Applied Computer Science, vol. 20, no. 2, pp. 60–74
doi: 10.35784/acs-2024-16

Submitted: 2024-02-14 | Revised: 2024-04-09 | Accepted: 2024-04-14

Keywords: Artificial intelligence, Natural language processing (NLP), Classification problem, International classification of diseases (ICD), Bidirectional Encoder Representations from Transformers (BERT), MIMIC-III

Dilek AYDOGAN-KILIC [0000-0002-9194-9400]*,
Deniz Kenan KILIC [0000-0002-6996-3425]*,
Izabela Ewa NIELSEN [0000-0002-3506-2741]*

EXAMINATION OF SUMMARIZED MEDICAL RECORDS FOR ICD CODE CLASSIFICATION VIA BERT

Abstract

The International Classification of Diseases (ICD) is utilized by member countries of the World Health Organization (WHO). It is a critical system to ensure worldwide standardization of diagnosis codes, which enables data comparison and analysis across various nations. The ICD system is essential in supporting payment systems, healthcare research, service planning, and quality and safety management. However, the sophisticated and intricate structure of the ICD system can sometimes cause issues such as longer examination times, increased training expenses, a greater need for human resources, problems with payment systems due to inaccurate coding, and unreliable data in health research. Additionally, machine learning models that use automated ICD systems face difficulties with lengthy medical notes. To tackle this challenge, the present study aims to utilize Medical Information Mart for Intensive Care (MIMIC-III) medical notes that have been summarized using the term frequency-inverse document frequency (TF-IDF) method. These notes are further analyzed using deep learning, specifically bidirectional encoder representations from transformers (BERT), to classify disease diagnoses based on ICD codes. Even though the proposed methodology using summarized data provides lower accuracy performance than state-of-the-art methods, the performance results obtained are promising in terms of continuing the study of extracting summary input and more important features, as it provides real-time ICD code classification and more explainable inputs.

1. INTRODUCTION AND BACKGROUND

This section introduces the background of the International Classification Diseases (ICD), emphasizing the importance of health statistics and the need for international comparability. Furthermore, the ICD code structure is explained, and the motivation for this study is clarified.

* Aalborg University, Department of Materials and Production, Operations Research Group, Denmark, dilekaydogankilic@gmail.com, denizkk@mp.aau.dk, izabela@mp.aau.dk

1.1. The role of the International Classification Diseases (ICD) in the International Health Statistics System

Health statistics provide numerical data that help experts study risk factors, track diseases, evaluate policies and programs, and assess healthcare quality.

International comparability of health systems is crucial for effective development and positive global interactions. World Health Organization (WHO), European Statistical Office (Eurostat), and Organization for Economic Co-operation and Development (OECD) collect internationally comparable data to monitor and compare health systems. Eurostat ensures standardization of health-related data and publishes the European Statistics Code of Practice (CoP) (European Commission, Eurostat, 2018) with 16 data quality and standardization principles. However, a significant amount of human resources is required to ensure these standards. In this sense, artificial intelligence can be an alternative solution and it can provide more objective approaches. The coronavirus disease 2019 (COVID-19) pandemic highlights the need to use the latest technology to produce timely statistical data (European Commission, 2020).

Eurostat's 2020-2024 strategic plan includes the use of artificial intelligence (AI) and machine learning (ML) to produce statistics, specifically in "Specific Objective 2" (European Commission, 2020). This initiative aims to incorporate new technologies and data sources into official statistics.

The goal of providing accessible and affordable health services to every individual and society is common among the European Union (EU) member states despite their differing health structures and financing methods. To evaluate healthcare systems, it is necessary to establish a standard health statistics system that measures the financial, human, and technical resources within the healthcare sector and their distribution among various healthcare activities, groups of healthcare providers, or healthcare professionals (Eurostat, 2012).

Eurostat has a vital place in the standardization of health statistics, as in other statistical fields. Moreover, to ensure worldwide partnership, joint studies (e.g., joint data collection, questionnaires, etc.) are carried out by WHO, OECD, and Eurostat on healthcare expenditure and non-expenditure statistics. Furthermore, Eurostat provides causes of death statistics, the European Health Interview Survey (EHIS), and health variables in EU statistics on income and living conditions (SILC).

ICD classification system has a very critical role in health statistics. It is a standardized disease categorization and coding system used by member countries of the WHO (Hsu et al., 2020), especially European countries, in the classification of diseases. On the other hand, it is a vital tool that helps provide information about diseases and deaths worldwide. Although widely used, the system can be challenging due to its complex structure and over seventy thousand unique codes. This complexity can result in assigning incorrect codes, which reduces data quality and causes errors in payment systems. The increased detail in the coding system may also cause patient visits to take longer and require extensive training on the new code set during the ICD transition phase. To overcome these issues, automatic ICD coding studies using machine learning have been explored since the 1990s (Yan et al., 2022).

The ICD is an essential tool used for diagnosis that standardizes the classifications of diseases globally. It is compiled and maintained by the WHO. Since the 19th century, 11 versions have been produced, and all member states of WHO use this coding system, even if not all of them are using the latest version. Therefore, it plays a critical role in providing

knowledge on the extent, causes, and consequences of disease and death worldwide. Additionally, it provides clinical term baselines for primary, secondary, and tertiary care and the cause of death certificates. This common usage worldwide ensures data standardization and enables comparable studies. Moreover, the data and statistics are used to assist payment systems, health services research, service planning, and quality and safety management.

