MDPI

Article

ICD prediction from MIMIIC-III clinical text using pre-trained clinicalBERT and NLP deep learning models achieving state-of-the-art.

Ilyas ADEN [1], Chris CHILD [2] and Carlos Reyes-Aldasoro CONSTANTINO [3]

[1] ilyas.aden@city.ac.uk
[2] cchild@city.ac.uk
[3] constantino-carlos.reyes-aldasoro@city.ac.uk
City, University of London
Department of Computer Science,
Northampton Square,
London EC1V 0HB, United Kingdom

Abstract: The International Classification of Diseases (ICD) is a commonly used system for assigning diagnosis codes to patients' electronic health records. These codes help summarize diagnoses and procedures performed during a patient's hospital admission [12]. This paper creates an ICD code prediction based on the MIMIIC-III clinical text dataset. We developed a pipeline using natural language processing and our deep learning models to extract useful information from MIMIIC-III: Medical Information Mart for Intensive Care III (MIMIC-III), which is a large dataset, de-identified and publicly available collection of medical records [13]. Our current system predicts diagnosis codes from unstructured information like discharge summaries including notes containing symptoms. We used state-of-the-art deep learning methods such as RNN, LSTM, BiLSTM [3] and BERT [11] models after tokenising the clinical test with the Bio_ClinicalBERT a pre-trained model from Hugging Face [20]. Our experiments used the MIMIC-III discharge dataset records to evaluate our approach. Employing the BERT model, our method accurately predicted the top 10 and top 50 diagnosis codes within the MIMIC-III data, achieving average accuracies of 88% and 80% respectively. In comparison to recent studies by Biseda and Kerang, as well as Gangavarapu, which reported F1 scores of 0.72 for predicting the top 10 ICD-10 codes [5], our model demonstrated superior performance with an F1 score of 0.87. Similarly, for predicting the top 50 ICD-10 codes, previous research achieved an F1 score of 0.75 [1] [4], whereas our method attained an F1 score of 0.81.These results further support that the deep learning models exhibit superior performance over conventional machine learning approaches in this domain, corroborating our findings. Predicting diagnoses early from notes could help doctors determine promising treatments, transforming traditional diagnosis-then-treatment care.

Keywords: ICD prediction, NLP, deep learning models(RNN, LSTM, BERT)

Citation: ADEN, I.; CHILD, C.; CONSTANTINO, C. Title. *Journal Not Specified* **2024**, *1*, 0. https://doi.org/

Received:
Revised:
Accepted:
Published:

Copyright: © 2024 by the authors. Submitted to *Journal Not Specified* for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

1.1. Background

The MIMIC-III database stands as a significant tool for researchers, clinicians, and students keen on delving into critical care medicine to enhance patient outcomes. It offers access to real-world data, enabling the examination and hypothesis testing concerning the treatment of critically ill patients. With its application in over 1,000 research studies and citations in more than 3,500 scientific papers, its impact on medical research is profound. A distinct aspect of the MIMIC-III database is its inclusion of detailed clinical notes. These notes, composed by healthcare providers, offer narrative accounts of patient care, presenting deep insights into the management of critically ill patients. These narratives are instrumental in uncovering trends and patterns in patient treatment, enriching the database's value for research purposes [14].

1.2. Data exploratory and analysis

The MIMIC-II dataset showcases a broad spectrum of patient demographics, notably featuring a predominance of older adults and males. It encompasses a wide array of clinical notes, diagnostic codes, and possibly additional pertinent details. Below images explain summarise a detailed exploratory data analysis:

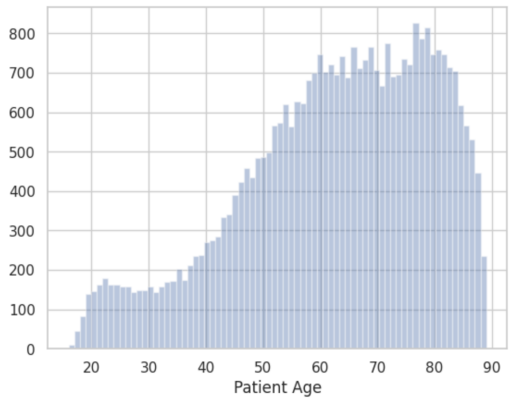

Figure 1. MIMIIC-III Patients Age Distribution

Figure 1 illustrates the age distribution of patients through a histogram, with a pronounced peak in the 60-70 age bracket. This suggests a predominant grouping of patients within this age interval. The data leans towards the right, indicating a larger share of older patients over younger ones.

