



Article

Clinical Screening Prediction in the Portuguese National Health Service: Data Analysis, Machine Learning Models, Explainability and Meta-Evaluation

Teresa Gonçalves ^{1,2,*} , Rute Veladas ¹, Hua Yang ^{1,3,*} , Renata Vieira ⁴ , Paulo Quaresma ^{1,2} , Paulo Infante ^{5,6} , Cátia Sousa Pinto ⁷, João Oliveira ⁷, Maria Cortes Ferreira ⁷, Jéssica Morais ⁷, Ana Raquel Pereira ⁷, Nuno Fernandes ⁷ and Carolina Gonçalves ⁷

¹ Department of Computer Science, University of Évora, 7000-671 Évora, Portugal

² Centro Algoritmi, Vista Lab, University of Évora, 7000-671 Évora, Portugal

³ Department of Computer Science, Zhongyuan University of Technology, Zhengzhou 450007, China

⁴ CIDEHUS, University of Évora, 7000-809 Évora, Portugal

⁵ Department of Mathematics, University of Évora, 7000-671 Évora, Portugal

⁶ CIMA—Centro de Investigação em Matemática e Aplicações, University of Évora, 7000-671 Évora, Portugal

⁷ Serviços Partilhados do Ministério da Saúde, 1050-099 Lisboa, Portugal

* Correspondence: tcg@uevora.pt (T.G.); huayangnho@hotmail.com (H.Y.)

Abstract: This paper presents an analysis of the calls made to the Portuguese National Health Contact Center (SNS24) during a three years period. The final goal was to develop a system to help nurse attendants select the appropriate clinical pathway (from 59 options) for each call. It examines several aspects of the calls distribution like age and gender of the user, date and time of the call and final referral, among others and presents comparative results for alternative classification models (SVM and CNN) and different data samples (three months, one and two years data models). For the task of selecting the appropriate pathway, the models, learned on the basis of the available data, achieved F1 values that range between 0.642 (3 months CNN model) and 0.783 (2 years CNN model), with SVM having a more stable performance (between 0.743 and 0.768 for the corresponding data samples). These results are discussed regarding error analysis and possibilities for explaining the system decisions. A final meta evaluation, based on a clinical expert overview, compares the different choices: the nurse attendants (reference ground truth), the expert and the automatic decisions (2 models), revealing a higher agreement between the ML models, followed by their agreement with the clinical expert, and minor agreement with the reference.

Keywords: clinical triage; clinical pathways; SNS24; data analysis; machine learning; support-vector machines; deep neural networks; explainability



Citation: Gonçalves, T.; Veladas, R.; Yang, H.; Vieira, R.; Quaresma, P.; Infante, P.; Sousa Pinto, C.; Oliveira, J.; Cortes Ferreira, M.; Morais, J.; et al. Clinical Screening Prediction in the Portuguese National Health Service: Data Analysis, Machine Learning Models, Explainability and Meta-Evaluation. *Future Internet* **2023**, *15*, 26. <https://doi.org/10.3390/fi15010026>

Academic Editor: Hamid Mcheick

Received: 10 November 2022

Revised: 8 December 2022

Accepted: 16 December 2022

Published: 3 January 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The Portuguese National Health Contact Center, called SNS24, is a national telephone and digital public service in Portugal, supplied by trained nurses, that delivers clinical services for citizens.

When receiving a call, and according to the citizen's medical history and self-reported symptoms, the nurse selects the most appropriate clinical pathway. This pathway leads to five possible final referrals: transference to Poison Information Center (PIC), transference to the National Medical Emergency Institute (INEM), clinical assessment in hospital emergency or at a primary health care unit, and self-care.

During the triage, nurses ensure that the main symptoms reported are registered and considered for the correct choice of the clinical pathway, which, in place, determine the final referral. As other triage protocols such as the Manchester Triage System [1], these pathways use a risk-averse system prioritisation. For this reason, SNS24 service has an

important role in the National Health Service since allows a close connection with citizens in their home places and, according to their symptoms, a specialist analysis upon their clinical situation and decision about the most appropriate referral.

Clinical pathways are developed by health professionals and approved by the General Directorate of Health (DGS). During the period of this study, the number of existent clinical pathways was 54, ranging from “Cough” or “Abdominal Pain” to “Allergy”.

The SNS24 Scout.AI project aims at applying Artificial Intelligence methodologies in the construction of decision support tools to help nurses in selecting the most appropriate clinical pathway and to support DGS in the process of optimising the design of clinical pathways and respective referrals.

This study presents the work developed during the second year of the project, which extends and elaborates on previous experiments using three years of receiving calls, in a total of around 2,577 millions of records. In [2], a preliminary study using different text representations and shallow ML classification algorithms is presented while in [3] a thorough analysis of Deep Neural Network (DNN) models is exposed. Both studies use a subset of data composed of 3 months calls (with around 270,000 records).

Differently from these previous works, the current one presents a detailed characterisation of the data, a comparison between both approaches along with a new comprehensive analysis of the results, regarding errors per class, model explainability and a meta-evaluation made by an clinical expert.

The main contributions of this paper are summarised as follows:

- a characterisation of 3 years SNS24 calls;
- a thorough comparison of different Machine Learning models to select the most appropriate clinical pathway;
- a comprehensive analysis of the results including classification and execution time performance, per class error analysis, explainability of decisions and experts’ meta-evaluation.

The rest of the paper is organised as follows: in Section 2, the related work is discussed, and, to support the choices of the conducted experiments, the machine learning models that have been usually employed and evaluated in clinical decision support systems are pointed out; Section 3 presents the materials and methods: the data, the features used, the prediction techniques and how experiments were organized; Section 4 presents the results, first giving an overview of the main findings regarding the data analysis to help better understand the problem, then evaluating the prediction models; Section 5 looks deeper into the results reported, presenting an error analysis, discussing how decisions are made for some specific examples through the use of an explainability method, and providing a meta evaluation performed by domain specialists; and finally, conclusions are presented in Section 6.

2. Related Work

Machine Learning (ML) and Natural Language Processing (NLP) have recently shown important progress in applications such as clinical decision support systems by improving the quality of information processing from clinical narratives [4–9].

In the literature we find approaches to health care data that are based on unsupervised ML models. Funkner et al. [10], propose a data-driven prediction model of clinical pathways for patient staying at hospitals: patients are clustered according to their movements (clinical paths) in a hospital; then, a decision tree is used to explore the existing diversity of patients’ clinical pathways. In [11], Simulation Modeling and ML are used for the purpose of designing pathways and evaluating the return on investment, having a elderly hip-fracture care scheme as use case. Almeida et al. [12] exploit the historical component of the patient trajectory to improve the performance of clinical decision support systems; by automatically extracting information from patient medical notes they aim to assimilate detailed relevant patient information and provide recommendations during clinical treatments.

Our approach, on the other hand, considers supervised models for the task of text classification. In this type of task, different paradigms and algorithms are often considered [13], and there is an objective way of measuring systems accuracy.

Traditional approaches usually represent text as a bag-of-words, using a specific weighting scheme, with the most used one being TF-IDF (Term Frequency—Inverse Document Frequency); a possible generalization is using bags of n-grams. More recently, word embeddings [8] are being used to represent text, where each word is represented by a vector (of a specific size) instead of a number.

As commonly known, each particular problem and dataset may have a preferred model, with no optimal general algorithm choice [14]. In this way, there are many works comparing approaches and experiment designs. One example is Mascio et al.'s work [15], where alternative classification algorithms and various word representations are compared regarding applications involving clinical texts.

Support Vector Machines (SVM) and Deep Learning (DL) architectures usually present the best evaluation results [6,7,16] on evaluations performed on clinical decision support problems [17]. One known difference between these two approaches is that using ML methods (such as SVM) often requires feature engineering efforts, as these features are not learned automatically in the process. On the other hand, DL based methods have shown powerful feature learning capabilities. The literature presents many comparisons of these two types of approaches. Baker et al. [18] deal with text classification on a cancer dataset. They showed that a Convolution Neural Network (CNN) model, with fine-tuned hyper-parameters, initialisation and training process, achieved better performance than an SVM model based on engineered features.

Flores et al. [19] present extensive comparisons of alternative models, such as active learning, SVM, Naïve Bayes, and a BERT classifier applied to biomedical datasets. They found that the active learning approach reduced the number of training examples necessary for achieving the same performance of the other classifiers. Other experimental results [20] show that Bidirectional Encoder Representations from Transformers (BERT) models did not achieve better performance when compared to CNN and Hierarchical Self Attention Networks (HiSAN) models.

Besides making efficient predictions, it is very important to analyse the data and explain the predictions made by the algorithm, particularly in this area of application [21,22]. In a different way from what was found in the literature, this work presents a detailed analysis of the data, since it is useful to better understand the characteristics and the issues regarding the learned models. Data analysis is also important for the clinical team to find (new) insights from the phone calls received along the years and the way those calls are being handled.

Regarding the classification task, and following the tendencies seen in the related work, this work presents a comparative study of the two most evident methods found in the literature: SVM and CNN. Moreover, a detailed analysis of the results, explainability of the models, and a meta evaluation of errors are presented.

3. Materials and Methods

This section introduces the materials, including a characterisation of datasets, and the methods: the conducted data analysis, the prediction task and the experiments organisation.

3.1. Materials

The datasets used in these experiments correspond to the information collected from the phone calls of triage type received by the SNS24 line throughout different periods of time. These records include personal data (such as age, gender, encrypted primary care unit, county and district of the call, between others), call data (start and end date/time, initial intention, comments, contact reason, clinical pathway and final disposition, between others) and triage agent encrypted id and decisions.

The data, provided by SPMS (Serviços Partilhados do Ministério da Saúde), the public business entity which provides specific shared services in the health area to establishments and services of the Portuguese National Health Service (SNS), was anonymised and the study protocol was approved by the competent ethics committee.

3.1.1. Three-Month Data

The three-month data, available in the beginning of the project and used on previous studies, included a total of 269,663 records dated from January to March 2018. It was composed of 18 different attributes and 53 clinical pathways.

For building the dataset, clinical pathways with less than 50 instances were removed from the original data, resulting in a dataset with 269,654 instances and 51 clinical pathways for classification.

3.1.2. Three-Year Data

The three-year data, a superset of the three-month data, was made available later and comprised records dated from January 2017 to December 2019. It was composed of a total of 64 attributes and examples from 54 pathways from the 59 possible ones. A similar procedure was performed by removing pathways with less than 100 instances, resulting in a dataset of 53 different pathways with a total of 2,577,517 instances.

The proportion of observations per clinical pathway is diverse, ranging from 10.68% for “Cough” to 0.01% for “Heat-related problems”; 4 clinical pathways have proportions above 5%, and 25 have less than 1%. The possible clinical pathways and the corresponding number of calls (in each of the three years of study) are listed in Table A1 in Appendix A.

