

International Journal of Simulation and Process Modelling

ISSN online: 1740-2131 - ISSN print: 1740-2123

<https://www.inderscience.com/ijspm>

**Optimising healthcare process efficiency with HyperBERT:
embedding ICD code hierarchy for national healthcare systems
simulation and improvement**

Dereje Weyessa, Berhanu Beyene, Teklu Urgessa, Tiruveeduula Gopikrishna, Satyasis Mishra, Bijay Kumar Paikaray

DOI: [10.1504/IJSPM.2025.10072798](https://doi.org/10.1504/IJSPM.2025.10072798)

Article History:

Received:	10 December 2024
Last revised:	19 March 2025
Accepted:	27 March 2025
Published online:	01 September 2025

Optimising healthcare process efficiency with HyperBERT: embedding ICD code hierarchy for national healthcare systems simulation and improvement

Dereje Weyessa

Department of Computer Science and Engineering,
Adama Science and Technology University,
Adama, Ethiopia
Email: derejeweyessaa@gmail.com

Berhanu Beyene

President of Ethiopian Cybersecurity Association,
Civil Service University of Ethiopia,
Ababa, Ethiopia
Email: bbeyene@e-cysa.org

Teklu Urgessa

Department of Computer Science and Engineering,
Adama Science and Technology University,
Adama, Ethiopia
Email: tekluurgessa2009@gmail.com

Tiruveduula Gopikrishna*

Department of Computer Science and Engineering,
Adama Science and Technology University,
Adama, Ethiopia
Email: gktiruveduula@gmail.com
*Corresponding author

Satyasis Mishra

Department of ECE,
Centurion University of Technology and Management,
Odisha, India
Email: s.mishra@cutm.ac.in

Bijay Kumar Paikaray

Centre for Data Science,
Department of Computer Science and Engineering,
Siksha 'O' Anusandhan (Deemed to be University),
Bhubaneswar, Odisha, India
Email: bijaypaikaray87@gmail.com

Abstract: This research study focuses on embedding the hierarchical structure of the International Classification of Diseases (ICD) codes for automatic code assignment in clinical texts. It explores: 1) incorporating diagnosis code hierarchy as auxiliary knowledge in a BERT training process; 2) how transformer-based models like BERT can serve as interpretable AI in Ethiopia's National Classification of Diseases (NCOD) and the MIMIC dataset. Utilising qualitative methods, including experimental and case study strategies, results indicate that a new model architecture called HyperBERT improved the macro average AUC by 0.003 and F1-score by 0.02. The model's interpretability was demonstrated through a code-wise label attention

mechanism using the bidirectional encoder representation for transformer (BERT) visualisation tool. Riemannian optimisation and Poincaré loss were employed to validate a new document representation model combining BERT with hyperbolic embeddings, enabling effective ICD code predictions from discharge summaries in Ethiopia's healthcare system.

Keywords: ICD coding; interpretable AI; code hierarchical; deep neural network; bidirectional encoder representation for transformer; BERT; hyperbolic embedding; healthcare systems.

Reference to this paper should be made as follows: Weyessa, D., Beyene, B., Urgessa, T., Gopikirishna, T., Mishra, S. and Paikaray, B.K. (2025) 'Optimising healthcare process efficiency with HyperBERT: embedding ICD code hierarchy for national healthcare systems simulation and improvement', *Int. J. Simulation and Process Modelling*, Vol. 22, Nos. 1/2, pp.1–17.

Biographical notes: Dereje Weyessa currently a PhD candidate with Program of Computer Science and Engineering at the School of Electrical Engineering and Computing from Computer Science Department at Adama Science and Technology University, Adama, Oromia region, Ethiopia. He has long work experience as a Lecturer at Hawassa University Institute of Technology at Computer Science Department, Hawassa, Ethiopia. He has an MSc degree in Software Engineering from HeLCoe Computer Science and Technology College and graduated in September 2015 and he also has a BSc in Information Science and Library from Addis Ababa University graduated in July 2003. He completed his high school from Butajera High School, SNNPR Region, Butajera Town in 1998. He designs publication model database for Addis Ababa City administration cadastral system in 2015.

Berhanu Beyene is a President of Ethiopian Cybersecurity Association (ECySA) Addis Ababa, Ethiopia. He is a researcher and educator with over 30 years of experience across Europe, Africa, and the Middle East. He holds a Doctorate in Computer Science from Hamburg University, along with degrees in Computer Science and Economics from Addis Ababa University, Ethiopia. He has supervised over 50 research theses. He is the former founding Director of Africa Institute of Governance and Development. His contributions include developing the blended e-learning policy for TVET in Ethiopia and serving as an IT Consultant for organisations like GIZ, UNESCO, and the East African Law Society. He is committed to advancing IT, education and development through sustainable networks and partnerships.

Teklu Urgessa is an Assistant Professor of Computer Science and Engineering and a Dean of the School of Electrical Engineering and Computing at Adama Science and Technology University (ASTU). He studied his PhD in Engineering in Seoul National University, South Korea from March 2013 to February 2018. He studied his MSc degree in Health Informatics at Addis Ababa University from November 2008 to July 2010. He attended his BSc degree in Information science from Addis Ababa University during September 1999 to July 2003. He has 20 years of teaching, research, community services and leadership at different academic positions at high education institutions. Among these, he had served as Department Head of the Department of Information Science from 2007 to 2008, Vice Dean of the School of Engineering and Information Technologies from 2011 to 2013. He had also served as a Dean of Continuing Education Institute of ASTU from 2018 to 2022. Currently, he is serving as a Dean of the School of Electrical Engineering and Computing (SoEEC) at ASTU.

Tiruveduula Gopikirishna is a seasoned professional with over 20+ years of experience in teaching and research in the field of computer science and engineering. He earned his PhD in CSE from Rayalaseema University in 2014, focusing on advanced data analysis techniques for extracting insights from large datasets. He also holds a Master of Engineering degree in Computer Science and Engineering (ME) (2004) from Anna University and a Master of Science in Computer Science (MSc) (2001) from Andhra University and a Bachelor's degree in Computer Science (BSc) (1997) from Andhra University. His profound understanding of machine learning and artificial intelligence, coupled with his teaching and research experience, positions him as a respected leader and authority in the field of computer science and engineering. Currently, he is working as an Associate Professor at Adama Science and Technology University, Ethiopia since 2017. His current research area is data science and big data analytics with AI/ML.

Satyasis Mishra completed PhD in (Electronics and Communication Engineering) from Biju Patnaik University of Technology, Odisha, India and an MTech (Electronics Systems and Communication) from NIT Rourkela, Odisha, India. He has published 85 national and international journal and conference papers. Currently, he is working as a Professor at the SoET, Electronics and Communication Department, Centurion University of Technology and Management, Odisha, India. He is currently working with artificial intelligence, machine learning, and optimisation technique applications to biomedical imaging and signal processing.

Bijay Kumar Paikaray currently serves as an Associate Professor at the Centre for Data Science, Department of Computer Science and Engineering, Siksha ‘O’ Anusandhan (Deemed to be University), Odisha, India. He earned his PhD and MTech in Computer Science and Engineering from Centurion University, Odisha, India. With 16 years of robust academic experience, he is deeply committed to both teaching and research. In addition to teaching, he has authored more than 60 research publications, holds patents, and has authored books. His research interests include information security, machine learning, data visualisation and IoT.