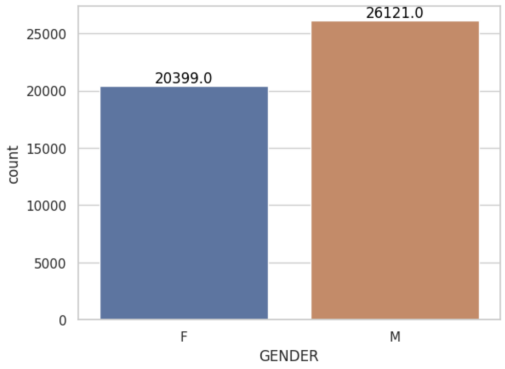

Figure 2. Patients Gender

Figure 2 showcases a bar chart detailing the gender distribution within the dataset, comparing male (M) and female (F) patients. The male patient count is noticeably higher, as seen in the taller bar for males, highlighting a gender disparity in the dataset.

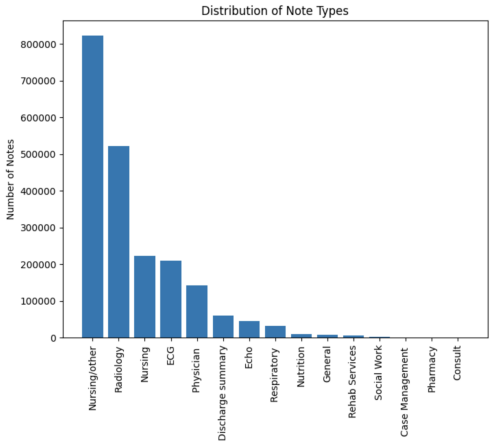

Figure 3. MIMIIC-III clinical notes catogories

Figure 3 offers a deeper dive into the dataset's notes categories, outlining a bar chart of the variety of note types, where "Nursing/other" emerges as the most frequent category.

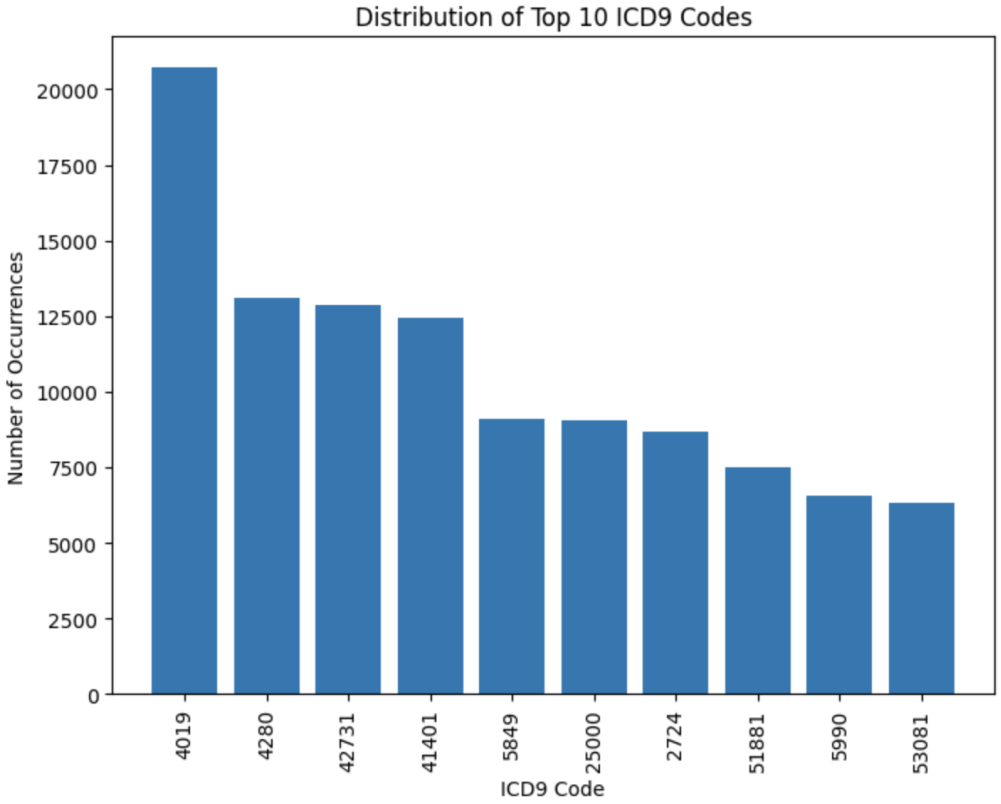

Figure 4. Example of top10 diseases

Figure 4 features a bar chart displaying the Top10 diseases or ten most common ICD-9 diagnosis code as example. The chart, with the y-axis for occurrence counts and the x-axis for the codes, it shows a clear standout with the code 401.9 marking a significantly higher occurrence than its counterparts. These visualisations and statistics can help us and any researchers or analysts better understand the characteristics and structure of the MIMIIC III dataset before conducting further analyses.