1.2. ICD code structure

In this part, the ICD system is briefly explained to both give a general idea about the coding system and to reveal the complexity mentioned in the previous part of the study. There are 11 versions of ICD. As the comprehensiveness of it becomes stronger with the development of each new version, the complexity inevitably increases.

The latest version, ICD-11, consists of 26 chapters. It has 17,000 codes and over 120,000 codable terms (Pezzella, 2022). The main chapters are listed in Table 1.

Tab. 1. Main chapters of ICD-11

Chapter	Title
01	Infectious diseases
02	Neoplasms
03	Diseases of the blood and blood-forming organs
04	Disorders of the immune system
05	Conditions related to sexual health
06	Endocrine, nutritional and metabolic diseases
07	Mental and behavioural disorders
08	Sleep-Wake disorders
09	Diseases of the nervous system
10	Diseases of the eye and adnexa
11	Diseases of the ear and mastoid process
12	Diseases of the circulatory system
13	Diseases of the respiratory system
14	Diseases of the digestive system
15	Diseases of the skin
16	Diseases of the musculoskeletal system and connective tissue
17	Diseases of the genitourinary system
18	Pregnancy, childbirth and the puerperium
19	Certain conditions originating in the perinatal period
20	Developmental anomalies
21	Symptoms, signs, clinical forms, and abnormal clinical and laboratory findings, not elsewhere classified
22	Injury, poisoning and certain other consequences of external causes
23	External causes of morbidity and mortality
24	Factors influencing health status and contact with health services
25	Codes for special purposes
26	Extension Codes

The ICD-11 codes are alphanumeric and fall within the range of 1A00.00 to ZZ9Z.ZZ. The initial character of a code is always associated with the chapter number and can be either

a letter or a number. Codes that begin with "X" signify an extension code. To avoid confusion with the numbers "0" and "1", the letters "O" and "I" are omitted. Additionally, the second place in a code is a letter, distinguishing it from ICD-10. Finally, the third place is a number, which prevents the possibility of "undesirable words" (World Health Organization, 2023).

The ICD-11 offers a complex navigation feature where a single term should have one or more, occasionally many, conceptual parents under its structure. This feature was not available in previous versions. The 11th version solves the problem by connecting the diagnosis of stomach cancer, which was previously only a child of the cancer tree, to the gastrointestinal illness parent as well (Chute & Çelik, 2021). The ICD-11 offers a mechanism to construct clusters by joining codes when more expressive explaining capacity is needed than any one category can offer. Any category can be coded with one or more stem codes. Moreover, stem codes can be qualified with one or more extension codes (Harrison et al., 2021). Harrison et al. (2021) present the code in Figure 1 to give an example of multiple stem code and extension code.

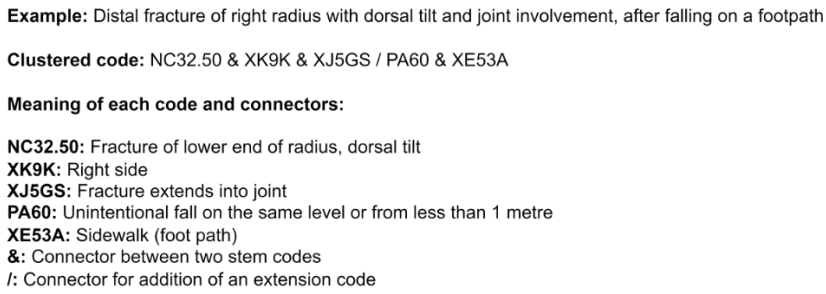

Fig. 1. An example of an ICD-11 code

1.3. Motivation

The complexity and intermittent updating of the codes make it necessary to automate the ICD. Automating the ICD coding system aims to enhance efficiency, accuracy, and consistency in the coding process, leading to improved healthcare data management, analysis, and overall patient care. Automation decreases the time and effort needed for human coding by accelerating the coding process. This is particularly beneficial in healthcare settings dealing with large volumes of patient data. Automated systems can reduce the possibility of human error associated with manual coding by using sophisticated algorithms and machine-learning approaches to improve coding accuracy. By simplifying the coding process and reducing the need for extensive manual labor, automation can lead to cost savings for healthcare providers. Automated solutions can be easily updated to include the most recent coding standards without requiring significant manual adjustments as the ICD coding system through upgrades and revisions.

BERT is frequently utilized in NLP tasks, and it enables understanding of complicated language patterns. It has the ability for transfer learning, a comprehensive understanding of document context on a global scale, and adaptability in managing diverse vocabularies. Additionally, its multilingual support and open-source characteristics enhance its usability and accessibility across various languages and contexts. Hence, this study uses BERT because of these NLP benefits to automate ICD coding.

Another issue is the length of medical texts. Medical notes can be extensive and detailed. Creating summaries while automating the system helps reduce the dimensionality of the data, making it more manageable for analysis and model training for ICD classification. Summarized data can contain the most relevant information, leading to more efficient and focused feature extraction. In addition, long medical notes can be computationally expensive to process, especially when dealing with large datasets. Summaries enable faster preprocessing and training of machine learning models, making them more practical for real-world applications. Moreover, summarizing the notes helps filter out noise. Finally, summaries may enhance the interpretability of models. Due to its presentation of the whole text before codes, it offers a better way to read discharge summaries. This increases the trust of physicians and coders in the recommended codes.

