scientific reports

Check for updates

OPEN

A labeled medical records corpus for the timely detection of rare diseases using machine learning approaches

Matias Rolando[1], Victor Raggio[2], Hugo Naya[1,3], Lucia Spangenberg[1,4]✉ & Leticia Cagnina[5,6]✉

Rare diseases (RDs) are a group of pathologies that individually affect less than 1 in 2000 people but collectively impact around 7% of the world's population. Most of them affect children, are chronic and progressive, and have no specific treatment. RD patients face diagnostic challenges, with an average diagnosis time of 5 years, multiple specialist visits, and invasive procedures. This 'diagnostic odyssey' can be detrimental to their health. Machine learning (ML) has the potential to improve healthcare by providing more personalized and accurate patient management, diagnoses, and in some cases, treatments. Leveraging the MIMIC-III database and additional medical notes from different sources such as in-house data, PubMed and chatGPT, we propose a labeled dataset for early RD detection in hospital settings. Applying various supervised ML methods, including logistic regression, decision trees, support vector machine (SVM), deep learning methods (LSTM and CNN), and Transformers (BERT), we validated the use of the proposed resource, achieving 92.7% F-measure and a 96% AUC using SVM. These findings highlight the potential of ML in redirecting RD patients towards more accurate diagnostic pathways and presents a corpus that can be used for future development and refinements.

Rare disease (RD) is pathology that affects less than 1 in 2000 people[1]. Although separately each disease is very rare, taken together, there are around 7000 different disorders, they involve a large number of patients (7% of the world population)[2]. Most of them affect children, have a high impact on quality of life and life expectancy, are generally chronic and progressive, and most of them have no specific treatment. Moreover, by their infrequency nature, they are a diagnostic challenge. On average, it takes five years from the onset of symptoms to diagnosis, and it takes a mean of about seven visits to different specialists and dozens of studies, some of them invasive or requiring general anesthesia[3]; this is generally referred to as the "diagnostic odyssey". According to a recent report by Globalgenes (https://globalgenes.org/rare-facts) 40% of general practitioners and 24% of specialists state that they do not have the time to work on these diagnoses, which contributes to the difficulty of a time-efficient diagnostic process. It is important to obtain a time-efficient diagnosis to avoid a detrimental disease progression.

RD patients are extremely vulnerable and neglected by the healthcare system since there is usually no global, nationwide strategy to tackle and fund this problem efficiently. Genomic approaches such as whole exome or genome sequencing have improved the diagnosis rate in RD patients[4,5]. Depending on the existing pathology and method used, the diagnosis rate can vary from 30 to 50%[6–8]. It is of imperative importance to include these types of tests at the right moment in the diagnostic algorithm.

Machine learning (ML) is a powerful tool in several fields of healthcare, allowing for the development of more personalized and accurate patient management, diagnoses and treatments[9–11]. The application of ML methods has the potential to improve healthcare and create more efficient, reliable, and cost-effective treatments[12–14].

The advent of large scale language models Transformer-based such as BERT[15], GPT-3[15,16] or T5[17] trained on a massive amount of text data has shown to outperform a wide range of natural language processing tasks

[1]Bioinformatics Unit, Institut Pasteur de Montevideo, Montevideo, Uruguay. [2]Departamento de Genética, Facultad de Medicina, Universidad de la República, Montevideo, Uruguay. [3]Departamento de producción animal y pasturas, Facultad de Agronomía, Universidad de la República, Montevideo, Uruguay. [4]Departamento Básico de Medicina, Hospital de Clínicas, Universidad de la República, Montevideo, Uruguay. [5]Universidad Nacional de San Luis, San Luis, Argentina. [6]Consejo Nacional de Investigaciones Científicas y Técnicas, CONICET, Buenos Aires, Argentina. ✉email: lucia@pasteur.edu.uy; lcagnina@unsl.edu.ar

including text summarization, sentiment analysis, question answering and language translation. The efforts of ML researchers have shifted towards the fine-tuning of freely available versions of these models with relatively small specialized datasets (few-shot learning). Hence, the importance of generating such smaller corpora related to particular fields of interest and making them available to the community.

In this study we aim to propose a resource to contribute to the research in the detection of RD patients in an early stage in their passage through a hospital. For this, we have made use of the large, freely-available, MIMIC-III database[18] of de-identified health-related data of over forty thousand patients who stayed in critical care units at the Beth Israel Deaconess Medical Center. MIMIC-III includes information of patients such as test results, medications, mortality and discharge summaries. Besides, we have collected medical notes from other sources such as case reports from Pubmed, our own records and diagnostics automatically generated by chatGPT.

ML techniques applied to several fields in medicine are becoming key in aiding diagnosis of different diseases. Models applied to biomedical images to properly diagnose or subclassify diseases[19–22] or to predict complications of common diseases[23–25] are being developed. In addition, more specifically to RD, ML models applied to phenotypic data[26] and to questionnaire-based and data mining-supported tools have been developed[27]. Moreover, ML models are being specifically applied to clinical records exploration to improve diagnosis of a specific rare disease such as AHP[28] or to RD in general[29]. Also, an extension of the large language model BERT, called RareBert[30], has been trained on administrative claims datasets and tailored specifically for improved diagnosis of RD. However, this model is not freely available.

Here, we explored a plethora of methods applied to the proposed (available) corpus, including classical supervised machine learning (SVM, decision trees (DT), logistic regression (LR)), deep learning (LSTM and CNN), and Transformers (BERT). Using SVM, we identified RD patients (any RD) with 92.7% of F-measure and an AUC of 96%. This information could be used to redirect them to a more accurate diagnostic algorithm.

Data and methods

Following, we describe the proposed dataset for the classification of medical records as containing signs of rare diseases or not. Then, we introduce the models used in the experimental study. Since the emergence of deep learning, models have become more accurate and reliable, outperforming classical machine learning techniques in many tasks. However, in many cases classical models are still more suitable than deep learning because they are simpler, obtain comparable results and require less computing power. For that reason, we explore several methods, from classical to deep learning approaches in order to show their performance in the particular task of detecting RDs.