For our study, two relevant tables will be considered: note-events and ICD-9 diagnosis.The note-events table has more than 2 million rows with columns for patient ID, admission ID, and discharge notes text. The notes contain details like medical history

including symptoms, medications, lab tests, hospital course, and final diagnosis including the ICD-9 code made by doctors. The ICD diagnosis table has 651,000 rows with columns for patient ID, admission ID, and ICD-9 diagnosis codes. There are 6,984 unique codes. Each time a patient is admitted, they may receive between 1 and 38 diagnosis codes, which indicate the order of importance of their conditions and reasons for their visit [2]. In summary, the two key tables contain patient admission records with unstructured discharge note text and structured ICD-9 diagnosis codes for analysis and mapping between text and codes. The below table describes the size of the dataset and their respective unique values in the initial dataset.

Table 1. MIMIC-III descriptive statistics

Category	Number of rows	Unique values
Notes-Events	2 083 180	2 023 185
Diagnosis	651047	6984

1.3. Data processing

The first step was to examine the list of ICD-9 diagnosis codes present in the MIMIC-III dataset. Subsequently, these codes were matched with their respective ICD-10 counterparts, and the accuracy of this mapping was validated using a Python script. After that, the notes and diagnosis tables from MIMIC-III were merged based on unique patient and hospital admission IDs. This created a unified dataset with each patient's admission ID, ICD-10 codes, and discharge summary text. The data was then filtered to create multiple datasets: one with the top 10 ICD-10 codes by frequency, one with the top 50, and one with all codes. The distributions across these datasets were compared. To mitigate potential out-of-memory issues when processing the full dataset, smaller randomized samples of the data were taken such as 30%, 70%, and 100% of the full dataset. This allows initial testing on smaller sizes before scaling up. The result of these steps was processed and sampled datasets containing patients' admission IDs, ICD-10 codes, and textual discharge summaries, ready for applying natural language processing and machine learning models to predict diagnosis codes from the text. The multiple sampled datasets allow testing model performance at different data volumes.

Table 2. Statistics for diagnosis tables with top 10 and top 50 prevalent codes

Category	Number of rows	Unique values	Note-events(%)
Top10 Diagnosis	677738	10	32.5
Top50 Diagnosis	1058988	50	52.8

2. Methodology

Our methodology consists of the following steps: data pre-processing, building language model and classifier model. Specifically, we use Python 3.10 for data pre-processing: Python, NumPy, Pandas, and Sklearn for feature extraction; PyTorch is the main framework for training and testing models. We used Jupyter Notebooks to run our experiments on a private cloud platform called Runpod.io. The below schema describes the methodology used:

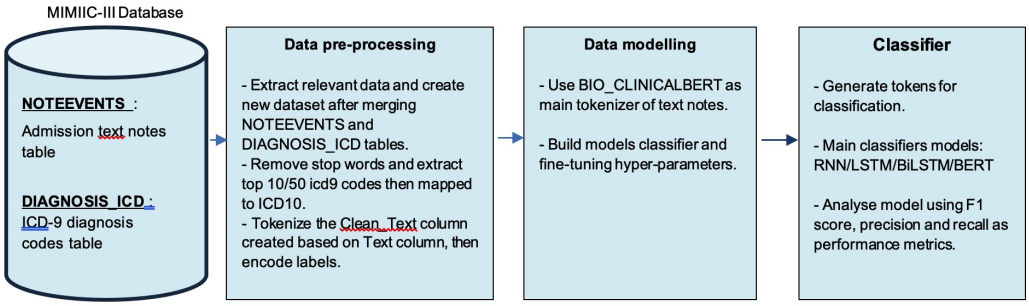

Figure 5. Methodology pipeline overview

The above figure depicts an overview of our methodology pipeline for processing and classifying medical text notes, likely from electronic health records, using machine learning models. Here is a brief explanation of all stages presented in our pipeline: **MIMIC-III Database:** This is a publicly available dataset that contains de-identified health-related data associated with over forty thousand patients who stayed in critical care units. The pipeline uses two main tables from this database:

- `NOTEEVENTS`: This table includes admission text notes, which are free-text descriptions of patient encounters.
- `DIAGNOSIS-ICD`: This table lists the ICD-9 diagnosis codes for the conditions diagnosed during the hospital stay.

