Date of publication 2024 10, date of current version 2024 09.
Digital Object Identifier 10.1109/ACCESS.2017.DOI

Model to prediction of diagnosis-related groups in the public hospital in San Bernardo

S. CHAVEZ[1], ()
[1]

Corresponding author: Chavez S. Author (e-mail: sebastian.chavez@gmail.com)
https://github.com/schavezh/UNAB_pronostico_consumo_electrico/tree/main/proyecto

ABSTRACT In this work, we focus on predicting Diagnosis-Related Groups (DRG) in a public hospital setting using advanced machine learning techniques. DRGs play a crucial role in hospital management, grouping patients based on their diagnoses and the resources required for treatment, facilitating better cost management and resource allocation. The dataset includes multiple diagnoses, procedures, and demographic details from over 14,000 patients. We implemented three predictive models using Long Short-Term Memory (LSTM) networks and embeddings, exploring various architectures. The first model processed all diagnoses and procedures in combined vectors, while the second model separated primary and secondary diagnoses and procedures before concatenating their respective embeddings. The third model incorporated Word2Vec to capture semantic relationships in the clinical text data, improving prediction accuracy. Results showed that the third model outperformed the others, achieving a weighted F1-score of 0.9259. The combination of LSTM, Word2Vec, and embeddings allowed the model to capture both temporal sequences and semantic relationships effectively. This approach demonstrates the potential of machine learning in predicting DRGs, optimizing hospital resources, and improving decision-making processes in public health institutions.

INDEX TERMS DRG, LSTM, recurrent neural networks (RNN), diagnosis-related groups

I. INTRODUCTION

DIAGNOSIS-Related Groups (DRGs) are a fundamental tool in hospital administration at a global level, including Chile, where they have been adopted to improve efficiency in the use of public resources. This system groups patients with similar clinical characteristics, allowing for better planning and control of hospital costs, which has been key in the management of public health in several countries. In Chile, its implementation began to deepen in 2018-2020, with the aim of encouraging a more cost-effective use of available resources, advancing significantly in recent years in high and medium complexity hospitals. [1]

At an international level, DRGs are used in health systems such as those in the United States and Europe, where they have proven to be effective in containing costs and promoting better clinical practices. [2] The American DRG system has played a crucial role in resource planning and hospital management, allowing hospitals to anticipate their needs and optimize treatments according to assigned diagnoses. Furthermore, advances in the use of artificial intelligence, such as natural language processing and machine learning models, have made it possible to predict DRGs more accurately, improving hospital management and reducing resource allocation times. [3]

In Chile, the DRG system has not only made it possible to improve clinical and financial management in public hospitals, but has also facilitated transparency and cost comparability between different institutions, promoting more effective and results-oriented hospital management. [1]

II. RELATED WORK

In recent work, the prediction of Diagnosis-Related Groups (DRGs) has been addressed with advanced natural language processing (NLP) techniques and machine learning (ML) models in order to optimize resource allocation in hospitals and improve the accuracy of patient classification.

The work of Liu et al. (2021) explored early prediction of DRGs using a deep learning-based NLP model applied to clinical texts, such as hospital admission notes. This approach allows patients to be classified into DRG groups before discharge, which represents an improvement over traditional methods that rely on post-discharge analysis. Using convolutional attention-based models, the authors were able to predict DRGs with an average AUC greater than 0.86 in two different cohorts (MS-DRG and APR-DRG), highlighting their ability to estimate hospital costs efficiently and early. Furthermore, they observed that using large amounts of textual clinical data allows identifying important patterns for classification, reducing the administrative burden of hospitals and facilitating better decision-making during hospitalization. [3]

On the other hand, Wang et al. (2024) presented the DRG-LLaMA model, an adjusted variant of LLaMA, optimized for DRG prediction using hospital discharge summaries. Using low-ranking adaptation (LoRA) techniques, they trained the model on a large amount of clinical texts from the MIMIC-IV database. This approach outperformed previous models, such as ClinicalBERT and CAML, achieving a top-1 prediction accuracy of 52% and an average macro AUC of 0.986. The findings highlight that the model not only improves the accuracy in predicting DRGs, but is also able to identify patients with complications or comorbidities, providing a more robust and accurate classification (Wang et al., 2024). [2]

Both works demonstrate that the use of NLP and deep learning in DRG prediction can significantly reduce the time and errors associated with traditional methods, improving both operational efficiency and hospital management capacity.

