Leveraging large language models for medical text classification: a hospital readmission prediction case

Nazyrova, Nodira, Chahed, Salma, Chausalet, Thierry and Dwek, Miriam

This is a copy of the author's accepted version of a paper subsequently published in the proceedings of the IEEE 14th International Conference on Pattern Recognition Systems, London, United Kingdom, 15 - 18 Jul 2024.

The final published version is available online at:

https://doi.org/10.1109/icprs62101.2024.10677826

Leveraging large language models for medical text classification: a hospital readmission prediction case

Nodira Nazyrova[1], Salma Chahed[1], Thierry Chaussalet[1], Miriam Dwek[2]
[1]*School of Computer Science and Engineering,* [2]*School of Life Sciences*
University of Westminster, London, United Kingdom
N.Nazyrova@westminster.ac.uk, S.Chahed@westminster.ac.uk,
chausst@westminster.ac.uk, M.V.Dwek@westminster.ac.uk,

***Abstract*—In recent years, the intersection of natural language processing (NLP) and healthcare informatics has witnessed a revolutionary transformation. One of the most groundbreaking developments in this realm is the advent of large language models (LLM), which have demonstrated remarkable capabilities in analysing clinical data. This paper aims to explore the potential of large language models in medical text classification, shedding light on their ability to discern subtle patterns, grasp domain-specific terminology, and adapt to the dynamic nature of medical information. This research focuses on the application of transformer-based models, such as Bidirectional Encoder Representations from Transformers (BERT), on hospital discharge summaries to predict 30-day readmissions among older adults. In particular, we explore the role of transfer learning in medical text classification and compare domain-specific transformer models, such as SciBERT, BioBERT and ClinicalBERT. We also analyse how data preprocessing techniques affect the performance of language models. Our comparative analysis shows that removing parts of text with a large proportion of out-of-vocabulary words improves the classification results. We also investigate how the input sequence length affects the model performance, varying sequence length from 128 to 512 for BERT-based models and 4096 sequence length for the Longformers.**

The results of the investigation showed that among compared models SciBERT yields the best performance when applied in the medical domain, improving current hospital readmission predictions using clinical notes on MIMIC data from 0.714 to 0.735 AUROC. Our next step is pretraining a model with a large corpus of clinical notes to potentially improve the adaptability of a language model in the medical domain and achieve better results in downstream tasks.

***Index Terms*—hospital readmission prediction, domain-specific transformer models, BERT, ClinicalBERT, SciBERT, BioBERT, large language models.**

I. Introduction

The advent of large language models has revolutionised the landscape of natural language processing and opened new possibilities for extracting meaningful insights from unstructured clinical data. Clinical narratives are a valuable source of insights; however, traditional methods make it challenging to utilise them effectively. Such clinical narratives can be used to improve the quality of care by identifying the risk of adverse clinical outcomes, such as hospital readmission and mortality.

Hospital readmissions hold significant importance in the realm of healthcare due to their profound impact on both healthcare costs and patient outcomes. Patients may face complications or unresolved health issues upon returning to the hospital after discharge. High readmission rates suggest poor quality care, indicating the initial hospitalisation may have failed to address the patient's medical problems effectively, or there were shortcomings in post-discharge care planning and follow-up [1]. Predicting hospital readmissions using traditional methods poses several challenges, often stemming from the limitations of relying on structured data and conventional statistical approaches. Traditional methods lack the depth and granularity needed to capture the complexity of a patient's health status and predict the likelihood of readmission accurately. Factors contributing to readmissions, such as social determinants of health, patient behaviour, and environmental factors, are often not fully accounted for in these models. This holds particular significance for older patients with multiple chronic conditions and complex health profiles. Addressing these challenges requires innovative approaches, and this is where the integration of large language models and advanced natural language processing techniques can play a transformative role in improving the accuracy and effectiveness of predicting hospital readmissions.

The potential of large language models in comprehending context, context shift, and semantics embedded in narrative text helps extract insights from unstructured clinical data. Unlike structured data, which often fails to capture the rich details of a patient's journey through the healthcare system, clinical notes provide a comprehensive and detailed summary that includes subjective observations, expert opinions, and contextual nuances that contribute significantly to a patient's healthcare trajectory.

While LLMs, in particular, transformer-based models (e.g. BERT), have demonstrated significant capabilities in various NLP tasks, including medical text classification, several limitations still exist. Such models can capture general language patterns, but they may lack the domain-specific knowledge required for accurate medical text classification. Therefore, in this study, we aim to use both general transformer-based model BERT [2], as well as biomedical domain-specific models, such as BioBERT [3], ClinicalBERT [4], [5] and SciBERT [6].