Proposed corpus

With the aim to get diversity, we used discharge summaries from MIMIC-III, some medical education notes from the School of Medicine in Uruguay, a few diagnosis reports from the URUGENOMES project (urugenomes.org), some clinical records scrapped from PubMed and a few clinical records "created" by chatGPT. Because all selected sources have different writing styles, we believe this is a positive aspect for the corpus. This variability provides a "more realistic" scenario in which medical records appear as unstructured data, with non-specific writing style (which depends on the health professional) and containing a variety of medical information. After collecting all documents, we checked for the label of each one (rare disease or common disease). Some of them such as the medical records from URUGENOMES project and case reports from PubMed had intrinsically the assigned category (rare disease or not). The rest of the records were revised and labeled by a medical geneticist to determine if each one corresponds to a medical note of a rare or a common disease. Below we describe in detail how the documents from the proposed corpus were selected.

MIMIC-III database has been used in previous studies related to the classification of diseases[18] and constitutes a valuable resource for research. This large and freely-available corpus has more than 2000 K de-identified data related to 46 K patients of critical care units of the Beth Israel Deaconess Medical Center (Boston, Massachusetts) between the years 2001 and 2012. The information contained is about demographics, vital sign measurements made at the bedside, test results of laboratories, procedures performed, medications, caregiver notes, imaging reports, data of mortality and discharge summaries. For our study, we used the discharge summaries of patients because these have interesting information about the diagnosis that doctors arrived at during their stay at the hospital. As we need to label some clinical notes as RDs for consideration in our proposed dataset, we considered a published article[29] in which the authors found RDs in a subset of discharge summaries taken from MIMIC-III. They used a two-steps approach. First, tokens appearing in the text are linked to medical concepts of the Unified Medical Language System (UMLS) with the SemEHR tool. The results were refined using particular rules for removing abbreviations and text-UMLS pairs with low mention frequencies in the clinical notes. Contextual representations of the pairs were obtained with BlueBERT and used to train a Logistic Regressor for a binary classification to confirm clinic mentions. Secondly, the UMLS concepts were linked to Orphanet Rare Disease Ontology (ORDO) and thus, the authors created a gold standard dataset with 1073 mentions: 146 of rare and 927 of common diseases. We searched for the discharge summary notes corresponding to these mentions and thus, obtained 65 notes containing clues of RD and the remaining 247, common diseases. We also randomly selected additional 100 clinical notes of MIMIC-III corresponding to the admission stage of patients. This is to introduce variability to the medical texts besides to augment the number of common diseases. We checked that those diagnoses were not labeled as rare (according to[29]). Finally, our proposed dataset has 412 records containing medical notes from MIMIC-III.

We also used medical education summaries collected from *Oficina del Libro* of the Medicine School at Universidad de la República (Uruguay)[31] which contain diagnostics related to cardiology, hematology, neurology and internal medicine. After a rigorous review performed by an expert (medical geneticist), we labeled the 98 notes in common (64) and rare (34) diseases. Those clinical texts were translated from spanish to english.

In order to balance both classes of diagnosis, we included 32 RDs obtained from clinical records of a previous project, URUGENOMES (urugenomes.org)[4,32] and 277 from PubMed. For the last, we scrapped the PubMed platform (https://pubmed.ncbi.nlm.nih.gov/) searching for case reports of RDs in free articles.

As obtaining medical notes on rare diseases is challenging, we generated some ones using chatGPT (https://chat.openai.com). Thus, we included 10 RD clinical records and 13 common diseases generated by the tool and these were manually curated by a clinical geneticist, adding 23 diagnostics to the corpus. Our complete corpus finally has 842 clinical records, 418 rare and 424 common. It is available under: https://sites.google.com/view/leticia-cagnina/research and can be free-used for research issues performing the corresponding citation. The interested reader is referred to the supplementary material (Methods Sect. 1) for further discussion of the limitations and challenges of using this dataset.

In neither case, no preprocessing nor normalization is done to preserve the data's genuine characteristics. The records collected constitute a balanced corpus which is more reliable to work with. Table 1 shows the main characteristics of the dataset. The vocabulary is large, with more than half of the unique words. This could be expected since the texts are related to specific medical issues (names of medications, lab tests, symptoms, diseases, abbreviations used by doctors, etc.). The diagnoses are written in 4 sentences on average although the variability of lengths is high: most clinical notes have only one sentence (possibly the diagnosis is written like a paragraph), just one with 340 (short sentences) and the rest with values oscillating the 2 and 70 sentences.

Table 1 (two bottom rows) shows the characteristics of the corpus separated by classes. As mentioned this corpus is balanced so it has a similar amount of records of RDs and common diseases (418 vs. 424). Although the size of vocabulary used is not very different, it seems that clinical notes of common diseases are much longer than ones for RDs (474 K vs. 312k). The same occurs with the number of sentences: RDs are written using one third of those included in common diagnostics. In fact, most of the notes of RDs use two sentences while those of common are slightly more extensive (6 on average). Only one record has a maximum number of sentences of 340 which corresponds to a common disease but the rest of this class are around 1 and 70. Moreover, the RD diagnosis has between 1 and 39 sentences as maximum. Examples of clinical notes are in Supplementary figure S1A (RD) and B (common diseases).

We also analyzed the contents of the clinical notes by using word clouds, which show the frequency of the words in each class. Most used words in common diseases notes are regular medical terms such as blood, pain and patient (Fig. 1A). Also some words related to non specific treatment: capsule, daily and hours. Interestingly the word cloud corresponding to RD notes (Fig. 1B) shows advanced studies and serious issues found in words such as tomography, examination, tumor and mass. Top 50 more representative words in the whole proposed dataset are in Supplementary Figure S2.