1 Introduction

The diagnostic categorisation standard used in the medical field for clinical and research reasons is the International Categorization of Diseases (ICD), which is endorsed by the World Health Organization. The ICD is a comprehensive, hierarchical list of all illnesses, injuries, and other associated medical conditions that are now recognised. ICD coding makes it easier to save, retrieve, and analyse health information for evidence-based decision making, share health information across hospitals, regions, settings, and countries, and compare data across time periods in the same area. ICD has undergone rewrites on occasion to reflect advancements in the medical domain. The ICD taxonomy has undergone eleven modifications to date, with ICD-9 and ICD-10 being the most researched in terms of automated assignment to medical documents (Moons et al., 2020). The significance of automating ICD coding in the context of healthcare process optimisation and simulation is to automatically assign the appropriate ICD codes to clinical notes. Automatic ICD coding can be viewed as a multi-label classification problem (Zhang et al., 2020). The complete page is used as the input to generate a text representation. Every ICD code is a label, and all labels are predicted at the same time.

Ethiopia’s Federal Ministry of Health (FMOH) acknowledged that ICT might help and revolutionise the country’s healthcare system (Ethiopia MOH, 2016). As part of the HMIS overhaul, the National Classification of Diseases (NCoD) was created to organise health information. An extensive nationwide analysis of samples from service centres and administrative units was the first step in the NCoD’s preparation. It is the system that is used nationally to define, track, and report health conditions and their results (Abate, 2017). It is far observed that the project of ICD codes in scientific texts is presently achieved entirely by hand. This has caused physicians in most Ethiopian hospitals to hesitate to use ICD encoding because it calls for lengthy clinical notes and is labour intensive, susceptible to error, blunders-susceptible, and calls for a large variety of highly trained workforce participants.

This study aims to design a novel model architecture that utilises ICD code hierarchy with hyperbolic embedding integrated into the BERT approach. Those researchers examined the technique for automatically assigning comprehensible ICD codes to clinical narratives. They centred on using ICD code hierarchical structure fashions primarily based on CNN, RNN, GNN, and BERT variations that were posted between 1 January 2010 and 1 July 2023, a factor comparison. The focus of the review is to summarise

the major challenges and opportunities of the task of automatic ICD code assignment, namely:

- 1 how to introduce diagnosis code hierarchy as auxiliary information knowledge in the training process
- 2 comparison of the techniques of explainability of the models.

The comparison is done based on architectural design goals, the dataset, the performance metrics, and the result obtained.

The transformers method is becoming more and more well-liked, and it performs better on many NLP tasks. BERT model (Devlin et al., 2019), draws inspiration from transformers and has been firmly established as the state-of-the-art approach in several NLP researches. The input length constraint imposed by BERT-based models restricts their suitability for processing multiple, longitudinal documents. Especially for hierarchical networks, hyperbolic embedding can maintain complicated linkages and graph distances in a relatively small number of dimensions (Tsai et al., 2021). Due to their great ability to model data and latent hierarchies, deep neural networks have recently been used for representation learning in hyperbolic spaces, which have a hierarchical structure (Nickel and Kiela, 2018; Ji, 2022). In literature review (Section 2.1) of this study, details about hyperbolic embeddings and hyperbolic ball properties are discussed.

1.1 Research questions

The following research issues are examined in this study as a result:

- How can we build a model that employs the BERT approach and hyperbolic embedding to automatically assign ICD codes to clinical narratives based on the hierarchical structure of ICD codes among diagnosis code labels?
- How can transformer-based models like BERT be applied as practice-oriented interpretable AI in the task of automatic clinical narratives assignment in the Ethiopian healthcare system?

The contributions of the research study are:

- Developed a novel model, HyperBERT, leveraging hyperbolic space to optimise ICD code hierarchy, overcoming limitations of existing BERT transformer architectures.

- Integrated BERT with hyperbolic embeddings to improve runtime performance while preserving ICD code hierarchical relationships, achieving a macro AUC improvement of 0.03 and an F1-score increase of 0.02 over state-of-the-art models.
- Introduced a technique to utilise ICD code hierarchy as external explicit knowledge, enhancing BERT's representation learning and generalisability.
- Theoretically validated and implemented Riemannian adaptive optimisation methods for the loss function, enabling better semantic similarity capture between document embeddings and ICD codes.
- Demonstrated the effectiveness of code-wise label attention using BERT visualisation on NCoD discharge summaries, providing an interpretable and accurate approach for ICD coding in healthcare decision making.
- Recommended the application of HyperBERT in Ethiopia's EMR systems, incorporating the Ethiopia NCoD.

The following parts of this research study are organised as follows: Section 2 provides literature review of related works, Section 3 presents research methods of the study, Section 4 presents results and discussion; finally, Section 5 provides conclusion, limitations, and future works.

2 Related work

Today's top NLP models rely on transformer architectures (Rothman, 2021), built on the groundbreaking self-attention mechanism in BERT (Devlin et al., 2019). This design, unlike LSTMs, excels at parallel processing and capturing distant text connections.

After early exploration of rule-based and conventional ML approaches in the '90s (Arkey, 1996; de Lima et al., 1998), deep learning models like CNNs, RNNs, and LSTMs held sway in automated ICD coding. But, the spotlight has now shifted decisively to transformer-based architectures, specifically pre-trained language fashions like BERT. This study highlights the main challenges identified by state-of-the-art models for task automatic ICD coding, which include clinical notes being longer than PLMs, class imbalance, and tokens in code use occasionally. Beside these in the task of automatic ICD coding our review paper also identified the lack of ICD code hierarchy usage as explicit external knowledge in deep neural network models is the potential cause of the underperformance in the next generation models, because of the way the codes' hierarchical structure affects how incorrect predictions can be understood. Understanding the model's forecast and giving external knowledge about the code hierarchy the model has learned priority are therefore crucial. These are because among the reviewed models of BERT approach for the task of automatic ICD coding task assignment of clinical text that not used code hierarchical structure as external

explicit knowledge usage in the representation learning processes include: BERT-ICD (Pascual et al., 2021), MIMIC-BERT (Bhavani Singh et al., 2020), KB-BERT (Remmer, 2021), G-BERT (Shang et al., 2019), HTN (Si and Roberts, 2021), TransICD (Biswas et al., 2021), DLAC (Feucht et al., 2021), HLAN (Dong et al., 2021), HiLAT (Liu and Xu, 2021), HTN (Si and Roberts, 2021). Models that used ICD code hierarchy usage as external explicit knowledge in the representation learning process such as Wikipedia, code description, and code hierarchy includes: C-MemNN, KSI, Hype core, DR_CAMEL, Hyper Core, G-BERT, Bi-GRU, JointLAAT, CAML Hierarchical, and MSATT-KG. Moreover, models that have been designed to exploit to capture the code hierarchy include HyperCore, G-BERT, Bi-GRU, JointLAAT, CAML Hierarchical, and MSATT-KG. But only two models out of the 28 models reviewed have applied hyperbolic geometry with the deep neural network, namely: the HyperCore model (Cao et al., 2020) which the encoder used CNN plus hyperbolic geometry, deep connections applied was stacking and decoders used LAN (label wise attention) plus GCN. The other model called MSATT-KG by Xie et al. (2019) encoder CNN plus attention, deep connections applied stacking, and the decoders used attention plus knowledge graph (KG). Unfortunately, none of the models in this research have taken into account the BERT approach when addressing the code hierarchical structure. Instead, this study has looked into how to introduce the diagnosis code hierarchy as explicit external knowledge during the training process, as well as which interpretability techniques are suitable for the models. When it comes to the mapping of clinical notes to ICD code assignments in the healthcare system, feature engineering should be investigated since the study of interpretability reveals how algorithms make conclusions and give evidence to support them. With the recent success of BERT (Devlin et al., 2019), transformer topologies are typically the foundation of state-of-the-art neural language models (Rothman, 2021). At comparison to recurrent neural networks like LSTMs, A multi-layer, multi-head self-attention method is introduced by the transformer model, which has demonstrated superiority in describing long-range dependencies in texts and making superior use of GPU-based parallel processing (Devlin et al., 2019). Automated ICD coding research dates back to the early '90s (de Lima et al., 1998; Arkey, 1996). Researchers have examined a wide range of rule-based, traditional machine learning and deep learning methodologies. Nevertheless, attention-based deep learning techniques like CNN, RNNs, and transformer-based techniques have lately surpassed them in performance. For a variety of NLP tasks, transformer-based architectures – specifically, the pre-trained transformer language models – have gained popularity (Devlin et al., 2019). Since transformer-based architectures (Lewis et al., 2020) can handle long-range dependencies, in practically every field of natural language processing, they have advanced to the state of the art (Qiu et al., 2020). On the other hand, deep learning models have dominated the automatic ICD coding new generation

