eXDiL: A Tool for Classifying and eXplaining Hospital Discharge Letters

Fabio Mercorio, Mario Mezzanzanica, Andrea Seveso

▶ **To cite this version:**

Fabio Mercorio, Mario Mezzanzanica, Andrea Seveso. eXDiL: A Tool for Classifying and eXplaining Hospital Discharge Letters. 4th International Cross-Domain Conference for Machine Learning and Knowledge Extraction (CD-MAKE), Aug 2020, Dublin, Ireland. pp.159-172, 10.1007/978-3-030-57321-8_9 . hal-03414738

HAL Id: hal-03414738

https://inria.hal.science/hal-03414738v1

Submitted on 4 Nov 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

eXDiL: A Tool for Classifying and eXplaining Hospital Discharge Letters

Fabio Mercorio[1][0000−0001−6864−2702], Mario Mezzanzanica[1][0000−0003−0399−2810], and Andrea Seveso*[2][0000−0001−7132−7703]

[1] Dept of Statistics and Quantitative Methods - CRISP Research Centre, University of Milano-Bicocca, Milan, Italy
[2] Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milan, Italy

Abstract. Discharge letters (DiL) are used within any hospital Information Systems to track diseases of patients during their hospitalisation. Such records are commonly classified over the standard taxonomy made by the World Health Organization, that is the International Statistical Classification of Diseases and Related Health Problems (ICD-10). Particularly, classifying DiLs on the right code is crucial to allow hospitals to be refunded by Public Administrations on the basis of the health service provided. In many practical cases the classification task is carried out by hospital operators, that often have to cope under pressure, making this task an error-prone and time-consuming activity. This process might be improved by applying machine learning techniques to empower the clinical staff. In this paper, we present a system, namely `eXDiL`, that uses a two-stage Machine Learning and XAI-based approach for classifying DiL data on the ICD-10 taxonomy. To skim the common cases, we first classify automatically the most frequent codes. The codes that are not automatically discovered will be classified into the appropriate chapter and given to an operator to assess the correct code, in addition to an extensive explanation to help the evaluation, comprising of an explainable local surrogate model and a word similarity task. We also show how our approach will be beneficial to healthcare operators, and in particular how it will speed up the process and potentially reduce human errors.

Keywords: eXplainable AI · Machine Learning · Healthcare · Text classification.

1 Introduction

A large amount of medical examinations are carried out every day at hospitals and emergency rooms. For every visit, many bureaucratic documents must be compiled. One of these is the *discharge letter* (DiL). A DiL is a document issued to the patient at time of discharge from a hospital. It is the summary of the information contained in the medical record - of which it is an integral part

* Corresponding Author

- and contains the advice for any checks or therapies to be carried out. The information contained in the document is therefore intended to be useful to the doctor who will follow the patient in the future.

To be correctly classified according to international standards, at least one International Statistical Classification of Diseases and Related Health Problems code (ICD [27]) must be associated with each visit. ICD-10 is a medical classification taxonomy created by the World Health Organization. It contains codes for diseases, signs and symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or diseases.

Responsibility for the correct completion of the letter of discharge lies with the doctor responsible for discharge. The ICD code must be assigned manually by the attending physician, who however is not always available to manually enter data, mainly due to the fast-paced environment of hospitals and first aid centers. Therefore, the code is often missing or entered as a placeholder.

We would like to investigate the following three research questions:

Q1: How can machine learning techniques support the clinical staff classifying the text data and help avoiding human errors?
Q2: Can such a system provide explanations to users for supporting transparency and trustworthiness?
Q3: Is there a way to discover lexical similarities and relationships between medical terms?

1.1 Motivating example

To elaborate on our research question, next we introduce a concrete example, which allows the reader to understand the problem and shows why and for whom the system is useful to. We investigate the scenario of a particular Italian medical center, where the number of visits without ICD is very high ($\frac{1}{3}$ of visits). Assigning the right ICD code to a DL is a crucial and beneficial task for the hospital, for many practical reasons: getting reimbursements from the regional health service provider, assessment of fund allocation and so on. If the code is not assigned, it is necessary to manually read the DiLs, which contain all the information necessary to trace the specific ICD of the visit. This process is error-prone and time-consuming; however, it could be improved by applying text mining techniques such as text classification.

This particular medical center only treats a subset of possible diseases, those that are specific to Italy and to the expertise of the center. As an application paper, we would like to focus on this specific domain for our analyses.