To benefit from both structured and unstructured clinical data, multimodal predictive models can be utilised. Such an approach can provide a more comprehensive and holistic view of a patient's health, enabling a more nuanced understanding of the factors contributing to clinical outcomes.

Hence, this research focuses on the application of BERT-based model on hospital discharge summaries to predict readmissions among older adults. We explore the role of transfer learning in medical text classification and compare domain-specific transformer models, such as SciBERT, BioBERT and ClinicalBERT. We investigate how text preprocessing techniques affect predictive performance and discuss the interpretation of the model output. Finally, we analyse how the integration of structured and unstructured data affects the classification performance.

II. RELATED WORK

Unplanned hospital readmission prediction models typically use structured clinical data such as demographics, diseases, medication, and vital signs [7]–[9]. While structured data provides valuable information about the diseases, treatment, and vital signs, it lacks context, missing critical details, such as patient behaviour, social determinants of health, and lifestyle factors. Clinical narratives may represent complex clinical scenarios, capturing the nuances, like the severity of a patient's emotional or psychological state [10], [11]. Among elderly patients such factors as decline in mobility, risks of falls, fatigue, and functional dependence will most likely be captured only in clinical narratives. Traditionally, rule-based approaches have been commonly used to analyse textual data, and it usually requires clinical knowledge to identify a set of explicit pattern-matching rules [12], [13]. NLP and text mining techniques have been used to derive social risk factors from clinical notes in the form of frequent features [14]–[16] or topics from clinical narratives [10].

Bag of Words remains a popular algorithm for clinical note analysis. However, most studies on predicting hospital readmission using this approach show relatively low performance, with AUROC values ranging between 0.55 and 0.65 [17]. On the contrary, word embeddings have been used to create more complex representations and achieve better results. Deep learning methods, such as recurrent neural networks, long-term short-term memory neural networks [16], and convolutional neural networks [17], [18], have been more successful in clinical note analysis tasks with AUROC varying from 0.65 to 0.77 for various sample size and cohorts of patients. Contextual embedding techniques such as BERT [2] and XLNet [19] have further advanced this approach by capturing the meaning of words depending on their context, capturing long-term dependencies, making them suitable for representing unstructured text and encouraging the development of domain-specific versions.

A 30-day hospital readmission classification task is, in general, a challenging task, and there are very few studies reporting high predictive performance. Moreover, a study [20] exploiting BERT for readmission classification shows that it is more difficult to predict 30-day hospital readmission than 48-hour or 7-day readmission using clinical notes. In their experiments, there was a drastic decrease in predictive performance from 0.81-0.86 AUROC for 48-hour readmission, 0.67-0.81 AUROC for 7-day readmission to 0.59-0.62 AUROC for 30-day readmission. These results are improved when domain-specific versions of BERT are used. Huang et al. [5] have shown that pre-training a BERT-based model, ClinicalBERT, on EHR doctor notes, specifically discharge summaries, followed by fine-tuning, can yield better results than other NLP models. The ClinicalBERT model outperformed the Bag of Words (AUROC 0.68) and BI-LSTM (AUROC 0.69) models in the hospital prediction task with AUROC 0.714. This pre-trained model was later used in [21] for the same 30-day hospital readmission classification task and achieved similar results of AUROC 0.721. Another biomedical version of BERT, BioBERT [3], was initialized from BERT and trained on the large biomedical corpus. In their study, they did not test BioBERT in hospital readmission classification task. This pre-trained model was used in [4], a clinicalBERT model, initialized from BioBERT and pre-trained on MIMIC-III dataset. However, the model was not tested in the hospital readmission prediction task either.

In the next section, various biomedical domain implementations of contextual embeddings will be discussed and their performance in hospital readmission classification in elderly patients will be analysed. To the best of our knowledge, there is no benchmarking study analysing the performance of several domain-specific BERT-like models in the clinical domain, as well as the general BERT base model. This study aims to address this knowledge gap.