model. A common tool for describing patient diagnosis data is the ICD. Automatic ICD coding is crucial since manually assigning codes is expensive, time-consuming and error-prone. Health-related texts, such as electronic health record notes, provide a range of understudied medical concept domains that can be analysed using natural language processing (NLP) technology (Newman-Griffis and Fosler-Lussier, 2021). Pascual et al. (2021) stated that to deploy automatic ICD coding systems, it is critical that their choices in the real world be comprehensible and explicable. Research on AI transformer models is presently ongoing.

2.1 Models of ICD coding in deep neural network

This paper focuses on the most recent models and methods for the automatic assignment of ICD codes to clinical narratives on the three categories. These methods are based on deep neural network types, notably variations of BERT, CNN and RNN. Firstly, the categories of based CNN approach for automatic ICD code assignment for clinical narratives reviewed include: HLAN (Dong et al., 2021), CAML (Mullenbach et al., 2018), DR-CAML (Mullenbach et al., 2018), MultiResCNN (Li and Yu, 2020), MVC-LDA (Sadoughi et al., 2018), C-MemNN, MSATT-KG (Xie et al., 2019), DCAN dilated CNN + N-gram (Kanwal, 2020), IHCE, MARN. Secondly, the category of models based on the RNN approach includes: JointLAAT (Vu et al., 2020), Bi-GRU and HA-GR, and attentive LSTM. Thirdly, category of models based BERT approach are BERT-ICD (Pascual et al., 2021), MIMIC-BERT (Bhavani Singh et al., 2020), Snip BERT (Huang et al., 2020), BERT-XML (Zhang et al., 2020), TransICD (Biswas et al., 2021), KB-BERT (Remmer, 2021), G-BERT, HTN (Si and Roberts, 2021), Hi-Transformer, DLAC, HiLAT (Liu and Xu, 2021). Out of the three approaches mentioned above, this study summed up the primary difficulties with the automatic ICD coding work as follows: clinical notes are longer than PLMs can be, there is an imbalance in class sizes, and code hierarchy is not being used during the training phase.

2.1.1 Deep neural network and hyperbolic geometry

Deep neural networks have recently been employed in representation learning in hyperbolic spaces, where we can add a linear layer on top of their hierarchical structure, because of their amazing ability to model data and latent hierarchies. The various geometric properties like manifold, tangent space, and Riemannian manifold and its metric make it more suitable for representing symbolic data with latent hierarchical structures than the Euclidean space (Nickel and Kiela, 2018). Hyperbolic space properties: hyperbolic space is a non-Euclidean space with unique geometric properties. In the context of hierarchical structures, hyperbolic space is particularly suitable for representing symbolic data due to its ability to capture latent hierarchical relationships (Nickel and Kiela, 2018). The Poincaré ball model is a common representation of

hyperbolic space. In this model, distances between points are measured using a hyperbolic metric. Embedding hierarchical structures in the Poincaré ball allows us to capture hierarchical relationships effectively. We aim to learn hyperbolic representations of ICD codes in the Poincaré ball. This involves embedding the hierarchical relationships between code pairs (Peng et al., 2021). These researchers have pointed out from the literature review as a takeaway capturing the ICD code hierarchical as explicit external knowledge with the BERT approach is a new research direction together with the hyperbolic geometry, and interpretability in an end-to-end manner as an opportunity in deed to be study. Therefore, this research study recognised the importance of using code hierarchy for diagnostic text understanding in the representation learning process of the BERT approach combined with hyperbolic geometry. So that as outcome to able to build a new model and theory. Thus, elucidating the challenges and opportunities inherent in this domain this study paper has tried to adequately address these key issues that have hindered the successful implementation as practice oriented automation of ICD coding processes within healthcare systems as case study of Ethiopia NCoD.

2.2 Ethiopia National Classification of Diseases

The Ethiopian Ministry of Health (MOH) has long since realised how ICT can help and alter Ethiopia's health system. As of April 2016, Ethiopia Health Management Information System (HMIS) indicator definitions and national health data dictionaries (NHDD) were created by the MOH. As part of the HMIS overhaul, the NCoD was created to organise health information. It is the system that is used nationally to define, track, and report health conditions and their results. An analysis of samples from service centres and administrative units conducted nationwide served as the foundation for the NCoD's preparation. In Ethiopia, disease recording and reporting are one of the four technical areas of the nation. The NCoD is made up of three documents that include the essential list of diseases based on the ability of each level of the health care system to diagnose them. The extended edition, which includes 2,055 disorders, was created out of the three NCoD components and is utilised as a reference by tertiary-level healthcare professionals. The NCoD is a condensed system of illness classification that is arranged using the ICD-10 framework. Each diagnostic or disease category has a specific set of ICD codes assigned to it. Moreover, it has a complete mapping to ICD10 (Abate, 2017). According to this study's early inquiry, ICD coding in Ethiopia is now done solely by manual assignment, which is labour-intensive, prone to error, and requires a large number of skilled encoder staff members. These researchers believed that, intelligent ICD coding assignment of clinical narratives using BERT and hyperbolic embedding model design is an essential technique that help to utilise the power of NLP along with interpretable AI model will play an important role in addressing the current and upcoming challenges. Since linking clinical narratives and coding

system (Newman-Griffis and Fosler-Lussier, 2021) is a crucial element in obtaining the information needed for medical text analysis in EMR software. This research study introduced a method for converting unstructured data into structured data on the task automatic and interpretable ICD coding assignment of clinical narratives with the context of NCoD. The topic is addressing ICD code assigning to clinical narratives of medical records – a vital problem in the healthcare sector that otherwise is manually processed which leads to medical records are error-prone and need high human labour (Pascual et al., 2021; Zhang et al., 2020). For automatic ICD code assignment of clinical narratives which uses the ICD code hierarchy as explicit external knowledge using hyperbolic embedding method to enhance the performance along with the BERT based approach is the main contribution of this study. Besides this the research investigates the importance of interpretability of the model approach in healthcare system.

In this study, researchers presented in Section 2 review of literature including concepts, principles and theories about ICD coding, hyperbolic geometry and empirical evidences of ICD coding in deep neural networks are reviewed and presented. In Section 3, methods part the research design approach, model building, mathematical proof equation as theory building, experiment settings, performance results gained from the new model, and case study on NCoD data and result obtained is presented. In results and discussion (Section 4), compared the base line models state of the art with our new model along with visualisation capability of interpretability on our model in predicting ICD codes from discharge summaries using BERT visualisation tool are described unlike the previous related works that are taken as a contribution. The discussion part analysed effectiveness of the new model by analysing the accuracy and the viability of the combined hyperbolic embedding and BERT representation learning as code aware document representation learning. Then the finding of this study are summarised and presented. Finally, the researchers have drawn in Section 5 the conclusion and recommendation for further study besides setting out limitations of this study.