1.2 Contribution

The contribution of this project is the following:

– Realisation of a XAI-based prototype to healthcare for classifying and explaining Discharge Letter (DiL) data, in order to answer the research questions Q1 and Q2. A demonstrative video of the prototype realised is also provided.

- The XAI module is beneficial to healthcare operators to understand the rationale behind the classification process as well as to speed-up the classification process as a whole;
- The trained italian Word Embeddings, specific to the healthcare domain, support word similarities and classification tasks and are required for the research task Q3.

We begin by discussing the technical backgrounds in Sec. 2. Then, we show the "eXplainable Discharge Letters" (`eXDiL`) system in Sec. 3 and its performances in Sec. 4. Sec. 5 concludes the paper with the discussion of pros and cons, conclusions and future work we intend to carry on.

2 Backgrounds and Related Work

In this section we introduce some background notions on word embeddings and XAI, as well as previous articles that explored this research area in the past.

Diagnosis code assignment is a well-known classification problem. In recent years it has been tackled with rule based methods as well as statistical and machine learning approaches.

One of the first to approach this task was [16] and [7] using rule based methods. Such methods are not easily created, and require extensive domain expertise in order to create the appropriate classification rules. However, the interpretability of this kind of models is the highest, as you can explain perfectly how a prediction was made.

In [17], the authors used machine learning methods such as Support Vector Machine (SVM) and Bayesian Ridge Regression (BRR). They include only the five most frequent ICD codes, and their classification performance is not very high. The upside of machine learning methods in classification tasks is that a lesser amount of domain expertise is needed, and the performance is in many cases higher than rule based systems.

More recent works include [26], who use a novel Convolutional Neural Network (CNN) model with attention. They select a subsection of the 50 most frequent codes, and perform a multilabel classification. They also conduct an human evaluation on the attention explanations. In [2] authors use multiple models; SVM, Continuous Bag of Words (CBOW), CNN and an hierarchical model, HA-GRU. The performance of more complex and deep models is superior to a model such as SVM. They interpret their results with an attention mechanism. Unlike other works presented, they include all the labels present in the dataset, using both the full 5 character codes and rolled up codes at 3 characters. The latter two works both use the publicly available MIMIC II and III datasets for training and testing their models.

2.1 Word Embeddings

Vector representation of words belongs to the family of neural language models [3], where each word of a given lexicon is mapped to a unique vector in the corresponding N-dimensional space (with a fixed N).

In our application, each word can be considered as the text content of an DiL. Here, an important contribution comes from the *Word2Vec* algorithm [22, 23], that computes the vector representations of words by looking at the context where these words are used. Intuitively, given a word w and its context k (i.e., m words in the neighbourhood of w), it uses k as a feature for predicting the word w. This task can be expressed as a machine learning problem, where the representation of m context words is fed into a neural network trained to predict the representation of w, according to the *Continuous Bag of Words* (CBOW) model proposed by [22][3].

Consider two different words w_1 and w_2 having very similar *contexts*, k_1 and k_2 (e.g., synonyms are likely to have similar though different contexts), a neural network builds an internal (abstract) representations of the input data in each internal network layer. If the two output words have similar input contexts (namely, k_1 and k_2) then, the neural network is motivated to learn similar internal representations for the output words w_1 and w_2. For more details, see [23].

After the Word2vec training on the lexicon, words with similar meanings are mapped to a similar position in the vector space. For example, "powerful" and "strong" are close to each other, whereas "powerful" and "Paris" are farther away. The word vector differences are also meaningful. For example, the word vectors can be used to answer analogy questions using simple vector algebra: "King" - "man" + "woman" $\approx$ "Queen" [24].

As one might note, this approach allows representing a specific word in the N-dimensional space, while our task is to compute the vector space of *documents* (i.e., research products), rather than words. We therefore apply the *Doc2Vec* approach [15], an unsupervised algorithm that learns fixed-length feature representations from variable-length pieces of texts, such as sentences, paragraphs, and documents as well. As a consequence, a vector is now the N-dimensional representation of documents.

In our framework, the use of word-embedding allows computing the similarities between DiLs, improving the effectiveness of the result.