III. STUDY OBJETIVE

The objective of this study is to develop a predictive model for the classification of Diagnosis Related Groups (DRG) in a public hospital in the metropolitan region of Santiago, Chile. Early prediction of DRGs is crucial to optimize resource allocation and improve budget management in hospitals, allowing to foresee the costs associated with patient treatment and improve operational efficiency. In addition, this model aims to predict the length of stay of patients in emergency rooms, which is essential for bed planning and continuous care in a high-demand context. Through the use of machine learning algorithms and historical clinical data, we seek to generate a system that supports real-time decision making, optimizing the quality of care and the efficiency of hospital resources.

IV. METHODOLOGY

A. DATASET DESCRIPTION

The dataset corresponds to two files with CSV extension, which contain training and test data separately.

The dataset to be used is a CSV file with 14,561 observations of patients who were in the hospital. The fields correspond to 35 diagnoses and 30 clinical procedures, and also contain two demographic data corresponding to the age and gender of the patient, ending with the assigned DRG code.

B. METHODOLOGY FOR DEVELOPMENT

In this work, the CRISP-DM methodology will be taken as a baseline, considering that we have a clear objective and a defined data set and we will carry out the following stages: Understanding the business, Understanding the data, Data preparation, Develop model, Evaluate models and Deployment [7].

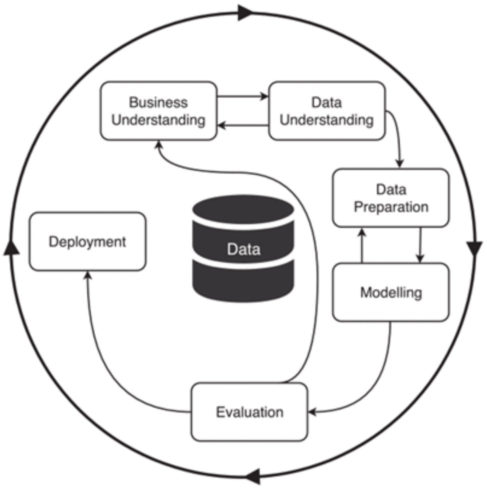


FIGURE 1. CRISP-DM data mining process model.

1) Business understanding

Diagnosis-Related Groups (DRG) prediction is of great importance both internationally and in Chile, as it is closely linked to the optimization of hospital resources, budget planning, and improvement in the quality of care.

Internationally, DRGs are widely used in countries such as the United States, Australia, and various European nations as a key mechanism in pay-for-performance systems. This system groups patients into categories based on their diagnosis and the resources they are likely to use, allowing hospitals to predict costs and improve efficiency in the administration of treatments. [1] In particular, early prediction of DRGs using artificial intelligence models has proven to be an effective tool to reduce operational costs and improve resource planning, which is vital in contexts of high hospital demand. [2] Accurate prediction of DRGs allows hospitals to adjust their operations, anticipate complications, and make informed decisions in real time, increasing transparency and efficiency in the use of hospital resources.

In Chile, the implementation of DRGs as a payment mechanism in public hospitals has been a significant advance in hospital management since 2018-2020. [1] Its introduction seeks to replace traditional financing methods that were based on historical budgets and payments for individual services. The ability to predict DRGs in Chile allows public hospitals to improve their control over hospital spending, optimize the use of beds and medical resources, and ensure quality care while reducing accumulated hospital debt. In addition, the DRG system facilitates comparison between different hospitals and promotes transparency in the allocation of funds, which is essential for the sustainability of the Chilean public health system. [1]

2) Data Preparation

Adding number of diagnoses and procedures: The number of secondary diagnoses and secondary procedures for each record is counted, storing this information in new columns. This process is important because diagnoses and procedures are key factors in the classification of DRGs. This approach is aligned with works such as Liu et al. (2021), which highlight the importance of structured clinical data to predict DRGs and calculate hospital costs. Additional diagnoses influence the complexity of the case and therefore the DRG category.