3 Methodology

3.1 Research design

This study used a qualitative research design that combined pragmatism and post-positivism in knowledge inquiry (Creswell, 2009). Moreover, it employed case study and experimental study strategies to explore the design of effective integrated hyperbolic embedding and BERT-based model in predicting ICD codes from discharge summaries.

3.2 Research area description

The question of how to use pre-trained contextualised language models for large-scale medical code prediction to encode extensive clinical notes has not been fully explored

in the field of clinical natural language processing (NLP). Pre-trained language models (PLM) that learn contextualised text representation, like BERT, have ushered in a new era of NLP (Zhang et al., 2020). Big-scale pre-training on big corpora is beneficial for NLP applications, and transfer learning has proven effective in using universal language representations from pre-PLMs in downstream tasks.

3.3 Tools for collecting data

Data collection techniques were used for the whole publicly available MIMIC III dataset and for the 50 ICD codes from <http://www.physionet.com>. The other data collection methods were used for the NCoD data types. We gathered NCoD data from the six hospitals in the IC unit, and the researchers' favourite way to collect data is to randomly pick patient department tally sheets. This makes it possible for several doctors to give the discharge summaries ICD codes. This method will probably produce a dataset with fewer coding errors, but it will require access to physicians who are willing to manually code. For this research, with just 1,200 records, manual annotation is thought to be viable because several clinical notes are required for classification tasks.

As a case study the researcher, observe phenomena in a representative subset of Ethiopia NCoD data collected from six comprehensive specialised hospitals namely: compressive specialised Hospital of Hawassa University College of Medicine, Yekatit 12 comprehensive specialised hospital, and St Peter specialised hospital. In addition to these, Bahar Dar and Gonder compressive specialised hospitals, and Dire Dawa compressive hospital after these researchers have taken ethical clearance letter from MoH and ASTU. Moreover, we used document analysis, along with the review of related literature.

3.4 Analysing and interpreting data

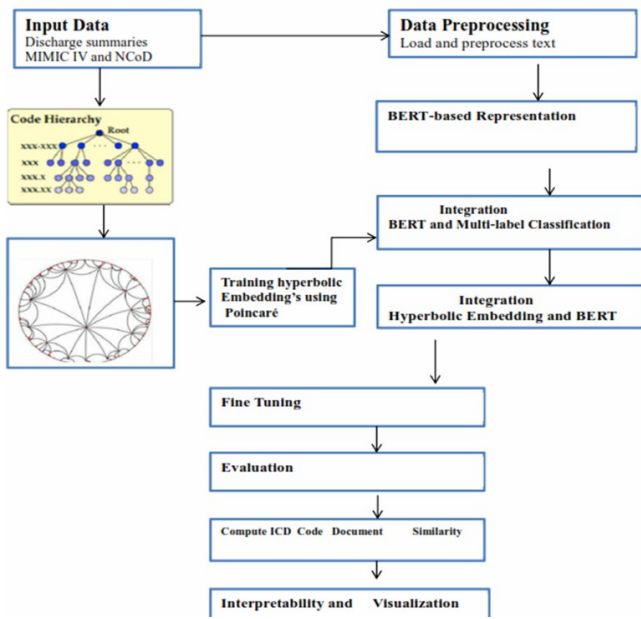
Explanatory analysis is the method of data analysis used in this study. The first example requires collecting a corpus of discharge summaries and their related ICD codes from the MIMIC III dataset, as the prediction task involves multi-label classification and the assignment of an ICD code to a discharge summary. For the Ethiopia NCoD case due to limitation of unavailable medical records and resource, only a total of 1,200 records of discharge summary and ICD codes are collected, 200 records from each six tertiary hospitals.

3.5 Model building

To build the new model as depicted in Figure 1, we identified in review of literature section that hyperbolic embedding can be used in conjunction with BERT to enhance the representation of hierarchical relationships between medical concepts encoded in ICD codes. By leveraging the properties of hyperbolic space, the embedding can capture the hierarchical structure of the ICD

code system more effectively. Moreover the BERT model assists in understanding the context and extracting relevant information from the text to accurately assign the appropriate ICD codes of hierarchical structure. Therefore hyperbolic embeddings can be learned by mapping the ICD codes into an exaggerated space, where the separations between codes represent the semantic connections between them. This allows the model to better understand the hierarchical relationships between different codes and make more informed predictions. By incorporating hyperbolic embedding into BERT, the model may be able to forecast ICD coding activities more accurately and better represent the hierarchical structure of the ICD code system. Thus, we have proposed a novel new model called HyperBERT.

Figure 1 Novel proposed model architecture called HyperBERT (see online version for colours)



3.6 Model architecture description

3.6.1 Input data

The input data are NOTEVENTS.csv for the BERT base representation learning, D_ICD_DIAGNOSES.csv for hyperbolic embedding, after data preprocessing and tokenisation from MIMIC III Full and top 50 dataset beside NCoD data.

3.6.2 Data preprocessing

The data preprocessing is focused on labelling the raw data to create a simplified binary multi label classification. The data preprocessing includes; tokenisation, loading dataset, filter columns and rows, drop rows with missing values and split the data in to train data 80%, test data 10%, validation set 10%.

3.6.3 Hierarchy of ICD codes

ICD codes are grouped in a manner resembling a hierarchical tree, with higher level nodes representing more general illness categories and lower level nodes representing more particular disorders. The sibling, mutual exclusion of ICD codes, and parent-to-child relationship can all be captured by the code hierarchy.

3.6.4 Hyperbolic embedding of the code hierarchy

We trained Poincaré embedding in Hyperbolic Embedding using the hierarchical structure of ICD codes. Hyperbolic embedding is used to map the ICD codes to hyperbolic space. The hierarchical link between the codes is captured by the model thanks to this transformation. We embedded the hierarchical structure into hyperbolic space using Poincaré embedding.

3.6.5 BERT base representation learning

The clinical narratives are encoded – or embedded, as it is called – into fixed-length representations using the pre-trained BERT model. These embedding's capture the contextual information of the text. The encoded clinical narratives and hyperbolic embedding of ICD codes are fed into a BERT model to determine which ICD code is most suited for a particular clinical narrative.

3.6.6 Combine BERT and Poincaré embedding

For ICD code assignment of clinical narratives the encoded clinical narratives and hyperbolic embedding of ICD codes are fed into a BERT model to determine which ICD code is most suited for a particular clinical narrative. We combine the BERT embedding of diagnostic code descriptions with their corresponding Poincaré embedding in the hyperbolic space. See below for detail description in theory building (Section 2.4) and experimental setting (Section 2.5).

3.6.7 Performance evaluation

The document-code similarity score was computed by comparing BERT embedding of documents with hyperbolic embedding of ICD codes for method comparison during the experiment study with known metrics of macro and micro F1-score, macro and micro AUC. Since BERT embeddings are in Euclidean space and ICD code embeddings are in hyperbolic space, an alignment mechanism is required. This could involve using a similarity measure that accommodates both spaces. The experimental setting and detail description are explained in Section 3.8.5. Since the performance evaluation of a model using macro and micro F1-scores and macro and micro AUC metrics is critical for assessing its ability to predict ICD code hierarchies from clinical notes. The macro and micro F1-scores are computed to assess the model's precision and recall performance across all ICD

codes. The macro and micro AUC metrics evaluate its ability to distinguish between correct and incorrect classifications at varying thresholds. As insights from evaluation, these researchers indicated that during experiment execution a higher macro F1/AUC indicates that the model performs well across all ICD codes, including less frequent ones. A high micro F1/AUC shows strong performance on more frequent ICD codes, reflecting the dataset's class distribution. The results highlight the effectiveness of the combining BERT's semantic strengths and hyperbolic geometry in handling hierarchical ICD code prediction. This evaluation ensures the model's robustness and effectiveness for real-world clinical applications.