Data Pre-processing: Relevant data from the NOTEEVENTS and DIAGNOSIS-ICD tables are merged to create a new dataset. Stop words (commonly used words that usually don't contain important meaning, like "the", "is", etc.) are removed from the text to reduce noise and focus on significant words. The most common ICD-9 codes (Top 10/50) are extracted and then mapped to ICD-10, which is a more current and detailed classification system for medical diagnoses. The text from the notes is tokenised, which means it's split into meaningful pieces (tokens) such as words or terms, and then these tokens are associated with the corresponding diagnostic labels (this process is called label encoding).

Data Modelling: BIO-CLINICALBERT is utilised as the primary tokenizer for the text notes. This is a version of the BERT model that has been pre-trained on biomedical and clinical text, making it more effective for understanding medical language. Classifier models are built, and their hyper-parameters are fine-tuned. Hyper-parameters are the settings for the algorithm that guide the training process and are set before the training starts.

Classifier: Tokens generated from the text are used for classification purposes. The main classifier models mentioned are Recurrent Neural Networks (RNN), Long Short-Term Memory Networks (LSTM), Bidirectional LSTM (BiLSTM), and BERT (Bidirectional Encoder Representations from Transformers). These are different neural network architectures commonly used in natural language processing tasks. The performance of these models is analysed using metrics like the F1 score (a harmonic mean of precision and recall that balances the two), precision (the number of true positive results divided by the number of all positive results), and recall (the number of true positive results divided by the number of positives that should have been retrieved). Overall, this pipeline is a structured approach to converting free-text medical notes into structured data that can be analysed and used for various purposes, such as predicting diagnoses, by leveraging advanced machine learning techniques.

3. Experimental Setup

Data Splitting: The dataset was split into 80% training data and 20% test data using the scikit-learn library in Python. This ensures we have sufficient data to train the models while holding out a subset to evaluate performance. The train-test split allows for an

unbiased assessment of the models.

Input Encoding: The text data was then encoded into numeric vectors suitable for machine learning using a pre-trained Bio-ClinicalBERT tokenizer from Hugging Face company. This state-of-the-art language representation model is designed specifically for the biomedical domain, allowing it to better handle medical terminology. The texts were tokenised and encoded into input vectors for the training and test sets.

Model Selection: Based on initial experiments, several model architectures were selected for comparison: recurrent neural networks (RNN), long short-term memory (LSTM), bi-directional LSTM, and BERT fine-tuning. These represent both traditional and cutting-edge deep learning approaches for NLP text classification tasks.

Evaluation Metric: The weighted average F1 score was chosen as the single metric to track during experiments. F1 score balances both precision and recall while weighting accounts for class imbalance. This offers a comprehensive assessment of performance. Additionally, various performance metrics such as precision, recall, and accuracy values are utilised to assess disparities in performance across different datasets and classifier models.

Model Optimisation: To improve results, various optimisation techniques were employed: · Hyper-parameter tuning to find optimal model configurations. · Error analysis to identify prediction pain points. · More aggressive data sampling strategies · Feature engineering such as text pre-processing · Regularization methods like dropout to prevent over-fitting. · Early stopping to halt training when the result is not improving. · Learning curves to determine if more training data is required.

Model Selection: Finally, the best-performing model architecture was selected based on the experiments. The top model was retrained on the full 80% training corpus and saved for future use. The pre-processed encodings were also retained for reuse in subsequent experiments.

4. Results and Discussions

The table below illustrates the performance of each model concerning their respective datasets, focusing on the top-10 and top-50 ICD-10 codes for diagnosis. The performance of the top-10 ICD-10 prediction using BERT is better with accuracy above 87% and 81% when using a single LSTM model. However, we slightly dropped performance when we tried predicting top-50 ICD-10 as we have an accuracy of 81% for the BERT model and 67% for a single LSTM model. The precision and recall scores for the top 10 are also relatively better than the top 50 data. In assessing these three metrics, our approach involves the calculation of average values rather than the examination of micro or macro-level data points.

Table 3. Summary results of our experiments.