3.1.3. Features under Analysis

Besides “Contact Reason”, previous studies considered other available information (age, time and day of the week of the call, and comments) for building classification models but no better results were obtained. Building upon those results, the classification models will consider only the “Contact Reason” attribute and the corresponding clinical pathway (selected by the nurse). “Contact Reason” is a medium length text attribute, written in Portuguese by the technician who answered the call, containing the most relevant information about the patient’s condition; it is composed of a maximum of 25 words and the average number of words is 8.27. Table 1 presents a few examples of the “Contact Reason” attribute (the original Portuguese text is written in italics following the corresponding translation to English) and the corresponding clinical pathway (one for each of the 10 most common pathways).

Table 1. One example of the “Contact Reason” for each of the 10 most common clinical pathways.

Contact Reason	Clinical Pathway
<i>Rinorreia transparente, tosse produtiva e febre há 3d.</i> Transparent rhinorrhea, productive cough and fever 3d.	Cough
<i>Dor Abdominal e nauseas há 15 dias, agravamento hoje</i> Abdominal pain and nausea for 15 days, worse today	Nausea and vomiting pr.
<i>Dor supraumbilical há cerca de alguns dias</i> Supraumbilical pain for about a few days	Abdominal pain
<i>Congestão nasal há 24 h.</i> Nasal congestion for 24 h.	Oropharynx problem
<i>Foliculite purulenta face, tronco, braços e pernas desde 3ª feira</i> Purulent folliculitis face, corpo since Tuesday	Rash

Table 1. *Cont.*

Contact Reason	Clinical Pathway
<i>desde dia 21 tosse seca provoca o vomito, mialgias</i> since the 21st dry cough causes vomiting, myalgias	Flu syndrome
<i>Sangue nas fezes e diarreia ha 5 h</i> Blood in stool and diarrhea for 5 h	Diarrhea
<i>Lombalgias e hematuria há 72 h.</i> Low back pain and hematuria 72 h ago.	Urinary problem
<i>Episódios de febre e extremidades cianosadas há 18 h</i> Episodes of fever and cyanotic extremities for 18 h	Body temp. change pr.
<i>Dor na região costal algumas vezes</i> Pain in the rib area sometimes	Chest pain

The following attributes were further considered to characterize the population using the SNS24 service: gender, age, date and time of the call, district of the call, and final referral.

3.2. Methods

As mentioned before, the present work builds upon the 3 years data, including a detailed characterisation of the data and a comparison between SVM and CNN enlarged models including a comprehensive analysis of the model performance, errors per class, model explainability and a meta-evaluation made by an clinical expert.

3.2.1. Data Characterisation

Carefully analysing the data is useful not only to understand the profile of the calling citizens but also for the SPMS team to get new insights over the phone calls received along the years and the way those are being handled. Thus, it was agreed to analyse the following information: distribution of calls per clinical pathway, per gender, per relation of the caller with the citizen, per district, per age of citizen (general and for the 10 most frequent pathways), per hour of the day, day of week and month of year (general and for the the 10 most frequent pathways). Moreover, an analysis of the distribution of referrals per pathway was also done.

3.2.2. Pathway Prediction

SNS24 calls are received by nurses that, after an initial assessment (pre-triage) that enables the detection of emergent situations, follow predefined clinical pathways. The outcome of a pathway is a referral (transference to PIC, transference to INEM, clinical assessment in hospital emergency or at a primary health care unit and self-care) that can be changed by the nurse.

Self reported symptoms and signs, as well as relevant information provided on medical history are considered for the selection of the most appropriate clinical pathway. This choice is extremely relevant since it should have the discriminant capability for not sending to Hospitals Emergency situations of low clinical risk and, most importantly, ensure safety by identifying all situations that demand urgent medical contact.

Since, in the course of a call, nurses need to choose a specific clinical pathway, this problem can be seen as a multi-class classification task where the “Clinical pathway” is the class to be predicted.

The “Contact Reason” text was processed to get 2 different representations aiming to weight the importance of a word in the prediction task: the traditional TF-IDF representation (which determines the importance of a word in the text within the corpus) and a locally trained word embedding (word embeddings map the words into a low-dimensional continuous space encoding their semantic and syntactic information by assuming that words in similar contexts should have similar meanings [23]).

The ML algorithms chosen to build the prediction models were SVM and CNN. This selection was made considering the previous work developed over a 3 month data period. In [2] Random Forests, SVM with linear and RBF kernels SVM and Multinomial Naïve Bayes were tested with different text representations (TF-IDF with different n-grams and BERT [24] and Flair [25] embeddings); linear SVM with Flair embedding model reached the highest performance. In [3], a comparison between several DNN architectures (CNN, RNN and Transformers, namely BERTimbau Base BERT model for Portuguese [26], based architectures) was made using locally trained word embeddings with a 200-dimension vector; the transformers based architecture reached the best performance. The RBF kernel SVM algorithm and the CNN based architecture were chosen for this work because they presented the best compromise between model performance and the time needed to build the models.

Following the setup from previous experiments, a SVM with standard RBF kernel with default parameters using TF-IDF to represent text [2] and a CNN based architecture with locally trained word embeddings [3] were used. The CNN architecture used was a classical convolutional neural network for text classification described by Yoon Kim [27] and was composed of 4 hidden layers: convolutional, max-pooling, fully connected and dropout layers. The convolution layer's settings were as follows: filters = 128, kernel_size = 3, activation = 'relu'; the dense layer's settings were as follows: batch_size = 128, activation = 'relu'; the settings of the dropout layer [28] were as follows: dropout_rate = 0.3; and finally, the ADAM optimizer [29] was selected with patience = 5 and learning_rate = 10^{-4} .

Scikit-learn (v1.2) [30], TensorFlow (v2.2) [31], Pytorch (v1.8) [32] and Python (v3.8) [33] were the libraries used to build the ML models.

3.2.3. Experiment Organisation

Experiments were set up to assess the impact of larger datasets and older models in the model performance. For that, the 2019 data was kept for testing and 3 training sets were built: one with 2017 data, one with 2018 data and one with 2017-2018 data. Moreover, the 3 months models obtained in the previous studies [2,3] were also tested over the 2019 data.

The prediction performance was assessed using balanced accuracy (recall, sensitivity), precision (positive predictive value) and F1 measures; an analysis of the prediction time of each model was also made.

4. Results

This section presents and discusses the results obtained concerning the characterisation of the population that use the SNS24 service and the clinical pathway prediction models.

4.1. Data Characterisation: Main Findings

The origin of the call was studied and Figure 1 presents the number of calls per district per 100,000 inhabitants. It can be observed that the great majority of calls are made from the greater Lisbon area (Lisbon and Setubal districts) with a number of calls between 2000 and 2500 per 100,000 inhabitants, followed by the coastal districts north of Lisbon with values between 1500 and 2000 calls. In the mainland, 3 inland districts (Bragança, Guarda and Portalegre) use less the line, with a number of calls ranging 500 and 1000 per 100,000 inhabitants. The citizens living in the islands (Azores and Madeira autonomous regions) use the phone line even less; one possible explanation for this minor use of the SNS24 phone line is the existence of a similar service run by the regional health authorities.

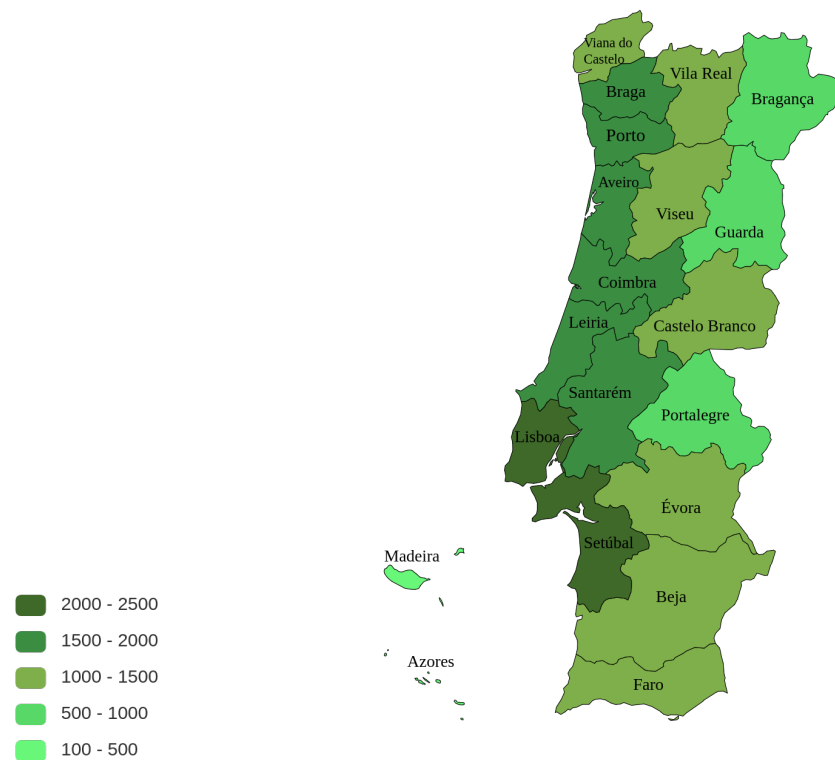


Figure 1. Total number of calls per district per 100,000 inhabitants (image generated using <https://paintmaps.com/>, accessed on 9 November 2022).

One important and needed data characterisation is the clinical pathway, being it the attribute the classification models will aim to predict (known as target class). The proportion of examples of each clinical pathway on the dataset is presented in Figure 2. As can be observed, the dataset is very unbalanced, which turns the learning of a classification model a more complex problem (the absolute numbers are presented in Table A1 on Appendix A).

The most frequent pathway is “Cough” with 10.68% of the calls, followed by “Nausea and vomiting problems”, “Abdominal pain”, “Oropharynx problem” and “Rash” with 7.44%, 6.26%, 5.08% and 4.86%, respectively. Moreover, the 23 most frequent pathways account for at least 80% of the calls, with the 7 most frequent ones accounting for around 42.4%; furthermore, there are 25 pathways for which the frequency is less than 1%.

Regarding the caller, Figure 3 shows that about half of the calls is to report a problem of the callers themselves, followed by calls from the mother of the patient (26%); father or husband of the patient account for 5% each, whereas caretakers account for 1% of the callers.

Concerning the citizen gender distribution, the majority is female representing 59% of the calls. For the age distribution it can be observed that the phone line is mostly used to attend patients in their first years (up to five years old), followed by young adults, with the number decreasing for the elderly; this trend is presented in Figure 4.