III. METHODOLOGY

A. Transformer-based language models

1) BERT: Bidirectional Encoder Representations from Transformers (BERT) is a language representation model, which employs a transformer architecture. BERT applies multiple self-attention mechanisms from transformer architecture to learn information, by assigning weights to each word in the sentence based on its relevance. Unlike previous language models, BERT is designed to pretrain deep bidirectional representations from unannotated text by simultaneously considering both the left and right context in all layers of the model [2]. This bidirectional approach has proven useful in capturing the intricacies of language semantics. BERT is pre-trained on an extensive corpus of text data through unsupervised learning, involving two primary tasks: masked language modelling and next-sentence prediction. The text embeddings and model parameters are fit using stochastic optimisation. The pre-trained BERT model can be fine-tuned with just one additional output layer to create models for a wide range of tasks, including text classification, without substantial task-specific architecture modifications. BERT has achieved state-of-the-art results in various NLP benchmarks and tasks, encompassing sentiment analysis (SA), text classification, named entity recognition (NER) and others.

BERT uses WordPiece embeddings [22] with a 30,000 token vocabulary. The first token of every sequence is always a special classification token ([CLS]]). The final hidden state corresponding to this token is used as the aggregate sequence representation for classification tasks.

TABLE I
LIST OF TEXT CORPORA USED FOR BIOBERT

Training Corpus	Number of words	Domain
English Wikipedia	2.5B	General
Books Corpus	0.8B	General
PubMed Abstracts	4.5B	Biomedical
PMC Full-text articles	13.5B	Biomedical

While BERT has shown remarkable performance in various NLP tasks, there are limitations and challenges associated with its application in healthcare and medical tasks. The medical domain often requires specialised knowledge, and obtaining large-scale labelled data for medical tasks can be challenging. Pre-training BERT on a general corpus may not capture domain-specific nuances. Medical texts contain numerous rare and specialised terms, including various medical conditions, drug names, and procedures. Moreover, BERT's contextual embeddings may not fully capture the intricacies of medical knowledge. In order to address these challenges, biomedical versions of BERT, such as BioBERT, SciBERT and ClinicalBERT were tested and compared to BERT.

2) BioBERT: BioBERT is a pre-trained language representation model for the biomedical domain. BioBERT was initialized with weights from BERT and further pre-trained on biomedical corpora (see Table I). For tokenization, BioBERT uses WordPiece tokenization with a mechanism to mitigate out-of-vocabulary words. Experimental results show that BioBERT performs superiorly over BERT on all benchmarking datasets in NER, relation extraction (RE), and question-answering (QA) tasks, outperforming most state-of-the-art models [3].

3) SciBERT: SciBERT model exploits its own SciVocab tokenization. SciVocab is a new WordPiece vocabulary constructed on scientific corpus. [6]. The resulting vocabulary size is set to 30K to match BaseVocab size, with the 42% overlap between BaseVocab and SciVocab. This shows a substantial difference between the words used in the general domain and scientific texts. SciBERT was trained on random scientific papers from Semantic Scholar, with 18% text from the computer science domain and 82% of text from biomedical domain. Experimental results show that SciBERT outperforms BERT-base on biomedical tasks. When compared to BioBERT, SciBERT outperforms BioBERT results on BC5CDR and ChemProt benchmarking datasets, and performs similarly on JNLPBA despite being trained on a substantially smaller biomedical corpus [6].

TABLE II
LIST OF TEXT CORPORA USED FOR SCIBERT

Training Corpus	Number of words	Domain
English Wikipedia	2.5B	General
Books Corpus	0.8B	General
Semantic Scholar	3.1B	Biomedical and Computer Science

4) ClinicalBERT: ClinicalBERT is an application of the BERT model to clinical corpora, using the same pretraining tasks as BERT. Two independent groups of researchers developed the ClinicalBERT model using MIMIC-III data. Alsentzer et al. [4] initialized ClinicalBERT from BERT-Base and BioBERT models and pre-trained it on MIMIC-III clinical notes. Experimental results showed an improvement in two clinical NER tasks and one medical natural language inference task.

Huang et al. [5] simultaneously developed the ClinicalBERT model also using MIMIC-III clinical notes and initialized from BERT. ClinicalBERT model was tested in clinical language modelling, clinical word similarity and hospital readmission prediction classification tasks. For the hospital readmission prediction task using discharge summaries ClinicalBERT model achieved 0.714 AUROC and 0.701 AUPRC.

As ClinicalBERT, like other BERT models, has a fixed length of input sequence, the clinical notes are split into several chunks. Predictions for patients with many notes are computed by binning the predictions on each sequence. Huang et al [5] proposed the following formula for calculating the probability of readmission using the predictions for each subsequence:

$$P(readmit = 1|h_{\text{patient}}) = \frac{P^n_{max} + P^n_{mean} n/c}{1 + n/c} \quad (1)$$

where **c** is a scaling factor that controls the amount of influence of the number of subsequences **n**, and $\mathbf{h_{patient}}$ is the implicit representation ClinicalBERT computes from the entirety of a patient's notes. $\mathbf{P^n_{max}}$ is the maximum of probability of readmission across the **n** subsequences, and $\mathbf{P^n_{mean}}$ is the mean of the probability of readmission across the **n** subsequences a patient's notes have been split into.