2.2 XAI

There is a growing interest in the use of AI algorithms in many real-life scenarios and applications. However, many AI algorithms - as in the case of machine learning - rarely provide explanations or justifications to allow users understanding what the system really learnt, and this might affect the reliability of the algorithms' outcomes when these are used for taking decisions. In order to engender trust in AI, humans must understand what an AI system is trying to achieve, and which criteria guided its decision. A way to overcome this problem requests the underlying AI process must produce explanations that are transparent and

[3] A similar (but reversed problem) is the *Skip-n-gram model* i.e., to train a neural network to predict the representation of n context words from the representation of w. The Skip-n-gram approach can be summarised as "predicting the context given a word" while the CBOW, in a nutshell, is "predicting the word given a context".

comprehensible to the final user, so that she/he can consider the outcome generated by the system as *believable*[4] taking decisions accordingly. Not surprisingly, an aspect that still plays a key role in machine learning relies on the quality of the data used for training the model. In essence, we may argue that the well-known principle *"garbage-in, garbage-out"* that characterizes the data quality research field, also applies to machine learning ,and AI in general, that is used to evaluated data quality on big data (see, e.g. [20, 1, 4, 21]) and perform cleaning tasks as well (see, e.g. [19, 18, 6]).

Given the success and spread of AI systems, all these concerns are becoming quite relevant enabling a wide branch of AI to emerge, with the aim of making AI algorithms explainable for getting an improved trustability and transparency (*aka* Explainable AI (XAI)). Though some research on explainable AI had already been published before DARPA's program that launched a call for XAI in 2016 (see, e.g., [31, 28]) XAI [8, 5, 25], effectively encouraged a large number of researchers to take up this challenge. In the last couple of years, several publications have appeared that investigate how to explain the different areas of AI, such as machine learning [12, 30], robotics and autonomous systems [11], constraint reasoning [10], and AI planning [9], just to cite a few. Furthermore, as recently argued in [5], a key element of an AI system relies on the ability to explain its decisions, recommendations, predictions or actions as well as the process through which they are made. Hence, explanation is closely related to the concept of interpretability: systems are interpretable if their operations can be understood by a human, either through introspection or through a produced explanation.

3 The `eXDiL` system

The "eXplainable Discharge Letters" system (`eXDiL`) intends to help the clinical staff identifying the main ICD code related to each visit in order to significantly lighten human workload. It would act at operational-level of the hospital to support operators in assigning the correct ICD code to each visit (i.e., operational-level information system). We propose a two-step system for semi-real time classification with an *human in the loop* approach. A visual representation of the workflow can be seen in Fig. 4.

In order to let the reader easily understand the workflow, we also present an operative example in this section. A video walkthrough of this example is available at https://youtu.be/u0UJnp4RyQQ. The following working example should clarify the matter.

Working Example. Let us consider the following DiL: "*Reason for visit*: Low back pain in patients with osteoporosis not under drug treatment. *Diagnosis*: lumbar pain in patient with osteoporosis, deformation in L1 with lowering of the limiting upper in outcomes, anterograde slipping of L3-L4 and L5-S1 with

[4] Here the term believability is inherited from the definition of [32] intended as "the extent to which data are accepted or regarded as true, real and credible"

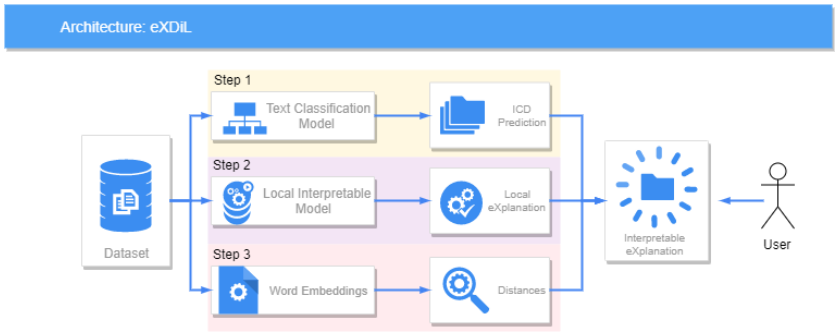


Fig. 1. a representation of the `eXDiL` workflow, highlighting the main modules.

reduction in amplitude of the interbody spaces." The true ICD class of this example is *Dorsalgia*. Assigning the correct ICD class to the letter is crucial to allow the hospitals to be refunded according to the health service really provided. To this end, in the following we discuss how `eXDiL` works.

3.1 Step 1: ICD Prediction

After the doctor finishes writing the letter, the system uses the text data to automatically classify the most common ICDs, as to perform a first skimming. If the classification is successful, then the clinician can accept or reject the suggestion.