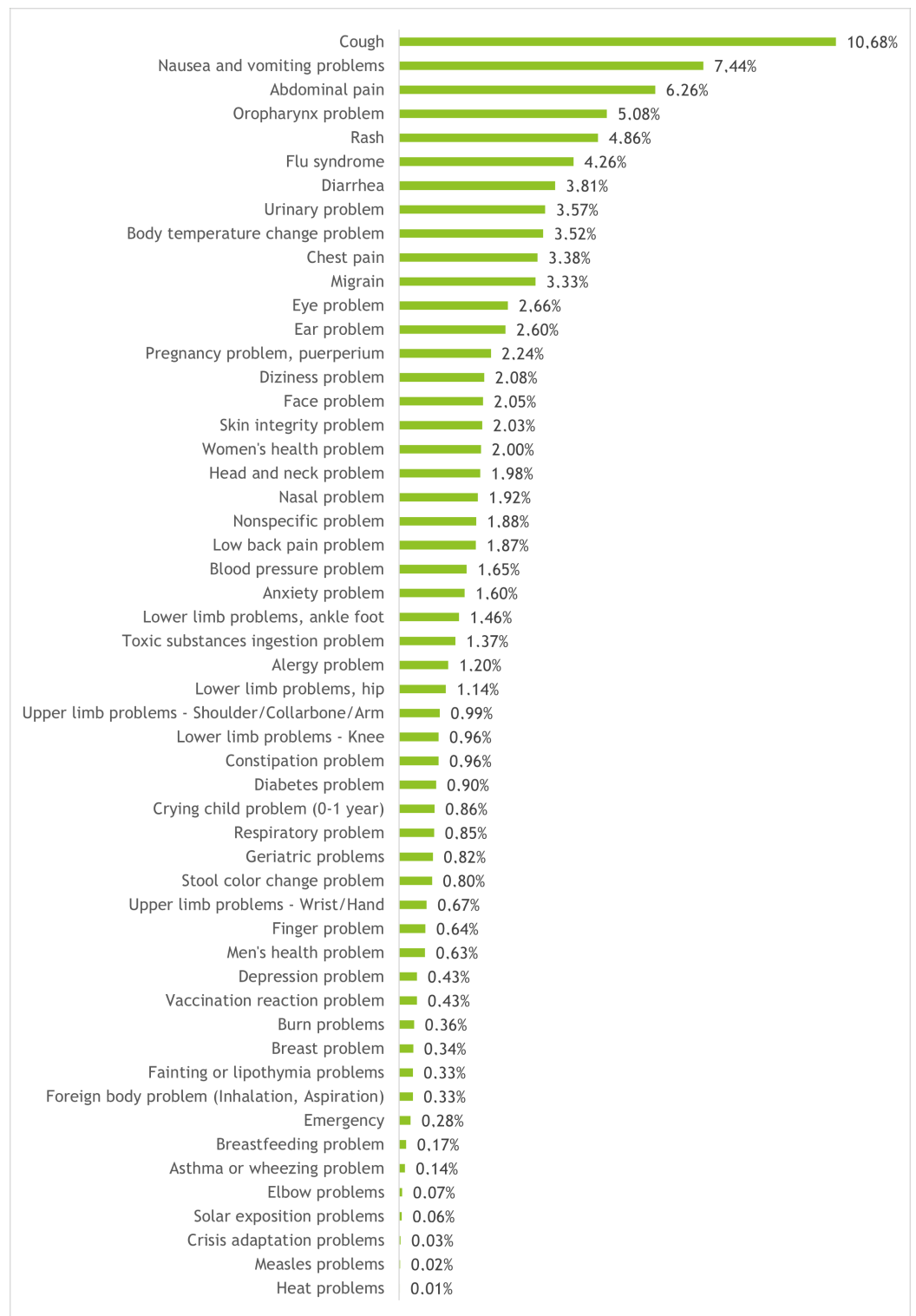


Figure 2. Distribution of calls per clinical pathway.

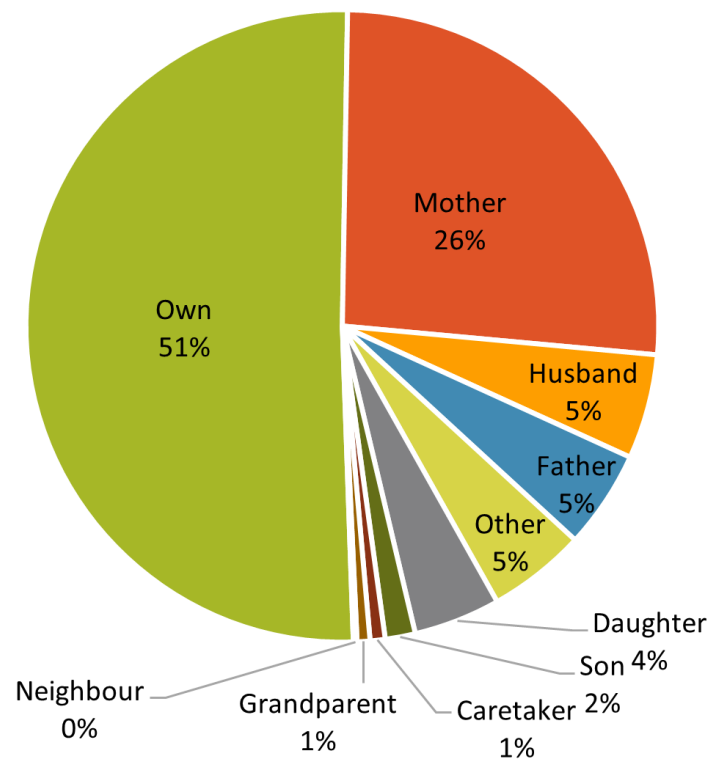


Figure 3. Distribution of calls by relation of the caller to the citizen.

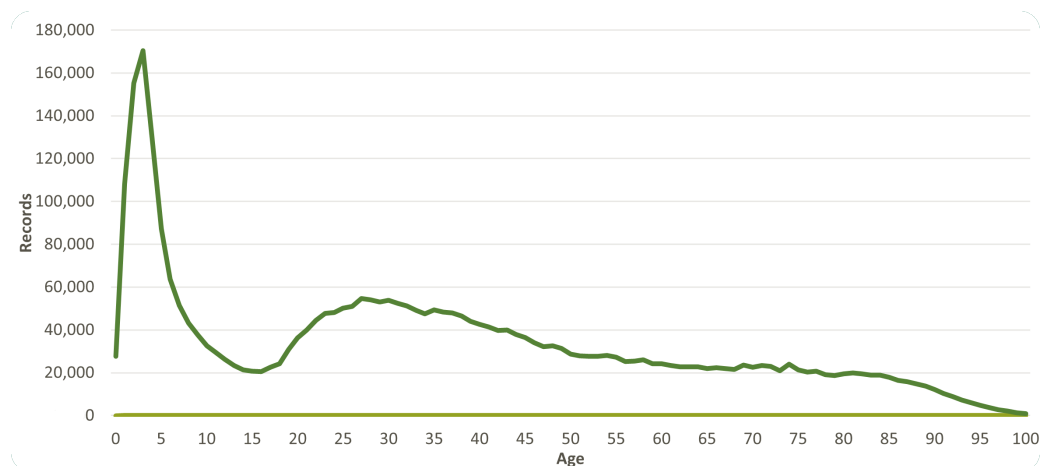


Figure 4. Distribution of calls per the age of the citizen.

A similar age distribution analysis was done for the 10 most common pathways; these were then empirically clustered by similar trends over patient age, having one cluster following the general trend and 3 others with some differences. The distribution for those clustered pathways can be observed in Figure 5. The first cluster (top-left) follows the general trend and includes four pathways: “Cough”, “Nausea and vomiting problems”, “Rash” and “Diarrhea”; the second cluster (top-right), shows the age distribution for the “Flu Syndrome”: younger patients are not so frequent, with an increased percentage for the adults between 30 and 40 years old. “Chest pain” and “Migraine” pathways represent the third cluster (bottom-left) and have the most dissimilar distribution when compared to the general age distribution. The final cluster (bottom-right) includes “Abdominal pain”, “Oropharynx problem” and “Urinary problem” pathways and follow, somehow in a smoother version, the general trend.



Figure 5. Distribution of calls per age for the 10 most frequent pathways (clustered by similarity).

A study of the distribution of calls over time, namely hour of the day, day of the week and month of the year was also made. Figure 6 presents the results (on top, the hour of the day; on bottom left and right, the day of the week and month of the year, respectively). It can be observed that dinner time (between 19:00 and 21:00) is the period summing around 20% of the calls; the value reaches its minimum at 5:00 and increases until 10:00, keeping more or less steady (with a small decrease) until 15:00, when it starts increasing again until 20:00.

In what concerns the day of the week, Monday and Friday are the weekdays with more and less calls, with 15.25% and 13.75%, respectively; while the values decrease between Monday and Friday, Saturday and Sunday receive 14.30% and 14.75% of the calls. The number of calls around the year ranges between less than 6% in September and 12% in January, with values around 10% from February to June and around 6% from July to November; December has a value of around 10.05%.

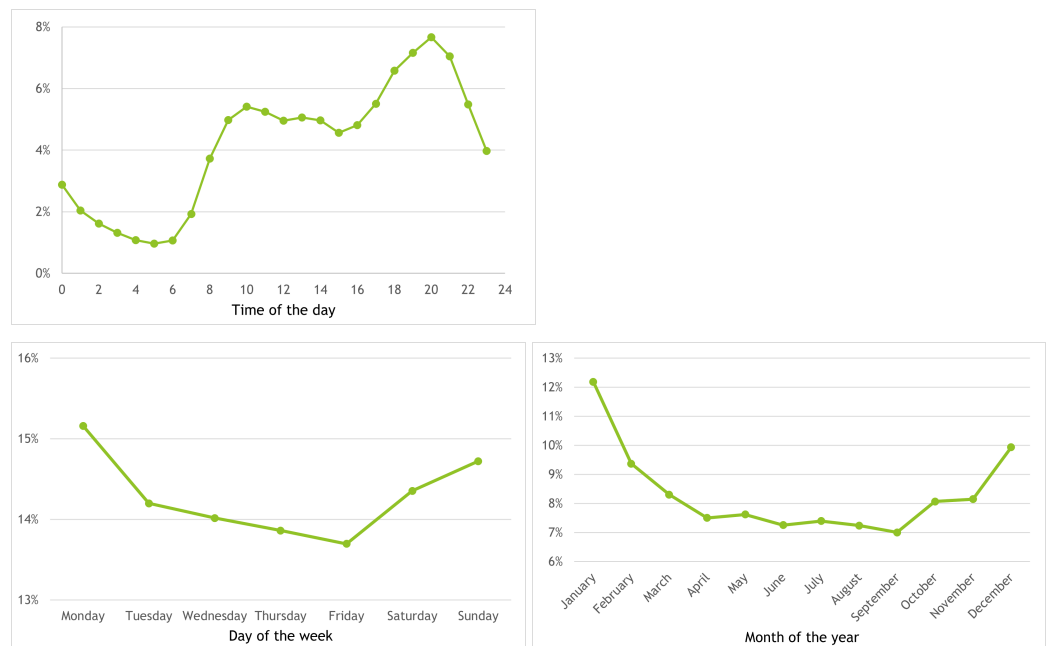


Figure 6. Distribution of calls per hour of the day (top), day of the week (bottom-left) and month of the year (bottom-right).

A study on the distribution of the calls per month per pathway for the 10 most common pathways was also made; these were empirically clustered by similar trends over time, having, again, one cluster following the general trend and 3 others with some differences. The distribution for those clustered pathways can be observed in Figure 7.