Models	**Diagnosis**	**Precision**(%)	**Recall/Accuracy**(%)	**F1 Score**(%)
RNN	**Top10**	24	26	25
LSTM		**81**	**81**	**81**
BiLSTM		78	78	78
BERT		**87**	**87**	**87**
RNN	**Top50**	8	8	5
LSTM		**68**	**68**	**66**
BiLSTM		65	65	65
BERT		**81**	**81**	**80**

Best results have been achieved using below hyperparameters after model-tuning:

This last table provides a summary of the best hyperparameters for different models, including RNN, LSTM, BiLSTM, and BERT, with both top 10 and top 50 diagnoses. The hyperparameters include batch size, number of epochs, embedding dimension, hidden dimension, optimizer, activation function, dropout rate, and learning rate.

Table 4. Summary of best hyperparameters values.

Models	Diagnosis	Hyperparameters
RNN	**Top10**	Batch_size=16, epochs=10, embedding_dim=128, hidden_dim=256, optimizer='AdamW', activation='relu', dropout=0.4, lr=0.00002
LSTM		**Batch_size=16, epochs=10, embedding_dim=128, hidden_dim=256,optimizer='AdamW', activation='relu',dropout=0.2,lr=0.00**
BiLSTM		Batch_size=16, epochs=10, embedding_dim=128, hidden_dim=256, optimizer='AdamW', activation='relu', dropout=0.2, lr=0.001
BERT		**Batch_size=16, epochs=10, embedding_dim=128, hidden_dim=256,optimizer='AdamW', activation='relu',dropout=0.4,lr=0.001**
RNN	**Top50**	Batch_size=16, epochs=10, embedding_dim=128, hidden_dim=256, optimizer='AdamW', activation='relu', dropout=0.4, lr=0.00002
LSTM		**Batch_size=16, epochs=10, embedding_dim=128, hidden_dim=256,optimizer='AdamW', activation='relu',dropout=0.2,lr=0.001**
BiLSTM		Batch_size=16, epochs=10, embedding_dim=128, hidden_dim=256, optimizer='AdamW', activation='relu', dropout=0.2, lr=0.001
BERT		**Batch_size=16, epochs=10, embedding_dim=128, hidden_dim=256,optimizer='AdamW', activation='relu',dropout=0.4,lr=0.001**

In our study, we found that models previously considered as having lower performance exhibited suboptimal results primarily because of inadequately chosen hyperparameters and the absence of fine-tuning the decision boundary. Through our updated comparison, we illustrated that when we trained our models using our configuration, it led to a reduction in the gap between the highest and lowest F1 scores. This confirms the results collected in the latest ICD-10 prediction research [3]. Additionally, Figure 6 and Figures 7, and 8 illustrate the precision, recall and F-1 score for LSTM/BERT classifiers built. Overall, the classifier with the top 10 diagnoses has higher scores when compared to the classifier with the top 50 diagnoses.

	precision	recall	f1-score	support
0	0.81	0.85	0.83	1953
1	0.77	0.74	0.76	705
2	0.83	0.85	0.84	826
3	0.86	0.89	0.87	333
4	0.96	0.96	0.96	364
5	0.67	0.76	0.71	560
6	0.68	0.71	0.70	436
7	0.83	0.72	0.77	1098
8	0.86	0.80	0.83	402
9	1.00	0.98	0.99	64
accuracy			0.81	6741
macro avg	0.83	0.83	0.83	6741
weighted avg	0.81	0.81	0.81	6741

Figure 6. Top 10 ICD predictions using LSTM model.

	precision	recall	f1-score	support
0	0.90	0.91	0.91	1953
1	0.84	0.84	0.84	705
2	0.86	0.88	0.87	826
3	0.89	0.86	0.88	333
4	0.96	0.98	0.97	364
5	0.79	0.80	0.80	560
6	0.70	0.78	0.74	436
7	0.89	0.84	0.87	1098
8	0.94	0.88	0.91	402
9	1.00	1.00	1.00	64
accuracy			0.87	6741
macro avg	0.88	0.88	0.88	6741
weighted avg	0.87	0.87	0.87	6741

Figure 7. Top 10 ICD predictions using BERT model.