Computing readmission probability using this equation outperforms predictions using the mean for each subsequence by 3-8% as authors suggest, and the results of our experiments confirm this claim.

5) Longformer: One of the biggest limitations of transformer-based models is their inability to process long sequences due to their self-attention operation, which scales quadratically with the sequence length. Longformer models were developed to address this limitation using an attention mechanism that scales linearly with sequence length, enabling the processing of long tokens. BERT and BERT-based models are limited to 512 token sequence length. Discharge summaries are usually longer; in MIMIC-IV, pre-processed notes vary from 1183 words to 1951 (25^{th} and 75^{th} percentiles) words. This poses additional challenges to model pretraining and can result in the loss of important cross-partition information, requiring complex architectures to address such issues. Longformers enable using sequences with up to 4096 tokens due to its updated transformer architecture. Longformer's attention mechanism is a combination of a windowed local-context self-attention and an end task motivated global attention that encodes inductive bias about the task [23]. For this study, Longformer implementation by AllenNLP longformer-base-4096 was used.

The full methodology of hospital readmission prediction using language models is presented in Fig. 1. The Bag of

Words model is used for comparison purposes. The model was implemented using scikit-learn CountVectorizer with 3000 features and logistic regression algorithm.

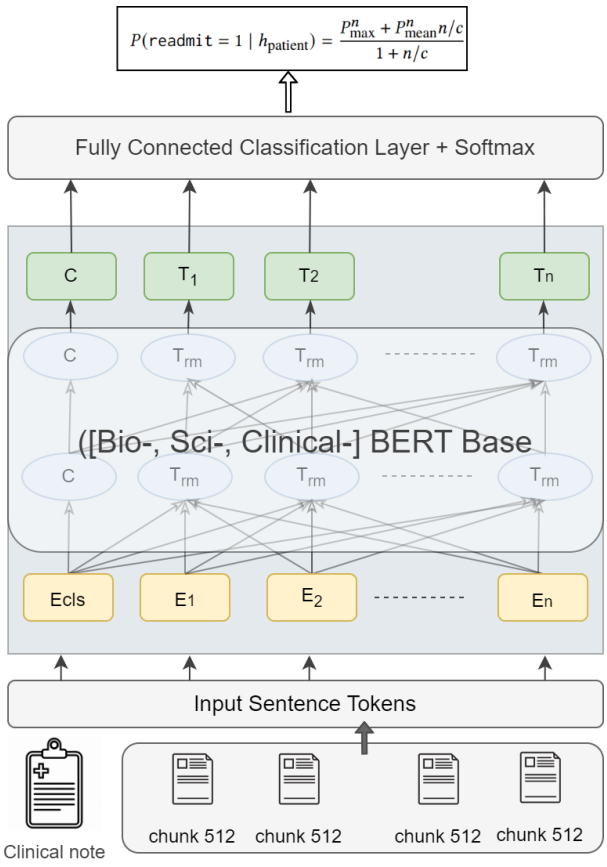

Fig. 1. Overview of hospital readmission classification model development using BERT, BioBERT, SciBERT and ClinicalBERT models, where E is an encoder layer, T is the transformer layer, and C is the classification layer. The figure was adapted from [2].

B. Data

We used the Medical Information Mart for Intensive Care IV (MIMIC-IV) dataset for the classification task [24]. MIMIC-IV Note contains 331,794 de-identified discharge summaries from 145,915 patients admitted to the hospital and emergency department at the Beth Israel Deaconess Medical Center in Boston, USA. The structure of the database was updated, and unlike MIMIC-III, MIMIC-IV no longer contains categorized clinical notes (e.g. nursing notes, physician notes, discharge summaries, radiology reports). All clinical notes are presented as discharge summaries. MIMIC-IV only contains discharge and radiology reports, and for this study, only discharge summaries were used. Discharge summaries are long narratives which describe the reason for a patient's admission to the hospital, their hospital course, family history, physical examination results, prescriptions and discharge instructions. For this analysis, a cohort of elderly patients over 65 years of age was selected. Admissions where a patient is readmitted to the hospital within 30 days after the index discharge are considered as readmitted patients. This includes only unplanned readmissions; hence, elective admission types are not considered as readmission cases. Admissions resulting in death in hospital or out of hospital within 30-days of discharge are excluded from the study. Data was split into 80%, 10%, and 10% for training, validation and test sets, respectively. Models were trained using the PyTorch and Huggingface libraries.