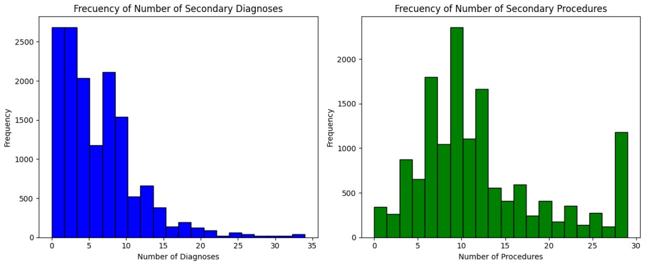


FIGURE 2. Frequency of secondary diagnoses and procedures

Creating a summary table by DRG: This step consists of calculating the cumulative percentage, specific percentage, and number of records for each DRG code in relation to the total number of records. This type of analysis is useful for understanding the distribution of DRGs in the dataset and was used in works such as Wang et al. (2024), who also classified DRGs by their frequency to analyze the performance of predictive models and optimize the use of hospital resources.

Data filtering by cumulative percentage: It is observed that the DRGs that represent 75% of the cases contain information from 97 different types of DRGs, which is a limited universe of codes that facilitates processing and improves the performance of the model by having less complexity. This dimensionality reduction technique is useful to simplify the model and focus the analysis on the most common DRGs, which was a similar approach in previous studies to improve the accuracy of predictive models in multi-category classification tasks, as suggested by Wang et al. (2024).

Cleaning of diagnosis and procedure codes: A cleaning of the diagnosis and procedure columns was applied, extracting only the codes and eliminating the descriptions. This action is crucial in models that use coded clinical data, as indicated by Liu et al. (2021), since standardized codes (such as ICD codes) are essential for the accurate classification of DRGs.

These preprocesses are aimed at optimizing the performance of the DRG predictive model. Counting diagnoses and procedures and cleaning codes allow the model to work with relevant and standardized information. Furthermore, the selection of more frequent DRGs facilitates more efficient prediction, which has also been highlighted in the literature as a crucial step to improve the accuracy and scalability of predictive models in the hospital context (Liu et al., 2021; Wang et al., 2024).

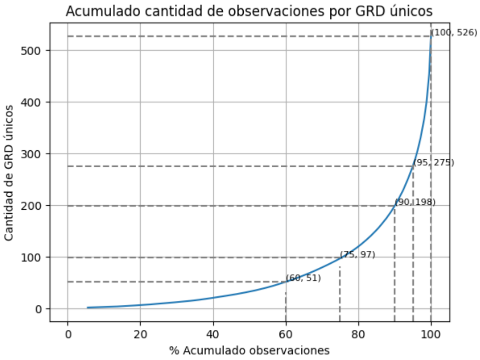


FIGURE 3. Enter Caption

3) Modeling

LSTM (Long Short-Term Memory): LSTM is a variant of recurrent neural networks (RNN) that is widely used for data sequences due to its ability to capture long-term dependencies. In the context of DRGs, LSTM is useful for processing sequences of medical events such as previous hospitalizations, diagnoses, and treatments. This architecture has been successfully employed to predict length of hospital stay, demonstrating good performance in predicting clinical outcomes in previous studies. [6] The main strength of LSTM is its ability to handle temporal dependency in data, which is crucial in predictive models such as DRG prediction, where medical history influences the final diagnosis. However, its weakness lies in its computational demand and the risk of overfitting when handling a large volume of complex clinical data. [3]

Word2Vec: It is a vector representation technique for words, which transforms words into feature vectors, allowing models to identify semantic patterns in medical texts. In our case, it can be used to analyze clinical notes and other descriptive texts that accompany diagnoses, as well as diagnosis and procedure codes along with their description. Word2Vecś ability to capture semantic relationships has been proven in natural language processing applications in the healthcare field, improving the extraction of relevant information. [7] Its main strength is the ability to efficiently learn semantic relationships with large textual data sets, which complements the temporal structure captured by the LSTM. However, its weakness is that it requires a significant amount of labeled textual data and its training can be expensive in terms of time and computational resources. [2]