	precision	recall	f1-score	support
0	0.57	0.82	0.67	160
1	0.54	0.73	0.62	196
2	0.60	0.75	0.66	186
3	0.72	0.83	0.77	140
4	0.51	0.70	0.59	278
5	0.65	0.58	0.62	194
6	0.53	0.76	0.63	100
7	0.70	0.44	0.54	1969
8	0.79	0.76	0.77	170
9	0.85	0.80	0.82	132
10	0.72	0.89	0.79	93
11	0.85	0.86	0.86	197
12	0.40	0.80	0.53	120
13	0.61	0.68	0.64	228
14	0.69	0.72	0.70	228
15	0.73	0.41	0.53	684
16	0.75	0.78	0.76	834
17	0.80	0.77	0.78	215
18	0.74	0.83	0.78	121
19	0.76	0.84	0.80	308
20	0.77	0.86	0.81	150
21	0.58	0.73	0.64	162
22	0.69	0.71	0.70	237
23	0.67	0.72	0.70	244
24	0.28	0.43	0.34	130
25	0.95	0.86	0.91	334
26	0.82	0.78	0.80	399
27	0.49	0.56	0.52	135
28	0.37	0.41	0.39	103
29	0.87	0.85	0.86	80
30	0.52	0.55	0.54	560
31	0.60	0.78	0.68	219
32	0.40	0.59	0.48	387
33	0.72	0.52	0.60	1079
34	0.68	0.63	0.65	297
35	0.64	0.88	0.74	77
36	0.70	0.75	0.72	134
37	0.77	0.66	0.71	88
38	0.71	0.86	0.78	140
39	0.78	0.84	0.81	428
40	0.59	0.72	0.65	237
41	0.52	0.60	0.56	294
42	0.74	0.84	0.79	174
43	0.71	0.83	0.77	222
44	0.64	0.82	0.72	155
45	0.53	0.72	0.61	141
46	0.57	0.64	0.60	118
47	0.98	0.87	0.92	46
48	0.88	0.96	0.92	54
49	0.97	0.97	0.97	30
accuracy			0.66	13407
macro avg	0.67	0.73	0.69	13407
weighted avg	0.68	0.66	0.66	13407

Figure 8. Top50 ICD prediction using LSTM model.

	precision	recall	f1-score	support
0	0.73	0.84	0.78	160
1	0.96	0.72	0.82	196
2	0.90	0.75	0.82	186
3	0.96	0.89	0.93	140
4	0.67	0.79	0.73	278
5	0.88	0.62	0.73	194
6	0.75	0.73	0.74	100
7	0.82	0.85	0.84	1969
8	0.89	0.86	0.88	170
9	0.86	0.82	0.84	132
10	0.98	0.85	0.91	93
11	0.87	0.95	0.91	197
12	0.86	0.78	0.82	120
13	0.84	0.70	0.77	228
14	0.70	0.86	0.77	228
15	0.72	0.70	0.71	684
16	0.86	0.80	0.83	834
17	0.88	0.86	0.87	215
18	0.84	0.90	0.87	121
19	0.87	0.90	0.88	308
20	0.85	0.88	0.87	150
21	0.84	0.82	0.83	162
22	0.81	0.71	0.76	237
23	0.74	0.80	0.77	244
24	0.41	0.38	0.39	130
25	0.99	0.90	0.94	334
26	0.85	0.94	0.90	399
27	0.56	0.80	0.66	135
28	0.62	0.24	0.35	103
29	0.89	0.93	0.91	80
30	0.80	0.69	0.74	560
31	0.89	0.63	0.73	219
32	0.63	0.78	0.70	387
33	0.75	0.88	0.81	1079
34	0.92	0.65	0.76	297
35	0.81	0.83	0.82	77
36	0.75	0.87	0.81	134
37	0.84	0.78	0.81	88
38	0.76	0.89	0.82	140
39	0.87	0.90	0.88	428
40	0.80	0.77	0.78	237
41	0.74	0.73	0.74	294
42	0.94	0.83	0.88	174
43	0.88	0.95	0.91	222
44	0.91	0.75	0.83	155
45	0.74	0.79	0.77	141
46	0.72	0.80	0.76	118
47	0.98	0.91	0.94	46
48	0.93	0.96	0.95	54
49	0.94	0.97	0.95	30
accuracy			0.81	13407
macro avg	0.82	0.80	0.80	13407
weighted avg	0.81	0.81	0.80	13407

Figure 9. Top50 ICD prediction using BERT model.

Previous studies have explored the feasibility of deep learning models for predicting ICD-10 codes. However, it is important to note that these deep learning models did not demonstrate high performance when applied to the MIMIC-III database.

The below table compares the main previous experiments and our results:

Table 5. Comparative evaluation of different studies from the literature review [15].