C. Data Preprocessing

For BERT-like models, minimal preprocessing is required. All text is converted to lowercase, line breaks and carriage returns are removed. Discharge summaries contain a large number of special characters, and those special characters that do not contain contextual meaning are removed (e.g., `'='`, `'*'`, `'_'`). Some sections of clinical notes contain a large proportion of out-of-vocabulary words, such as the 'Pertinent results' section, see below:

```
___ 10:09AM BLOOD cTropnT-<0.01
___ 05:25AM BLOOD Calcium-9.4 Phos-4.1 Mg-2.2
___ 05:02AM BLOOD Calcium-9.0 Phos-4.2 Mg-2.1
___ 09:05PM BLOOD Cholest-220*
___ 09:05PM BLOOD Triglyc-70 HDL-53 CHOL/HD-4.2
    LDLcalc-153* LDLmeas-155*
___ 04:50PM BLOOD TSH-0.20*
___ 11:00AM BLOOD T4-11.7 T3-99
```

As BERT models rely on a fixed-size vocabulary, out-of-vocabulary words are tokenized into sub-words or special tokens. Sections of discharge summaries with pure lab events result in inaccurate text representations and noisy embeddings, therefore degrading the model performance. The results of the experiments demonstrated that removing lab events data from clinical notes results in improved accuracy and AUROC and better model convergence. Therefore, such sections with purely cryptic data were removed. Our experimental results are corroborated by similar findings in the existing literature [25], [26]. The study shows a significant degradation in the performance of BERT on fundamental NLP tasks like sentiment analysis and textual similarity in the presence of noise on benchmark datasets [25].

As the average length of clinical notes is times bigger than the maximum sequence length that BERT-based models can analyse, each clinical note was split into chunks of 128, 256, and 512 tokens for testing. Full-length clinical notes were used with the longformer model only.

D. Fine-tuning

For fine-tuning, different model hyperparameters for learning rate, batch size and weight decay were experimented. For BERT, BioBERT, SciBERT and ClinicalBERT models the best performance was achieved with a learning rate of 2e-6, weight decay = 0.01, 10 epochs with early stopping and batch size of 8 for models with 512 sequence length, 16 for models with 256 token sequence length and 32 for models with 128 token sequence lengths. The longformer model was trained on full-length clinical notes with 5 epochs.

IV. RESULTS

Table III shows the test performance of models built with balanced dataset settings for elderly patients. We can see that longer sequence lengths result in higher AUROC and AUPRC values for all BERT-based models. Longformer that uses the longest sequence length, however, does not perform very well compared to other transformer-based models. Longformer models are computationally expensive and require large GPU and running time, so it was not possible to optimise it, given the computing resource constraints. Therefore, the Longformer model did not converge in 5 epochs, as we could observe from training and validation loss. The probabilities of readmission for BERT-based models were computed using the predictions for each subsequence using the formula in [5]. Such an approach was significantly more efficient than using simple means of probabilities.

Among all tested models SciBERT has the highest predictive performance with 0.7351 AUROC and 0.5361 AUPRC, followed by BioBERT model with 0.7298 AUROC and 0.4866 AUPRC. Among BERT-based models original BERT-base has the lowest AUROC. Experimenting with tokenizers also demonstrated that SciBERT tokenizers had a better vocabulary for medical text. Tokenization with BioBERT resulted in the highest number of out-of-vocabulary words. On average, with SciBERT, 83% of words had the corresponding token in the vocabulary, with BERT 77%, with ClinicalBERT 75%, and with BioBERT only 74% of words in the vocabulary. The best-performing model, SciBERT, was further used with structured data, including patient demographics, administrative data, diseases and medication-related data. Integrating structured and unstructured clinical positively affected classification results. The classification model with structured data was designed as described in [27] had AUROC of 0.7401 on the same sample. Combining this model with SciBERT helped to increase this metric to 0.7561 as shown in Fig. 2.