3.6.8 Fine-tuning process

The process of fine tuning the BERT model for specific task of ICD coding of medical text as downstream task of inputs and outputs are fed in to the model and all parameters are fine-tuned in an end to end manner.

3.6.9 Compute document-code similarities

To compute document embedding we used the BERT-based representation to encode the discharge summaries into dense vector representations along with hyperbolic embedding using known metrics of AUC and F1-score.

3.6.10 Interpretability of our model

In this study attention mechanisms were applied code wise label attention for interpretability of our new model. Moreover, we visualised the embeddings and their relationships to understand the hierarchical structure as code-aware document representation.

3.7 Mathematical analysis of the proposed research

This section discusses building theory along with mathematical proof. BERT plays a significant role for contextual text representations and it is based on Euclidian space. This transformer-based model is designed to learn contextual embeddings through pretraining on extensive datasets, utilising techniques like MLM and NSP (Devlin et al., 2019). Hyperbolic geometry provides a non-Euclidean space with constant negative curvature, embedding data with hierarchical structures, such as ICD codes, presents a valuable opportunity for enhanced data organisation (Peng et al., 2021).

Thus, in this research study, hyperbolic embeddings and BERT approaches are complementary techniques in natural language processing (NLP). Combining them involves leveraging the geometry of hyperbolic space for representing hierarchical relationships in medical ICD codes while utilising BERT for contextual text representations.

Prior research has employed a standard method of embedding hierarchical data structures using the hyperbolic space. In hyperbolic embedding's, hyperbolic spaces are

often used to model hierarchical data structures due to their ability to capture tree-like relationships efficiently. The Poincaré ball model is a popular representation of hyperbolic space. In this model a hyperbolic metric is used to estimate the distances between locations. Embedding hierarchical structures in the Poincaré ball allows us to capture hierarchical relationships effectively (Peng et al., 2021).

3.8 Mathematical equations description

The detailed explanation of the mathematical proof equations are described follows:

A loss function or the Poincaré ball space $D(X, Y)$ was established to quantify the difference between the hyperbolic embedding and the original data's hierarchical structure. One commonly used loss function for hyperbolic embedding's the Poincaré ball distance. It calculates the separation in hyperbolic space between points.

The equation involving in the context of ICD codes hierarchy embedding using Poincaré distance refers to the distance metric used to represent how 'far apart' two ICD codes are in the hierarchical structure.

As a result, the Poincaré distance in hyperbolic space between two points, X_i and Y_j , is given by equation (1), which is the total of the Poincaré distances between points and their corresponding embedding's.

The Poincaré distance, given two points (x, y) in the hyperbolic space, is defined as follows.

$$D(X, Y) = \text{arcosh} \left(1 + \frac{2(\|X - Y\|)}{(1 - \|X\|^2) - (1 - \|Y\|^2)} \right) \quad (1)$$

where

- X, Y are points in the Poincaré ball (a bounded space with radius $\|X\| < 1$ and $\|Y\| < 1$)
- $\|X\|$ and $\|Y\|$ are the Euclidean distances between them
- arcosh is the inverse hyperbolic cosine function.

This equation is used to represent a distance function involving the inverse hyperbolic cosine (arcosh), a common function used in metrics related to hyperbolic geometry. The terms involving absolute values, squares, and the denominators suggest a relationship tied to similarity search between ICD coding and clinical narrative contexts. The reason we used Poincaré embeddings for ICD codes includes: ICD codes have a hierarchical structure, where broad disease categories branch into finer subcategories.

Hyperbolic space naturally represents hierarchical data, as distances between nodes increase exponentially, and Poincaré embeddings efficiently capture hierarchy, meaning parent-child relationships in ICD codes are well preserved.

To optimise hyperbolic embeddings in a Riemannian space, we need to use Riemannian optimisation techniques. Riemannian optimisation methods take into account the underlying geometry of the space to update the parameters (Bonnabel, 2013).

Adaptive moment estimation (Adam) is a commonly used optimiser in deep learning, combining the benefits of adaptive gradient or AdaGrad and root mean square propagation or RMSprop. However, Adam can suffer from instability, especially in early stages. Riemannian adaptive optimisation can be thought of as adaptive methods that respect the non-Euclidean geometries in embedding spaces. Many datasets, particularly in NLP or hierarchical data, benefit from embedding points on manifolds that respect their inherent structure. Euclidean embeddings, for instance, may not capture the true ‘distances’ between points in a hierarchical structure

In Riemannian optimisation, one typically considers curvature constraints to ensure the embeddings remain consistent with the underlying geometry (often hyperbolic). This can involve Poincaré embeddings.

The optimisation process involves updating the parameters of the hyperbolic embeddings using the Riemannian gradients.

Here is a simplified version of the update equation mention in model build in Section 3.4:

$$\theta_{\{t+1\}} = \{\text{RiemannianAdamUpdate}\} \quad (2)$$

In this equation, the parameters at time step (t) are represented by (θ_{-t}) . g_{-t} represents the Riemannian gradients at time step (t).

RiemannianAdamUpdate is the update function specific to RAdam of Riemannian optimisation algorithm.

In learning hyperbolic representation, we aim to learn hyperbolic representations of ICD codes in the Poincaré ball. This involves embedding the hierarchical relationships between code pairs. By learning embeddings in the hyperbolic space and computing similarity scores based on distances between ICD code and the documents.

Therefore integration of hyperbolic embeddings and BERT embeddings is defined as: let U be the input data representing hierarchical information encoded in BERT.

For the hyperbolic embedding derived from the Poincaré model, let V be used.

The goal is to find Z which is the optimal hyperbolic embeddings Y that capture the hierarchical relationships encoded in X .

$$Z = \{\text{similarityscoreweight}(U + V)\}(\theta_{\{t+1\}}) \quad (3)$$

In this equation, V is the Poincaré distance defined in equation (1), U is the vector representation of input text in the BERT model, and $\theta_{\{t+1\}}$ is the RAdam learning rate defined in equation (2).

Therefore, in this research study, the researchers identified using the above function of X and Y by incorporating the Poincaré distance as the loss function and using Riemannian adaptive optimisation techniques, called RAdam learning rate, we effectively optimise hyperbolic embeddings to capture hierarchical relationships within the hyperbolic space is the contribution of this study as theory building.

3.9 Experimental study

This section provides a comprehensive description of the execution of our experiment that is done to implement the model architecture described in Section 3.6 and theory building which is described in Section 3.8 which includes: data sources, data preprocessing, model training procedures, fine-tuning process, performance evaluations and metrics along with interpretability of our model in case study of NCoD. The conducted experiment is to predict using the input text of discharge summaries from MIMIC III full, top 50 codes dataset, and the case of Ethiopia NCoD data, as the ICD coding task is a multi-label classification problem. To capture the hierarchical structure of the diagnosis ICD code during the representation learning process of the transformer-based approach of BERT base model combined with hyperbolic embedding, the prediction tasks are evaluated on the proposed model to the prediction of ICD codes from discharge summary.

