Yenokyan, G., France, D., et al. (2016). An electronic emergency triage system to improve patient distribution by critical outcomes. *The Journal of Emergency Medicine*, *50*(6), 910–918.

Farrohknia, N., Castrén, M., Ehrenberg, A., Lind, L., Oredsson, S., Jonsson, H., et al. (2011). Emergency department triage scales and their components: a systematic review of the scientific evidence. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, *19*(1), 42.

Gilboy, N., Tanabe, P., Travers, D., Rosenau, A., et al. (2012). *Emergency severity index (esi): a triage tool for emergency department care, version 4. Implementation handbook 2012 Edition* (pp. 12–0014). Agency for Healthcare Research and Quality, AHRQ Publication No. 12-0014.

Goto, T., Camargo, C. A., Faridi, M. K., Freishtat, R. J., & Hasegawa, K. (2019). Machine learning-based prediction of clinical outcomes for children during emergency department triage. *JAMA Network Open*, *2*(1), e186937.

Graham, B., Bond, R., Quinn, M., & Mulvenna, M. (2018). Using data mining to predict hospital admissions from the emergency department. *IEEE Access*, *6*, 10458–10469.

Hao, T., Chen, X., Li, G., & Yan, J. (2018). A bibliometric analysis of text mining in medical research. *Soft Computing*, *22*(23), 7875–7892.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, [ISSN: 0899-7667] *9*(8), 1735–1780.

Hong, W. S., Haimovich, A. D., & Taylor, R. A. (2018). Predicting hospital admission at emergency department triage using machine learning. *PloS One*, *13*(7), e0201016.

Leegon, J., Jones, I., Lanaghan, K., & Aronsky, D. (2006). Predicting hospital admission in a pediatric emergency department using an artificial neural network. In *AMIA annual symposium proceedings, Vol. 2006* (p. 1004). American Medical Informatics Association.

Levin, S., Toerper, M., Hamrock, E., Hinson, J. S., Barnes, S., Gardner, H., et al. (2018). Machine-learning-based electronic triage more accurately differentiates patients with respect to clinical outcomes compared with the emergency severity index. *Annals of Emergency Medicine*, *71*(5), 565–574.

Lucini, F. R., Fogliatto, F. S., da Silveira, G. J., Neyeloff, J. L., Anzanello, M. J., Kuchenbecker, R. d. S., et al. (2017). Text mining approach to predict hospital admissions using early medical records from the emergency department. *International Journal of Medical Informatics*, *100*, 1–8.

Lucke, J. A., de Gelder, J., Clarijs, F., Heringhaus, C., de Craen, A. J., Fogteloo, A. J., et al. (2018). Early prediction of hospital admission for emergency department patients: a comparison between patients younger or older than 70 years. *Emergency Medical Journal*, *35*(1), 18–27.

Mackway-Jones, K., Marsden, J., & Windle, J. (2013). *Emergency triage: Manchester triage group*. John Wiley & Sons.

Meystre, S., & Haug, P. J. (2005). Automation of a problem list using natural language processing. *BMC Medical Informatics and Decision Making*, *5*(1), 30.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *NIPS'13, Proceedings of the 26th international conference on neural information processing systems - volume 2* (pp. 3111–3119). USA: Curran Associates Inc..

Mujtaba, G., Shuib, L., Idris, N., Hoo, W. L., Raj, R. G., Khowaja, K., et al. (2019). Clinical text classification research trends: Systematic literature review and open issues. *Expert Systems with Applications*, [ISSN: 0957-4174] *116*, 494–520.

Parker, C. A., Liu, N., Wu, S. X., Shen, Y., Lam, S. S. W., & Ong, M. E. H. (2019). Predicting hospital admission at the emergency department triage: a novel prediction model. *The American Journal of Emergency Medicine*, *37*(8), 1498–1504.

Peck, J. S., Benneyan, J. C., Nightingale, D. J., & Gaehde, S. A. (2012). Predicting emergency department inpatient admissions to improve same-day patient flow. *Academic Emergency Medicine*, *19*(9), 1045–1054.

Peck, J. S., Gaehde, S. A., Nightingale, D. J., Gelman, D. Y., Huckins, D. S., Lemons, M. F., et al. (2013). Generalizability of a simple approach for predicting hospital admission from an emergency department. *Academic Emergency Medicine*, *20*(11), 1156–1163.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). Catboost: unbiased boosting with categorical features. In *Advances in neural information processing systems* (pp. 6638–6648).

Raita, Y., Goto, T., Faridi, M. K., Brown, D. F., Camargo, C. A., & Hasegawa, K. (2019). Emergency department triage prediction of clinical outcomes using machine learning models. *Critical Care*, *23*(1), 64.

Schisterman, E. F., Perkins, N. J., Liu, A., & Bondell, H. (2005). Optimal cut-point and its corresponding Youden index to discriminate individuals using pooled blood samples. *Epidemiology*, *16*(1), 73–81.

Scikit-learn. org (2019). 3.1. cross-validation: evaluating estimator performance — scikit-learn 0.22.1 documentation. Available at: https://scikit-learn.org/stable/modules/cross_validation.html#time-series-split. Accessed on 2019-12-26.

Sun, Y., Heng, B. H., Tay, S. Y., & Seow, E. (2011). Predicting hospital admissions at emergency department triage using routine administrative data. *Academic Emergency Medicine*, *18*(8), 844–850.

van Veen, M., & Moll, H. A. (2009). Reliability and validity of triage systems in paediatric emergency care. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, *17*(1), 38.

Xgboost (2020). Xgboost parameters. Available at: https://xgboost.readthedocs.io/en/latest/parameter.html. Accessed on 2020-01-05.

Zachariasse, J. M., Seiger, N., Rood, P. P., Alves, C. F., Freitas, P., Smit, F. J., et al. (2017). Validity of the manchester triage system in emergency care: A prospective observational study. *PLoS One*, *12*(2), e0170811.

Zhang, X., Kim, J., Patzer, R. E., Pitts, S. R., Patzer, A., & Schrager, J. D. (2017). Prediction of emergency department hospital admission based on natural language processing and neural networks. *Methods of Information in Medicine*, *56*(5), 377–389.