Embeddings: In general, whether based on Word2Vec or on more recent models, such as context embeddings generated by LLMs (Large Language Models), they are essential to reduce dimensionality and preserve semantic information in unstructured clinical data. These have proven useful in multiple clinical studies, where they are used to classify medical texts, improve the accuracy of diagnoses and reduce errors. [8] Embeddings are particularly useful for handling sparse textual data, such as discharge notes, where they can identify hidden relationships between diagnoses and treatments. One limitation is that, although they are effective for handling texts, they do not always capture the full complexity of more structured clinical data, which requires their integration with other models such as LSTM to obtain better results. [9]

LSTM in concatenation with embedding: By using LSTM, important sequences of medical events are retained, which is key for tasks such as DRG prediction, where past events influence future ones, while embeddings allow the model to have a richer representation of medical concepts, which improves the understanding of clinical text or its description. By combining networks specialized in different aspects of the data (temporal sequences and textual semantics), the model's ability to generate more accurate and robust predictions in complex clinical scenarios is improved [9]. The combination of these components allows to take advantage of the strengths of each LSTM to capture temporal relationships, Word2Vec to process medical text and embeddings to reduce the dimensionality of the feature space.

4) Evaluation

Precision: Precision measures the proportion of true positives (correct predictions of a DRG) out of the total number of positive predictions (true positives + false positives). Precision is critical when the cost of false positives is high. In the context of DRG prediction, a false positive means that the model incorrectly predicted a DRG that does not correspond to the patient, which could result in inappropriate treatment or misuse of hospital resources. Therefore, having high precision is important to minimize these errors, especially in medical triage where false positives can lead to misdiagnosis or unnecessary treatments [6].

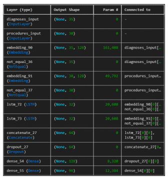
Recall: Recall measures the proportion of true positives correctly identified out of all true positive cases (true positives /+ false negatives). In a DRG prediction system, recall is critical when it is crucial to detect all positive cases, i.e. when you do not want the model to miss patients with certain diagnoses. In the hospital context, false negatives(when the model fails to identify a real DRG) can be more dangerous, as it means that the patient would not receive the appropriate treatment. A model with high recall ensures that these errors are minimized, which is vital to providing the correct medical care [7].

F1-score: The F1-score is the harmonic mean between precision and recall, and provides a balanced measure of both metrics. The F1-score is particularly useful when there is an imbalance between classes in the dataset, as may be the case for DRGs. In many cases, certain categories of diagnoses may be underrepresented, making metrics such as precision or recall alone insufficient to assess performance. The F1-score allows you to find a balance between the two. If the model has high precision but low recall, or vice versa, the F1-score will be low, indicating the need to improve that balance [2].

V. EXPERIMENTS (RESULTS AND THEIR DISCUSSION)

A. DIAGNOSIS + PROCEDURES + EMBEDDING + LSTM

In this first experiment, the dataset filtered to 75% of the records with the most frequent GRDs was used. All the diagnoses (including the main one) were taken and placed in a vector to later transform them into numerical data. In the case of the procedures, all their dimensions were also taken and placed in a vector. After this, they were passed through a model with the following architecture:



```
Layer (type)             Output Shape        Param #   Connected to
diagnoses_input          (None, 35)                0   -
(InputLayer)
procedures_input         (None, 30)                0   -
(InputLayer)
embedding_90             (None, 35, 128)     161,408   diagnoses_input[…
(Embedding)
not_equal_36             (None, 35)                0   diagnoses_input[…
(NotEqual)
embedding_91             (None, 30, 128)      49,792   procedures_input…
(Embedding)
not_equal_37             (None, 30)                0   procedures_input…
(NotEqual)
lstm_72 (LSTM)           (None, 32)           20,608   embedding_90[0][…
                                                       not_equal_36[0][…
lstm_73 (LSTM)           (None, 32)           20,608   embedding_91[0][…
                                                       not_equal_37[0][…
concatenate_27           (None, 64)                0   lstm_72[0][0],
(Concatenate)                                          lstm_73[0][0]
dropout_27               (None, 64)                0   concatenate_27[0…
(Dropout)
dense_54 (Dense)         (None, 128)           8,320   dropout_27[0][0]
dense_55 (Dense)         (None, 96)           12,384   dense_54[0][0]
```

FIGURE 4. Model 1

The overall metrics indicate moderately poor performance in terms of overall classification.