Figure 1C shows the proportion of medical notes collected from the different sources. Note that no requirement was stated in the time to collect the record and thus, the diversity of the distribution of words (log-scaled) is high. The distribution corresponding to RDs is approximately normal which is suitable for classification models (Fig. 1D). However, some outliers above 3500 words can be observed for this class. The kernel density estimation is shown with a narrow kurtosis (leptokurtic) and central tendency around 500. The distribution for the class of common diseases seems to be bimodal although the central tendency is around 100 without major outliers (see Fig. 1D).

Patient consent, data privacy and ethics

As was previously reported[18], *data in MIMIC-III* was desidentified according to Health Insurance Portability and Accountability Act (HIPAA) standards and dates shifted to avoid possible identification of patients. Prior to the assembly of the MIMIC-III, the project was approved by the Institutional Review Boards of Beth Israel Deaconess Medical Center (Boston, MA) and the Massachusetts Institute of Technology (Cambridge, MA). For that reason, the authors state that individual patient consents were not required since there was no impact in the clinical care, and protected health information was desidentified. Beyond that, the ethical use of this valuable resource requires to complete a specific course related to protecting human research participants including the HIPAA requirements and to sign a data use agreement.

Regarding *medical education summaries* collected from the Medicine School (Universidad de la República), these are public learning material created by the Faculty for medical students. These records were already desidentified by removing any personal data (names, birthdate, social security number, address and telephone number) before publication to the students and preserving some clinically relevant data, such as age, physical conditions, ethnic and hereditary background. In that way, we ensure confidentiality and privacy of this portion of the data.

All *medical records from the URUGENOMES* project have signed written informed consents and ethics committee approval under the name "Proyecto URUGENOMES - Fortalecimiento de las capacidades técnicas y humanas en el proyecto Genoma Humano Uruguay".

	Diagnosis	Vocabulary	Words	Word average	Sentences	Sentence average	Sentence min	Sentence max
Corpus	842	41,765	787,277	935	3576	4	1	340
Common	424	29,671	474,648	1119	2753	6	1	340
Rare	418	21,686	312,629	747	823	1	1	39

Table 1. Dataset statistics.

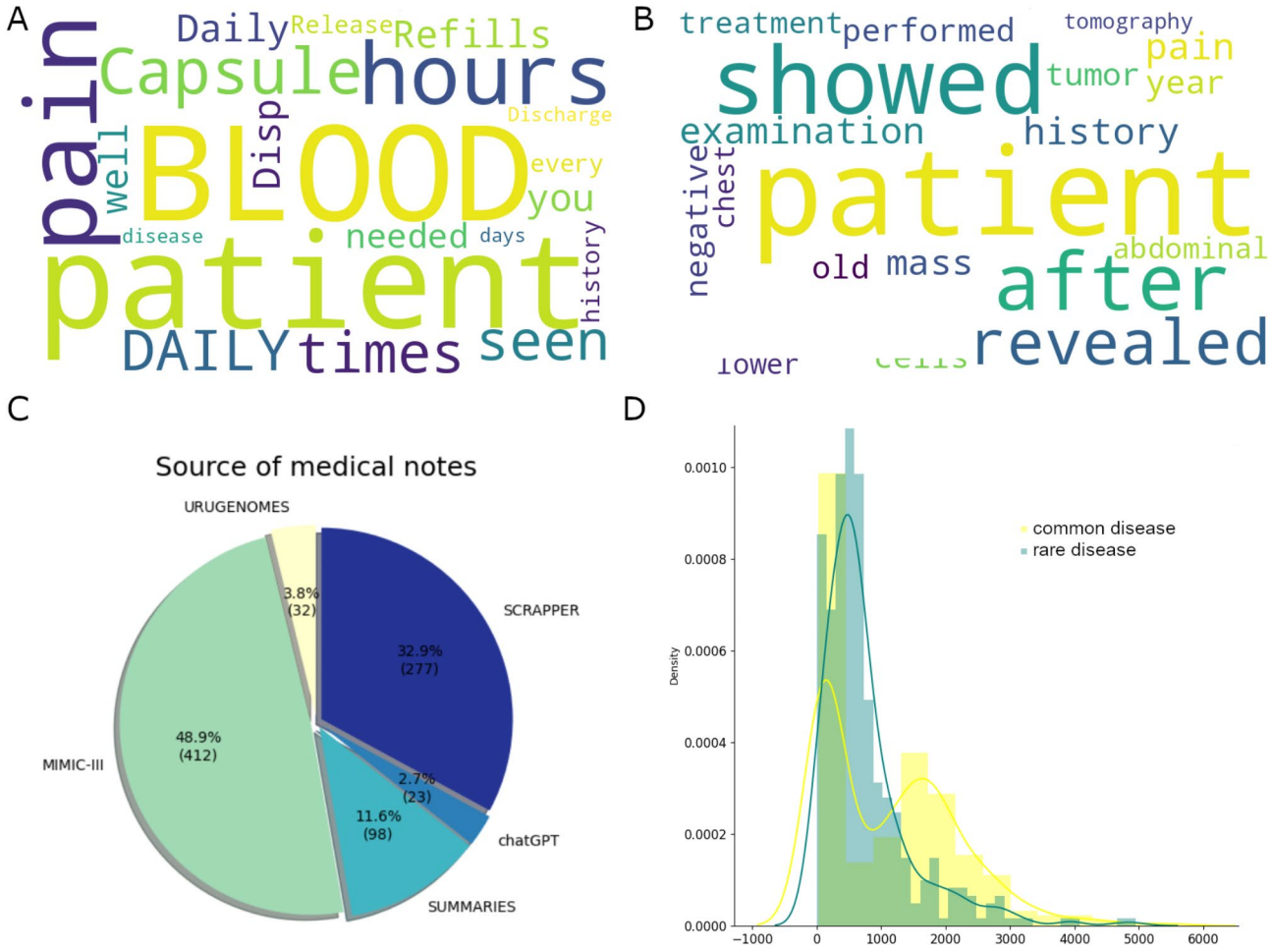

Fig. 1. Corpus description. (**A**) Word cloud that represents the frequency of each word for the Common class. (**B**) Word cloud for the RD class. (**C**) Number and proportion of records collected from different sources for the proposed dataset. (**D**) Distribution of text length (in log number of words units) for each class.