Work	Data	Method	Target Variable	Performance Measures
Hsu et al.[8]	Discharge summary	Deep Learning	(I) 19 distinct ICD-9 chapter codes, (II) Top 50 ICD-9 codes, (III) Top 100 ICD-9 codes	(I) Micro F1 score of 0.76, (II) Micro F1 score of 0.57, (III) Micro F1 score of 0.51
Gangavarapu et al.[6]	Nursing notes	Deep Learning	19 distinct ICD-9 chapter codes	Accuracy of 0.833
Samonte et al.[15]	Discharge summary	Deep Learning	10 distinct ICD-9 codes	Precision of 0.780, Recall of 0.620, F1 score of 0.678
Obeid et al.[10]	Clinical notes	Deep Learning	ICD-9 code from E950-E959	Area under the ROC curve score of 0.882, F1 score of 0.769
Hsu et al.[8]	Subjective component	Deep Learning	(I) 17 distinct ICD-9 chapter codes, (II) 2017 distinct ICD-9 codes	(I) Accuracy of 0.580, (II) Accuracy of 0.409
Xie et al.[17]	Diagnosis description	Deep Learning	2833 ICD-9 codes	Sensitivity score of 0.29, Specificity score of 0.33
Singaravelan et al.[16]	Subjective component	Deep Learning	1871 ICD-9 codes	Recall score for chapter code is 0.57, Recall score for block is 0.49, Recall score for three-digit code is 0.43, Recall score for full code is 0.45
Zeng et al.[18]	Discharge summary	Deep Learning	6984 ICD-9 codes	F1 score of 0.42
Huang et al.[7]	Discharge summary	Deep Learning	(I) 10 ICD-9 codes, (II) 10 blocks 1131 ICD-10 codes	(I) F1 score of 0.69, (II) F1 score of 0.72
Current study	Discharge summary	Deep Learning	(I) Top 10 ICD-10 codes, (II) Top 50 ICD-10 codes	(I) Precision of 0.88, Recall of 0.88, F1-score of 0.88, (II) Precision of 0.81, Recall of 0.81, F1 score of 0.80

5. Limitation and future work

One of the main challenges we faced during work was a lack of resources to run the high-end operations. Indeed, handling 7 GB of MIMIIC-III data (26 tables) demands a quantity number of resources and time. The RunPods.io platform RTX A6000, a new GPU enabled me to move forward from any limited-resources environment. Future work relies on the prediction of ICD or diagnosis, using an ensemble model unlike that of the single models predicting diagnosis distinctly. Also, some refinements are to be made to enhance the accuracy and performance of the model when the top 20, top 50 or even more than 100 diagnoses would be used. Additionally, we can consider a k-fold cross-validation approach instead of using the classical approach of an 80%-20% split of training/test datasets.

6. Conclusion

In conclusion, this research examines the efficacy of deep learning models such as LSTM and BERT architectures, specifically the BERT model, for automated extraction of medical concepts from clinical notes in the MIMIC-III database. Empirical results demonstrate that deep learning natural language processing techniques can effectively encode clinical texts and assign appropriate ICD codes without manual supervision. The proposed methodology establishes a competitive baseline for concept extraction, achieving strong diagnostic code prediction from discharge summaries. Compared to the Top10 ICD code prediction [6] with an F1 score of 0.72, we achieved a better F1 score of 0.87. Also similarly, in comparison to the Top-50 ICD code prediction [1][4] with an F1 score of 0.75, we achieve a final F1 score of 0.81. Moreover, the generalisability of the current LSTM/BERT models creates promise for holistic, unified systems that can extract multiple data types such as diagnosis codes, simultaneously from unstructured electronic health records. This research thereby underscores the capability of artificial intelligence methods to unlock clinical knowledge from textual data sources and meaningfully impact healthcare delivery. Furthermore, Large language models (LLMs) have shown the potential to accelerate clinical curation via few-shot in-context learning. Indeed, in the latest paper of Zelalem et al.[19], self-verification represents a crucial milestone in harnessing the capabilities of Large Language Models (LLMs) within healthcare contexts. As LLMs consistently enhance their overall performance, the use of LLMs in clinical data extraction combined with self-verification (LLMs + SV) is poised to see notable improvements.