In the prototype, the doctor must type a reason for visit and diagnosis in a free text format, or choose an example from a predefined list. Then, the `eXDiL` system will attempt to classify the data using the workflow described in Fig. 4.

In the other case, if the reason for the visit is not found among the most common cases, then the most relevant chapter is proposed, but the single ICD is not provided. The doctor can use this suggestion to input manually the correct code.

3.2 Step 2: Local eXplanation

After the prediction, a visual explanation of the result is displayed in order to make the clinician aware of the main reasons why certain classes are assigned to certain visits. Moreover, this part is fundamental to establish a relationship of trust between man and algorithm.

Using the LIME [30] approach, we propose a first visualization that explains quantitatively how much a particular term is in favour or against w.r.t. the aforementioned classification (see Fig. 2).

This visualization is accompanied by a second one, in which the terms in question are highlighted directly in the text in such a way as to easily identify the context of use and determine whether or not they are significant according to the judgment of the domain expert, who in this case is the doctor of reference.

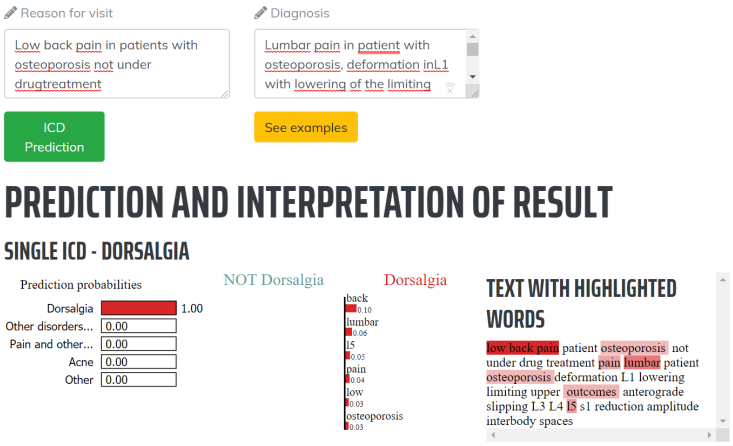
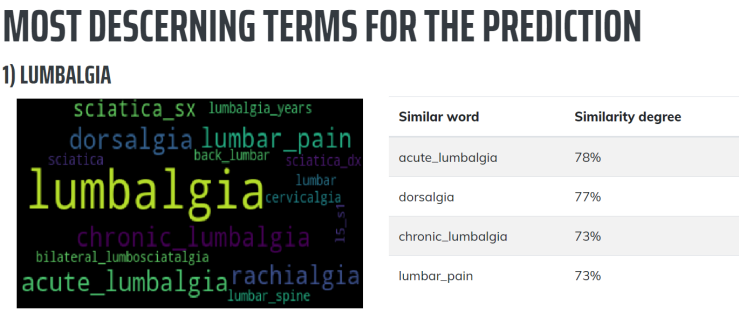


Fig. 2. Example of explainable output from the `eXDiL` prototype.

3.3 Step 3: Word Similarity

In addition, a Word2Vec [24, 22, 23] system is offered to suggest words similar to those already inserted, in order make the clinician more aware of similar cases. A word cloud generated by the model is printed on the screen in order to generate immediate cues. The size of the related words is related to the similarity degree. The aforementioned is accompanied by a more detailed outline of the similarity of the suggested words w.r.t. the starting one. An example of the visualization can be seen in Fig. 3.



Similar word	Similarity degree
acute_lumbalgia	78%
dorsalgia	77%
chronic_lumbalgia	73%
lumbar_pain	73%

Fig. 3. Example of most similar words from the `eXDiL` prototype.