From *PubMed we extract interesting case reports* related to rare diseases. Because the documents are published in medical/biomedical journals, most of them (if not all) required ethics statements which imply that the study must have signed written informed consents from patients/guardians and the approval of an ethics committee (which ensures patient privacy).

Finally, the *medical notes artificially generated with chatGPT* are synthetic data and therefore there is no need to ensure consent or anonymize data. We quickly checked that the obtained records are correct from an ethical perspective.

Machine learning models

We employed TF-IDF (Term Frequency-Inverse Document Frequency) weighting and boolean schemes to transform text data into numerical feature representations. For TF-IDF, each term's weight was computed by multiplying its frequency in a document by the inverse of its frequency across the corpus. Boolean schemes convert term presence into binary values (0 or 1). Before vectorization, the text data was preprocessed by converting all text to lowercase, removing punctuation, and filtering out stop words.

For feature selection, we limited the vocabulary size to the top 5000 or 10,000 terms with the highest TF-IDF scores, ensuring a balance between model complexity and computational efficiency. We considered two types of features: individual words (unigrams) and trigrams (sequences of three consecutive words).

After vectorization, we applied normalization to the resulting feature matrix to standardize the data for the machine learning models. L2 normalization scaled each feature vector to have a unit norm (Euclidean distance of 1), while L1 normalization scaled feature vectors such that the sum of their absolute values equaled 1. These normalization steps ensured that the feature magnitudes were comparable across different samples, which is critical for algorithms sensitive to feature scaling, such as SVM and LR.

Finally, the processed feature matrices were used to train several classification models, including SVM, LR, and DT. For SVM and LR, hyperparameters such as the regularization strength and kernel types (for SVM) were tuned using grid search with cross-validation. Decision Trees were optimized by varying tree depth, minimum samples per leaf, and splitting criteria to minimize overfitting and improve generalization. For each baseline

model we performed 30 independent runs with different dataset partitions considering 80% for training and 20% to test. The best results were obtained with the following hyperparameters: LR used Solver Newton-CG, TF-IDF norm L2 representation and max features 10,000; SVM included a sigmoid kernel, TFIDF norm L1 representation; maximum depth of DT was 23 and considered TFIDF norm L2 representation.

We also include models based on deep neural networks such as the Long-Short Term Memory (LSTM)[33] and Convolutional Neural Network (CNN)[34], model architectures are shown in Fig. 2. The input of these models are static dense representations of words, that is, embeddings. After several experiments, we decided to use a combination of our own pre-trained vectors (obtained from signs and symptoms of rare diseases) and others specific for the task to solve in our study (obtained from https://github.com/yao8839836/obesity). We used two embedding layers with the pre-trained word vectors which are concatenated to obtain a 400 dimensional higher-level representation of each input text. After processing the input with the specific architecture of each model, the output is obtained as the result of a fully connected softmax layer to perform the classification using the probability distribution over the output labels (RD vs. common). The LSTM includes a layer with 64 units (Fig. 2A) while the CNN only 4 1D-convolutional layers for extracting 70 filters with different sizes of kernels (varying between 1 and 4) (Fig. 2B). After applying max pooling operation to each feature map, the outputs are concatenated, flatted and passed to a dense 64-unit layer. Previous to the output layer, a dropout operation is performed to reduce overfitting (rate 0.2). The only change introduced to the previous architecture to construct the CNN + LSTM model is the inclusion of a LSTM layer between the concatenation of the 1D-convolutional and the dense layer (Fig. 2C). The number of units in the dense layer and filters is lower (here 50). Finally, the architecture of the LSTM + CNN is obtained by embedding the CNN between the LSTM and the output layer (Fig. 2D). We removed the 64-unit dense layer and the dropout operation to simplify the ended model. We also reduced the size of the filters and the units in the LSTM (10 and 16 respectively).

For each deep neural network architecture the input layer is the text to classify. All architectures are shown in Fig. 2A-D).These architectures were obtained as the best after testing several models with different configurations of hyperparameters.

Large language models are pre-trained using large amounts of data (in an unsupervised way) and usually fine-tuned (in a supervised way) with specific data depending on the task to solve. An example of such models is the Bidirectional Encoder Representations from Transformers[15] (BERT). Unlike the models we proposed before, BERT uses contextualized word representation of the input which feeds several stacked Transformer encoders. We select BERT for our experiments because, beyond its good performance in various NLP tasks, there are several models pre-trained with biomedical texts. We compare a base version of BERT with Bio_ClinicalBERT. The latter is a fine-tuned (with medical conditions data) version of the first domain-specific BERT based model pre-trained on large scale biomedical corpora, named BioBert[35]. The model is available in *sid321axn/Bio_ClinicalBERT-finetuned-medicalcondition · Hugging Face*. BERT was pre-trained on English Wikipedia and General BooksCorpus while Bio_ClinicalBERT, besides the same as BERT, was pre-trained with PubMed Abstracts and PMC Full-text articles (that is, biomedical domain-specific texts) and fine-tuned with clinical conditions of diseases.