Metric	Valor
Presicion (weighted)	0.444903
Recall (weighted)	0.483974
Fscore (weighted)	0.444434

TABLE 1. Metrics results of Model 1

The F1-Score per Class graph shows that out of the 96 total classes, 89 are below the 0.8 threshold, indicating that most classes are being classified with a low F1-Score and the loss graph shows typical overfitting behavior.

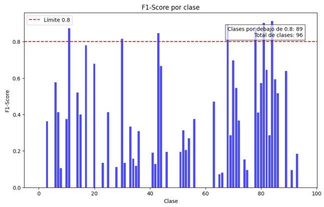
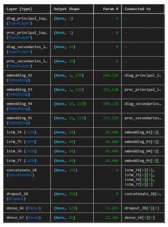


FIGURE 5. Fscore by class model 1

B. PRIMARY DIAGNOSIS + SECONDARY DIAGNOSIS + PRIMARY PROCEDURE + SECONDARY PROCEDURES + EMBEDDING + LSTM

In this second experiment, we keep 75% of the data but separate the main diagnosis and main procedure from the secondary ones to enter them independently into an embedding and then into an LSTM layer to finally concatenate all the vectors.



Layer (type)	Output Shape	Param #	Connected to
diag_principal_inp… (InputLayer)	(None, 1)	0	-
proc_principal_inp… (InputLayer)	(None, 1)	0	-
diag_secundarios_i… (InputLayer)	(None, 34)	0	-
proc_secundarios_i… (InputLayer)	(None, 29)	0	-
embedding_92 (Embedding)	(None, 1, 128)	640,128	diag_principal_i…
embedding_93 (Embedding)	(None, 1, 128)	256,128	proc_principal_i…
embedding_94 (Embedding)	(None, 34, 128)	640,128	diag_secundarios…
embedding_95 (Embedding)	(None, 29, 128)	256,128	proc_secundarios…
lstm_74 (LSTM)	(None, 64)	49,408	embedding_92[0][…
lstm_75 (LSTM)	(None, 64)	49,408	embedding_93[0][…
lstm_76 (LSTM)	(None, 64)	49,408	embedding_94[0][…
lstm_77 (LSTM)	(None, 64)	49,408	embedding_95[0][…
concatenate_28 (Concatenate)	(None, 256)	0	lstm_74[0][0], lstm_75[0][0], lstm_76[0][0], lstm_77[0][0]
dropout_28 (Dropout)	(None, 256)	0	concatenate_28[0…
dense_56 (Dense)	(None, 128)	32,896	dropout_28[0][0]
dense_57 (Dense)	(None, 96)	12,384	dense_56[0][0]

FIGURE 6. Model 2

Comparing these results to Model 1, this second model performs much better. The metrics have considerably higher values, indicating that the model is handling both frequent and less frequent classes better.

Metric	Valor
Presicion (weighted)	0.782863
Recall (weighted)	0.778846
Fscore (weighted)	0.770883

TABLE 2. Metrics results of Model 2

Higher F1-Score suggests that there is a good balance between precision and recall for most classes, however, it shows that 52 out of 96 classes have an F1-Score below the 0.8 cutoff indicating that there are still many classes that are below the threshold.