3.9.1 Data sources

3.9.1.1 MIMIC III dataset

The Medical Information Mart for Intensive Care III (MIMIC-III) dataset is a large, freely accessible database containing de-identified health-related data of over 40,000 critical care patients admitted to the Beth Israel Deaconess Medical Center between 2001 and 2012. It is one of the benchmark datasets for ICU medical records is MIMIC, which is frequently used in ICD coding prediction. There is a sizable collection of de-identified, openly accessible medical records for individuals who were admitted to the intensive care unit (Johnson et al., 2016). A discharge summary that lists the diagnosis and treatments that took place during a patient’s stay is included in every one of its records. Skilled coders assign an ICD code to each record. MIMIC III data collection includes, among other things, discharge summaries, radiological reports, laboratory measurements, microbiology cultures, prescription drug orders, vital signs, and other numerical and textual data. ICD-9 codes are used to identify all of the diagnoses in the MIMIC dataset (Johnson et al., 2016). MIMIC-III-50 is a subset of MIMIC-III-Full that includes EMRs that have been assigned at least one of the top 50 most frequently used codes. We have created three common configurations of the dataset: MIMIC-III, top 50 codes, and NCoD, building on earlier efforts on BERT_XML and BERT_ICD. The MIMIC III dataset of discharge summary notes file contains with IDs for the subject, note, and hospital admission and raw texts as NoteEvent.csv file. Using the IDs the notes can be combined with diagnosis code in another file called Code_dignosiscode.csv. The vocabulary size of this MIMIC dataset is 51,917. The MIMIC III dataset of discharge summary notes file contains with IDs for the subject, note, and hospital admission and raw texts as NoteEvent.csv file. Using the IDs the notes can be combined with diagnosis code in another file called Code_dignosiscode.csv.

3.9.1.2 NCoD dataset

The raw data that is used for fine-tuning the model is discharge summaries from the database of EMR software from selected six compressive specialised hospitals from Ethiopia are collected. NCoD_description.csv with 200 from each hospital total 1,200 rows is applied with the model to check the AUC, and F1-scores. We have applied augment it on these 12,000 rows and split them training, test and validation set of 80%, 10% and 10%, respectively. The NCoD data were entered, cleaned, and coded and checked for missing value and inconsistencies. Data obtained from electronic medical records were cross-matched for availability in NCoD and ICD.

3.9.2 Data preprocessing

In data preprocessing is focused on labelling the raw data to create a simplified binary multi label classification. The major tasks includes: load the data and preprocess the three datasets described above. Then extract relevant features, including discharge summaries and associated diagnostic codes.

We used to extract discharge summary along with diagnostic ICD codes of each patient ICU unit dataset of MIMIC III for the training data after data preprocessing and tokenisation process are done. These activities includes such as loading MIMIC III dataset, filter columns and rows as ('SUBJECT_ID', 'HADM_ID', 'TEXT', 'ICD_CODES'), drop rows with missing values, and split the data as train data 80%, test data 10%, validation set 10%. By training the complete discharge summary set for MIMIC III datasets and the Top 50 ICD code, we were able to use BERT word pieces to create word embedding's of size, $d = 254$ for each sample of discharge summary notes.

For each discharge summary notes sample in NCoD dataset, in addition, we tokenise the content, lowercase it, and eliminate all punctuation, digits, and English stop words. We note that over 88% of the discharge summaries are constrained within 2,500 tokens from the resulting distribution of token count per record. Therefore, the maximum length of token sequence used for training by these researchers was 25,000. Ultimately, we take an 8037 token vocabulary out of the training set and add 'PAD' and 'UNK' tokens to it for padding and out-of-vocabulary terms, respectively.

3.9.3 Model training

We have defined the Poincaré ball space as described theory building and mathematical equations description (Section 3.7) of this study together with RAdam optimiser. The model was then trained on the training set of NOTEVENTS.csv, D_ICD_DIAGNOSES.csv from MIMIC III Full and top 50 data and the researchers verified it on the validation set. We adjusted the pre-trained BERT model for the ICD code assignment task to enhance the model's performance. Then, using established metrics for the multi-label text classification problem such as the AUC

and macro and micro F1-scores, which are discussed in Section 3.9.6, model was evaluated on the testing set from NCoD, and also test set from NOTEVENTS.csv, D_ICD_DIAGNOSES.csv from MIMIC III full and top 50 data.

3.9.4 Training processes

After looking for the ideal hyperparameter, we settle on the following value set: two encoder layers, 12 attention heads, 32 epochs, dropout rate: 0.1 for both the beta 1 and beta 2 learning rates of the Ramdam optimisation. The models are trained by these researchers using Colab Google NVIDIA Tesla P100 GPU assistance. At optimal settings, an epoch lasts around 168 seconds.

3.9.5 Detail description of performance evaluation

Numerous assessment criteria have been put out to assess the methods' efficacy from various angles. In this study's technique comparisons, the F1-score and the area under the receiver operating characteristic curve (AUC) were used to report the predictive performance in the automatic ICD code assignment of discharge summaries. AUC evaluates a classifier's capacity for class distinction. The true positive (TPR) and false positive (FPR) rates are plotted against one other. Standard metrics such as AUC and F-score can be measured micro- or macro-averaged to assess performance in multi-label classification issues described on article (Mullenbach et al., 2018).

In this study following state of the art model by Yan et al. (2022) macro- and micro-average methods of F1-score and AUC are used to determine each. Metrics computed for each label are averaged to get macro- average metrics. Since the task at hand involves multiple imbalanced classes, the prediction of rare labels is given priority by macro-average metrics, which view this as a stronger signal of the model's performance in tiny classes.

The model is trained on clinical notes, with the ICD code hierarchy incorporated using hyperbolic embeddings and contextual representations provided by BERT. The macro and micro F1-scores are computed to assess the model's precision and recall performance across all ICD codes. The macro and micro AUC metrics evaluate its ability to distinguish between correct and incorrect classifications at varying thresholds. We have taken the following actions to calculate document-code similarity scores by comparing BERT embeddings with hyperbolic embeddings, as outlined in the theory. Building portion of Sections 3.7 and 3.8, and to assess the effectiveness of our new model using F1-scores and AUC performance evaluation using F1-score, we calculated F1-scores for the predictions and ground truth labels using Scikit-learn's `f1_score` function of macro and micro F1-scores. For calculating the AUC score we also calculated the macro and micro average weight using Scikit-learn's `roc_auc_score` function. For further explanation see source code in <https://github.com/derejeweyessaa/HyperBERT/blob/main/HyperBERT.ipynb>.

3.9.6 Interpretability of the model

We used code-wise label attention mechanism and implemented it in our experiment that focuses on specific ICD codes during the prediction process. By using the discharge summaries' content as a guide, train the model to dynamically attend to pertinent ICD codes. For visualising attention weights and embeddings the researchers used tools like BERT visualisation libraries to examine attention weights and embeddings. Analyse the distribution of attention and determine which sections of the input text the model is focusing on by visualising the attention heads.

4 Results

4.1 HyperBERT model results

This study aims to address the research questions mentioned in Section 1.3 of this study and we in this section discuss, describe and present the results from the experimental study and case study. Our new model HyperBERT is fine-tuned from the pertained BERT based uncase model from the transformers library. AutoModelSequenceClassification and trainers' classes from the transformer library of BERT model and hyperbolic embedding using PyTorch and Hanging Face libraries have used to train and evaluate the performance during the fine tuning. The results gained from the known metrics of F1-score and AUC metrics are shown in Table 1. The results from the integrated HyperBERT approach show the model focuses on relevant words while classifying the input text of discharge summaries as code document aware representation learning. Table 1 presents the comparison of our proposed HyperBERT with the base lines of previous related works on the full and top-50 ICD codes for NCOD with the MIMIC-III dataset. Using the previously described hyperparameter setting, the scores are calculated on the held-out test set. The dataset used for evaluation of performance on the experimental our new model HyperBERT takes discharge summary text of NOTEVENT.csv D-DIGNOSISCODE.csv from MIMIC III dataset of full, top 50 codes and NCOD data. The F1-score AUC (Yan et al., 2022) is the primary evaluation measure used. It is computed using both macro- and micro-average methods for the categorisation or prediction of an ICD code based on the discharge summary.