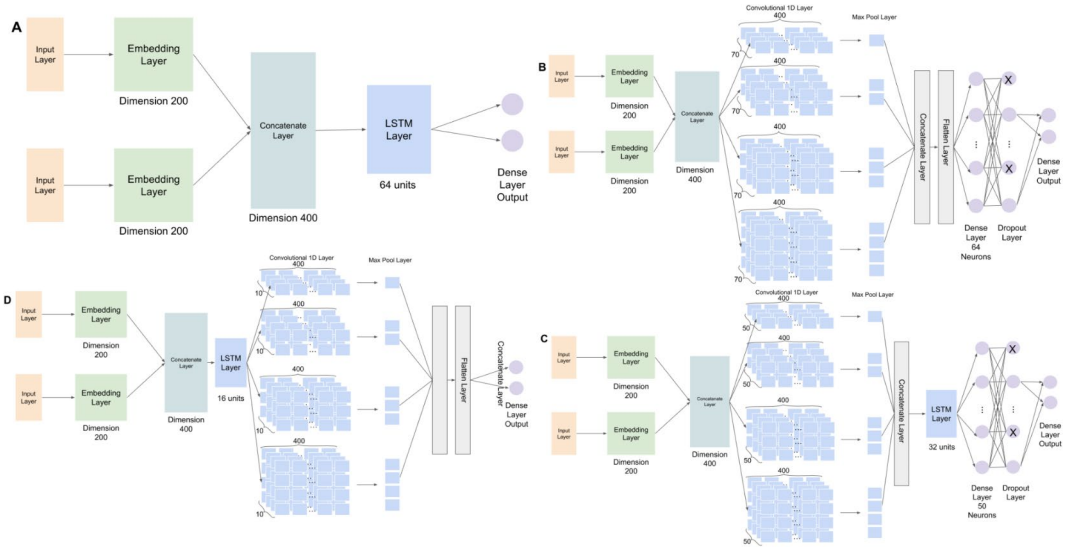

Fig. 2. Model architecture description. Input layer is the medical record to be processed. The output (dense layer) is the probability that the input is RD or common. The architectures vary in number and type of units involved. (**A**). LSTM model. (**B**). CNN model. (**C**). CNN + LSTM model. (**D**). LSTM + CNN model.

	Accuracy	F-measure	AUC
LR	0.923	0.922	0.925
DT	0.888	0.893	0.887
SVM	0.929	0.927	0.960
CNN	0.893	0.906	0.944
LSTM	0.899	0.898	0.952
CNN + LSTM	0.893	0.877	0.922
LSTM + CNN	0.888	0.887	0.920
BioBERT	0.882	0.873	0.881
BERT	0.858	0.852	0.858

Table 2. Best results obtained with each method for the proposed corpus.

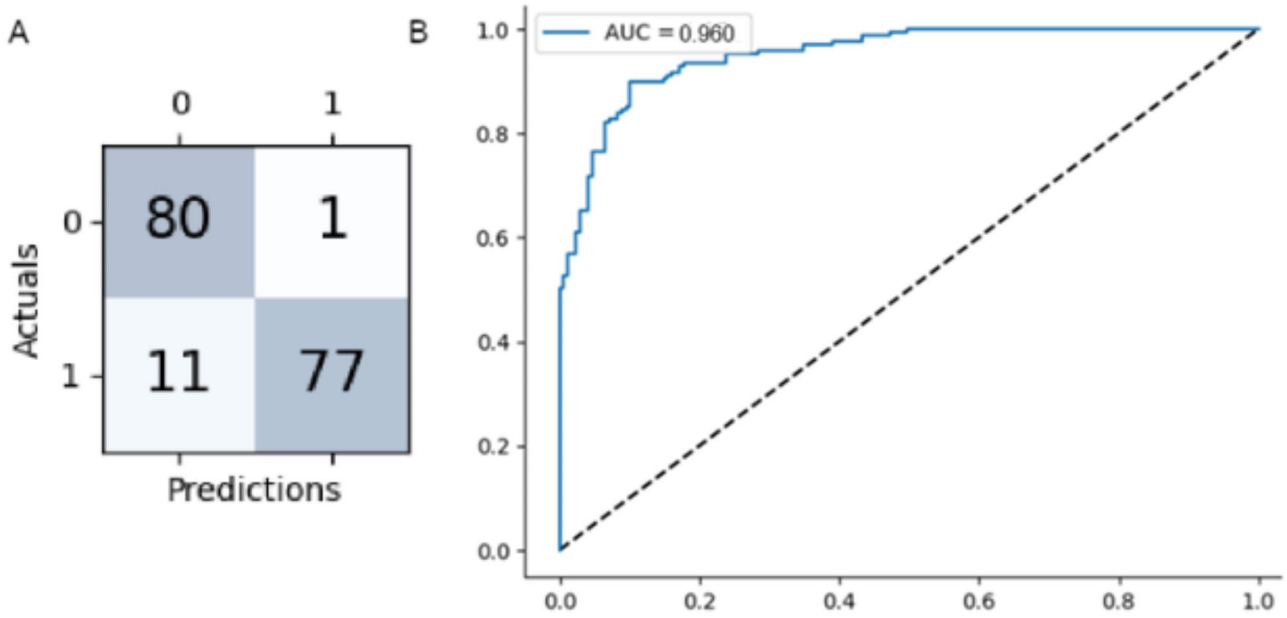

Fig. 3. SVM model metrics. (**A**) confusion matrix with predictions of common diseases (label 0) and RDs (label 1). (**B**) ROC curve.

Results

We analyze the results of the different models implemented and show the one that obtained the best performance in detecting RDs from medical notes. Then, we perform an error analysis on the systematically misclassified records to highlight some characteristics about the proposed corpus.

The SVM better predicts RD patients on discharge summary texts

Table 2 summarizes the results of applying all models and considering the metrics Accuracy (useful when the problem has balanced classes), the standard F-measure and Area Under Curve Receiver Operator Characteristic (AUC) for complementing the evaluation of the performance.

First rows in Table 2 shows for each one of the baseline models the results obtained with the best configuration. The results suggest that SVM outperforms the models obtained with the baseline classifiers but also the advanced ones (LSTM, CNN and BERT) when F-measure is considered (0.927).

Regarding the results of deep models (Table 2, rows 4–7), CNN and LSTM perform similarly with a small difference between metrics: CNN is better than LSTM in F-measure (0.906 and 0.898, respectively) but LSTM is better than CNN in accuracy and AUC (0.899 and 0.893 in accuracy, respectively). It is interesting to note that LSTM outperforms the combination of LSTM and CNN models which would indicate that LSTM can learn long-term dependencies from sequences of higher-level representations without help of convolutional operations. Figure 3A shows the confusion matrix of the best model (SVM). The class 0 corresponds to common disease and class 1 to RD. Only one common disease was misclassified and 11 RDs were wrongly classified as common. Figure 3B the ROC curve obtained from the data.