The Bag of Words model was only used for benchmarking purposes and enabling the interpretation of clinical notes in hospital readmission tasks. When vector representations are used as feature inputs of the model, feature importance can be used for analysing the contributors to hospital readmissions. Hence, feature importance analysis demonstrated that such words as 'chemotherapy', 'recurrent', 'transplant', 'biopsy', 'multiple', 'ulcer', 'unchanged', 'confusion', 'previous' are highly linked with readmissions. For BERT-based models, self-attention mechanism was used as an indicator of word importance. The higher attention weight indicates the relative importance or relevance of specific words or tokens in the input sequence to the model's understanding of the context and semantics of the text [2]. The model places greater emphasis on tokens with a higher attention weight, making it indicative of readmission. In the example of the sentence: "patient has had multiple admissions, most recently two weeks prior to admission, for recurrent right pleural effusion thought to be malignancy related", we can see the words with higher attention weights in the heatmap in Fig. 3. 'Recurrent', 'pleural effusion', 'malignancy' are the most important words for SciBERT classifier, which aligns with Bag of Words model feature importance, where oncology-related words, and words signifying the reoccurring events were highly correlated with hospital readmissions.

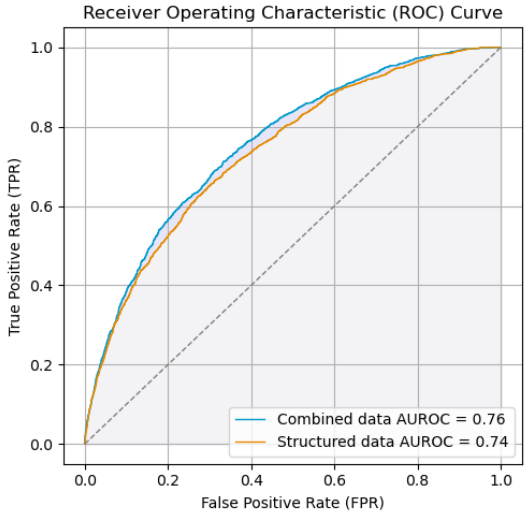

Fig. 2. Comparison of classification performance of two models: structured data from MIMIC-IV (demographics, diseases, medications, procedures, vital signs) with the XGBoost algorithm; and SciBERT model using unstructured clinical notes using the same sample from MIMIC-IV dataset.

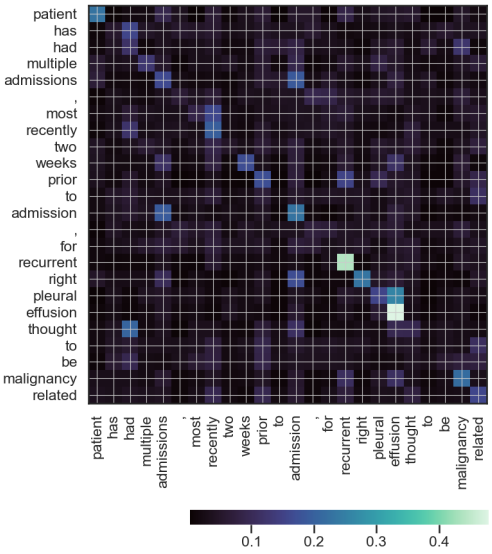

Fig. 3. The self-attention mechanism is used to interpret model predictions on clinical notes as in [5]. In this example, SciBERT model was used for interpretation. The x-axis labels are query tokens and the y-axis labels are key tokens. High attention weights indicates token importance for the output of the model (e.g., 'recurrent', 'effusion', 'pleural', 'pleural effusion', 'malignancy').

V. DISCUSSION

The experimental outcomes underscore the significance of employing domain-specific transformer-based models within the clinical domain. As expected, all domain-specific BERT

TABLE III
CLASSIFICATION PERFORMANCE METRICS FOR ALL MODELS.

Model	***AUROC***					***AUPRC***				
	128	256	512	4096	3000*	128	256	512	4096	3000*
BERT-base	0.7088	0.7161	0.7260			0.5293	0.5331	0.5345		
BioBERT	0.7100	0.7195	0.7298			0.5112	0.5019	0.4866		
ClinicalBERT	0.7152	0.7202	0.7291			0.5347	0.5328	0.5216		
SciBERT	0.7154	0.7171	0.7351			0.5291	0.5355	0.5361		
Longformer				0.6768					0.4989	
Bag of Words					0.6600					0.4843

*For the Bag of Words model, the value indicates the number of features used in the model.

models have performed better than general BERT. A longformer model that is not pre-trained on the biomedical domain training corpora also demonstrated poor predictive performance. However, it should be noted that the Longformer model requires large computational resources, so it has not been fully optimised and did not converge in 5 epochs. From the models in comparison SciBERT achieved the best results in predicting elderly patients 30-day hospital readmissions with 512 maximum sequence length. When compared to existing similar studies, this result outperforms the classification results in [5] with 0.714 and results in [21] with AUROC of 0.721. A better performance of the SciBERT model could be attributed to its tokenization methodology. Unlike other analysed BERT models which use WordPiece tokenizer, SciBERT uses SentencePiece tokenizer. Among distinctive features of the result of tokenization, we can highlight a substantially larger amount of tokens with digits, which could be useful for medical data. Moreover, SciBERT tokenization was the most efficient having the least number of out-of-vocabulary words.