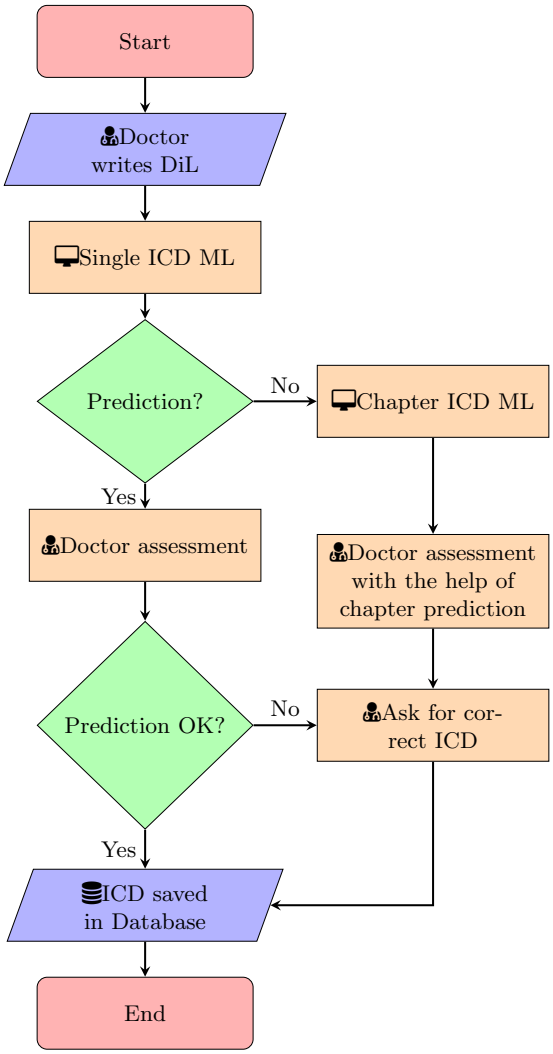


Fig. 4. Semi-real time ML assisted ICD classification workflow.

4 Experimental Results

4.1 Dataset Description and Characteristics

The dataset is composed by 168.925 individual visits, collected from 2011 to 2018 in an Italian Hospital.[5] Notice that the ICD taxonomy is built to include *any disease*, including eradicated ones. For this reason and due to the massive amount

[5] The name of the Hospital is omitted due to non-disclosure agreements

of existing classes (around 10,000), the classification task should concentrate only on diseases that are more likely to be treated by a given hospital.

Following the example of [14], the codes are truncated at the three character level; as an example, code M54.1 (*Radiculopathy*) is converted to M54 (*Dorsalgia*). We also chose to restrict the dataset to the 5 most common specialities described in Tab. 1. For each speciality, the top-2 most common ICD classes were chosen for the single ICD classifier, as seen in Tab. 2.

Table 1. Count of selected ICD chapters.

Chapter	Chapter label	N	Percent
XIII	Diseases of the musculoskeletal system and connective tissue	32051	38.9%
XIV	Diseases of the genitourinary system	14650	17.8%
XII	Diseases of the skin, subcutaneous tissue	12400	15.0%
V	Mental and behavioural disorders	11907	14.4%
X	Diseases of the respiratory system	11431	13.9%

Table 2. Count of most frequent ICD classes.

Code	Label	N	Percent
M54	Dorsalgia	9009	29.9%
J30	Vasomotor and allergic rhinitis	3705	12.3%
M77	Other enthesopathies	3002	10.0%
N76	Other inflammation of vagina and vulva	2806	9.3%
F41	Other anxiety disorders	2692	8.9%
N94	Pain and other conditions associated with female genital organs and menstrual cycle	2692	8.9%
F60	Specific personality disorders	2496	8.3%
L70	Acne	1484	4.9%
L50	Urticaria	1115	3.7%
J45	Asthma	1085	3.6%

After filtering for only the five chosen specialities and applying the preprocessing pipeline, 82.439 letters remain (49% of total data). For the single ICD prediction task, the two most common ICD classes were chosen for each speciality, obtaining 30.086 letters (36% of the 5 specialities, 18% of total data).

For each visit, the DiL describes many features of the patient, such as: reason for the visit, free text doctor diagnosis, medical history, therapy and clinical tests to be carried out, specialization of the practitioner and other information related to the DiL, such as clinical indications, allergies, follow-up instructions.

Out of these features, reason for the visit and free text doctor diagnosis have been chosen to make a prediction, as they are most informative for this task.

4.2 Classification Pipelines

Free text data is not eligible to use as-is. A pre-processing step is required in order to clean the data, according to the following steps: (i) fix character encoding issues, (ii) remove punctuation, isolated numbers and stop words, (iii) lower casing, (iv) remove domain-specific common words, such as "visit" and "control". After pre-processing, we can use the text data to train the classifiers. The mean word count of each note was 16 words is of 15.6 words, with a high standard deviation of 28.7. The longest note in the dataset contains 1124 words. The data set was then separated in two sets: the training set, with 67% of the data, and the test set, with 33% of the data.

In order to create a classification model, a text representation, classifier and set of hyperparameters must be chosen. We have considered the following text representations: (i) Bag of Words (BoW), (ii) Tf-Idf with only Unigrams, (iii) Tf-Idf with Bigrams.