The last 2 rows of Table 2 show the results obtained with the Transformer-based models. BioBert obtained better performance than BERT considering F-measure (0.873 vs. 0.852) and the same behavior with the other

metrics. The reason is probably because BioBERT was trained with biomedical data. BioBert performs quite similarly to all models, demonstrating that for this particular task, the complexity of transformed-based models is not synonymous with better results.

We chose the AUC metric in model analysis as it provides valuable insights into the model's ability to distinguish between classes and its impact on specific errors like false positives and false negatives, as demonstrated in previous studies[36]. In our case, higher AUC means how better the model is at distinguishing between patients with a RD and not. LSTM, CNN and SVM classifiers obtained the highest AUC (0.95, 0.94 and 0.96, respectively) indicating their ability to correctly classify diagnosis with RDs with relatively small models (in comparison to the transformer-based ones).

Systematically misclassified clinical records

The most frequent error of our model is that RDs are misclassified as common diseases. In almost all models (baseline and not) a median of ~87.25% of all misclassified records correspond to RD wrongly classified as common. A median of 80.95% of misclassified RD corresponds to MIMIC-III clinical records that were previously classified by another study[29]. The rest (mostly) correspond to common diseases that were translated from Spanish. Focusing on the SVM classification, all misclassified records (11) were in fact RDs (a RD classified as common); of the 11 texts "misclassified" 10 of them correspond to the MIMIC-III discharge summaries labeled as RD by Dong et al.[29]. After careful consideration of an expert we found that most of them were actually common diseases, but mislabeled as RDs in the corpus since the clinical records describe a large amount of complications of common diseases probably in elderly patients. For instance, aortic dilation though rare, is a complication of thoracic aortic aneurysm (TAA), a relatively common condition in older people. Similarly, cases involving multiple common diseases, such as chronic pancreatitis, chronic kidney disease, or complications from alcoholism, can mimic rare diseases due to overlapping symptoms. Infections in immunocompromised patients, such as those with HIV, also illustrate this issue, as rare pathogens may complicate a common condition. Hence, the clinical text becomes very long, complicated with several interactions with procedures, medical specialties staff, drugs, interventions, and so on, which might be an explanation of the misclassification. The expert noted that the complexity of these cases, combined with messy medical records and overlapping features, leads to these errors. Training data correctness for building the model is an important issue to be considered because wrong data for training derive in a wrong model to test. Even though the corpus might have some noise regarding the labels (which is a realistic scenario in the context of several applications) the classifiers are able to perform fairly well in practice.

Discussion

RDs are difficult to detect, to diagnose and to treat. Patients with RDs have to navigate the healthcare system patiently, inefficiently and with economical (and psychological) costs. Timely diagnosis, hence early strategies to assess the disease, might be of great importance to control the impact on the patient and the family.

The diagnosis pipeline of RDs is different from other diseases and very frequently includes consultation with geneticists and molecular studies for proper diagnosis. An early detection of the presence of a RD might be a substantial improvement in many cases. Because we daily observe the positive impact of the application of machine-learning-based technologies in many areas, we think that the detection of RDs could benefit from these.

Our method aims to timely detect RD patients from medical records (discharge summaries) obtained from many sources, specially from an emergency unit. When the discharge summary of a patient classifies as a potential RD a flag could be raised in the hospital system and measurements could be set in place, such as consultation with geneticist and other specialists, molecular analysis, improving the time until diagnosis. Figure 4 illustrates the diagnosis pipeline in a medical center, involving ML models to help the early detection of RDs. After admission, medical notes will be processed with our best model (SVM, for example) trained with the proposed corpus. The answer will determine if there is a warning for a possible RD, then the patient should be referred to the geneticist to deepen with further studies. Otherwise, the patient will be treated as suffering a common disease, saving thus, valuable and expensive resources involved in the analysis of rare diseases. A pseudocode of the complete process can be observed in Supplementary material (Methods Sect. 2).

Our best ML model (SVM) achieved an AUC of 0.96 and F-measure of 0.927, while previous studies based on similar data (clinical records) such as RareBert[30] reported an AUC of approximately 0.80 and Dong et al.[29] achieved a F-measure value of 0.702. Although the training and test data are not the same (Pakash et al.[30] used medical notes with particular comorbid conditions such as disorder of phosphorus metabolism, rickets, muscle weakness and bone spurs, while Dong et al.[29] considered some MIMIC-III discharge summaries), these metrics give us an idea of how different models perform in detecting RDs. Our dataset comprises a variety of diseases collected from different sources providing a more general corpus for ML methods. Regarding the performance of SVM, the good values for the metrics were obtained thanks to the well-tuned configurations and preprocessing techniques used for the proposed model. We considered a sigmoid kernel, which effectively captured non-linear relationships in the input data by transforming the input space into a more separable feature space. Additionally, the TF-IDF vectorizer with L1 normalization was employed to vectorize the input data of the model, ensuring that features are uniformly scaled and the model remains stable. During this process we also removed English stop words, further refining the input data by eliminating irrelevant words. Another configuration that was fine-tuned was the regularization parameter C, which was set to 2, striking an optimal balance between underfitting and overfitting ensuring the model generalized well to unseen data. These configurations, along with precise preprocessing, enabled the SVM model to achieve a good accuracy (92.9%) and strong balance between precision and recall while maintaining high discriminative power, as reflected in the AUC score.