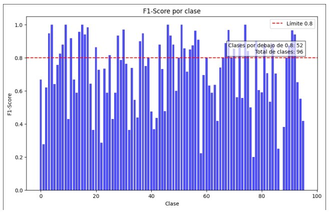
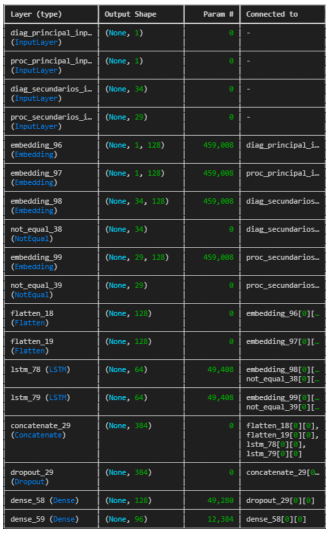


FIGURE 7. Fscore by class model 2

C. PRIMARY DIAGNOSIS + SECONDARY DIAGNOSIS + PRIMARY PROCEDURE + SECONDARY PROCEDURES + EMBEDDING + WORD2VEC + LSTM

In the third model, LSTM, embeddings and Word2Vec were implemented due to their ability to handle sequential and textual data in the clinical setting. In previous studies, the use of LSTM proved to be highly effective in predicting hospital stays based on comorbidity networks, capturing temporal information about patients [6]. On the other hand, embeddings provide vector representations of medical concepts, allowing the model to relate diagnoses and treatments semantically and finally, Word2Vec is used to transform clinical text into numerical vectors, which facilitates its integration into networks such as LSTM. This approach has been successful in extracting information from medical records, improving the accuracy of predictive models in the healthcare context [7]. Furthermore, the combined use of embeddings and LSTM allows handling both unstructured text and temporal sequences, optimizing prediction in complex clinical settings [2].



Layer (type)	Output Shape	Param #	Connected to
diag_principal_inp… (InputLayer)	(None, 1)	0	-
proc_principal_inp… (InputLayer)	(None, 1)	0	-
diag_secundarios_i… (InputLayer)	(None, 34)	0	-
proc_secundarios_i… (InputLayer)	(None, 29)	0	-
embedding_96 (Embedding)	(None, 1, 128)	459,008	diag_principal_i…
embedding_97 (Embedding)	(None, 1, 128)	459,008	proc_principal_i…
embedding_98 (Embedding)	(None, 34, 128)	459,008	diag_secundarios…
not_equal_38 (NotEqual)	(None, 34)	0	diag_secundarios…
embedding_99 (Embedding)	(None, 29, 128)	459,008	proc_secundarios…
not_equal_39 (NotEqual)	(None, 29)	0	proc_secundarios…
flatten_18 (Flatten)	(None, 128)	0	embedding_96[0][…
flatten_19 (Flatten)	(None, 128)	0	embedding_97[0][…
lstm_78 (LSTM)	(None, 64)	49,408	embedding_98[0][… not_equal_38[0][…
lstm_79 (LSTM)	(None, 64)	49,408	embedding_99[0][… not_equal_39[0][…
concatenate_29 (Concatenate)	(None, 384)	0	flatten_18[0][0], flatten_19[0][0], lstm_78[0][0], lstm_79[0][0]
dropout_29 (Dropout)	(None, 384)	0	concatenate_29[0…
dense_58 (Dense)	(None, 128)	49,280	dropout_29[0][0]
dense_59 (Dense)	(None, 96)	12,384	dense_58[0][0]

FIGURE 8. Model 3

Finally we can see a comparison of the metrics of the 3 implemented models, with model 3 being the one that achieved the best performance with 75% of the GRD data with the greatest number of observations.

Métrica	Model 1	Model 2	Model 3
Presicion (weighted)	0.444903	0.782863	0.930939
Recall (weighted)	0.483974	0.778846	0.926740
Fscore (weighted)	0.444434	0.770883	0.925922

TABLE 3. Comparison of the metrics of the 3 implemented model

The F1-Score per class graph shows that only 14 out of 96 classes are below the 0.8 cutoff, which is a clear improvement over previous models. Most classes have an F1-Score above 0.8, showing that the model is accurately classifying a large portion of the dataset. This significant improvement suggests that the model has efficiently captured the relationships between diagnoses and procedures, likely thanks to the combination of LSTM and Word2Vec, which allows for a better handling of the sequence and interrelationships of clinical events.