Shorter, more focused input can make it easier for clinicians and researchers to understand and trust the decisions made by machine learning models. This is particularly important in healthcare, where interpretability and explainability are critical. Minh et al. (2022), state that data summarization is an important model for pre-modeling explainability. They point out that summaries can represent big data and can be used instead of the entire data set in classification analyses. For these reasons, this article investigates the effect of shortening medical notes on performance before the ICD codes-based classification of diseases. The novelty of the article is employing summary techniques that have not been used in BERT models for ICD code automation.

The research question of the article is "Can summarized medical notes and the BERT language model achieve comparable disease classification results with ICD codes, compared to state-of-the-art methods using long texts?".

The rest of this article is organized as follows: Section 2 examines the related works. Section 3 explains the methods and data. Results are illustrated in Section 4. Finally, Section 5 discusses the results and concludes the article.

2. LITERATURE REVIEW

Studies on automating the ICD coding system began decades ago (Li et al., 2019; Chen et al., 2021). More specifically, these applications are called natural language processing (NLP), which is a branch of AI that focuses on teaching computers to comprehend spoken and written language like that of humans. Although deep learning methods are preferred mostly in the current NLP studies, in the early stages several studies used rule-based and traditional machine learning-based methods (Marafino et al., 2014; Perotte et al., 2014; Scheurwegs et al., 2016; Kaur & Ginige, 2018; Teng et al., 2022).

Rule-based methods use rules and expert experiences (Nawalkar et al., 2022); and are manually adjusted for the datasets. Farkas & Szarvas (2008) is one of the studies using rule-based methods to automate ICD coding. They get high accuracy results with 88.93%. However, rule-based methods have some disadvantages since they are not flexible and portable. Therefore, they can cause conflicts for the higher number of codes (Yan et al., 2022). In addition, manual coding is costly and time-consuming (Li et al., 2019; Zeng et al., 2019).

Most traditional machine learning methods are implemented by training separate classifiers for each code. These methods are successful when the number of codes is low,

but today ICD codes are in the tens of thousands (Yan et al., 2022). Accordingly, deep learning has led to an increase in the popularity of research on automated ICD coding that uses neural networks (NNs) (Teng et al., 2022).

Shi et al. (2017) employ a hierarchical NN model with a two-level long short-term memory (LSTM) to automate the ICD system. They only select diagnosis descriptions to shorten the texts.

Baumel et al. (2018) is another study that utilizes recurrent neural network (RNN)-based models, but it also provides interpretability by providing relevant sentences of each code. They use a bidirectional gated recurrent unit (BiGRU) with sentence-level attention.

Mullenbach et al. (2018) introduce another state-of-the-art model, named convolutional attention for multi-label classification (CAML). They apply a per-label attention mechanism after automatically extracting characteristics from the discharge summaries using convolutional neural networks (CNNs).

Moreover, Vu et al. (2020) explain a label attention model to deal with the problem of CNNs using fixed window sizes and the model uses a joint learning mechanism to perform better for infrequent codes. The model is named LAAT and it has four layers (embedding, bidirectional LSTM (BiLSTM), label attention, and output layer).

Cao et al. (2020a) follow the logic behind Mullenbach et al. (2018) but employ a dilated CNN with holes in the filters.

Li & Yu (2020) utilize also the CNN model and add a residual convolutional layer to it to expand the receptive field.

In (Wang et al., 2018), labels and words are both embedded into the same vector space, and the labels are predicted using the cosine similarity between them.

In addition, there are several graph neural network (GNN)-based studies that represent the ICD in a graph structure and search for the relationship between words and the ICD codes (Rios & Kavuluru, 2018; Cao et al., 2020b; Du et al., 2020).

Current studies in this area generally prefer to include pre-trained language models (PLMs) in their approaches. Zhang et al. (2020) apply BERT to the medical texts to encode the data. Since BERT has a 512-token limitation, they extend the model to allow 1,024 tokens. However, as Pascual et al. (2021) point out, it is still not effective as medical texts are much longer than 1,024 tokens. Therefore, Pascual et al. (2021) try to split the text in various ways and apply the PubMedBERT model afterward. However, it shows that this strategy does not provide as high accuracy results as other deep learning methods.

Huang et al. (2022) propose a PLM-ICD model and employ a robustly optimized BERT pretraining approach based on PubMed data (RoBERTa-PubMed). They use segment pooling to surpass the maximum length limitation and label-aware attention to solve the large label set problem.

Ponthongmak et al. (2023) examine the CNN-PubMedBERT model and use PubmedBERT for tokenization and matrix representation of the texts. Then CNN is used to make ICD code predictions.

In NLP, the BERT approach has several benefits. BERT generates contextualized word embeddings during pre-training by collecting bidirectional context, which enables it to understand complicated language patterns and successfully resolve ambiguities. Because of its adaptability, it may be adjusted for different NLP tasks. The reason for BERT's extensive use is its capacity for transfer learning, global comprehension of document context, and

flexibility in handling different vocabularies. Its multilingualism and open-source nature further increase its usability and accessibility in a variety of languages and contexts.

In this article, the BERT model based on the attention mechanism is utilized to classify diseases based on ICD codes by using a summarized Medical Information Mart for Intensive Care (MIMIC-III) data set (Johnson et al., 2016a; Johnson et al., 2016b). Before classifying, the term frequency-inverse document frequency (TF-IDF) summary approach is used to overcome long text problems.