We also considered the Word2Vec Skipgram, Continuous Bag Of Words (CBoW) and GloVe [29] text representations. However, these embeddings did not provide sufficient performances on the classification tasks. As such, they were not included in the results, and the embeddings were used exclusively for the word similarity task.

The following models and hyperparameters sets were considered:

1. SVC with $C \in \{1, 0\}$
2. Random Forest with number of estimators $\in \{10, 100\}$
3. Naive Bayes with $\alpha \in \{0.1, 0.01, 0.001\}$
4. Logistic regression with mode $\in$ {multinomial, sag} $\times$ $C \in \{1, 10\}$

A 5-fold cross-validated grid search on the text representations, models and hyperparameters was conducted in order to find the best models for classification of single and chapter ICD. Also, in order to test the hypothesis that the system will improve with additional data from human input, we started by training the models with only 50% of the available training data, then increasing to 67% and finally using all the training data. The test set was not changed between tests to ensure results consistency.

Similarly to [2], the main metric for model evaluation was chosen as the F1 micro average score.

4.3 Evaluation of Single ICD Model

Fig. 5 shows the performances of the models on the test set. Each point in the violin plot represents the F1 score for a tuple (model type, text representation, hyperparameters). The y axis represents the F1 score, ranging between [0.7, 1.0], while the X axis represents the ICD class. For a complete list of the ICD classes, see Tab. 2. The average performance increases slightly when increasing the available training data. This suggests that increasing the training set size might lead to better performances, however the impact of this increase might is not statistically significant.

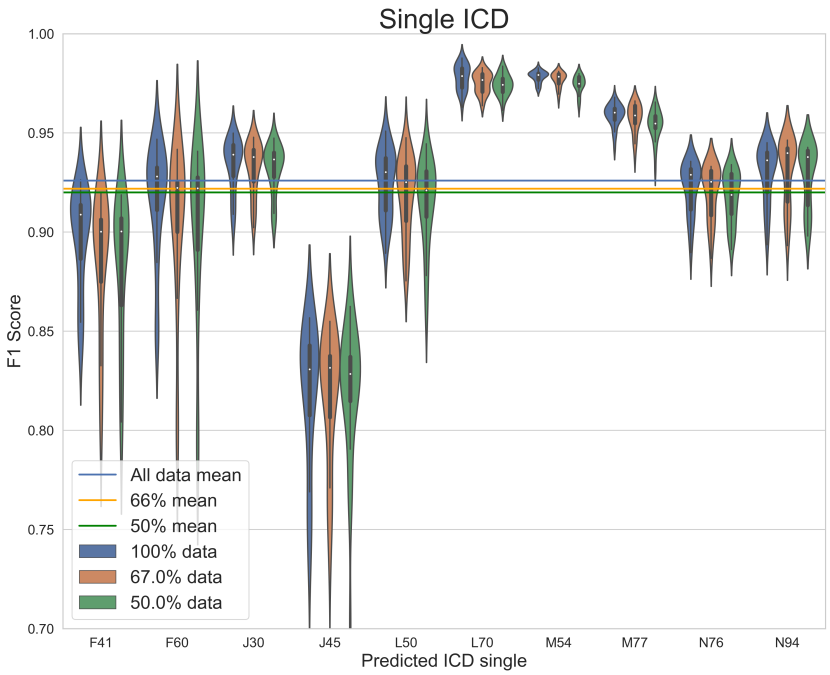


Fig. 5. Visualization of single ICD classification models performance.

The best model out of all the possible combinations is a logistic regression with mode = OVR and C = 1 on the BoW representation. This model reaches an average F1 score of 0.952. The highest percentage of errors is between ICD classes *vasomotor and allergic rhinitis* and *asthma*. This might be induced by the semantic similarity between the two concepts, and also the fact that a patient might be affected by both conditions at the same time.

4.4 Evaluation of Chapter ICD Model

In this higher granularity of classification the performance is higher compared to the third character level. The classes are semantically different between each other, and this distance is reflected on the more accurate results in the classifiers. In Fig. 6 we show the results for each classifier trained. In order to properly show the differences between classes, we restricted the y axis between [0.8, 1.0]. In this case, the models performances do not decrease nor increase with different amounts of training data.

The best performing model is a SVC classifier with $C = 1$ on Tf-Idf Bigrams text representation. This model reaches an average F1 score of 0.983.