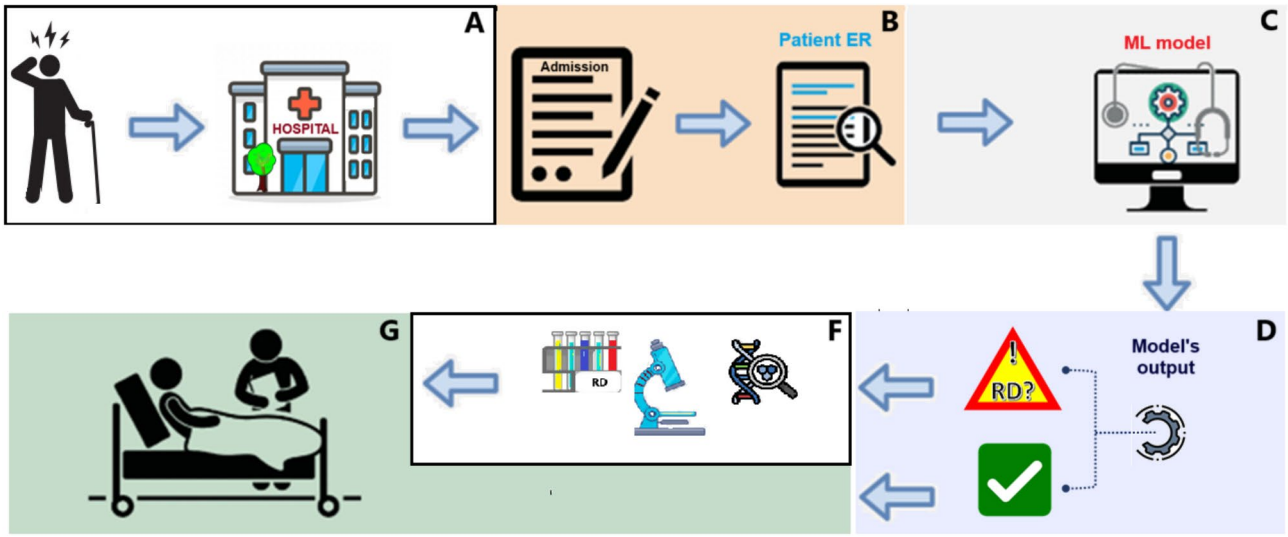

Fig. 4. Diagnosis pipeline of RDs. (**A**) Sick person goes to the hospital. (**B**) After admission, the patient is assigned an electronic record (ER). (**C**) Machine learning (ML) model is executed with the ER to find possible RD in medical notes. (**D**) If the MLmodel detects signs of RD, the patient is referred to specialists to confirm or reject the fact (**F**,**G**). After the patient has been diagnosed, he or she receives appropriate treatment.

An accuracy around 90% implies that in 90% of the cases the classification is correct, and the RD flag should be raised. The remaining 10% of cases correspond to individuals that have a common disease but they were classified as RD. The impact of such an error would be mostly an economical loss for a public hospital, since additional unnecessary consultations and/or laboratory examination might be done, however, the savings generated by the remaining 90% most likely outweigh the cost of these additional consultations. Besides, upon manual reexamination most of these false calls would be easily detected. On the other hand, an error misclassifying RD as common would have a higher impact on patients well-being (patients would go through the standard algorithmic path, hence probably a diagnostic odyssey) and costs would be even bigger.

Future work relies on the fine-tuning of models that are already close to the clinical aspects of the problem, such as Bio_ClinicalBERT (freely available) or even generative open-source large language models such as LLama2, with this corpus. The idea is to use a partition of our proposed dataset to obtain models performing better for the specific tasks of detection RD clinical notes. This will be achieved by adding a classification-RD-specific layer after the encoder/decoder stack for adjusting the model to our data and thus, improving our results by better understanding the technical words and their contexts.

Also, the tested models are not strong in the explicativeness. In some models, we are not able to understand why a specific clinical record or discharge summary is classified as RD. Understanding the results of the classification process would improve our knowledge on RDs in general, and also, how to write discharge summaries so that models would work properly.

Additionally, we intend to expand the corpus by adding several case reports that we are currently generating in the context of new projects (spin-off of previous URUGENOMES project, where we are analyzing hundreds of RD patients in the next few years). Common and possible additional rare diseases anonymized clinical records are planned to be obtained from different services from a University public hospital in Uruguay.

The inclusion of more reliable RD clinical records and manual curation of those already included, are going to improve downstream results.

Finally, we believe that this valuable corpus is in line with the trend of few-shot learning for classification above all in the biomedical domain and we would try other transformer-based methods for few-shot identification of rare diseases.

Conclusions

We presented a corpus for the classification of rare diseases from clinical notes. We showed a detailed exploratory analysis of the data collected and concluded with a balanced dataset with a similar number of notes labeled as containing RD or not.

To test the proposed resource, we performed a comparative study of different models for the classification of rare diseases, the classical ones SVM, logistic regression and decision tree, the artificial neural networks LSTM and CNN and, the recent transformer-based BERT. SVM performs the best with a F-measure of 0.927.

Thus, we conclude that the SVM-based model is able to accurately predict rare diseases based on the clinical record of the patients, hence enabling the possibility to be included as a warning and a lead to a more accurate diagnostic path.

By making the corpus available we encourage future applications to be developed and refined. In addition to helping mitigate the lack of annotated data for the identification of RDs, this resource can be safely used for few-shot machine learning algorithms in classification as well as other tasks.

Data availability

The complete corpus that support the findings of this study is freely available under https://sites.google.com/view/leticia-cagnina/research.