3. METHODS AND DATA

3.1. Bidirectional encoder representations from transformers (BERT)

In NLP, PLMs have shown remarkable success. The idea is to provide the ability to use them in other tasks, after pre-training in one task, simply by fine-tuning them. BERT and generative pre-trained transformers (GPT) are the most popular types of PLMs. The main difference between BERT and GPT is that, while GPT uses the decoder part of the transformers, BERT utilizes the encoder part.

BERT is a state-of-the-art transform-based PLM developed by Google in 2018 (Devlin et al., 2018) and it is preferred for use in NLP tasks such as text representation, text classification, question and answering systems, machine translation, etc. It addresses some LSTM problems: since they are generated sequentially it can take significant time to learn. Transformer models illustrated in Figure 2 (Vaswani et al., 2017) are faster since the words can be processed simultaneously. The context of words is better learned since they can learn from both directions simultaneously.

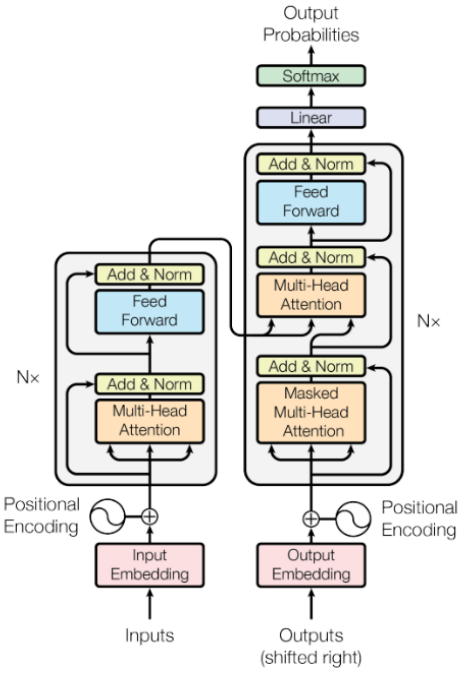

Fig. 2. The transformer model architecture (Vaswani et al., 2017)

BERT consists of two main phases: pre-training and fine-tuning. These phases are illustrated in Figure 3 (Devlin et al., 2018). During the pre-training phase, two tasks are performed: Masked Language Model (MLM) and Next Sentence Prediction (NSP). In MLM, a certain percentage of input tokens are masked and then predicted. To train a model that understands sentence relationships, NSP is also pre-trained. In the fine-tuning phase, the pre-trained BERT model can be fine-tuned by changing inputs and outputs. Compared to the pre-training phase, the fine-tuning process takes much less time.

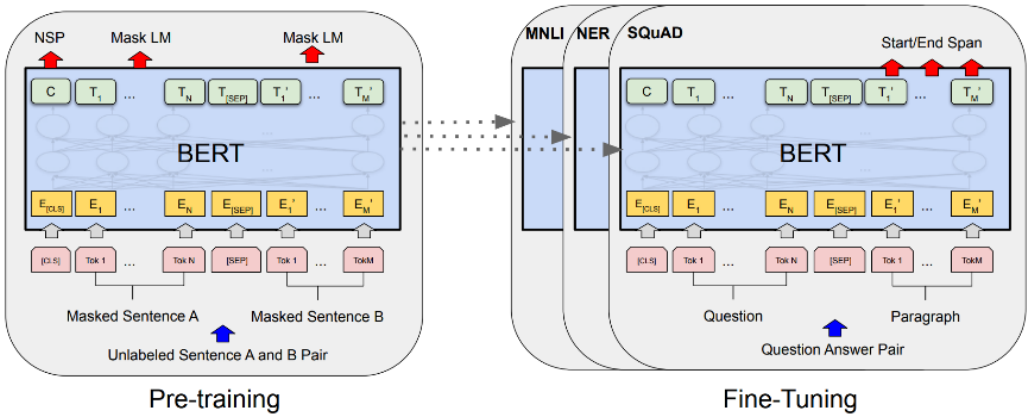

Fig. 3. BERT model architecture (Devlin et al., 2018)

3.2. Term frequency-inverse document frequency (TF-IDF)

To overcome long text problems, a summarizing method TF-IDF is performed before making the classification. TF-IDF works with the principle of determining the importance of a word in the related document. While doing this, it uses two statistics: term frequency (TF) and inverse document frequency (IDF). It compares the frequency of the word in a specific document to the inverse proportion of that word over the entire document corpus. Hence, it gives much importance to frequent words other than naturally frequent words such as articles and propositions with the help of its inverse mechanism. It is the product of TF and IDF as mathematically described in Equation 1.

$$w_d = f_{w,d} * \log\left(|D|/f_{w,D}\right) \tag{1}$$

where D is a document collection, w is a word, $d \in D$ is an individual document, $f_{w,d}$ and $f_{w,D}$ are the functions that are equal to the number of times w appears in d and D, respectively. High w_d implies that the word w is important in document d.

3.3. Data

In this study, a frequently used dataset, MIMIC-III (Moons et al., 2020; Singaravelan et al., 2021; Goldberger et al., 2000; Johnson et al., 2016a; Johnson et al., 2016b), is used to see the overall performance of the model among other studies in the literature. It is a comprehensive data of intensive care unit (ICU) patients and covers over forty thousand

patients who stayed in critical care units of the Beth Israel Deaconess Medical Center (BIDMC) between 2001 and 2012. It includes discharge summaries, radiology reports, laboratory measurements, microbiology cultures, medication prescriptions, vital signs, caregiver notes, and other data. To avoid any confusion that may occur in the rest of the study, it is important to clarify that the discharge summaries referenced in this study are not the same as the summaries generated by TF-IDF. A discharge summary is the medical note that physicians have recorded regarding the relevant case.

