City Research Online

City, University of London Institutional Repository

Citation: Aden, I., Child, C. H. T. & Reyes-Aldasoro, C. C. (2024). International Classification of Diseases Prediction from MIMIIC-III Clinical Text Using Pre-Trained ClinicalBERT and NLP Deep Learning Models Achieving State of the Art. Big Data and Cognitive Computing, 8(5), 47. doi: 10.3390/bdcc8050047

This is the published version of the paper.

This version of the publication may differ from the final published version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/33037/

Link to published version: https://doi.org/10.3390/bdcc8050047

Copyright: City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

Reuse: Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

City Research Online: http://openaccess.city.ac.uk/ publications@city.ac.uk

big data and cognitive computing

MDPI

Article

International Classification of Diseases Prediction from MIMIC-III Clinical Text Using Pre-Trained ClinicalBERT and NLP Deep Learning Models Achieving State of the Art

Ilyas Aden *, Christopher H. T. Child and Constantino Carlos Reyes-Aldasoro

Department of Computer Science, City, University of London, Northampton Square, London EC1V 0HB, UK; c.child@city.ac.uk (C.H.T.C.); constantino-carlos.reyes-aldasoro@city.ac.uk (C.C.R.-A.)
* Correspondence: ilyas.aden@city.ac.uk

Abstract: The International Classification of Diseases (ICD) serves as a widely employed framework for assigning diagnosis codes to electronic health records of patients. These codes facilitate the encapsulation of diagnoses and procedures conducted during a patient's hospitalisation. This study aims to devise a predictive model for ICD codes based on the MIMIC-III clinical text dataset. Leveraging natural language processing techniques and deep learning architectures, we constructed a pipeline to distill pertinent information from the MIMIC-III dataset: the Medical Information Mart for Intensive Care III (MIMIC-III), a sizable, de-identified, and publicly accessible repository of medical records. Our method entails predicting diagnosis codes from unstructured data, such as discharge summaries and notes encompassing symptoms. We used state-of-the-art deep learning algorithms, such as recurrent neural networks (RNNs), long short-term memory (LSTM) networks, bidirectional LSTM (BiLSTM) and BERT models after tokenizing the clinical test with Bio-ClinicalBERT, a pre-trained model from Hugging Face. To evaluate the efficacy of our approach, we conducted experiments utilizing the discharge dataset within MIMIC-III. Employing the BERT model, our methodology exhibited commendable accuracy in predicting the top 10 and top 50 diagnosis codes within the MIMIC-III dataset, 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 in predicting the top 10 ICD-10 codes, our model demonstrated better performance, with an F1 score of 0.87. Similarly, in predicting the top 50 ICD-10 codes, previous research achieved an F1 score of 0.75, whereas our method attained an F1 score of 0.81. These results underscore the better performance of deep learning models over conventional machine learning approaches in this domain, thus validating our findings. The ability to predict diagnoses early from clinical notes holds promise in assisting doctors or physicians in determining effective treatments, thereby reshaping the conventional paradigm of diagnosis-then-treatment care. Our code is available online.

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

check for updates

Citation: Aden, I.; Child, C.H.T; Reyes-Aldasoro, C.C. International Classification of Diseases Prediction from MIMIC-III Clinical Text Using Pre-Trained ClinicalBERT and NLP Deep Learning Models Achieving State of the Art. *Big Data Cogn. Comput.* **2024**, *8*, 47. https://doi.org/10.3390/bdcc8050047

Academic Editors: Tim Schlippe and Matthias Wölfel

Received: 27 March 2024
Revised: 27 April 2024
Accepted: 30 April 2024
Published: 10 May 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