4.2 Comparison of experimental result and baselines

According to the literature study in Section 2.2, these researchers compared the proposed model with the best score results from state-of-the-art models, specifically BERT-ICD (Pascual et al., 2021), BERT-XML (Zhang et al., 2020), and Snip BERT (Huang et al., 2020; Mullenbach et al., 2018) based method. BERT-ICD investigated ICD coding using PubMed BERT and the length of discharge summary is the bottleneck to BERT and CNN approach is better for this task. BERT-XML model applied by adding XML to the BERT architecture with self attention module to accommodate larger input length and recommend hierarchal based methods are promising. While Snip-BERT model proposed a novel fine tuning to obtain efficient interpretable from long clinical text by identify crucial snips and truncated BERT in hierarchy manner using Bio BERT. MIMIC BERT model proposed five strategies to split the long text different fixed length chunks using PubMed BERT for ICD coding task.

From Table 1, we can observe that unlike the baselines, our mode HyperBERT achieved best result in AUC macro and micro with 0.937 and 0.966 score, respectively in MIMIC full dataset. And macro F1-score 0.75 and micro F1-score of 0.94 in MIMIC III full dataset.

HyperCore AUC micro result was 0.929 and 0.897, respectively. Moreover, in MIMIC top ICD codes our model HyperBERT achieved best result on AUC micro 0.97 and micro F1-score of 0.62 other than best result of the BERT_XML 0.933 F1 micro score. Unlike CAML model that used CNN approach that tried to solve the class imbalance problem of ICD coding task but ignored capturing ICD code hierarchy during training process, and achieved 0.876 and 0.912 of AUC macro and micro while our new model result increase in performance of 0.15 AUC and 0.184 AUC macro and micro respectively. Unlike HyperCore model with 0.609 and 0.663 AUC macro and micro in full MIMIC dataset while this model utilised joint approach CNN approach and hyperbolic embedding in graph approach in capturing ICD the code hierarchy during the representation learning. Our model achieved best result in F1-score of macro 0.75 and micro 0.684, respectively.

Table 1 Comparison of our proposed HyperBERT with the base lines of previous related works on the full and top-50 ICD codes for NCOD with the MIMIC-III dataset

Models	MIMIC III Full AUC macro and micro	MIMIC III full F1-score macro and micro	MIMIC III Top 50 AUC macro and micro	MIMIC III top 50 F1-score macro and micro	NCOD top 50 AUC macro and micro	NCOD top 50 F1 macro and micro
CAML	0.875 and 0.919	0.532 and 0.614	0.895 and 0.936	0.888 and 0.539		
Hyper_Core	0.895 and 0.929	0.609 and 0.663	0.93 and 0.929	0.9 and 0.55		
BERT-ICD			0.887 and 0.845			
BERT XML			0.933 and 0.825			
Snip-BERT			0.994 mic	0.86 mic		
New model HyperBERT	0.937 and 0.966	0.75 and 0.684	0.89 and 0.97	0.542 and 0.62	0.92 and 0.95	0.542 and 0.822

With regards to top 50 MIMIC dataset our model did not show significant improvement unlike the baseline models but only in AUC micro of 0.97 score, and might be due to small data usage and/or infrequent tokens not sufficiently considered during in the fine tuning process and we believe there is a need to study capturing infrequent ICD codes with hierarchical ICD code embedding. The use of hyperbolic embedding in conjunction with the integrated model architecture of HyperBERT, as shown in Figure 1 in Section 3.4 of this study to capture the ICD code hierarchy from the ICD_Diagnosis.csv file is most likely the cause of the above enhanced performance over the previous state of the art models along with the usage of optimisation techniques of Riemannian adaptive update function besides, theoretically verified concept and loss function utilisation so that our model resulting in increase 0.003 AUC macro and AUC micro with 0.02 in MIMIC Full, NCOD data. The Ethiopian NCoD dataset offers a localised alternative to the MIMIC III dataset for integrating hyperbolic embeddings with BERT for ICD code prediction. This has relevance to Ethiopia’s healthcare context since NCoD reflects the unique disease burden, coding standards, and linguistic context of Ethiopia, making it more applicable for developing models tailored to the Ethiopian healthcare system. It includes specific disease codes and classifications relevant to the region, which may differ from those in MIMIC III. The advantages of the case study approach are included: improved ICD code prediction accuracy by aligning local clinical notes with the NCoD hierarchy. Enhanced interpretability through code-wise label attention, aiding in healthcare decision making within Ethiopia. A scalable framework that can be adapted to other regional healthcare systems. The results gained from the experimental execution of this study shown in Table 1 indicate that our new proposed model achieved best result from NCOD data in AUC macro and micro 0.92 and 0.95, respectively. Moreover, the F1-score mac 0.542 and micro 0.822 respectively indicate that there is improvement in AUC micro by 0.02. Therefore, we believe that intelligent ICD coding assignment of clinical narratives using BERT and hyperbolic embedding model design is an essential technique that helps to utilise the power of NLP along with interpretable AI model, and will play an important role in achieving automatic and interpretable NCOD clinical narratives in predicting ICD codes with the context of since connecting clinical narratives with the EMR software’s coding. This system is essential to releasing the information contained in medical text analysis. Ethiopia’s healthcare system is tackling both present and future obstacles.

4.3 Interpretability of the HyperBERT model result

To enable the characterisation of model correctness, fairness, transparency, and outcome in AI-powered decision making, the autonomous ICD coding decision-making process needs to be explicable. We implemented code-wise label attention mechanism that focuses on specific ICD codes while processing discharge summaries. The attention

mechanism helps to weigh the importance of different ICD codes during the prediction process. To visualise attention weights, extract the attention weights for each code during inference. BERT visualisation tool is applied to plot the graph. These attention weights help to interpret how the model attends to different ICD codes plot the embeddings to observe clusters and relationships between discharge summaries and ICD codes so that assist in understanding the task for clinicians during decision making. In the implementation of our model, we used for each input the average classification token embedding for both class 1 (relevant tokens are presented from lung disease chapters NCOD dataset) and class 2 (relevant tokens are not available of lung disease chapter NCOD dataset) of binary label text classification problem, can be calculated using the code wise label attention score. During implementation all tokens of discharge summary the difference of the embedding vectors calculated and are used to plot the graph using BERT visualisations tools. The results from the above-described method shown in Figures 2–4 indicate that our model is reliable to objectively predict the ICD codes while classifying the input text from the NCOD data. One or more attention heads in the same layer are visualised in the head view. It is predicated on excellence. An overview of attention across all layers and heads is provided by the model view.