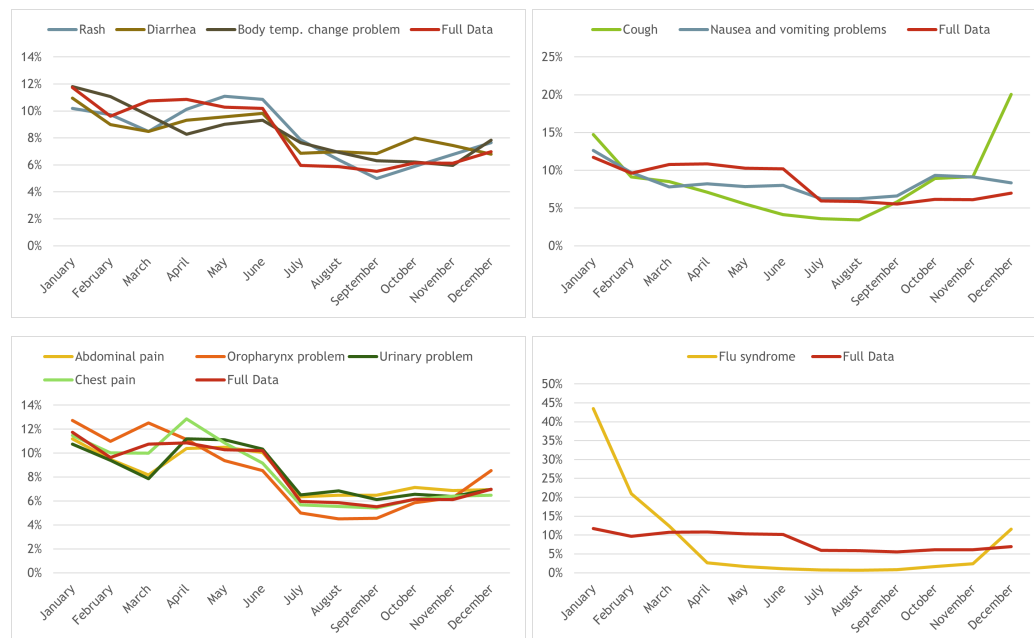


Figure 7. Distribution of calls per age for the 10 most frequent pathways (clustered by similarity over time).

The first cluster is composed by the “Rash”, “Diarrhea” and “Body temperature change problem” pathways and it follows, more or less, the general trend: decreases until March, raises until June, decreases in July, keeping steady until September and slowly increasing until December (except “Diarrhea” that decreases from October to December). “Cough” and “Nausea and vomiting problems” compose the second cluster and show a somehow similar behaviour over the year (top-right): the number of calls drops until August, then the first raises until December while the second has the highest value in October slowly decreasing until December.

The third cluster (bottom-left) is composed of four pathways: “Abdominal pain”, “Chest pain”, “Oropharynx problem” and “Urinary problem” and has the following trend: starting from January the values drop until March, raise until April, decreases until July and stays steady until December. Finally, “Flu syndrome” composes the fourth cluster (bottom-right): it has a huge peak in January (with almost 45% of the calls) and drops to 2.5% in April reaching almost 0% in August, starting to raise slowly until November and having a steep raise in December.

Finally, a study of the distributions of referrals (self-care, primary care unit, emergency room, INEM, others) per pathway was also made. Figure 8 presents the results for all pathways. “Chest pain”, “Blood pressure problem” and “Fainting or lipothymia problems” (not considering “Emergency”) are the pathways with higher percentage of INEM referral with values around 25% of the calls, while “Vaccination reaction problem”, “Body temperature change problem”, “Nasal problem”, “Flu syndrome” and “Diarrhea” (not considering “Asymptomatic contact tracing”) are the ones with higher percentage of self-care referral, with percentages between around 60% and 75%.

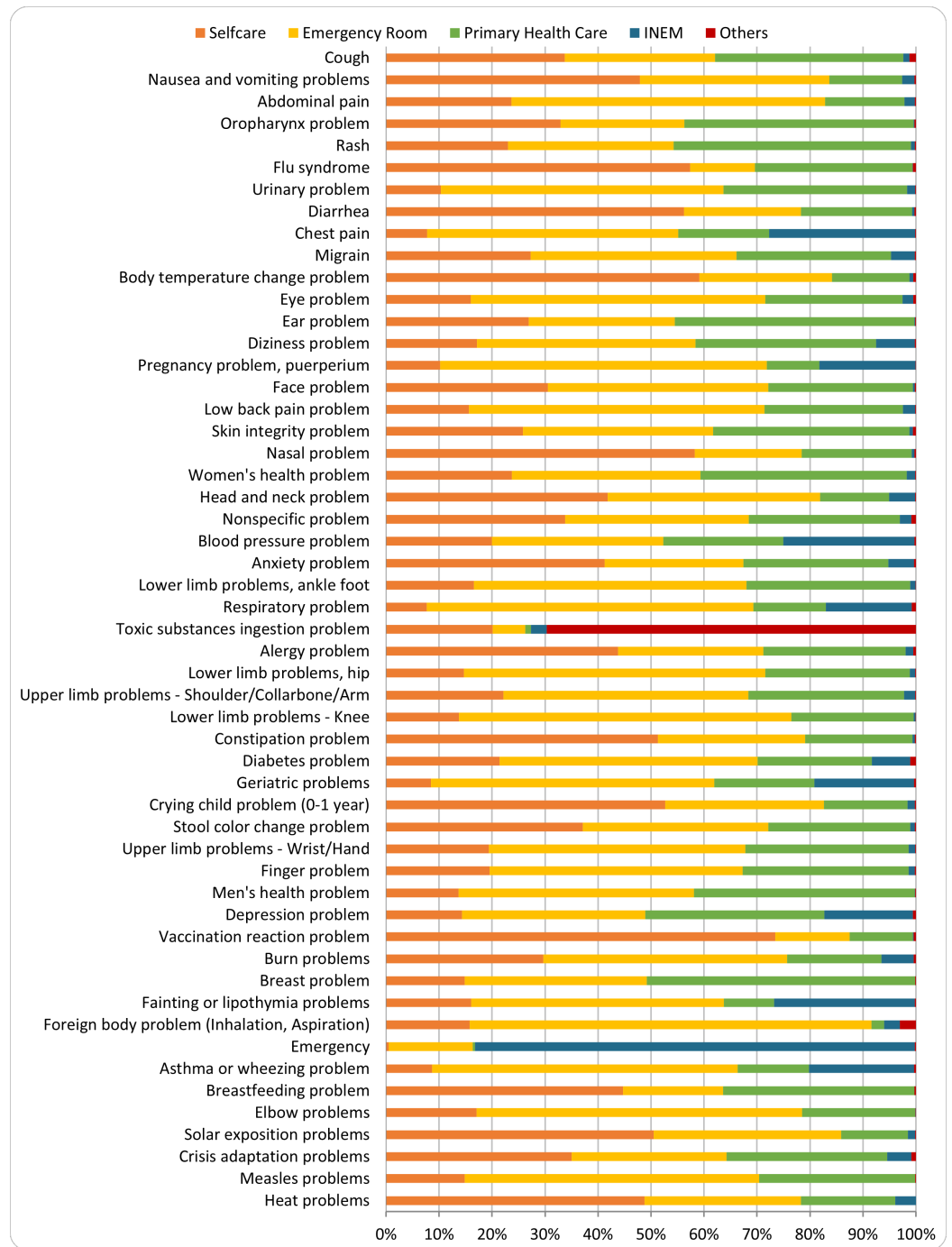


Figure 8. Distribution of referrals per pathway.

Figure 9 further presents the same analysis but for the 10 most common pathways, arranging them by the percentage of selfcare referral: “Body temperature change”, “Diarrhea” and “Flu syndrome” have percentages higher than 50%; “Cough”, “Oropharynx problems” and “Nausea and vomiting problems” have percentages higher than 30% and lower than 50%; “Abdominal pain” and “Rash” have a similar selfcare of around 23%; “Chest pain” and “Urinary problems” have selfcare percentages lower than 10%. On the other end, around 28% of “Chest pain” are referred to INEM and around 60% of “Abdominal pain” and 50% of “Urinary problems” and “Chest pain” calls are referred to the emergency room.

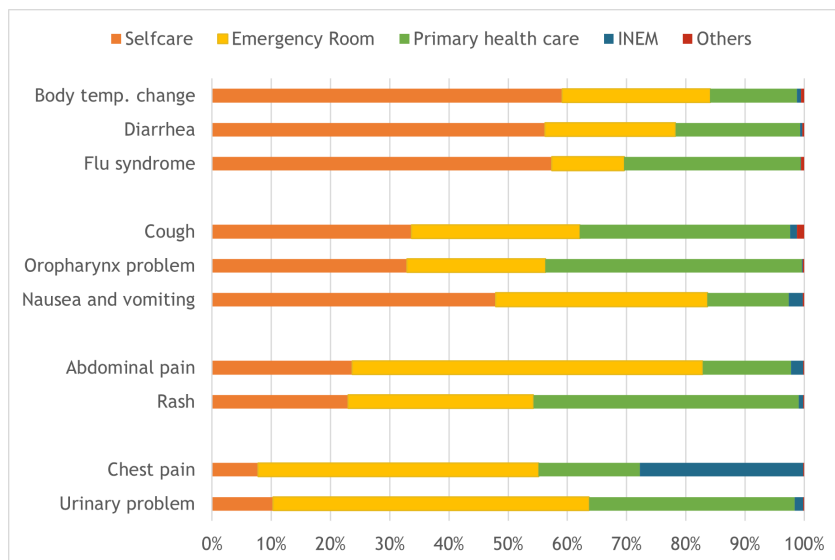


Figure 9. Distribution of referrals for the 10 most common pathways.

4.2. Evaluation of Prediction Models

As previously mentioned models were built to assess the impact of larger datasets and older models in the model performance. Table 2 presents the balanced accuracy of the models when considering the 1, 3 and 5 most probable clinical pathways (top-1, top-3, top-5) along with the average time to predict 1000 examples (over 10 runs). The 3 months models were built and assessed in previous works [2,3].

Table 2. Test set balanced accuracy (top 1, top 3 and top 5 most probable clinical pathways) and average prediction time (for 1000 examples) of SVM and CNN models with different training sets.

Training Set	Accuracy						Pred Time (ms)	
	SVM			CNN			SVM	CNN
	Top 1	Top 3	Top 5	Top 1	Top 3	Top 5		
3 months	0.743	0.927	0.959	0.631	0.775	0.804	14.82	12.31
2017	0.755	0.932	0.963	0.772	0.931	0.961	54.32	12.41
2018	0.766	0.938	0.968	0.780	0.937	0.964	54.57	12.50
2017–2018	0.768	0.940	0.969	0.782	0.940	0.967	92.48	14.19

Observing the top-1 accuracies, they range from 63.1% and 78.3% for CNN model with 3 months and 2 years data, respectively. It is curious to observe that the lowest and highest performance were obtained with a deep neural network architecture, which is in accordance with previous studies: deep architectures are able to achieve higher performances than other algorithms when a large amount of training data exists, overfitting easily when there is not enough data. We can also observe that SVM models with different sizes of data are more stable than CNN ones: the difference in accuracy between the 3 months and the 2017-2018 SVM and CNN data models is 2.5% and 15.1%, respectively.