The Bag of Words model did not perform well. However, it can contain useful information for the interpretability of the results. In fact, top words identified with a Bag of Words and the logistic regression model were often associated with the higher attention weight in the SciBERT model. Hence, words like 'recurrent', 'multiple', 'effusion', 'transplant', 'cancer',' haemodialysis', chronic, and 'tracheostomy' were often associated with the risk of readmission using the Bag of Words model and had high attention weight, indicating their importance for BERT-like models.

The attention mechanism in models like BERT is used to determine the importance of each token in the input sequence relative to others when generating the output. High attention weight values can indicate tokens that are important for the classification task; they do not directly specify whether a word is important for a positive or negative class. So mixing the attention weight evaluation approach with feature importance analysis with simpler more interpretable models can be useful for the understanding of classification results.

This confirms that unstructured clinical notes can bring additional contextual information and help enhance hospital readmission predictions. Combining unstructured clinical notes with structured data such as demographics, diseases, medications and some important vital signs resulted in higher AUROC than the model exploiting structured or unstructured data separately. Models exploiting only structured data achieved an AUROC of 0.740, the SciBERT model with clinical notes achieved an AUROC of 0.735, and a combination of both models resulted in an AUROC of 0.757. For the classification task of high uncertainty it is a substantial increase in performance.

Examining the contents of clinical notes reveals that most of the information available in discharge summaries can be found in the MIMIC-IV dataset in a structured format, such as diseases of the patient, demographics, lab events or vital signs. Thus, clinical notes contain a lot of redundant information. However, parts of the summary describing the patient's family history, current social status of the patient, psychological state of patients or refusal to follow the doctor's instructions are only available in discharge summaries. Such information should be the most interesting in analysing clinical notes. In future, we will consider deriving and studying such psychosociological factors in more detail.

Among the limitations of the study is the insufficiency of data. It is not possible to identify patients readmitted to different hospitals. The date of death is only available for patients who died within a year of hospitalization. More complete information about the patient pathways would enhance the performance of classifiers.

Also, most of the developed models had a tendency to overemphasize a positive class due to the subsampling technique for battling class imbalance. As a positive class bears more risks for patients, bias towards a positive class was deemed appropriate.

VI. CONCLUSION

This study provided a comparative overview of the biomedical domain and general BERT-based models, as well as the Longformer and Bag of Words models to predict hospital readmission among older adults. We conclude that all domain-specific models perform better than a general model for our task, and among them, SciBERT shows the best results due to its different approach to tokenization. We also conclude that integrating both structured and unstructured data results in improved predictions. In future, we are aiming to develop a new clinicalBERT model initialized from SciBERT and using preprocessed MIMIC-IV data.

REFERENCES

[1] NHS Digital. Hospital Admitted Patient Care Activity 2019-20, https://digital.nhs.uk/dataand-

information/publications/statistical/hospital-admitted-patient-care-activity/2019-20; 2020, last accessed 2024/2/12

[2] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[3] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234-1240.

[4] Alsentzer, E., Murphy, J. R., Boag, W., Weng, W. H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323.

[5] Huang, K., Altosaar, J., & Ranganath, R. (2019). Clinicalbert: Modeling clinical notes and predicting hospital readmission. arXiv preprint arXiv:1904.05342.

[6] Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150.

[7] Jones, C. D., Falvey, J., Hess, E., Levy, C. R., Nuccio, E., Barón, A. E., Masoudi, F. A., & Stevens-Lapsley, J. (2019). Predicting Hospital Readmissions from Home Healthcare in Medicare Beneficiaries. Journal of the American Geriatrics Society, 67(12), 2505–2510. https://doi.org/10.1111/jgs.16153

[8] Low, L. L., Liu, N., Wang, S., Thumboo, J., Ong, M. E., & Lee, K. H. (2016). Predicting 30-Day Readmissions in an Asian Population: Building a Predictive Model by Incorporating Markers of Hospitalisation Severity. PloS one, 11(12), e0167413. DOI: 10.1371/journal.pone.0167413

[9] Glans, M., Kragh Ekstam, A., Jakobsson, U., Bondesson, Å., & Midlöv, P. (2020). Risk factors for hospital readmission in older adults within 30-days of discharge - a comparative retrospective study. BMC geriatrics, 20(1), 467. https://doi.org/10.1186/s12877-020-01867-3