Although MIMIC-III is a comprehensive and frequently used database, it contains ICD-9 code correspondences of diseases. Finding a comprehensive database for the most current versions in this field is quite challenging, but it is assumed that it will not be difficult to adapt a method that works for the 9th version to data sets containing codes in the current version. Especially in pre-trained models such as BERT, this type of adaptation can be done in a short time with the fine-tuning process.

The challenging features of medical text data for MIMIC-III are:

- the large label space,
- the unbalanced label distribution,
- the long text of documents, and,
- the interpretability of coding.

To solve the first two challenging features, MIMIC-III-50 is used. For the third and fourth challenging features, summarization is applied to the medical text before employing the BERT model.

MIMIC-III-50 is a subset of MIMIC-III, which is utilized in various studies that involve electronic medical records (EMR) labeled by at least one of the top 50 most frequently used codes.

MIMIC-III-50 has 11,368 discharge summaries. To have a comparable study the training, validation, and test set are determined (Mullenbach et al., 2018). Namely, there are 8,066 discharge summaries in the training part, 1,573 in validation, and 1,729 in the test part.

The data is provided from "*Physionet.org*" (Goldberger et al., 2000; Johnson et al., 2016a; Johnson et al., 2016b) by fulfilling the requirements of being a credentialed user, completing Massachusetts Institute of Technology (MIT) courses, taking the "CITI Data or Specimens Only Research" certificate, and signing the data usage agreement for the project.

After selecting the discharge summaries from the medical notes, non-alphabetical terms are deleted from the texts, and all the letters are reduced to lowercase.

4. RESULTS

Analyses are performed in Python using important libraries such as the natural language toolkit (NLTK), transformers (Hugging Face), PyTorch, TensorFlow, Keras, and Scikit-learn.

Discharge summaries are extremely long documents. In the MIMIC-III dataset, the length of the discharge summaries after tokenization ranges from 78 to 18,429 tokens with a mean of 2,740 and a median of 2,500. However, the BERT model has a token limit of 512. Therefore, although BERT is state-of-the-art in many NLP tasks, the transformer architecture is not in assigning ICD codes. Pascual et al. (2021) examine this fact in their study.

TF-IDF summarizing approach is used in the model to deal with long text issues. The full texts are shortened to make them less than 512 tokens by using TF-IDF. Then, the summaries are processed using the BERT model. Among the BERT models, BioClinicalBERT (Alsentzer et al., 2019) is the preferred one as it is initialized with BioBERT and trained on all MIMIC-III notes. The model's hyperparameters are selected by grid search and they are given as follows:

- Maximum length = 512
- Training batch size = 1
- Epoch number = 11
- Learning rate = 1e-05

The model uses binary cross-entropy loss with logits. The model output provides probabilities for each label. A threshold of 0.20 is chosen to select the labels accurately.

Figure 4 shows the training and validation losses. The model seems to overfit leading to a lower performance on validation data.

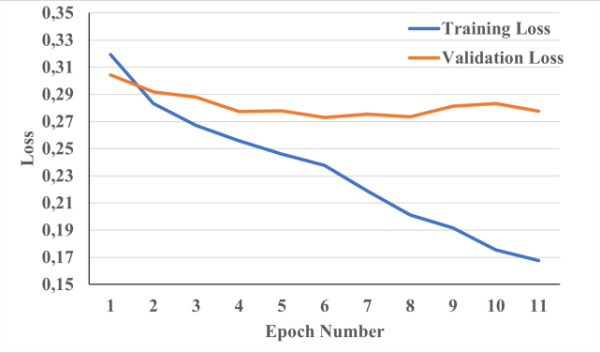

Fig. 4. Training and validation losses of the BioClinicalBERT-TF-IDF approach

The performance of the model is evaluated using Micro-AUC and Macro-AUC metrics. The Micro-AUC score is 74.78%, while the Macro-AUC score is 71.67%. Although these scores are satisfactory, they are not better than the state-of-the-art models in the literature, as shown in Table 2 (Pascual et al., 2021). In Table 2, BioClinicalBERT-TF-IDF is the proposed methodology in this article, and the rest of the results are taken from the same MIMIC-III-50 data preprocessed with the same methods used in the (Pascual et al., 2021).

Tab. 2. Comparison of different state-of-the-art models for ICD code classification

Model	**Macro-AUC (%)**	**Micro-AUC (%)**
Label Attention	92.1	94.6
MSATT-KG	91.4	93.6
DR-CAML	88.0	90.2
CAML	87.5	90.9
BERT-ICD	84.45	88.65
BioClinicalBERT-TF-IDF	71.67	74.78

TF-IDF is mainly used in the summarization part of this study due to its fast working structure. However, since the AUC results are low compared to other studies, the transformer-based Longformer model is also utilized as an alternative method to investigate more clearly whether these low performances are related to the summarization method. Longformer is an attention-based model designed for long texts.

In addition, experiments are done for PubMedBERT (Gu et al., 2021), BERT-base-uncased (Devlin et al., 2018), and BERT-tiny (Bhargava et al., 2021; Turc et al., 2019), without adhering to a single BERT model. PubMedBERT is pre-trained from scratch using abstracts from PubMed and full-text articles from PubMedCentral. It is currently also named BiomedBERT. On the other hand, BERT-base-uncased is an uncased transformers model that was self-supervised and pre-trained on a sizable corpus of English data using MLM. It was trained using the English Wikipedia dataset (excluding lists, tables, and headings) plus the dataset from BookCorpus, which contains 11,038 unpublished books. Lastly, BERT-tiny is one of the smaller pre-trained BERT variants.