It is also possible to conclude that adding more data consistently improves the models and that models built with more recent data are also better. For example, using the 2 previous years increments the accuracy by 0.2% when compared with a model built with only the previous year (2017–2018 vs. 2018); using a model built with data from the previous year increments the accuracy by 0.8% (CNN) to 1.1% (SVM) when compared to a model built with the data from 2 years apart (2018 vs. 2017).

When looking at the model performance considering the 3 most probable pathways (top-3), SVM is substantially better than CNN for the 3 months model (15.2% higher) presenting very similar accuracies for the other models (with SVM surpassing CNN in 2017

and 2018 models by 0.1% and having an equal accuracy for 2017-2018 model). This trend is also true considering the 5 most probable pathways: SVM is 15.5%, 0.2%, 0.4% and 0.2% higher than CNN for the 3 months, 2017, 2018 and 2017–2018 models.

Finally, observing the prediction time, and as expected, SVM takes more time to predict when models are built with more data, with CNN maintaining an almost fixed time independently of the size of the training set.

Table 3 presents the weighted precision, recall (values are repeated here as they are usually presented along with precision and F1) and F1 values over the test set (2019 data) of the 8 trained models (2 algorithms: SVM and CNN; 3 datasets: 3 months, 2017 data, 2018 data and 2017–2018 data).

Table 3. Test set precision, recall and F1 (weighted average) of SVM and CNN models with different training sets.

Training Set	SVM			CNN		
	Prec	Rec	F1	Prec	Rec	F1
3 months	0.754	0.743	0.745	0.657	0.632	0.642
2017	0.761	0.755	0.755	0.776	0.772	0.772
2018	0.768	0.766	0.765	0.782	0.780	0.779
2017–2018	0.771	0.768	0.768	0.785	0.783	0.782

Similarly to balanced accuracy (recall), the minimum and maximum precision and F1 values were obtained for the CNN model with 3 months and 2 years data, respectively; precision range between 65.7% and 78.5%, and F1 between 64.2% and 78.2%. We can also observe, when comparing CNN and SVM models, the same trend for these measures (as observed in accuracy): the performance difference between the 3 months and 2 years models varies between 1.7% (precision) and 2.3% (F1) for SVM while, for CNN, it varies between 12.8% (precision) and 14% (F1).

5. Error Analysis, Explainability and Meta-Evaluation

This section analyses, from different points of view, the predictions made by the developed ML models. Namely, it looks into the errors by class, exemplifies how a model makes its decisions and evaluates the similarity between the clinical pathways chosen initially by the nurses (and used as ground truth to build and evaluate the models), the models and a clinical expert.

5.1. Error Distribution by Class

Some disadvantageous facets of the problem at hands are: the unbalanced nature of the dataset, the considerable number of clinical pathways with pathways sharing common symptoms as referred by the healthcare experts (for instance, “Abdominal pain”, “Nausea and vomiting problems” and “Diarrhea”, or “Crying child problem” and “Body temperature change problem”) and the existence of very short text sequences describing the contact reason of the call. Upon these conditions, we consider that the results achieved are very satisfactory.

Table 4 shows the individual precision, recall and F1 values for the 10 most and least common pathways along with minimum, maximum, average and standard deviation values obtained for the SVM and CNN 2017-2018 models (Table A2, in Appendix A, presents the results for all pathways sorted, in descending order, by the pathway support). As can be observed, there is a huge performance difference between classes: F1 ranges between 0.000 and 0.931 for CNN and 0.013 and 0.919 for SVM (precision and recall have similar ranges). As expected, the average performance is higher for 10 most frequent pathways when compared with the 10 least frequent ones, with F1 equal or above 0.792 and equal or below 0.395, respectively. Moreover, standard deviation is much higher for the least common pathways; for example, and for the SVM model, the least common pathways

have a value of 0.260 while most common pathways present a value of 0.082 (similar trends are valid for the CNN model and for precision and recall).

Nonetheless, there are pathways, with similar support values, that seem to be much more difficult to classify than others: for example, while “Emergency” presents an F1 value of 0.140, “Foreign body problem” has an F1 of 0.679 (with support of 3201 and 3220, respectively); the case is similar, but with a smaller difference, between “Flu syndrome” and “Diarrhea”, with the F1 values of 0.609 and 0.810 and support of 41827 and 37202, respectively.

Table 4. Test set precision, recall and F1 of 2017–2018 SVM and CNN models for the 10 most and least frequent pathways.

Clinical Pathway	SVM			CNN			Sup
	Prec	Rec	F1	Prec	Rec	F1	
Cough	0.818	0.835	0.826	0.843	0.829	0.836	122,598
Nausea and vomiting	0.779	0.831	0.804	0.803	0.864	0.832	72,311
Abdominal pain	0.771	0.781	0.776	0.807	0.807	0.807	65,419
Oropharynx problem	0.769	0.712	0.740	0.793	0.741	0.766	56,327
Rash	0.875	0.906	0.890	0.888	0.897	0.892	47,756
Flu syndrome	0.586	0.634	0.609	0.562	0.663	0.608	41,827
Diarrhea	0.842	0.780	0.810	0.846	0.826	0.836	37,202
Urinary problem	0.897	0.873	0.885	0.897	0.880	0.888	37,757
Body temp change pr.	0.730	0.764	0.747	0.742	0.762	0.752	34,364
Chest pain	0.836	0.836	0.836	0.847	0.853	0.850	35,393
average (10 most freq)	0.790	0.795	0.792	0.803	0.812	0.807	
stdev (10 most freq)	0.088	0.080	0.082	0.096	0.072	0.083	
Fainting or lipothymia	0.647	0.688	0.667	0.701	0.720	0.710	3684
Foreign body problem	0.675	0.684	0.679	0.675	0.651	0.663	3220
Emergency	0.437	0.083	0.140	0.413	0.114	0.179	3201
Breastfeeding problem	0.595	0.505	0.547	0.597	0.457	0.518	1579
Asthma or wheezing	0.576	0.257	0.355	0.577	0.275	0.372	1519
Elbow problem	0.565	0.621	0.592	0.664	0.599	0.630	766
Solar exposure pr.	0.565	0.759	0.648	0.585	0.582	0.584	601
Crisis adaptation pr.	1.000	0.006	0.013	0.000	0.000	0.000	313
Measles problem	0.706	0.107	0.186	0.800	0.071	0.131	112
Heat problem	0.333	0.079	0.128	0.000	0.000	0.000	76
average (10 least freq)	0.610	0.379	0.395	0.501	0.347	0.379	
stdev (10 least freq)	0.176	0.300	0.260	0.282	0.287	0.280	
minimum (full data)	0.333	0.006	0.013	0.000	0.000	0.000	
maximum (full data)	1.000	0.922	0.919	0.929	0.932	0.931	
average (full data)	0.725	0.671	0.677	0.712	0.674	0.683	
stdev (full data)	0.136	0.213	0.199	0.185	0.224	0.212	

Figure 10 shows the F1 values vs. the support of the class over the test set for the 2017–2018 SVM and CNN models. It can be observed that, as expected, the class prediction performance tends to increase with a higher number of examples.

In a few cases, this tendency is interrupted (see Table A2, in Appendix A for the pathways names): “Geriatric problems” presents a sharp decrease when compared to pathways with similar support such as “Diabetes problem” or “Finger problem”, with F1 values of 0.460, 0.711, and 0.697 for the SVM model, respectively; the same occurs with “Nonspecific problem” when compared to “Blood pressure problem” or “Anxiety problem” (F1 values of 0.386, 0.780, and 0.735 for the SVM model, respectively) or “Flu syndrome” when compared to “Diarrhea” or “Urinary problem” (F1 values of 0.609, 0.810, and 0.885 for the SVM model, respectively). On the other hand, “Solar exposure problem” and “Elbow problem” present higher performance than pathways with few similar cases. This analysis

allows us to conclude that despite the number of examples per class, some pathways are more difficult to predict than others.

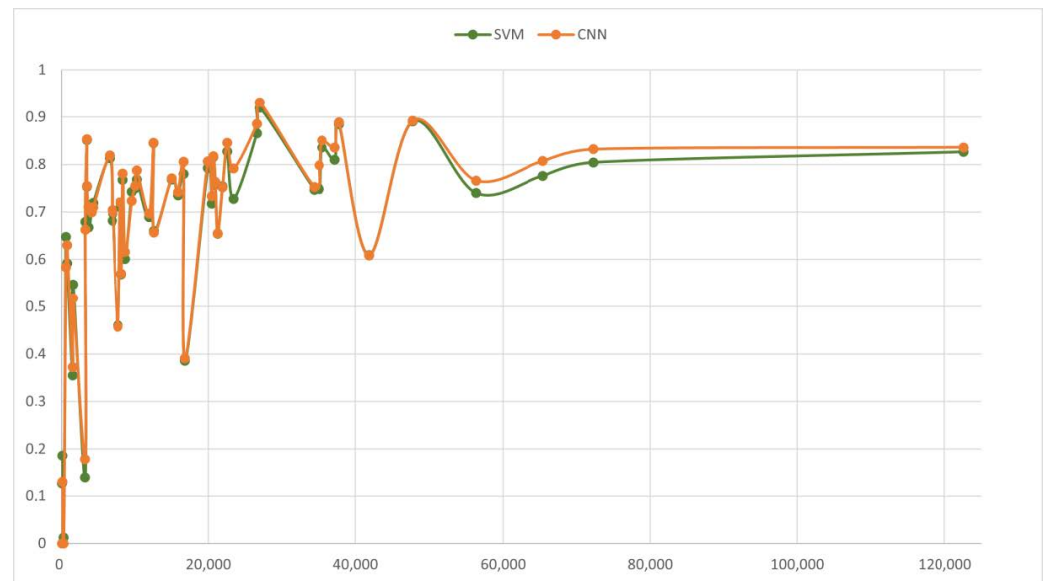


Figure 10. Per class performance: F1 value vs. support.

Table 5 presents the confusion matrix between “Cough” and “Flu syndrome”, two problems that share symptoms using the 2017–2018 models. It can be observed that, for the CNN model, the number of “Cough” examples misclassified as “Flu syndrome” surpasses the total for all other pathways (10,636 vs. 10,359), being also true for “Flu syndrome” examples classified as “Cough” (7157 vs. 6922); for the SVM model this trend is also true for “Flu syndrome” examples classified as “Cough” (7821 vs. 7484), but there are less “Cough” examples misclassified as “Flu syndrome” when compared to all other classes (9271 vs. 10,988). Moreover, while SVM classifies more “Flu syndrome” cases as “Cough” (7821 vs. 7157), CNN classifies more “Cough” cases as “Flu syndrome” (10,636 vs. 9271).

Table 5. Confusion matrix for “Cough” and “Flu syndrome” pathways using 2017–2018 models.