[10] Goh, K. H., Wang, L., Yeow, A. Y. K., Ding, Y. Y., Au, L. S. Y., Poh, H. M. N., Li, K., Yeow, J. J. L., & Tan, G. Y. H. (2021). Prediction of Readmission in Geriatric Patients From Clinical Notes: Retrospective Text Mining Study. Journal of medical Internet research, 23(10), e26486. https://doi.org/10.2196/26486

[11] Navathe, A. S., Zhong, F., Lei, V. J., Chang, F. Y., Sordo, M., Topaz, M., Navathe, S. B., Rocha, R. A., & Zhou, L. (2018). Hospital Readmission and Social Risk Factors Identified from Physician Notes. Health services research, 53(2), 1110–1136. https://doi.org/10.1111/1475-6773.12670

[12] Spasic, I., Nenadic, G. (2020). Clinical Text Data in Machine Learning: Systematic Review. JMIR Med Inform. 8(3):e17984. Published 2020 Mar 31.

[13] Wasfy, J. H., Singal, G., O'Brien, C., Blumenthal, D. M., Kennedy, K. F., Strom, J. B., Spertus, J. A., Mauri, L., Normand, S. L. T., and Yeh, R. W. (2015). Enhancing the Prediction of 30-Day Readmission after Percutaneous Coronary Intervention Using Data Extracted by Querying of the Electronic Health Record. Circ. Cardiovasc. Qual. Outcomes, vol. 8, no. 5, pp. 477–485.

[14] Davis, S., Zhang, J., Lee, I. et al. (2022). Effective hospital readmission prediction models using machine-learned features. BMC Health Serv Res 22, 1415. https://doi.org/10.1186/s12913-022-08748-y

[15] Agarwal, C. Baechle, R. Behara and X. Zhu. (2018). A Natural Language Processing Framework for Assessing Hospital Readmissions for Patients With COPD. IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 2, pp. 588-596, doi: 10.1109/JBHI.2017.2684121.

[16] Ashfaq, A., Sant'Anna, A., Lingman, M., Nowaczyk, S. (2019). Readmission prediction using deep learning on electronic health records. J. Biomed. Inform., 97

[17] Liu, X., Chen, Y., Bae, J., Li, H., Johnston, J., & Sanger, T. (2019). Predicting Heart Failure Readmission from Clinical Notes Using Deep Learning. 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). doi: 10.1109/bibm47256.2019.8983095.

[18] Craig, E., Arias, C., & Gillman, D. (2017). Predicting readmission risk from doctors' notes. arXiv preprint arXiv:1711.10663.

[19] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems, 32.

[20] Kovoor, J. G., Bacchi, S., Gupta, A. K., Stretton, B., Nann, S., Aujayeb, N., Lu, A., Nathin, K., Lam, L., Jiang, M., Lee, S., To, M., Ovenden, C. D., Hewitt, J. N., Goh, R., Gluck, S., Reid, J., Khurana, S., Dobbins, C., Maddern, G. J. (2023). Surgery's Rosetta Stone: Natural language processing to predict discharge and readmission after general surgery. Surgery, 174(6), 1309–1314. https://doi.org/10.1016/j.surg.2023.08.021

[21] Thapa, N. B., Seifollahi, S., Taheri, S. 2022. Hospital Readmission Prediction Using Clinical Admission Notes. In Proceedings of the 2022 Australasian Computer Science Week (ACSW '22). Association for Computing Machinery, New York, NY, USA, 193–199. https://doi.org/10.1145/3511616.3513115

[22] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... & Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.

[23] Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. arXiv preprint arXiv:1903.10676.

[24] Johnson, A.E.W., Bulgarelli, L., Shen, L. et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data 10, 1 (2023). https://doi.org/10.1038/s41597-022-01899-x

[25] Kumar, A., Makhija, P., Gupta, A. (2020). Noisy Text Data: Achilles' Heel of BERT. WNUT.

[26] Schick, T., Schütze, H. (2019). Attentive mimicking: Better word embeddings by attending to informative contexts. arXiv preprint arXiv:1904.01617.

[27] Nazyrova, N., Chaussalet, T.J. and Chahed, S. (2022). Machine Learning models for predicting 30-day readmission of elderly patients using custom target encoding approach. International Conference on Computational Science ICCS 2022. London, UK 21 - 23 Jun 2022 Springer. https://doi.org/10.1007/978-3-031-08757-8_12