Table 3 compares the performance results of different BERT models concerning TF-IDF and Longformer summarization methods. However, none of these experiments give better results than the main experiment. Despite several combination efforts, it is seen that the accuracy results do not improve. In other words, all the approaches produce similar Micro-AUC and Macro-AUC results between 65-75% compared to the BioClinicalBERT-TF-IDF application. The potential reason is that BERT models get confused when learning ICD codes from the summaries. This confusion may be caused by the fact that there is other information in the summaries that is not related to the diagnosis codes. Indeed, this is a possible risk of the study since the summarization and code assignment parts are performed separately. This reveals the need to try the use of feature extraction methods to obtain more precise summaries.

Tab. 3. Comparison of different BERT models using TF-IDF and Longformer summarization methods

Models	**TF-IDF**		**Longformer**	
	Macro-AUC (%)	Micro-AUC (%)	Macro-AUC (%)	Micro-AUC (%)
BioClinicalBERT	71.67	74.78	69.32	72.70
PubMedBERT	68.43	72.10	67.12	71.94
BERT-base-uncased	70.52	73.63	66.72	71.20
BERT-tiny	65.59	69.36	63.81	69.02

5. DISCUSSION AND CONCLUSION

Classification systems play a crucial role in creating a useful structure and ensuring international comparability in statistics. Two types of classification systems exist: monetary and non-monetary health statistics. The importance of the ICD is particularly evident in non-monetary health statistics as it provides a source for many health data. The ICD system is highly detailed and demands significant workforce and time investment.

Recent studies show that applying state-of-the-art AI methods to MIMIC-III data results in an accuracy rate of over 80%. In this study, BERT (more specifically BioClinicalBERT), which is a state-of-the-art model in many other NLP tasks, is chosen for the classification part of the study. Moreover, it is decided to employ a summarization method to create an

interpretable model and overcome the token limit in BERT models. In addition, another reason for the summarization is that it will provide more brief information about the cases to both physicians and patients.

The large number of patients in hospitals requires physicians to get results faster. In this case, real-time calculations need to be made and as a result, classification calculation time is also very important. Therefore, it is crucial to summarize lengthy medical notes by using the most informative and relevant parts. For this purpose, it is consequential to use improved summary medical notes.

It is important to note that the proposed BioClinicalBERT-TF-IDF method provides a better solution for interpreting discharge summaries as it presents a summary of the entire text before providing codes. This boosts the confidence of physicians and coders about the suggested codes. Although the accuracy results of this method do not surpass some state-of-the-art studies in the literature, it offers a way to improve interpretability. Wang et al. (2020) state that it is very important to classify disease diagnoses according to ICD codes in medical notes, which are shorter and easier to process.

Although the suggested approach improves interpretability, the AUC values are not as good as those of the state-of-the-art. These performance outcomes could be the consequence of running the classification and summarizing parts independently. The summaries may contain information that is not required to derive the diagnostic codes. Nevertheless, combining these two components will result in lengthy processing times.

Therefore, future work will focus on improving the summaries to increase the performance results. Research will be conducted on not keeping the information in the summary that causes performance loss when summarizing long medical notes, in a way that does not cause long processing time. Moreover, various feature extraction methods (e.g., CountVectorizer, word embeddings (Huang et al., 2019), bag of words (Tabassum & Patil, 2020), bag of n-grams (Huang et al., 2019), HashingVectorizer, latent Dirichlet allocation (LDA), non-negative matrix factorization (NMF), principle component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), part-of-speech tagging (Kaur et al., 2021), etc.) for medical notes will be tested for different machine learning methods.

Author Contributions

***Dilek AYDOGAN-KILIC**: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing - review & editing, Visualization.* ***Deniz Kenan KILIC**: Software, Validation, Formal analysis, Writing – original draft, Writing - review & editing.* ***Izabela Ewa NIELSEN**: Conceptualization, Writing - review & editing, Supervision, Project administration, Funding acquisition.*

Funding

This research has been partly funded by the Jean Monnet Scholarship Programme 2022-2023 Academic Year.

Conflicts of Interest

The authors declare no conflict of interest.

Data Availability

The MIMIC-III dataset is freely available. Researchers seeking to use the database must formally request access.

REFERENCES

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W. H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical BERT embeddings. *ArXiv*, *abs/1904.03323*. https://doi.org/10.48550/arXiv.1904.03323

Baumel, T., Nassour-Kassis, J., Cohen, R., Elhadad, M., & Elhadad, N. (2018). Multi-label classification of patient notes: case study on ICD code assignment. In Workshops at the thirty-second AAAI conference on artificial intelligence. *ArXiv*, *abs/1709.09587*. https://doi.org/10.48550/arXiv.1709.09587

Bhargava, P., Drozd, A., & Rogers, A. (2021). Generalization in NLI: Ways (not) to go beyond simple heuristics. arXiv preprint. *ArXiv*, *abs/2110.01518*. https://doi.org/10.48550/arXiv.2110.01518