Received: 24 July 2024; Accepted: 13 February 2025
Published online: 26 February 2025

References

1. *The Voice of 12,000 Patients. Experiences and Expectations of Rare Disease Patients on Diagnosis and Care in Europe*. (EURORDIS - Rare Diseases Eu, 2009).
2. Sireau, N. *Rare Diseases: Challenges and Opportunities for Social Entrepreneurs* (Routledge, 2017).
3. Yan, X., He, S. & Dong, D. Determining how far an adult rare disease patient needs to travel for a definitive diagnosis: A cross-sectional examination of the 2018 National Rare Disease Survey in China. *Int. J. Environ. Res. Public Health* **17**, (2020).
4. Raggio, V. et al. Whole genome sequencing reveals a frameshift mutation and a large deletion in YY1AP1 in a girl with a panvascular artery disease. *Hum. Genomics* **15**, 28 (2021).
5. Meyer, E. J. et al. CBG Montevideo: A clinically novel mutation leading to haploinsufficiency of corticosteroid-binding globulin. *J. Endocr. Soc.* **5**, bvab115 (2021).
6. Della Mina, E. et al. Improving molecular diagnosis in epilepsy by a dedicated high-throughput sequencing platform. *Eur. J. Hum. Genet.* **23**, 354–362 (2015).
7. Liu, H.-Y. et al. Diagnostic and clinical utility of whole genome sequencing in a cohort of undiagnosed Chinese families with rare diseases. *Sci. Rep.* **9**, 19365 (2019).
8. Clark, M. M. et al. Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. *NPJ Genom. Med.* **3**, 16 (2018).
9. Esteva, A. et al. A guide to deep learning in healthcare. *Nat. Med.* **25**, 24–29 (2019).
10. Hwang, S. & Lee, B. Machine learning-based prediction of critical illness in children visiting the emergency department. *PLoS One* **17**, e0264184 (2022).
11. Hatachi, T. et al. Machine learning-based prediction of hospital admission among children in an emergency care center. *Pediatr. Emerg. Care* **39**, 80–86 (2023).
12. Gulshan, V. et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* **316**, 2402–2410 (2016).
13. Golden, J. A. Deep learning algorithms for detection of lymph node metastases from breast cancer: Helping artificial intelligence be seen. *JAMA* **318**, 2184–2186 (2017).
14. Doshi-Velez, F., Ge, Y. & Kohane, I. Comorbidity clusters in autism spectrum disorders: An electronic health record time-series analysis. *Pediatrics* **133**, e54-63 (2014).
15. Kenton, J. D. M. W. C. & Lee, K. T. BERT: Pre-training of deep bidirectional transformers for language understanding. in *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.4171–4186.
16. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. Language models are few-shot learners. (2020).
17. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. & Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **140**, (2020).
18. Johnson, A. E. W. et al. MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 160035 (2016).
19. Abidin, A. Z., Jameson, J., Molthen, R. & Wismüller, A. Classification of micro-CT images using 3D characterization of bone canal patterns in human osteogenesis imperfecta. *Proc. SPIE* **10134**, (2017).
20. Mahmood, T. et al. Breast lesions classifications of mammographic images using a deep convolutional neural network-based approach. *PLoS One* **17**, e0263126 (2022).
21. *Cancer Unveiled: A Deep Dive Into Breast Tumor Detection Using Cutting-Edge Deep Learning Models*.
22. Masumoto, H. et al. Accuracy of a deep convolutional neural network in detection of retinitis pigmentosa on ultrawide-field images. *PeerJ* **7**, e6900 (2019).
23. Bilal, A. et al. Improved support vector machine based on CNN-SVD for vision-threatening diabetic retinopathy detection and classification. *PLoS One* **19**, e0295951 (2024).
24. Bilal, A., Liu, X., Shafiq, M., Ahmed, Z. & Long, H. NIMEQ-SACNet: A novel self-attention precision medicine model for vision-threatening diabetic retinopathy using image data. *Comput. Biol. Med.* **171**, 108099 (2024).
25. Bilal, A., Liu, X., Baig, T. I., Long, H. & Shafiq, M. EdgeSVDNet: 5G-enabled detection and classification of vision-threatening diabetic retinopathy in retinal fundus images. *Electronics (Basel)* **12**, 4094 (2023).
26. Jia, J. et al. RDAD: A machine learning system to support phenotype-based rare disease diagnosis. *Front. Genet.* **9**, 587 (2018).
27. *Diagnostic Support for Selected Paediatric Pulmonary Diseases Using Answer-Pattern Recognition in Questionnaires Based on Combined Data Mining Applications-A Monocentric Observational*.
28. Hersh, W. R., Cohen, A. M., Nguyen, M. M., Bensching, K. L. & Deloughery, T. G. Clinical study applying machine learning to detect a rare disease: Results and lessons learned. *JAMIA Open* **5**, ooac053 (2022).
29. Dong, H. et al. Rare disease identification from clinical notes with ontologies and weak supervision. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* **2021**, 2294–2298 (2021).
30. Prakash, P. K. S., Chilukuri, S., Ranade, N. & Viswanathan, S. RareBERT: Transformer architecture for rare disease patient identification using administrative claims. *Proc. Conf. AAAI Artif. Intell.* **35**, 453–460 (2021).
31. Bello, F. L., Naya, H., Raggio, V. & Rosá, A. From medical records to research papers: A literature analysis pipeline for supporting medical genomic diagnosis processes. *Inform. Med. Unlocked* **15**, 100181 (2019).
32. Spangenberg, L. et al. Novel frameshift mutation in PURA gene causes severe encephalopathy of unclear cause. *Mol. Genet. Genomic Med.* **9**, e1622 (2021).
33. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780.
34. LeCun, Y. et al. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **1**, 541–551.
35. Lee, J. et al. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).
36. Fawcett, P. F. Robust classification for imprecise environments. **42**, 203–231.

Acknowledgements

This study was partially funded by BID (Banco Iberomericano de desarrollo) in the context of the URUGENOMES Project (Proyecto ATN/KK-L4584-"Fortalecimiento de las capacidades técnicas y humanas para las exportaciones de servicios genómicos"). Additionally, support was obtained from the CONICET, Short Research

Stages program given to Leticia Cagnina. PEDECIBA under Grant Number: AlicuotasINNOVA II under Grant Number: DCI-ALA /2011/23-502.

Author contributions

MR: Data curation, formal analysis, Roles/Writing - original draftVR: Data curation, Writing - review & editingHN: Validation, Methodology, Writing - review & editingLC: Supervision, Funding acquisition, Roles/Writing - original draft, MethodologyLS: Supervision, Roles/Writing - original draft, Project administration, Data curation.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-90450-0.

Correspondence and requests for materials should be addressed to L.S. or L.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.