	SVM			CNN			
	Cough	Flu	Other	Cough	Flu	Other	
Cough	102,339	9271	10,988	Cough	101,603	10,636	10,359
Flu	7821	26,522	7484	Flu	7157	27,748	6922
Other	14,989	9432	991,171	Other	11,728	11,015	994,040

5.2. Explainability

It is well known that explaining a ML model decision is not a straightforward task (except in very simple models). Presently, there is a considerable effort in the community to overcome this issue; in that sense, several tools are becoming available.

Some of these tools intend to show the most relevant features contributing to the decision process, thus providing ways to explain black-box models. ELI5 (v0.13) [34] is one such tool and was used here to illustrate how we can explain the output produced by the SVM model. Figure 11 shows the ELI5 output for two examples corresponding to the following contact reasons:

1. *derrame ocular à direita há 2 semanas aprox. e prurido ocular desde ontem*
right eye effusion 2 weeks ago approx. and itchy eyes since yesterday

2. *Congestão nasal, tosse com expectoração não eficaz e secreções oculares amarelas há 24 horas*
Nasal congestion, coughing ineffective sputum and yellow eye secretions for 24 h

Both texts above were assigned by the nurse to the pathway “Ocular problem”; they were classified by the SVM model as “Ocular problem” in the first case and as “Cough” in the second (with “Ocular problem” being the third guess).

Figure 11 shows, for each example, the input words and corresponding weights for two pathways. For the first example, it shows the pathway chosen by the classifier the second most probable pathway; for the second example, it shows the pathways chosen by the classifier and by the nurse (the 3rd most probable one). The green colour signals words contributing in favour of the pathway; the red colour, against.

OCULAR PROBLEM		RASH		COUGH		OCULAR PROBLEM	
Weight	Feature	Weight	Feature	Weight	Feature	Weight	Feature
+11.727	ocular	+5.078	prurido	+9.108	tosse	+9.600	oculares
+3.259	derrame	+1.099	<BIAS>	+3.176	expectoração	+4.763	secreções
+2.585	direita	+0.406	ontem	+2.919	nasal	+2.568	amarelas
+2.299	prurido	+0.338	desde	+2.766	secreções	+0.712	horas
+0.760	há	+0.279	agravar	+1.363	<BIAS>	+0.634	há
+0.615	ontem	+0.271	há	+1.247	eficaz	+0.630	com
+0.537	<BIAS>	+0.112	hoje	+1.118	congestão	+0.618	<BIAS>
+0.485	desde	-0.337	semanas	+1.098	com	+0.374	congestão
+0.387	hoje	-0.340	direita	+1.077	não	+0.287	eficaz
-0.120	semanas	-0.484	ocular	+0.748	24	-0.334	expectoração
-0.847	agravar	-0.824	aprox	+0.532	horas	-0.338	24
-1.494	aprox	-1.215	derrame	+0.379	há	-0.350	não
				-1.052	amarelas	-0.729	nasal
				-1.494	oculares	-1.118	tosse

Figure 11. ELI5 output for two “Ocular problem” pathway examples: the left example was classified as “Ocular problem”; the right example was classified as “Cough”.

Looking at the examples and their explanations, it is noticeable that sometimes the most contributing words have similarities with the pathway name. Nonetheless, there are also cases in which different terms are chosen as important contributors to the decision.

For the first example (on the left), the most relevant features for classifying it as “Ocular problem” are: “ocular” (being related to the algorithm denomination), followed by “secretion” (*derrame*), “right” (*direita*) and “itching” (*prurido*); only 3 words count negatively. For the second most probable pathway, “Rash”, “itching” (*prurido*) is the word contributing the most to the class, with six words contributing negatively (being “ocular” one of them). ELI5 also states that the example has a belonging probability of 99.2% to “Ocular problem” and of 0.1% to “Rash”.

For the second example (on the right), classified as “Cough” by the ML model with a probability of 70.1%, “cough” (*tosse*) is the word with the highest positive weight, followed by “expectoration” (*expectoração*) and “nasal”. The pathway chosen by the nurse, “Ocular problem”, was the 3rd most probable one, with a probability of 4.3% (after “Flu syndrome” with a probability of 12.6%); for this pathway, “oculars” (*oculares*), “secretions” (*secreções*) and “yellow” (*amarelas*) have considerable positive weights, but there are 5 words contributing negatively, being “cough” (*tosse*) and “nasal” the most important ones. Although contributing negatively, their weights in this pathway are not as large (absolute value) as the weights contributing positively to pathway “Cough” (for example, -1.118 vs. $+9.108$ for word “cough”).

From a clinical perspective, the ML model leads to the correct assessment of the clinical symptoms priority when choosing the algorithm. In the specific case of the second example, the opinion of clinical expert is that the ocular involvement is most likely due to respiratory symptoms and consequently, the ML model suggestion of “Cough” is the most adequate one, despite the choice of the nurse.

When it comes to practice, this kind of analysis should be incorporated in the system with the explanations about the decisions being presented to the user whenever required.

Here, one such explainability tool and the explanation for two specific contact reasons were presented as illustration.

5.3. Meta-Evaluation

This meta evaluation aims at assessing the differences and similarities between the clinical pathways chosen initially by the nurses (and used as ground truth to build and evaluate the models), the ML models and a clinical expert.

For this study a total of 200 examples were chosen randomly, 10 for each of the 10 most frequent and the 10 least frequent clinical pathways. These examples were presented to a clinical expert and to the 2017-2018 SVM and CNN models. For some examples (totalling 14), and looking only to the “Contact Reason” information, the clinical expert was not able to decide the correct pathway, so these examples were not considered for the analysis. It is important to stress out that by having access only to the “Contact Reason” text, the clinical expert was not able to consider any other information the nurse had when attending the call. It is also important to have in mind that, although being specifically trained for choosing the most suitable clinical pathway, the choice, seen as ground truth by the ML algorithms, is made by different nurses (for the 3 years it totals 1888 different nurses) and with a diversity of clinical practice.

Table 6 resumes, per clinical pathway, the agreement between each pair of decisions (clinical expert, ground truth, SVM and CNN).

Table 6. Agreement between experienced analyst (e), ground truth (g) and SVM (s) and CNN (c) models.

Clinical Pathway	Total	e-g	e-s	e-c	g-s	g-c	s-c
Cough	10	2	4	5	5	5	9
Nausea and vomiting problem	9	2	4	3	3	3	8
Abdominal pain	10	5	5	6	6	7	9
Oropharynx problem	9	4	6	7	3	4	9
Rash	8	1	2	3	4	4	7
Flu syndrome	10	3	7	8	4	3	8
Diarrhea	9	1	2	3	4	4	9
Urinary problem	10	7	8	7	9	10	9
Body temperature change problem	10	3	7	7	5	3	7
Chest pain	9	5	7	8	7	6	9
Sub-total (10 most frequent)	94	33	52	57	50	49	84
Fainting or lipothymia	10	3	6	5	5	4	7
Foreign body problem	10	8	6	5	8	7	8
Emergency	7	3	2	2	2	0	6
Breastfeeding problem	10	0	6	7	3	2	9
Asthma or wheezing problem	9	0	6	6	0	0	8
Elbow problem	10	4	2	3	2	3	9
Solar exposure problem	8	6	5	6	4	5	8
Crisis adaptation problem	8	0	6	6	0	0	9
Measles problem	10	0	6	6	0	0	10
Heat problem	10	1	7	6	0	0	8
Sub-total (10 least frequent)	92	25	52	52	24	21	82
TOTAL	186	58	104	109	74	70	166

Observing the total number of agreements one can conclude that, while the agreement of the ML models is the highest (166), the clinical expert agrees the less with the pathway chosen by the SNS24 working nurses (58). Moreover, the number of agreements between both ML models and the nurses or the clinical expert is similar (74 and 70 with the nurses; 104 and 109 with the clinical expert, for SVM and CNN, respectively) but they are higher with the clinical expert by 30 cases or more.

Looking at the agreements with the clinical expert (columns e-g, e-s, e-c), a conclusion that stands out is that the agreement with the ML models (columns e-s, e-c) is much higher

when compared with the agreement with the SNS24 working nurses (column e-g); this is true for the least and for the most frequent pathways, with the number of agreements more than doubling for the least frequent pathways (52 for ML models vs. 25 for the nurses). Moreover, the agreement with the ML models is similar for the most and least frequent pathways (between 52 and 57). On the other hand, the agreement with the nurses is higher for the most frequent pathways when compared to the least frequent ones (33 vs. 25).

Although not as important, it is interesting to observe the agreements of the working nurses with the ML models (columns s-g, c-g). Even if the number of agreements is similar for both ML models, it more than doubles for the most frequent pathways when compared to the least frequent ones (50 and 49 vs. 24 and 21 for SVM and CNN, respectively).

Similarly to the agreement with the clinical expert, the agreement between the two ML models (column s-c) is similar for the most and least common pathways (84 vs. 82).

Looking at individual pathways, it seems that some are easier to distinguish than others. From a total of 10 examples tagged by the nurses as “Urinary problem”, the clinical expert agreed on 7 and the CNN on 10; for the 10 examples tagged as “Cough”, the clinical expert agreed on 2 and both ML models agreed on 5. A similar conclusion can be drawn from the set of the least common pathways: “Foreign body problem” seems to be a much easier pathway to be detected (7 and 8 agreements with the ML models and clinical expert) than the “Elbow problem” (2 to 4 agreements with the ML models and clinical expert).

Finally, it is also interesting to observe that, for the 14 examples where the clinical expert was not able to choose a pathway, SVM and CNN models always disagreed with the ground truth but agreed with each other in all but one example. Following, there are two examples where the ML models although agreeing with each other, disagreed with the working nurses:

1. *Palpitações, hiperventilação há 1 hora*
Palpitations, hyperventilation 1 h ago
2. *refere prurido anal há 1 hora*
reports anal itching 1 h ago

For first example, nurses chose “Chest pain” and the ML models predicted as being an “Anxiety problem”; for second example, nurses chose “Rash” pathway, the SVM model predicted as being a “Stool color change problem” and the CNN as being a “Skin integrity problem”.

These findings stress the importance of using such a decision support system, since it can help the attending nurses, when in doubt, to choose a clinical pathway.

6. Conclusions and Future Work

This work presents the analysis of a dataset containing calls received by the Portuguese National Health Contact Center and a comparison of different Machine Learning approaches to the task of classifying these calls according to their contact reason.

To help further understand the problem, the data was characterised, with its distribution regarding the classes of health problems and the customer gender and age and living district being reported. The service seems to have a great impact in child care and young adults, mainly on coastal districts.

The contact reason of the reported situation serves to help in selecting the pathway algorithm to be followed by the attendant to reach the appropriate referral between self-care, clinical assessment at primary care unit or at a hospital emergency if not transferred to the INEM. The pathway is to be chosen from 59 options so Artificial Intelligence techniques are applied to accelerate the finding of the best choice.

A comparative analysis among the best options according to the literature was presented and, on the basis of three years data, different configurations of training sets were tested. The conclusion is that date proximity along with larger training data are relevant for producing models with better performance.

The unbalanced nature of the classes normally impacts the generation of the models; this was confirmed for this problem when analysing performance and frequency. Another

issue found in the study is the great similarity among symptoms of some pathways, such as “Cough” and “Flu syndrome”, having a clear negative impact in the performance due to the difficulty to generalise the fine distinctions among these situations.