Cao, P., Chen, Y., Liu, K., Zhao, J., Liu, S., & Chong, W. (2020a). HyperCore: Hyperbolic and co-graph representation for automatic ICD coding. *58th Annual Meeting of the Association for Computational Linguistics* (pp. 3105-3114). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.282

Cao, P., Yan, C., Fu, X., Chen, Y., Liu, K., Zhao, J., Liu, S., & Chong, W. (2020b). Clinical-coder: Assigning interpretable ICD-10 codes to Chinese clinical notes. *58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 294-301). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-demos.33

Chen, P. F., Wang, S. M., Liao, W. C., Kuo, L. C., Chen, K. C., Lin, Y. C., Yang, C., Chiu, C., Chang, S., & Lai, F. (2021). Automatic ICD-10 coding and training system: deep neural network based on supervised learning. *JMIR Medical Informatics*, *9*(8), e23230. https://doi.org/10.2196/23230

Chute, C. G., & Çelik, C. (2021). Overview of ICD-11 architecture and structure. *BMC Medical Informatics and Decision Making*, *21*(6), 378. https://doi.org/10.1186/s12911-021-01539-1

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, *abs/1810.04805*. https://doi.org/10.48550/arXiv.1810.04805

Du, Y., Xu, T., Ma, J., Cen, E., Zheng, Y., Liu, T., & Tong, G. (2020). An automatic ICD coding method for clinical records based on deep neural network. *Big Data Research*, *6*(5), 3-15. https://doi.org/10.11959/j.issn.2096-0271.2020040

European Commission, Eurostat, (2018). *European statistics code of practice: for the national statistical authorities and Eurostat (EU statistical authority), Publications Office of the European Union*. https://data.europa.eu/doi/10.2785/798269

European Commission. (2020). *Strategic plan 2020-2024*. https://commission.europa.eu/system/files/2020-10/eac_sp_2020_2024_en.pdf

Eurostat. (2012). *Healthcare statistics*. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Healthcare_statistics&oldid=86497

Farkas, R., & Szarvas, G. (2008). Automatic construction of rule-based ICD-9-CM coding systems. *BMC Bioinformatics*, *9*(3), S10. https://doi.org/10.1186/1471-2105-9-S3-S10

Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., Mietus, J. E., Moody, G. B., Peng, C. K., & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, *101*(23), e215–e220. https://doi.org/10.1161/01.cir.101.23.e215

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J. & Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, *3*(1), 1-23. https://doi.org/10.1145/3458754

Harrison, J. E., Weber, S., Jakob, R., & Chute, C. G. (2021). ICD-11: an international classification of diseases for the twenty-first century. *BMC Medical Informatics and Decision Making*, *21*(6), 206. https://doi.org/10.1186/s12911-021-01534-6

Hsu, J. L., Hsu, T. J., Hsieh, C. H., & Singaravelan, A. (2020). Applying convolutional neural networks to predict the ICD-9 codes of medical records. *Sensors*, *20*(24), 7116. https://doi.org/10.3390/s20247116

Huang, C. W., Tsai, S. C., & Chen, Y. N. (2022). PLM-ICD: automatic ICD coding with pretrained language models. *ArXiv*, *abs/2207.05289*. https://doi.org/10.48550/arXiv.2207.05289

Huang, J., Osorio, C., & Sy, L. W. (2019). An empirical evaluation of deep learning for ICD-9 code assignment using MIMIC-III clinical notes. *Computer Methods and Programs in Biomedicine*, *177*, 141–153. https://doi.org/10.1016/j.cmpb.2019.05.024

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L. W. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016b). MIMIC-III, a freely accessible critical care database. *Scientific data*, *3*, 160035. https://doi.org/10.1038/sdata.2016.35

Johnson, A., Pollard, T., & Mark, R. (2016a). *MIMIC-III Clinical Database (version 1.4)*. PhysioNet. https://doi.org/10.13026/C2XW26

Kaur, R., & Ginige, J. A. (2018). Comparative analysis of algorithmic approaches for auto-coding with ICD-10-AM and ACHI. *IOS Press*, *252*, 73-79. https://doi.org/10.3233/978-1-61499-890-7-73

Kaur, R., Ginige, J. A., & Obst, O. (2021). A systematic literature review of automated ICD coding and classification systems using discharge summaries. *ArXiv*, *abs/2107.10652*. https://doi.org/10.48550/arXiv.2107.10652

Li, F., & Yu, H. (2020). ICD coding from clinical text using multi-filter residual convolutional neural network. *AAAI conference on artificial intelligence* (pp. 8180-8187). https://doi.org/10.1609/aaai.v34i05.6331

Li, M., Fei, Z., Zeng, M., Wu, F. X., Li, Y., Pan, Y., & Wang, J. (2019). Automated ICD-9 coding via a deep learning approach. *IEEE/ACM transactions on computational biology and bioinformatics*, *16*(4), 1193-1202. https://doi.org/10.1109/TCBB.2018.2817488

Marafino, B. J., Davies, J. M., Bardach, N. S., Dean, M. L., & Dudley, R. A. (2014). N-gram support vector machines for scalable procedure and diagnosis classification, with applications to clinical free text data from the intensive care unit. *Journal of the American Medical Informatics Association*, *21*(5), 871-875. https://doi.org/10.1136/amiajnl-2014-002694

Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2022). Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*, *55*, 3503-3568. https://doi.org/10.1007/s10462-021-10088-y

Moons, E., Khanna, A., Akkasi, A., & Moens, M. F. (2020). A comparison of deep learning methods for ICD coding of clinical records. *Applied Sciences*, *10*(15), 5262. https://doi.org/10.3390/app10155262