References

1. Brent Biseda, Gaurav Desai, Haifeng Lin, and Anish Philip. Prediction of ICD Codes with Clinical BERT Embeddings and Text Augmentation with Label-Balancing-using-MIMIC-III. *arXiv:2008.10492* **2020**.
2. Choi, E., Schuetz, A., Stewart, W. F., Sun, J. Using recurrent neural network models for early detection of heart failure onset. *J Am Med Inform Assoc,* 24(2), 361-370.**2017***.doi:10.1093/jamia/ocw112*
3. Choi, Y., Kang, S. A systematic review of deep learning-based automated diagnosis of neurologic disorders using EEG signals. *BMC Medical Informatics and Decision Making,* 22(1), 1-18.**2022**
4. Edin, Joakim Junge, Alexander Drachmann Havtorn, Jakob Borgholt, Lasse Maistro, Maria Ruotsalo, Tuukka Maaløe, Lars. **2023** *Automated Medical Coding on MIMIC-III and MIMIC-IV: A Critical Review and Replicability Study,* 2572-2582. 10.1145/3539618.3591918.
5. Keyang Xu, et. al. Multimodal Machine Learning for Automated ICD Coding. *Proceedings of Machine Learning Research.* 106:1-17. **2019**
6. Gangavarapu, T.; Krishnan, G.S.; Kamath, S.; Jeganathan, J. FarSight: Long-Term Disease Prediction Using Unstructured Clinical Nursing Notes. *IEEE Trans. Emerg. Top. Comput.* **2020**, 9, 1151–1169.
7. Huang, J.; Osorio, C.; Sy, L.W. An empirical evaluation of deep learning for ICD-9 code assignment using MIMIC-III clinical notes. *Comput. Methods Programs Biomed.* **2019**, 177, 141–153.
8. Hsu, C.C.; Chang, P.C.; Chang, A. Multi-Label Classification of ICD Coding Using Deep Learning. *In Proceedings of the International Symposium on Community-Centric Systems (CcS), Tokyo, Japan,* **23–26 September 2020;** pp. 1–6.
9. Hsu, J.L.; Hsu, T.J.; Hsieh, C.H.; Singaravelan, A. Applying Convolutional Neural Networks to Predict the ICD-9 Codes of Medical Records. *Sensors* **2020**, 20, 7116.

10. Obeid, J.S.; Dahne, J.; Christensen, S.; Howard, S.; Crawford, T.; Frey, L.J.; Stecker, T.; Bunnell, B.E. Identifying and Predicting intentional self-harm in electronic health record clinical notes: Deep learning approach. *JMIR Med. Inform.* **202**0, 8, e17784.
11. Lee, J., Shin, H., Kim, Y. **2020**. The Effects of Hyperparameters in Deep Learning on Medical Dataset: A Case Study on EMR. *arXiv preprint arXiv:2009.05451.* https://arxiv.org/abs/2009.05451
12. Masud, J.H.B.; Kuo, C.-C.; Yeh, C.-Y.; Yang, H.-C.; Lin, M.-C. Applying Deep Learning Model to Predict Diagnosis Code of Medical Records. *Diagnostics* **2023**, 13,2297.-https://doi.org/10.3390/diagnostics13132297
13. National Health Service. **2022**. International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD-10), 5th Edition. *Retrieved from: https://classbrowser.nhs.uk/ref$_b$ooks/IC* $10_2022_5th_Ed_NCCS.pdf$
14. PhysioNet. **2016**. MIMIC-III Clinical Database (version 1.4). *Retrieved from: https://physionet.org/conten*
15. Samonte, M.J.C.; Gerardo, B.D.; Fajardo, A.C.; Medina, R.P. ICD-9 tagging of clinical notes using topical word embedding. *In Proceedings of the 2018 International Conference on Internet and e-Business, Taipei, Taiwan,* **16–18 May 2018;** pp. 118–123.
16. Singaravelan, A.; Hsieh, C.-H.; Liao, Y.-K.; Hsu, J.L. *Predicting ICD-9 Codes Using Self-Report of Patients. Appl. Sci.* **2021**, 11, 10046.
17. Xie, P.; Xing, E. A Neural Architecture for Automated ICD Coding. *In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia,* **15–20 July 2018**; Association for Computational Linguistics: Melbourne, Australia, 2018; pp. 1066–1076.
18. Zeng, M.; Li, M.; Fei, Z.; Yu, Y.; Pan, Y.; Wang, J. Automatic ICD-9 coding via deep transfer learning. *Neurocomputing* **2019**, 324, 43–50.
19. Zelalem Gero, Chandan Singh, Hao Cheng, Tristan Naumann, Michel Galley, Jianfeng Gao, Hoifung Poon, **May 2023**, Self-Verification Improves Few-Shot Clinical Information Extraction. *https://arxiv.org/abs/2306.00024*
20. Zhao, B., Qiu, T., Wu, Q., Wang, W. **2019**. Attention-Based Multi-Modal Fusion for Explainable Recommendation. *arXiv preprint arXiv:1904.03323.* https://arxiv.org/abs/1904.03323

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