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 [1]. 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 1000 research studies and citations in more than 3500 scientific papers, its impact on medical research is profound. A distinct aspect of the MIMIC-III database is its inclusion of detailed clinical notes [2]. These notes, produced 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. Perhaps the most well-known work on ICD prediction using the MIMIC-III dataset is the 2018 study by James Mullenbach et al., entitled "Explainable Prediction of Medical Codes from Clinical Text". This paper is renowned for introducing and applying the CAML (Convolutional Attention for Multi-Label Classification) model, which combines convolutional neural networks (CNNs) with an attention mechanism to predict ICD codes from clinical text. This work was among the first to incorporate explainability into ICD prediction models—a crucial advancement for fostering trust and understanding in healthcare applications that depend on AI predictions. Although not the top performer in terms of raw accuracy metrics—owing to continual improvements in the field—the CAML model demonstrated competitive results on MIMIC-III at the time. Its explainability features underscored its significance [3]. This work has spurred further research in the domain of medical code prediction from clinical texts, influencing methodologies across both academic and practical healthcare settings. Notably, the study titled "An Empirical Evaluation of Deep Learning for ICD-9 Code Assignment Using MIMIC-III Clinical Notes" has become a cornerstone in the field, comparing various deep learning approaches for ICD-9 code prediction and establishing a benchmark for future research. It underscores the potential of deep learning to automate ICD coding tasks [4]. Another influential subsequent work is the 2021 paper "TransICD: Transformer Based Code-wise Attention Model for Explainable ICD Coding" [5]. This study introduces a transformer-based architecture named TransICD, which employs a code-wise attention mechanism. This mechanism enables the model to concentrate on specific segments of the clinical notes that are pertinent to each ICD code prediction. The paper is notable for its high micro-AUC score of 0.923, although it does not detail the exact F1 scores for the top 10 and top 50 ICD predictions. It is also important to acknowledge the foundational work in NLP deep learning models, particularly the transformer architecture introduced by Vaswani et al. in their landmark 2017 paper, "Attention is All You Need" [6]. While this paper did not focus on clinical applications, it has significantly influenced NLP through its effectiveness in tasks like machine translation and text summarization. Understanding transformers is essential for grasping many clinical NLP studies that utilize this architecture. Subsequently, Google AI's development of BERT in 2018, based on the transformer's encoder mechanism, marked a pivotal shift in how contextual information is processed in NLP models. Moreover, recent research has explored further innovations based on the transformer architecture, such as the Transformer-in-Transformer (TNT) model, which offers a novel approach to visual recognition tasks. Although the TNT model is primarily designed for visual tasks, its methodological innovations provide useful parallels for text-based applications like ICD prediction [7]. Similarly, the Multi-Generator Orthogonal GAN (MGO-GAN) introduces a novel approach utilizing multiple generators to enhance output diversity. This method could analogously enhance the diversity in ICD code prediction from clinical texts, potentially capturing a broader array of diagnoses from complex medical narratives [8].

In this context, our current paper utilizes various deep learning models, including RNN, LSTM, and BERT, to predict ICD codes from clinical text data in the MIMIC-III dataset. Our focus is on comparing its performance, particularly with the transformer-based BERT model, which remains a benchmark in many NLP tasks [9].

1.2. Data Exploratory and Analysis

The MIMIC-III 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. The below images explain and summarize the details of the exploratory data analysis:

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 lean towards the right, indicating a larger share of older patients over younger ones.

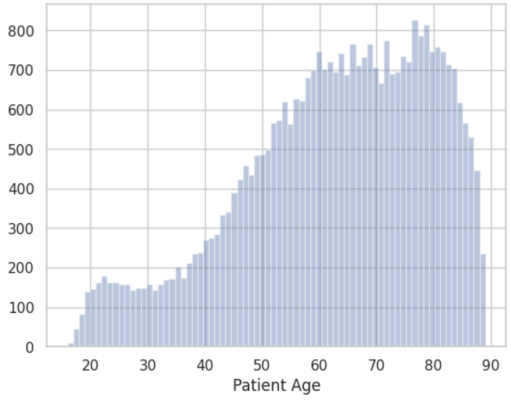

Figure 1. MIMIIC-III patient age distribution.

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.

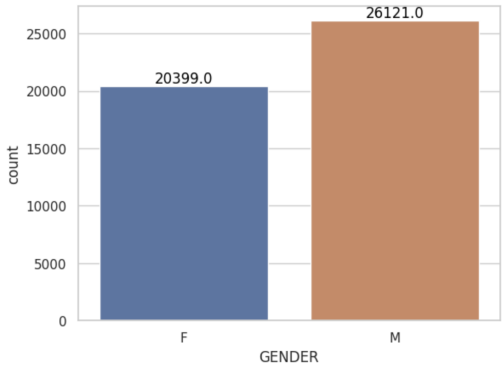

Figure 2. Gender of patients.