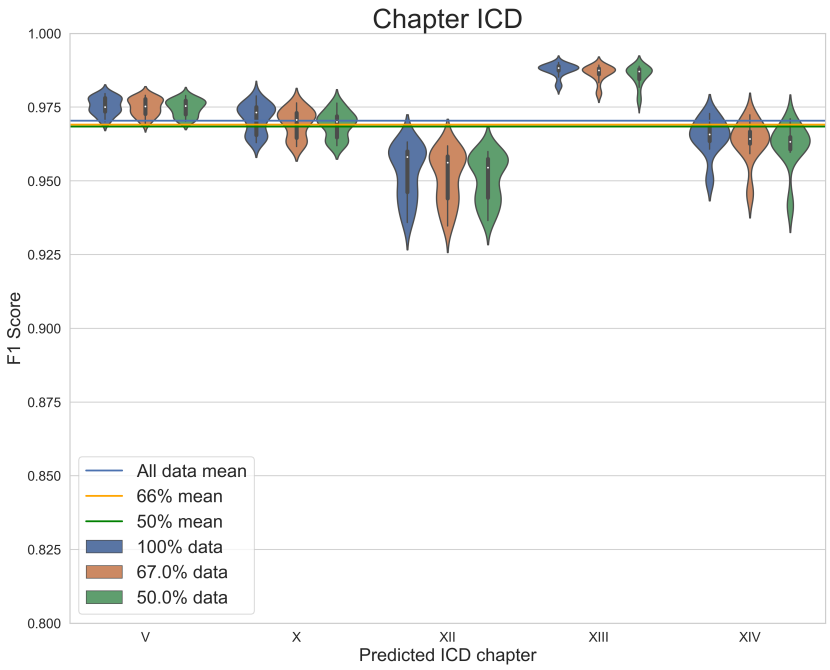


Fig. 6. Visualization of chapter ICD classifications models performance.

5 Discussion and Conclusion

Considering the *Pros*, the resulting classifications have excellent performances on both the chapter and single ICD levels. In particular, the chapter level can be trusted with high confidence (F1 Micro=0.983). It seems that increasing the collected data helps improve performance by a small amount, however the increase does not seem to be significant. The system may therefore help the doctor save time and better classify the DiLs.

Evaluating the *Cons*, a major issue is that this procedure cannot be done on all ICD classes at once. It is firstly advised to choose specific specialties, since without filtering, classifying most of the over 10,000 classes would be infeasible with our dataset. Therefore the scope must be restricted to certain specialties, and, as shown in [14], the granularity should be set at the third character level, as in our case it is not possible to distinguish accurately between the subtle differences at the fourth character level.

In conclusion, we have shown that the `eXDiL` system is an accurate XAI system for classifying hospital discharge letters. Future work can be conducted to improve and assess the whole procedure to finally bring it to life in a real world environment providing an hopefully useful service. Firstly, it has to be tested in the hospital field to check for real world usefulness. Secondly, it is to be understood if the "human in the loop" works as it has been conceived.

To date, `eXDiL` has been trained on Italian DLs provided by an Italian Hospital. We have been working on applying `eXDiL` on the well-known benchmark MIMIC-III [13], a widely known and used dataset. It comprises a larger amount of data, with more labels and variety, and importantly it is in english, meaning it would create a classifier with a broader use case.

DEMO. A demonstration video of the system has been also provided.[6]