Moreover, it was shown that current techniques aiming at explaining ML models decisions may be applied by providing users with an explanation of the system decision, an essential feature for AI applications in the health domain.

It is also important to note that the ground truth was considered the right decision, assuming that the attendant is always correct regarding the chosen algorithm, a situation that was questioned during the process of meta evaluation. The presented error analysis enables to stress the importance of having a decision support system to help the attending nurses decide the best clinical pathway when in doubt.

Regarding the comparison of different classification methods, the best results were obtained with CNN, with an F1 of 0.782 against a maximum of 0.768 obtained with SVM. It is known, however, that the computational cost for training DNNs are much higher than the cost of using SVM. Moreover, the experiments showed that, although both ML algorithms give an answer within the reasonable time for having the system running in real time, the SVM model classification time is higher especially for bigger models. There are issues regarding these choices and their associated carbon footprint [35,36]. No specific carbon footprint measurements were done in this study but it is known that there is a large difference in resources consumption between these two approaches. Considering that the application of these solutions on a daily practice will require a periodical re-training, that is a relevant factor indicating the one year SVM model as a fair alternative, if a 1.4% reduction in performance (accuracy, precision, recall and F1) is not crucial.

As future work we intend to investigate other ML algorithms, namely a boosting ensemble approach like Gradient Tree Boosting [37], since these models generally provide top accuracies if the parameters are set correctly, and update the models with the calls from 2020 and 2021 when new clinical pathways related to COVID were designed.

On a different line of research, we intend to study the consistency between the clinical pathway final referrals and the follow-up screening at the hospital emergency or primary health care unit. Moreover, and since the number and practice of nurses attending the calls is very diverse, we intend to investigate trends in the nurses selection of clinical pathways.

Author Contributions: Conceptualization, T.G. and P.Q.; methodology, T.G. and R.V. (Renata Vieira); software, R.V. (Rute Veladas) and H.Y.; validation, P.I., A.R.P., N.F. and C.G.; investigation, R.V. (Renata Vieira); resources, C.S.P., J.O., M.C.F. and J.M.; data curation, T.G., H.Y., R.V. (Rute Veladas), C.S.P., J.O., A.R.P., N.F. and C.G.; writing, T.G. and R.V. (Renata Vieira); visualization, T.G. and P.I.; supervision, T.G.; project administration, T.G.; funding acquisition, P.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This research work was funded by FCT—Fundação para a Ciência e a Tecnologia, I.P, within the project SNS24.Scout.IA—Aplicação de Metodologias de Inteligência Artificial e Processamento de Linguagem Natural no Serviço de Triagem, Aconselhamento e Encaminhamento do SNS24 (ref. DSAIPA/AI/0040/2019).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Ethics Committee of University of Évora (Documento 20044, 14 April 2020).

Informed Consent Statement: Patient consent was waived due to the explicit acceptance for registry and record of calls by the users of the SNS24 line and the full anonymisation of the records.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank Pedro Salgueiro for the help provided in setting up the experimental environment.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area Under the Learning Curve
BERT	Bidirectional Encoder Representations from Transformers
CNN	Convolution Neural Network
DGS	Directorate General of Health
DL	Deep Learning
DNN	Deep Neural Networks
HiSAN	Hierarchical Self-Attention Network
INEM	National Medical Emergency Institute
ML	Machine Learning
NLP	Natural Language Processing
NN	Neural Network
PIC	Poison Information Center
SNS	Serviço Nacional de Saúde (Portuguese National Health Service)
SNS24	Portuguese National Health Line
SPMS	Serviços Partilhados do Ministério da Saúde
SVM	Support-Vector Machines

Appendix A. Clinical Pathways and Performance of 2017–2018 Models

This Appendix presents details the data made available for this work, namely the distribution of examples per clinical pathway (Table A1, along with the performance results per pathway for 2 years models (Table A2).

Table A1. Total and per year clinical pathway distribution.

Clinical Pathway	3 Years		2017		2018		2019	
	abs.	%	abs.	%	abs.	%	abs.	%
Cough	275,291	10.68	62,986	9.22	89,707	10.43	122,598	11.86
Nausea and vomiting problem	191,716	7.44	60,116	8.80	59,289	6.89	72,311	6.99
Abdominal pain	161,461	6.26	43,430	6.36	52,612	6.12	65,419	6.33
Oropharynx problem	130,999	5.08	32,850	4.81	41,822	4.86	56,327	5.45
Rash	125,142	4.86	35,691	5.22	41,695	4.85	47,756	4.62
Flu syndrome	109,880	4.26	24,486	3.58	43,567	5.07	41,827	4.04
Diarrhea	98,092	3.81	30,020	4.39	30,870	3.59	37,202	3.60
Urinary problem	91,910	3.57	23,457	3.43	30,696	3.57	37,757	3.65
Body temperature change problem	90,630	3.52	25,335	3.71	30,931	3.60	34,364	3.32
Chest pain	87,098	3.38	22,587	3.31	29,118	3.39	35,393	3.42
Migrain	85,846	3.33	22,799	3.34	28,008	3.26	35,039	3.39
Eye problem	68,546	2.66	18,272	2.67	23,313	2.71	26,961	2.61
Ear problem	66,928	2.60	18,322	2.68	22,058	2.56	26,548	2.57
Pregnancy problem, puerperium	57,781	2.24	15,932	2.33	19,942	2.32	21,907	2.12
Dizziness problem	53,564	2.08	10,848	1.59	19,306	2.24	23,410	2.26
Face problem	52,851	2.05	14,356	2.10	17,555	2.04	20,940	2.03
Skin integrity problem	52,327	2.03	13,396	1.96	17,689	2.06	21,242	2.05
Women's health problem	51,568	2.00	13,635	2.00	17,251	2.01	20,682	2.00
Head and neck problem	50,916	1.98	13,844	2.03	17,142	1.99	19,930	1.93
Nasal problem	49,464	1.92	12,430	1.82	16,572	1.93	20,462	1.98
Nonspecific problem	48,585	1.88	16,876	2.47	14,881	1.73	16,828	1.63

Table A1. *Cont.*

Clinical Pathway	3 Years		2017		2018		2019	
	abs.	%	abs.	%	abs.	%	abs.	%
Low back pain problem	48,222	1.87	8766	1.28	16,930	1.97	22,526	2.18
Blood pressure problem	42,491	1.65	11,075	1.62	14,782	1.72	16,634	1.61
Anxiety problem	41,300	1.60	11,321	1.66	14,086	1.64	15,893	1.54
Lower limb—ankle foot	37,583	1.46	10,004	1.46	12,556	1.46	15,023	1.45
Toxic sub ingestion problem	35,355	1.37	10,549	1.54	12,221	1.42	12,585	1.22
Allergy problem	31,034	1.20	7976	1.17	10,435	1.21	12,623	1.22
Lower limb—hip	29,300	1.14	7968	1.17	9350	1.09	11,982	1.16
Upper limb—shoulder collarbone arm	25,469	0.99	6779	0.99	8380	0.97	10,310	1.00
Constipation problem	24,851	0.96	8181	1.20	8260	0.96	8410	0.81
Lower limb—knee	24,700	0.96	6327	0.93	8289	0.96	10,084	0.98
Diabetes problem	23,157	0.90	7139	1.04	7982	0.93	8036	0.78
Crying child problem (0–1 year)	22,068	0.86	6545	0.96	7332	0.85	8191	0.79
Respiratory problem	21,821	0.85	5585	0.82	7528	0.88	8708	0.84
Geriatric problem	21,241	0.82	6688	0.98	6879	0.80	7674	0.74
Stool color change problem	20,723	0.80	3359	0.49	7727	0.90	9637	0.93
Upper limb—wrist hand	17,374	0.67	4665	0.68	5783	0.67	6926	0.67
Finger problem	16,599	0.64	4020	0.59	5593	0.65	6986	0.68
Men’s health problem	16,272	0.63	4258	0.62	5435	0.63	6579	0.64
Vaccination reaction problem	11,073	0.43	3234	0.47	3531	0.41	4308	0.42
Depression problem	10,980	0.43	3315	0.49	3588	0.42	4077	0.39
Burn problem	9287	0.36	2682	0.39	3164	0.37	3441	0.33
Breast problem	8775	0.34	2338	0.34	2980	0.35	3457	0.33
Fainting or lipothymia problem	8500	0.33	1686	0.25	3130	0.36	3684	0.36
Foreign body problem	8479	0.33	2394	0.35	2865	0.33	3220	0.31
Emergency	7272	0.28	1236	0.18	2835	0.33	3201	0.31
Breastfeeding problem	4436	0.17	1323	0.19	1534	0.18	1579	0.15
Asthma or wheezing problem	3618	0.14	904	0.13	1195	0.14	1519	0.15
Elbow problem	1819	0.07	445	0.07	608	0.07	766	0.07
Solar exposure problem	1550	0.06	447	0.07	502	0.06	601	0.06
Crisis adaptation problem	663	0.03	124	0.02	226	0.03	313	0.03
Measles problem	576	0.02	257	0.04	207	0.02	112	0.01
Heat problem	334	0.01	102	0.01	156	0.02	76	0.01
TOTAL	2,577,517	–	683,360	–	860,093	–	1,034,064	–

Table A2. Performance values (precision, recall and F1) for the 2017–2018 SVM and CNN models.

Clinical Pathway	SVM			CNN			Sup
	Prec	Rec	F1	Prec	Rec	F1	
Cough	0.818	0.835	0.826	0.843	0.829	0.836	122,598
Nausea and vomiting problem	0.779	0.831	0.804	0.803	0.864	0.832	72,311
Abdominal pain	0.771	0.781	0.776	0.807	0.807	0.807	65,419
Oropharynx problem	0.769	0.712	0.740	0.793	0.741	0.766	56,327
Rash	0.875	0.906	0.890	0.888	0.897	0.892	47,756
Flu syndrome	0.586	0.634	0.609	0.562	0.663	0.608	41,827
Diarrhea	0.842	0.780	0.810	0.846	0.826	0.836	37,202
Urinary problem	0.897	0.873	0.885	0.897	0.880	0.888	37,757
Body temperature change problem	0.730	0.764	0.747	0.742	0.762	0.752	34,364
Chest pain	0.836	0.836	0.836	0.847	0.853	0.850	35,393
Migrain	0.735	0.761	0.748	0.801	0.796	0.798	35,039
Eye problem	0.916	0.922	0.919	0.929	0.932	0.931	26,961
Ear problem	0.889	0.844	0.866	0.905	0.867	0.886	26,548
Dizziness problem	0.728	0.727	0.727	0.777	0.807	0.791	23,410

Table A2. Cont.