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

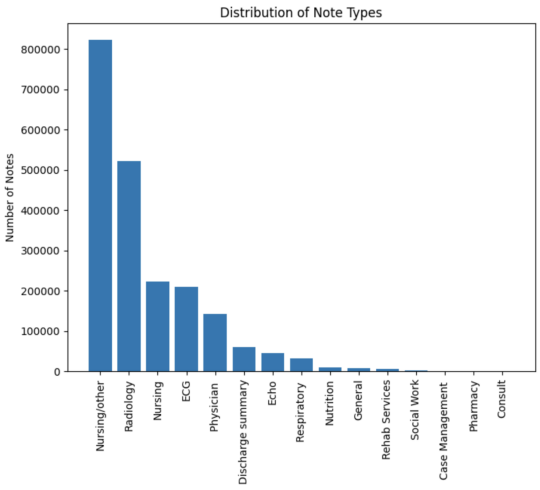

Figure 3. MIMIIC-III clinical note categories.

Figure 4 features a bar chart displaying the top 10 diseases or ten most common ICD-9 diagnosis codes as an example. The chart, with the y-axis for occurrence counts and the x-axis for the codes, shows a clear standout with the code 401.9 marking a significantly

higher occurrence than its counterparts. These visualizations 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.

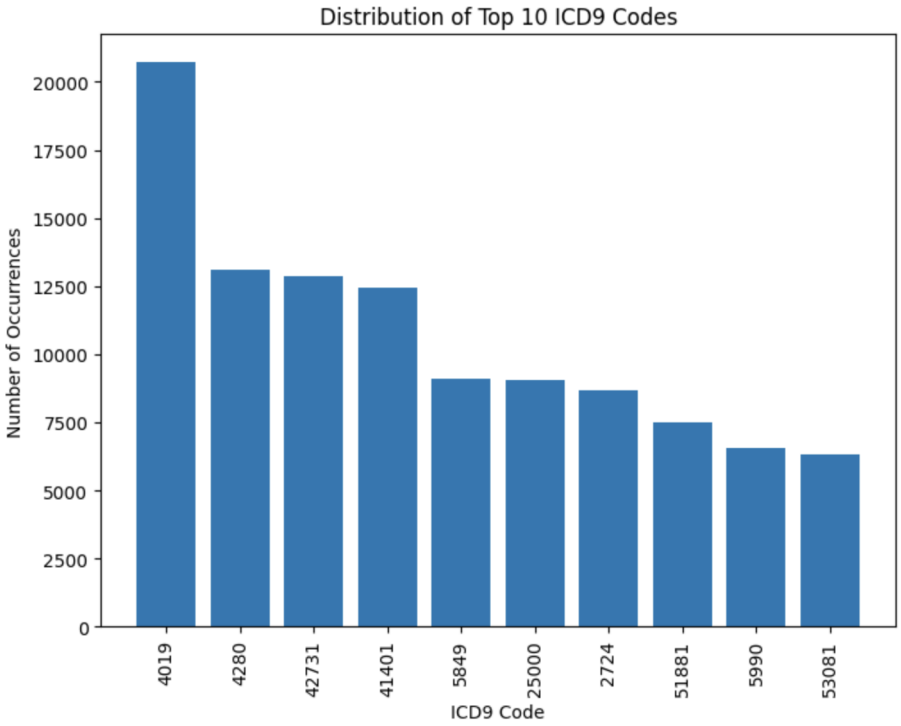

Figure 4. Example of top 10 diseases.

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 and columns for patient ID, admission ID, and discharge note text. The notes contain details like medical history, including symptoms, medications, lab tests, hospital course, and final diagnosis, including the ICD-9 code given by doctors. The ICD diagnosis table has 651,000 rows and columns for patient ID, admission ID, and ICD-9 diagnosis codes. There are 6984 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. 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. Table 1 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
Note events	2,083,180	2,023,185
Diagnosis	651,047	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 were 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 results of these steps were processed and sampled datasets containing patients' admission IDs, ICD-10 codes, and textual discharge summaries, ready for the application of natural language processing and machine learning models to predict diagnosis codes from the text. The multiple sampled datasets allow model performance to be tested at different data volumes. Below Table 2 describe diagnosis statistics.