References

1. Amato, F., Castiglione, A., Mercorio, F., Mezzanzanica, M., Moscato, V., Picariello, A., Sperlì, G.: Multimedia story creation on social networks. Future Generation Computer Systems **86**, 412 – 420 (2018)
2. Baumel, T., Nassour-Kassis, J., Cohen, R., Elhadad, M., Elhadad, N.: Multi-label classification of patient notes: case study on icd code assignment. In: Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence (2018)
3. Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A neural probabilistic language model. Journal of machine learning research **3**(Feb), 1137–1155 (2003)
4. Bergamaschi, S., Carlini, E., Ceci, M., Furletti, B., Giannotti, F., Malerba, D., Mezzanzanica, M., Monreale, A., Pasi, G., Pedreschi, D., Perego, R., Ruggieri, S.: Big data research in italy: A perspective. Engineering **2**(2), 163 – 170 (2016)
5. Biran, O., Cotton, C.: Explanation and justification in machine learning: A survey. In: IJCAI-17 Workshop on Explainable AI (XAI). p. 8 (2017)
6. Boselli, R., Cesarini, M., Mercorio, F., Mezzanzanica, M.: Inconsistency knowledge discovery for longitudinal data management: A model-based approach. In: Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data - Third International Workshop, HCI-KDD. pp. 183–194 (2013)
7. Crammer, K., Dredze, M., Ganchev, K., Talukdar, P.P., Carroll, S.: Automatic code assignment to medical text. In: Proceedings of the workshop on bionlp 2007: Biological, translational, and clinical language processing. pp. 129–136. Association for Computational Linguistics (2007)
8. DARPA: Explainable artificial intelligence (xai) program. http://www.darpa.mil/program/explainable-artificial-intelligence. full solicitation at http://www.darpa.mil/ attachments/DARPA-BAA-16-53.pdf (2016)
9. Fox, M., Long, D., Magazzeni, D.: Explainable planning. arXiv preprint arXiv:1709.10256 (2017)
10. Freuder, E.C.: Explaining ourselves: Human-aware constraint reasoning. In: AAAI. pp. 4858–4862 (2017)
11. Hayes, B., Shah, J.A.: Improving robot controller transparency through autonomous policy explanation. In: 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI). pp. 303–312. IEEE (2017)
12. Hendricks, L.A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., Darrell, T.: Generating visual explanations. In: European Conference on Computer Vision. pp. 3–19. Springer (2016)
13. Johnson, A.E., Pollard, T.J., Shen, L., Li-wei, H.L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G.: Mimic-iii, a freely accessible critical care database. Scientific data **3**, 160035 (2016)

[6] https://youtu.be/u0UJnp4RyQQ

14. Koopman, B., Zuccon, G., Nguyen, A., Bergheim, A., Grayson, N.: Automatic icd-10 classification of cancers from free-text death certificates. International journal of medical informatics **84**(11), 956–965 (2015)
15. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: International Conference on Machine Learning. pp. 1188–1196 (2014)
16. de Lima, L.R., Laender, A.H., Ribeiro-Neto, B.A.: A hierarchical approach to the automatic categorization of medical documents. In: International conference on Information and knowledge management. pp. 132–139. ACM (1998)
17. Lita, L.V., Yu, S., Niculescu, S., Bi, J.: Large scale diagnostic code classification for medical patient records. In: Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II (2008)
18. Mezzanzanica, M., Boselli, R., Cesarini, M., Mercorio, F.: Data quality through model checking techniques. In: Advances in Intelligent Data Analysis X - 10th International Symposium, IDA. pp. 270–281 (2011)
19. Mezzanzanica, M., Boselli, R., Cesarini, M., Mercorio, F.: Data quality sensitivity analysis on aggregate indicators. In: International Conference on Data Technologies and Applications. pp. 97–108 (2012)
20. Mezzanzanica, M., Boselli, R., Cesarini, M., Mercorio, F.: Automatic synthesis of data cleansing activities. In: Proceedings of the 2nd International Conference on Data Technologies and Applications. pp. 138–149 (2013)
21. Mezzanzanica, M., Boselli, R., Cesarini, M., Mercorio, F.: A model-based approach for developing data cleansing solutions. J. Data and Information Quality **5**(4), 13:1–13:28 (2015)
22. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
23. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
24. Mikolov, T., Yih, W.t., Zweig, G.: Linguistic regularities in continuous space word representations. In: Hlt-naacl. vol. 13, pp. 746–751 (2013)
25. Miller, T., Howe, P., Sonenberg, L.: Explainable ai: Beware of inmates running the asylum. In: IJCAI-17 Workshop on Explainable AI (XAI). p. 36 (2017)
26. Mullenbach, J., Wiegreffe, S., Duke, J., Sun, J., Eisenstein, J.: Explainable prediction of medical codes from clinical text. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 1101–1111 (2018)
27. Organization, W.H., et al.: The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines. Geneva: World Health Organization (1992)
28. Papadimitriou, A., Symeonidis, P., Manolopoulos, Y.: A generalized taxonomy of explanations styles for traditional and social recommender systems. Data Mining and Knowledge Discovery **24**(3), 555–583 (2012)
29. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
30. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you?: Explaining the predictions of any classifier. In: ACM SIGKDD. pp. 1135–1144. ACM (2016)
31. Swartout, W., Paris, C., Moore, J.: Explanations in knowledge systems: Design for explainable expert systems. IEEE Expert **6**(3), 58–64 (1991)
32. Wang, R.Y., Strong, D.M.: Beyond accuracy: What data quality means to data consumers. Journal of management information systems **12**(4), 5–33 (1996)