Mullenbach, J., Wiegreffe, S., Duke, J., Sun, J., & Eisenstein, J. (2018). Explainable prediction of medical codes from clinical text. *2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1101-1111). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1100

Nawalkar, N., Attar, V. Z., & Kalamkar, S. P. (2022). Automated icd-9 medical code assignment from given free text using deep learning approach. In S. Tiwari, M. C. Trivedi, M. L. Kolhe, K. K. Mishra, & B. K. Singh (Eds.), *Advances in Data and Information Sciences* (Vol. 318, pp. 317–327). Springer Singapore. https://doi.org/10.1007/978-981-16-5689-7_28

Pascual, D., Luck, S., & Wattenhofer, R. (2021). Towards BERT-based automatic ICD coding: Limitations and opportunities. *ArXiv*, *abs/2104.06709*. https://doi.org/10.48550/arXiv.2104.06709

Perotte, A., Pivovarov, R., Natarajan, K., Weiskopf, N., Wood, F., & Elhadad, N. (2014). Diagnosis code assignment: models and evaluation metrics. *Journal of the American Medical Informatics Association*, *21*(2), 231-237. https://doi.org/10.1136/amiajnl-2013-002159

Pezzella, P. (2022). The ICD-11 is now officially in effect. *World Psychiatry*, *21*(2), 331-332. https://doi.org/10.1002/wps.20982

Ponthongmak, W., Thammasudjarit, R., McKay, G. J., Attia, J., Theera-Ampornpunt, N., & Thakkinstian, A. (2023). Development and external validation of automated ICD-10 coding from discharge summaries using deep learning approaches. *Informatics in Medicine Unlocked*, *38*, 101227. https://doi.org/10.1016/j.imu.2023.101227

Rios, A., & Kavuluru, R. (2018). Few-shot and zero-shot multi-label learning for structured label spaces. *Conference on Empirical Methods in Natural Language Processing* (pp. 3132-3142). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1352

Scheurwegs, E., Luyckx, K., Luyten, L., Daelemans, W., & Van den Bulcke, T. (2016). Data integration of structured and unstructured sources for assigning clinical codes to patient stays. *Journal of the American Medical Informatics Association*, *23*(e1), e11-e19. https://doi.org/10.1093/jamia/ocv115

Shi, H., Xie, P., Hu, Z., Zhang, M., & Xing, E. P. (2017). Towards automated ICD coding using deep learning. *ArXiv*, *abs/1711.04075*. https://doi.org/10.48550/arXiv.1711.04075

Singaravelan, A., Hsieh, C. H., Liao, Y. K., & Hsu, J. L. (2021). Predicting icd-9 codes using self-report of patients. *Applied Sciences*, *11*(21), 10046. https://doi.org/10.3390/app112110046

Tabassum, A., & Patil, R. R. (2020). A survey on text pre-processing & feature extraction techniques in natural language processing. *International Research Journal of Engineering and Technology (IRJET)*, *7*(06), 4864-4867.

Teng, F., Liu, Y., Li, T., Zhang, Y., Li, S., & Zhao, Y. (2022). A review on deep neural networks for ICD coding. *IEEE Transactions on Knowledge and Data Engineering*, *35*(5), 4357-4375. https://doi.org/10.1109/TKDE.2022.3148267

Turc, I., Chang, M. W., Lee, K., & Toutanova, K. (2019). Well-read students learn better: On the importance of pre-training compact models. *ArXiv*, *abs/1908.08962*. https://doi.org/10.48550/arXiv.1908.08962

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *ArXiv*, *abs/1706.03762*. https://doi.org/10.48550/arXiv.1706.03762

Vu, T., Nguyen, D. Q., & Nguyen, A. (2020). A label attention model for ICD coding from clinical text. *ArXiv*, *abs/2007.06351*. https://doi.org/10.48550/arXiv.2007.06351

Wang, D., Su, J., & Yu, H. (2020). Feature extraction and analysis of natural language processing for Deep Learning english language. *IEEE Access*, *8*, 46335-46345. https://doi.org/10.1109/ACCESS.2020.2974101

Wang, G., Li, C., Wang, W., Zhang, Y., Shen, D., Zhang, X., Henao, R., & Carin, L. (2018). Joint embedding of words and labels for text classification. *ArXiv*, *abs/1805.04174*. https://doi.org/10.48550/arXiv.1805.04174

World Health Organization, (2023). *International Classification of Diseases for Mortality and Morbidity Statistics Eleventh Revision (ICD-11)*. https://icdcdn.who.int/icd11referenceguide/en/html/index.html

Yan, C., Fu, X., Liu, X., Zhang, Y., Gao, Y., Wu, J., & Li, Q. (2022). A survey of automated International Classification of Diseases coding: development, challenges, and applications. *Intelligent Medicine*, *2*(3), 161-173. https://doi.org/10.1016/j.imed.2022.03.003

Zeng, M., Li, M., Fei, Z., Yu, Y., Pan, Y., & Wang, J. (2019). Automatic ICD-9 coding via deep transfer learning. *Neurocomputing*, *324*, 43-50. https://doi.org/10.1016/j.neucom.2018.04.081

Zhang, Z., Liu, J., & Razavian, N. (2020). BERT-XML: Large scale automated ICD coding using BERT pretraining. *ArXiv*, *abs/2006.03685*. https://doi.org/10.48550/arXiv.2006.03685