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

Category	Number of Rows	Unique Values	Note Events (%)
Top 10 Diagnosis	677,738	10	32.5
Top 50 Diagnosis	1,058,988	50	52.8

2. Methodology

Our methodology consists of the following steps: data pre-processing and building the language model and classifier model. Specifically, we use Python 3.10 for data pre-processing and 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. Figure 5 below describes the methodology used:

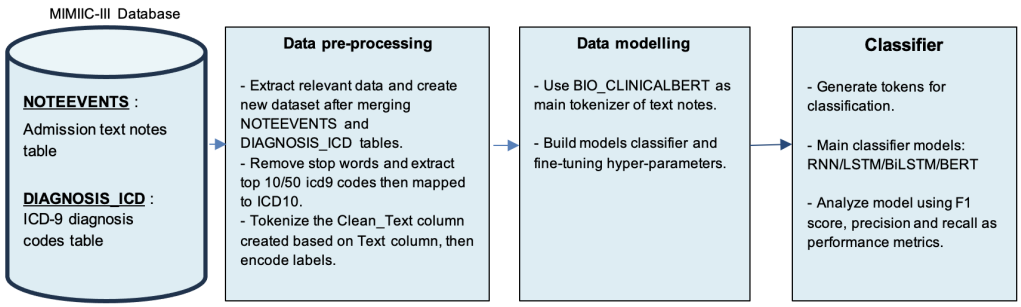

Figure 5. Methodology pipeline overview.

This above diagram illustrate 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 do not 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 tokenized, which means it is 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 Modeling: Bio-ClinicalBERT [10] is utilized 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 hyperparameters are fine-tuned. Hyperparameters 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 (RNNs), long short-term memory (LSTM) networks, bidirectional LSTM (BiLSTM) [11], 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 analyzed 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 analyzed 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 ensured that we had sufficient data to train the models while retaining a subset to evaluate performance. The train–test split allows for an unbiased assessment of the models.

Input Encoding: The text data were then encoded into numeric vectors suitable for machine learning using a pre-trained Bio-ClinicalBERT tokenizer from the 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 tokenized 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 (RNNs), 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 utilized to assess disparities in performance across different datasets and classifier models.

Model Optimization: To improve results, various optimization techniques were employed:

- Hyperparameter 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 does not improve.
- Learning curves to determine whether more training data are 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

Table 3 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, performance decreased slightly when we tried predicting top 50 ICD-10 as we obtained 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 better than those for

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 of results of our experiments.

Model	Diagnosis	Precision (%)	Recall/Accuracy (%)	F1 Score (%)
RNN	**Top 10**	24	26	25
LSTM		**81**	**81**	**81**
BiLSTM		78	78	78
BERT		**87**	**87**	**87**
RNN	**Top 50**	8	8	5
LSTM		**68**	**68**	**66**
BiLSTM		65	65	65
BERT		**81**	**81**	**80**

The best results were achieved using the hyperparameters below after model-tuning. Table 4 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 the batch size, number of epochs, embedding dimension, hidden dimension, optimizer, activation function, dropout rate, and learning rate.

Table 4. Summary of best hyperparameter values.

Model	Diagnosis	Hyperparameters
RNN	**Top 10**	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	**Top 50**	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**

Table 5 below presents a comparison between the main findings of previous studies and the results we obtained.

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

Work	Data	Method	Target Variable	Performance Measures
Hsu et al. [12]	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. [13]	Nursing notes	Deep learning	19 distinct ICD-9 chapter codes	Accuracy of 0.833
Samonte et al. [14]	Discharge summary	Deep learning	10 distinct ICD-9 codes	Precision of 0.780, Recall of 0.620, F1 score of 0.678

Table 5. *Cont.*

Work	Data	Method	Target Variable	Performance Measures
Obeid et al. [15]	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. [16]	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. [18]	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. [19]	Discharge summary	Deep learning	6984 ICD-9 codes	F1 score of 0.42
Huang et al.	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

Indeed, our research indicates that models previously considered as having lower performance exhibited suboptimal results primarily because of inadequately chosen hyperparameters and the absence of a fine-tuned 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 [20]. Additionally, Figures A1–A4 in Appendix A illustrate the precision, recall, and F-1 score for the 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.