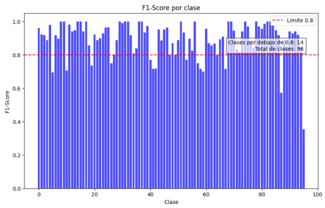


FIGURE 9. Fscore by class model 3

The loss plot shows a very stable and consistent fit. From epoch 5 onwards, both the validation and training loss converge and stabilize, indicating that the model is not overfitting. This suggests an excellent generalization ability of the model. The validation loss does not have large increases, which is indicative of a well-fitted model.

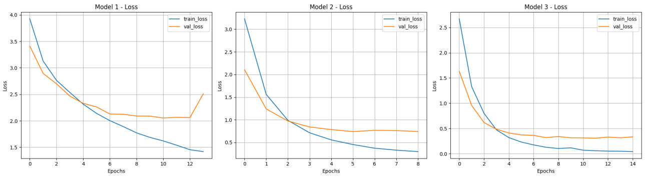


FIGURE 10. Comparative grafic loss

VI. CONCLUSION

The conclusions of the work highlight the success of implementing models based on LSTM, Embeddings, and Word2Vec to predict Diagnosis Related Groups (DRGs). After comparing three different architectures, the third model, which combined the processing of primary and secondary diagnoses and procedures using LSTM and Word2Vec, achieved the best results, with a weighted F1-Score of 0.9259 and a stable loss with no signs of overfitting.

The use of Embeddings improved the representation of medical codes, while LSTMs captured the sequential relationships between diagnoses and procedures. Although the performance in some minority classes is still low, this model showed a significant improvement compared to simpler approaches.

It is concluded that techniques based on sequences and vector representations are suitable to address prediction problems in complex clinical data. Next steps should include the use of attention mechanisms to improve performance in difficult classes and fine-tuning hyperparameters to further optimize the model.

REFERENCES

[1] Peña-Torres, J., & Kaufmann, J. (2023). Gestión Hospitalaria Pública en Chile y el Mecanismo de Pago GRD. CLAPES UC, Documento de Trabajo N°126.

[2] Wang, H., Gao, C., Dantona, C., Hull, B., & Sun, J. (2024). DRG-LLaMA: tuning LLaMA model to predict diagnosis-related group for hospitalized patients. npj Digital Medicine, 7(16), 1-16. https://doi.org/10.1038/s41746-023-00989-3

[3] Liu, J., Capurro, D., Nguyen, A., & Verspoor, K. (2021). Early prediction of diagnostic-related groups and estimation of hospital cost by processing clinical notes. npj Digital Medicine, 4(103), 1-13. https://doi.org/10.1038/s41746-021-00474-9

[4] Lai, G., Chang, W.-C., Yang, Y., & Liu, H. (2018). Modeling Long- and Short-Term Temporal Patterns with Deep Neural Networks. In Proceedings of ACM Conference (SIGIR'18), 11 pages.

[5] Rick, R., & Berton, L. (2022). Energy forecasting model based on CNN-LSTM-AE for many time series with unequal lengths. Engineering Applications of Artificial Intelligence, 113, 104998.

[6] Kalgotra, P., & Sharda, R. (2021). When will I get out of the hospital? Modeling length of stay using comorbidity networks. Journal of Management Information Systems, 38(4), 1150-1184.

[7] Hu, J., Jiang, C., Ma, G., Ding, J., Wang, Y., Xu, J., & Wang, Y. (2021, October). Power entity information recognition method based on Bi-LSTM+ CRF. In 2021 International Conference on Advanced Electrical Equipment and Reliable Operation (AEERO) (pp. 1-5). IEEE.

[8] Chandru, A. S., & Seetharam, K. (2022). Processing of clinical notes for efficient diagnosis with dual LSTM. International Journal of Advanced Computer Science and Applications, 13(2).

[9] Andreu-Mateu, C., Andreu-Vilarroig, C., Sánchez-Bermejo, N., Santamaría, C., & Tosca-Segura, R. (2024). Analysis and prediction of long-term survival using a clinically applicable risk score based on the Electronic Health Record. International Journal of Medical Informatics, 187, 105470.

• • •