Clinical Pathway	SVM			CNN			Sup
	Prec	Rec	F1	Prec	Rec	F1	
Low back pain problem	0.843	0.812	0.827	0.849	0.842	0.846	22,526
Pregnancy problem, puerperium	0.827	0.693	0.754	0.811	0.700	0.752	21,907
Skin integrity problem	0.665	0.645	0.655	0.656	0.654	0.655	21,242
Face problem	0.736	0.780	0.757	0.725	0.804	0.763	20,940
Women's health problem	0.796	0.836	0.816	0.804	0.831	0.817	20,682
Nasal problem	0.706	0.730	0.717	0.710	0.757	0.733	20,462
Head and neck problem	0.775	0.811	0.793	0.788	0.826	0.806	19,930
Nonspecific problem	0.346	0.437	0.386	0.375	0.408	0.391	16,828
Blood pressure problem	0.788	0.773	0.780	0.834	0.780	0.806	16,634
Anxiety problem	0.761	0.711	0.735	0.738	0.745	0.742	15,893
Lower limb—ankle foot	0.777	0.760	0.768	0.755	0.787	0.771	15,023
Allergy problem	0.766	0.580	0.660	0.769	0.573	0.657	12,623
Toxic sub ingestion problem	0.836	0.852	0.844	0.842	0.849	0.846	12,585
Lower limb—hip	0.648	0.735	0.689	0.654	0.745	0.696	11,982
Upper limb—shoulder collarbone arm	0.738	0.801	0.768	0.759	0.819	0.788	10,310
Lower limb—knee	0.819	0.692	0.750	0.830	0.692	0.755	10,084
Stool color change problem	0.731	0.754	0.742	0.745	0.703	0.723	9637
Respiratory problem	0.571	0.633	0.601	0.582	0.653	0.616	8708
Constipation problem	0.757	0.779	0.768	0.735	0.832	0.780	8410
Crying child problem (0–1 Year)	0.671	0.493	0.568	0.667	0.498	0.570	8191
Diabetes problem	0.845	0.613	0.711	0.821	0.641	0.720	8036
Geriatric problem	0.447	0.474	0.460	0.453	0.462	0.457	7674
Finger problem	0.664	0.733	0.697	0.689	0.711	0.700	6986
Upper limb—wrist hand	0.724	0.644	0.682	0.757	0.658	0.704	6926
Men's health problem	0.831	0.797	0.813	0.845	0.794	0.819	6579
Vaccination reaction problem	0.743	0.697	0.719	0.747	0.676	0.710	4308
Depression problem	0.767	0.647	0.702	0.783	0.632	0.700	4077
Breast problem	0.702	0.812	0.753	0.714	0.800	0.755	3457
Burn problems	0.862	0.840	0.851	0.853	0.853	0.853	3441
Fainting or lipothymia problem	0.647	0.688	0.667	0.701	0.720	0.710	3684
Foreign body problem	0.675	0.684	0.679	0.675	0.651	0.663	3220
Emergency	0.437	0.083	0.140	0.413	0.114	0.179	3201
Breastfeeding problem	0.595	0.505	0.547	0.597	0.457	0.518	1579
Asthma or wheezing problem	0.576	0.257	0.355	0.577	0.275	0.372	1519
Elbow problem	0.565	0.621	0.592	0.664	0.599	0.630	766
Solar exposure problem	0.565	0.759	0.648	0.585	0.582	0.584	601
Crisis adaptation problem	1.000	0.006	0.013	0.000	0.000	0.000	313
Measles problems	0.706	0.107	0.186	0.800	0.071	0.131	112
Heat problems	0.333	0.079	0.128	0.000	0.000	0.000	76
minimum	0.333	0.006	0.013	0.000	0.000	0.000	
maximum	1.000	0.922	0.919	0.929	0.932	0.931	
average	0.725	0.671	0.677	0.712	0.674	0.683	
stdev	0.136	0.213	0.199	0.185	0.224	0.212	

References

1. Mackway-Jones, K.; Marsden, J.; Windle, J. *Emergency Triage: Manchester Triage Group*; John Wiley & Sons: New York, NY, USA, 2013.
2. Veladas, R.; Yang, H.; Quaresma, P.; Gonçalves, T.; Vieira, R.; Sousa Pinto, C.; Martins, J.P.; Oliveira, J.; Cortes Ferreira, M. Aiding Clinical Triage with Text Classification. In *Proceedings of the EPIA Conference on Artificial Intelligence*; Springer: New York, NY, USA, 2021; pp. 83–96.
3. Yang, H.; Gonçalves, T.; Quaresma, P.; Vieira, R.; Veladas, R.; Pinto, C.S.; Oliveira, J.; Ferreira, M.C.; Morais, J.; Pereira, A.R.; et al. Clinical Trial Classification of SNS24 Calls with Neural Networks. *Future Internet* **2022**, *14*, 130. [[CrossRef](#)]
4. 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*, 155–166. [[CrossRef](#)] [[PubMed](#)]

5. Marafino, B.J.; Boscardin, W.J.; Dudley, R.A. Efficient and sparse feature selection for biomedical text classification via the elastic net: Application to ICU risk stratification from nursing notes. *J. Biomed. Inform.* **2015**, *54*, 114–120. [[CrossRef](#)] [[PubMed](#)]
6. Mujtaba, G.; Shuib, L.; Idris, N.; Hoo, W.L.; Raj, R.G.; Khawaja, K.; Shaikh, K.; Nweke, H.F. Clinical text classification research trends: Systematic literature review and open issues. *Expert Syst. Appl.* **2019**, *116*, 494–520. [[CrossRef](#)]
7. Shickel, B.; Tighe, P.J.; Bihorac, A.; Rashidi, P. Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE J. Biomed. Health Inform.* **2017**, *22*, 1589–1604. [[CrossRef](#)] [[PubMed](#)]
8. Shao, Y.; Taylor, S.; Marshall, N.; Morioka, C.; Zeng-Treitler, Q. Clinical text classification with word embedding features vs. bag-of-words features. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 2874–2878.
9. Mustafa, A.; Rahimi Azghadi, M. Automated Machine Learning for Healthcare and Clinical Notes Analysis. *Computers* **2021**, *10*, 24. [[CrossRef](#)]
10. Funkner, A.A.; Yakovlev, A.N.; Kovalchuk, S.V. Data-driven modeling of clinical pathways using electronic health records. *Procedia Comput. Sci.* **2017**, *121*, 835–842.
11. Elbattah, M.; Molloy, O.; Zeigler, B.P. Designing Care Pathways Using Simulation Modeling and Machine Learning. In Proceedings of the 2018 Winter Simulation Conference (WSC), Gothenburg, Sweden, 9–12 December 2018; pp. 1452–1463. [[CrossRef](#)]
12. Almeida, J.R.; Silva, J.F.; Sierra, A.P.; Matos, S.; Oliveira, J.L. Leveraging Clinical Notes for Enhancing Decision-Making Systems with Relevant Patient Information. In *Proceedings of the Biomedical Engineering Systems and Technologies*; Ye, X., Soares, F., De Maria, E., Gómez Vilda, P., Cabitza, F., Fred, A., Gamboa, H., Eds.; Springer International Publishing: Cham, Switzerland, 2021; pp. 521–540.
13. Aggarwal, C.C.; Clustering, C.R.D. *Algorithms and Applications*; Springer: New York, NY, USA, 2014.
14. Gómez, D.; Rojas, A. An empirical overview of the no free lunch theorem and its effect on real-world machine learning classification. *Neural Comput.* **2016**, *28*, 216–228. [[CrossRef](#)] [[PubMed](#)]
15. Mascio, A.; Kraljevic, Z.; Bean, D.; Dobson, R.; Stewart, R.; Bendayan, R.; Roberts, A. Comparative analysis of text classification approaches in electronic health records. *arXiv* **2020**, arXiv:2005.06624.
16. Stein, R.A.; Jaques, P.A.; Valiati, J.F. An analysis of hierarchical text classification using word embeddings. *Inf. Sci.* **2019**, *471*, 216–232. [[CrossRef](#)]
17. Kadhim, A.I. Survey on supervised machine learning techniques for automatic text classification. *Artif. Intell. Rev.* **2019**, *52*, 273–292. [[CrossRef](#)]
18. Baker, S.; Korhonen, A.L.; Pysalo, S. Cancer Hallmark Text Classification Using Convolutional Neural Networks. In Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2016), Osaka, Japan, 11–16 December 2016. [[CrossRef](#)]
19. Flores, C.A.; Figueroa, R.L.; Pezoa, J.E. Active Learning for Biomedical Text Classification Based on Automatically Generated Regular Expressions. *IEEE Access* **2021**, *9*, 38767–38777. [[CrossRef](#)]
20. Gao, S.; Alawad, M.; Young, M.T.; Gounley, J.; Schaefferkoetter, N.; Yoon, H.J.; Wu, X.C.; Durbin, E.B.; Doherty, J.; Stroup, A.; et al. Limitations of Transformers on Clinical Text Classification. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 3596–3607. [[CrossRef](#)] [[PubMed](#)]
21. Markus, A.F.; Kors, J.A.; Rijnbeek, P.R. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *J. Biomed. Inform.* **2021**, *113*, 103655. [[CrossRef](#)]
22. Amann, J.; Blasimme, A.; Vayena, E.; Frey, D.; Madai, V.I. Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 1–9. [[CrossRef](#)] [[PubMed](#)]
23. Li, Y.; Yang, T. Word Embedding for Understanding Natural Language: A Survey. In *Guide to Big Data Applications*; Springer International Publishing: Cham, Switzerland, 2018; pp. 83–104. [[CrossRef](#)]
24. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. 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, NAACL-HLT 2019, Minneapolis, MN, USA, 2–7 June 2019, Volume 1 (Long and Short Papers)*; Burstein, J., Doran, C., Solorio, T., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 4171–4186. [[CrossRef](#)]
25. Akbik, A.; Blythe, D.; Vollgraf, R. Contextual String Embeddings for Sequence Labeling. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; Association for Computational Linguistics: Santa Fe, NM, USA, 2018; pp. 1638–1649.
26. Souza, F.; Nogueira, R.; Lotufo, R. BERTimbau: Pretrained BERT models for Brazilian Portuguese. In Proceedings of the 9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, 20–23 October 2020.
27. Kim, Y. Convolutional Neural Networks for Sentence Classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; Association for Computational Linguistics: Doha, Qatar, 2014; pp. 1746–1751. [[CrossRef](#)]
28. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
29. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.

30. Scikit-learn. Available online: <https://scikit-learn.org/stable/> (accessed on 7 December 2022).
31. TensorFlow. Available online: <https://www.tensorflow.org/> (accessed on 7 December 2022).
32. Pytorch. Available online: <https://pytorch.org/> (accessed on 7 December 2022).
33. Python. Available online: <https://www.python.org/> (accessed on 7 December 2022).
34. ELI5. Available online: <https://eli5.readthedocs.io/en/latest/index.html> (accessed on 7 December 2022).
35. Dhar, P. The carbon impact of artificial intelligence. *Nat. Mach. Intell.* **2020**, *2*, 423–425. [CrossRef]
36. Bannour, N.; Ghannay, S.; Névéol, A.; Ligozat, A.L. Evaluating the carbon footprint of NLP methods: A survey and analysis of existing tools. In Proceedings of the EMNLP, Workshop SustainNLP, Online, 7–11 November 2021.
37. Friedman, J.H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **2002**, *38*, 367–378. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.