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.

To summarize, this experimentation utilizes a diverse array of deep learning models, including RNN, LSTM, BiLSTM, and BERT, with a specific emphasis on the Bio-ClinicalBERT model, which is pre-trained for biomedical texts. The study takes advantage of various neural network architectures, particularly focusing on a specialized version of BERT pre-trained for biomedical contexts. This approach enhances the model's ability to interpret clinical language effectively. Furthermore, these results showcase significant advancements in the automation of ICD coding and present the most comprehensive F1 score metrics available to date. These scores are internationally recognized for evaluating the balance between precision and recall in classification tasks.

5. Limitation and Future Work

One of the main challenges we faced during our work was a lack of computational resources to execute high-end operations necessary for training and optimizing complex models like RNN, LSTM, and BERT. Indeed, handling the extraction of 7 GB from the MIMIC-III dataset, which has an initial total size of 3 TB and consists of 26 tables, demands significant computational resources and time. Using a private cloud server such as the RTX A6000 GPU helped us overcome environments limited by resources, enabling more efficient data processing and model training.

During the training of RNNs, LSTM, and particularly BERT models using the MIMIC-III dataset, we encountered several additional challenges. Firstly, the complexity and

heterogeneity of healthcare data present in MIMIC-III can lead to issues such as imbalanced classes and missing values, which significantly affect the performance of predictive models. Addressing these data quality issues required sophisticated preprocessing steps, which themselves are resource-intensive.

Moreover, the temporal dependencies and high dimensionality of the data make RNNs and LSTMs computationally expensive to train. These models also suffer from issues like vanishing and exploding gradients, making it challenging to train deep networks effectively without careful tuning of hyperparameters and the adoption of techniques like gradient clipping and batch normalization. BERT and other transformer-based models, while powerful in capturing contextual information from clinical notes, demand even more computational resources due to their attention mechanisms and large number of parameters. Training these models from scratch on a dataset like MIMIC-III can be prohibitively expensive, often necessitating the use of pre-trained models followed by fine-tuning on specific tasks. However, the adaptation of these models to domain-specific medical language and tasks requires careful calibration and validation to ensure that the models do not perpetuate biases or errors inherent in the training data.

Future work will focus on enhancing prediction models for ICD codes or diagnoses by using an ensemble approach rather than relying on single models. Such an approach may leverage the strengths of various model architectures to improve accuracy and robustness. Additionally, refinements are necessary to boost the performance and accuracy of models when predicting a larger number of diagnoses, such as the top 20, top 50, or even more than 100 diagnoses.

Furthermore, adopting more advanced validation techniques, such as k-fold cross-validation, will be explored to ensure the robustness and generalizability of the models. Unlike the traditional approach of splitting the dataset into a fixed training and test set, k-fold cross-validation provides a more comprehensive evaluation of model performance by partitioning the data into multiple subsets for training and validation. This helps in assessing the model's performance across different subsets of the data and provides a more accurate estimate of its true performance on unseen data.

Lastly, addressing the limitations in explainability and transparency of these complex models is crucial, especially in a high-stakes field like healthcare. Developing methods to interpret model decisions and ensure they align with clinical reasoning will be critical in future work, enabling clinicians to trust and effectively use AI-driven tools in their decision-making processes.

6. Conclusions

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 top 10 ICD code prediction with an F1 score of 0.72 [21], we achieved a better F1 score of 0.87. Furthermore, similar to the top 50 ICD code prediction with an F1 score of 0.75 [22,23], we achieved a final F1 score of 0.81. Moreover, the generalizability 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. [24], 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.

Author Contributions: Conceptualization, I.A.; methodology, I.A.; software, I.A.; validation, I.A.; formal analysis, I.A.; investigation, I.A.; resources, I.A.; data curation, I.A.; writing—original draft preparation, all authors; writing—review and editing, all authors; visualization, I.A.; supervision, C.H.T.C. and C.C.R.-A.; project administration, I.A.; funding acquisition, I.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The research data is available here: https://mimic.mit.edu/ (accessed on 31 January 2024) and the code source is available here: https://github.com/userilyo/BDCC-Paper-Work (accessed on 25 March 2024).

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

Appendix A

	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 A1. 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 A2. 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 A3. Top 50 ICD predictions 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 A4. Top 50 ICD predictions using BERT model.

References

1. National Health Service. International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD-10), 5th Edition. 2022. Available online: https://classbrowser.nhs.uk/ref_books/ICD-10_2022_5th_Ed_NCCS.pdf (accessed on 10 January 2024).
2. PhysioNet. MIMIC-III Clinical Database (Version 1.4). 2016. Available online: https://physionet.org/content/mimiciii/1.4/ (accessed on 12 January 2024).
3. Mullenbach, J.; Wiegreffe, S.; Duke, J.; Sun, J.; Eisenstein, J. Explainable Prediction of Medical Codes from Clinical Text. *arXiv* **2018**, arXiv:1802.05695.
4. 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. [CrossRef]
5. Biswas, B.; Pham, T.-H.; Zhang, P. TransICD: Transformer Based Code-wise Attention Model for Explainable ICD Coding. *arXiv* **2021**, arXiv:2104.10652.
6. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
7. Han, K.; Xiao, A.; Wu, E.; Guo, J.; Xu, C.; Wang, Y. Transformer in transformer. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 15908–15919.
8. Li, W.; Fan, L.; Wang, Z.; Ma, C.; Cui, X. Tackling mode collapse in multi-generator GANs with orthogonal vectors. *Pattern Recognit.* **2021**, *110*, 107646. [CrossRef]
9. Lee, J.; Shin, H.; Kim, Y. The Effects of Hyperparameters in Deep Learning on Medical Dataset: A Case Study on EMR. *arXiv* **2020**, arXiv:2009.05451.
10. Alsentzer, E.; Murphy, J.R.; Boag, W.; Weng, W.; Jin, D.; Naumann, T.; McDermott, M.B.A. Publicly Available Clinical BERT Embeddings. *arXiv* **2019**, arXiv:1904.03323.
11. Choi, Y.; Kang, S. A systematic review of deep learning-based automated diagnosis of neurologic disorders using EEG signals. *BMC Med. Inform. Decis. Mak.* **2022**, *22*, 1–18.
12. 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
13. 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. [CrossRef]
14. 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.
15. 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.* **2020**, *8*, e17784. [CrossRef] [PubMed]
16. 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. [CrossRef] [PubMed]
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. 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. [CrossRef]
19. 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. [CrossRef]
20. 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. [CrossRef] [PubMed]
21. Xu, K.; Lam, M.; Pang, J.; Gao, X.; Band, C.; Mathur, P.; Papay, F.; Khanna, A.K.; Cywinski, J.B.; Maheshwari, K.; et al. Multimodal Machine Learning for Automated ICD Coding. In Proceedings of the Machine Learning Research, Ann Arbor, MI, USA, 9–10 August 2019; Volume 106, pp. 1–17
22. Biseda, B.; Desai, G.; Lin, H.; Philip, A. Prediction of ICD Codes with Clinical BERT Embeddings and Text Augmentation with Label-Balancing-using-MIMIC-III. *arXiv* **2020**, arXiv:2008.10492.
23. Edin, J.; Junge, A.; Havtorn, J.D.; Borgholt, L.; Maistro, M.; Ruotsalo, T.; Maaløe, L. Automated Medical Coding on MIMIC-III and MIMIC-IV: A Critical Review and Replicability Study. In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information, Taipei, Taiwan, 23–27 July 2023; pp. 2572–2582. [CrossRef]
24. Gero, Z.; Singh, C.; Cheng, H.; Naumann, T.; Galley, M.; Gao, J.; Poon, H. Self-Verification Improves Few-Shot Clinical Information Extraction. *arXiv* **2023**, arXiv:2306.00024.

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.