Figure 2 Model view of our model visualisation tool (see online version for colours)

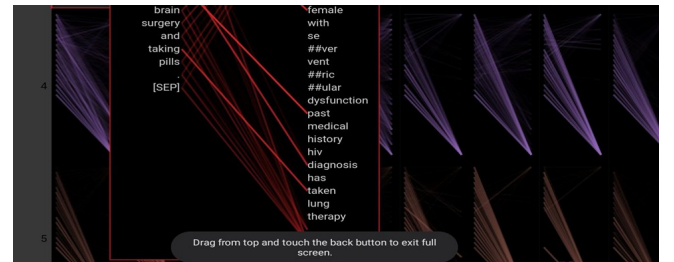
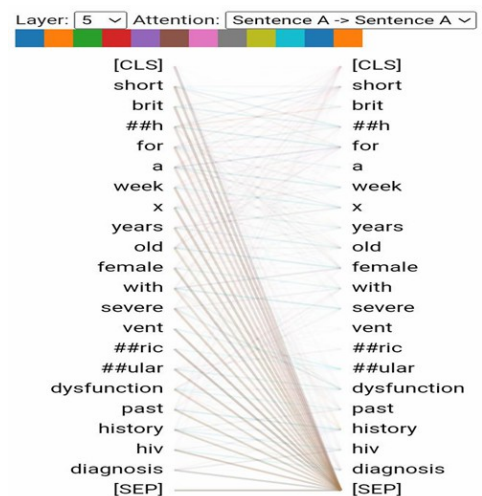
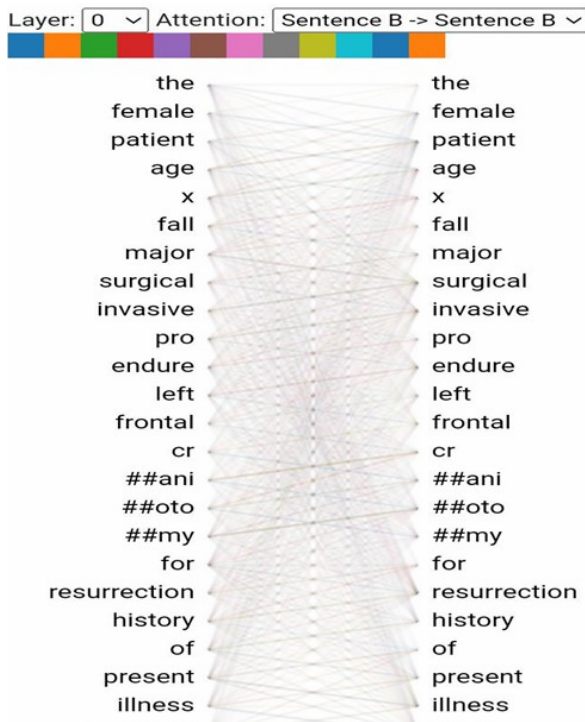


Figure 3 Head view of token embedding using BERT vis tool. Head view (attention, tokens, sentence_a_start) (see online version for colours)



An example of the above described interpretability approach is depicted in model view (Figure 2), head view lung relevant tokens (Figure 3), and head view visualisation of weight of each token (Figure 4) in predicting ICD code. The view portrays the attention weights between tokens in the fifth layer of the model.

Figure 4 Head view visualisation using BERT vis tool. Head view (attention, tokens, sentence_a_start) (see online version for colours)



4.3.1 Class 1 – discharge summary note relevant with lung disease

Short birth for a week X years of old female with severe ventricular dysfunction past medical history HIV diagnosis, has taken heart therapy, stopped intermittently taking her pills three months ago because she stated she had foamy vomit every time she took them.

4.3.2 Class 0 – discharge summary note with not relevant text for lung disease.

The female patient aged x years old fall major surgical or invasive procedure. Left frontal crab into my for-resection history of present illnesses. This patient was bored to have weakness underwent CT head, c spine and orbital CT scan which showed a C1 fracture imaging also showed a large brain mass and the patient was transferred to for neuro surgical intervention.

4.4 Analysis of the research results

This study set out to explore how we can design a model that able to capture ICD code hierarchy using hyperbolic embedding's and using BERT approach for understanding

the context discharge summary during classifying ICD codes from medical texts.

4.4.1 Analysis of the hyperbolic code embedding

The idea of modelling ICD hierarchy in hyperbolic space is taken from similar pervious work called HyperCore model (Cao et al., 2020) which utilised CNN approach and hyperbolic embedding. Prior research has employed a standard method of embedding hierarchical data structures using the hyperbolic space. Hyperbolic embeddings and BERT approaches are complementary techniques in natural language processing (NLP). Combining them involves leveraging the geometry of hyperbolic space for representing hierarchical relationships in medical ICD codes while utilising BERT for contextual text representations. When applied to tasks like ICD coding MIMIC-III dataset, these methods improve hierarchical understanding and semantic retrieval. Theory: hyperbolic geometry provides a non-Euclidean space with constant negative curvature, embedding data with hierarchical structures, such as ICD codes, presents a valuable opportunity for enhanced data organisation. Principle: the nature of hyperbolic space, where distances increase exponentially, is particularly advantageous for representing hierarchical relationships, as it allows for parent-child distances to diminish logarithmically compared to Euclidean spaces (Peng et al., 2021). BERT plays a significant role in this context. Theory: this transformer-based model is designed to learn contextual embeddings through pretraining on extensive datasets, utilising techniques like masked language modelling (MLM) and next sentence prediction (NSP). Principle, by capturing deep contextual relationships in text, BERT generates effective embeddings for words and sentences, which can be leveraged for various applications (Devlin et al., 2019; Newman-Griffis and Fosler-Lussier, 2021; Pascual et al., 2021). Unlike HyperCore (Cao et al., 2020), BERT-ICD (Pascual et al., 2021), Snip BERT (Huang et al., 2020), BERT-XML (Zhang et al., 2020), model utilised using CNN and BERT without approach hyperbolic embedding, in this study by integrating these strategies, incorporating RAdam optimisation, and emphasising semantic similarity analysis, code-aware representations, and attention mechanisms, we can able to potentially boost the AUC micro metric and F1-macro average in HyperBERT mode as described in Section 4.1 of this study. These researchers combined hyperbolic embeddings with BERT-based representations and optimise the embeddings to capture hierarchical relationships within the hyperbolic space, these researchers used Riemannian update optimisation methods as described in Section 2.6 and of this study, explained the model architecture process in Section 3.6 and Section 3.8 about mathematical proof equations involved as theory building.

4.4.2 *Effectiveness of contextual representation learning on BERT*

For analysing semantic similarity between discharge summary document and ICD code, we used pre-trained models of BERT to obtain the learned representation of discharge summaries as code-aware document representations that incorporate information about ICD codes into the encoding of discharge summaries by combining embeddings of relevant ICD codes with discharge summary embeddings to create comprehensive representations (Newman-Griffis and Fosler-Lussier, 2021; Pascual et al., 2021). We examined feature importance on BERT visualise attention heads to determine which sections of the input text the model is focusing on and to examine the distribution of attention. Subsequently, we evaluated the trained model's semantic similarity in order to evaluate the efficacy of contextual representation learning on BERT's base representation. The better experiment results gained in Section 4.1 of this study using metrics F1-score and AUC to indicate the accuracy of the model's performance in predicting ICD codes from discharge summaries.

4.4.3 *Effectiveness of interpretability of the model*

When comparing the interpretability methods for an ICD coding task, self-attention enables models like BERT to weigh input tokens differently during encoding, providing insight into which words contribute most to predictions. The label-wise attention focuses on how specific labels or classes are attended to in the model, shedding light on the importance of different labels in the context of ICD coding. With regards to hierarchical label-wise attention extends label-wise attention to hierarchical structures, potentially capturing dependencies between labels at different levels of abstraction. The gradient and attribution propagation techniques such as integrated gradients or SHAP can help understand the impact of input features on model predictions by attributing importance scores to each input feature (Huang et al., 2020; Yan et al., 2022).

However, in our interpretability method which is called code-wise label attention approach would likely involve focusing on how the model attends to specific ICD codes during prediction, offering insights into the model's decision process and how it utilises code-specific information. We have defined a function for attention visualisation for the code wise label attention analysis. These attention visualisation analyses take input tokeniser from discharge summary and return tensor, text with each input ID then record the attention mask from BERT approach, and output input IDs and attention mask. Then we have compared the successfulness of each method in the review of literature demonstrated by visualising the attention weights of code wise label attention plotting in graphs, the results in our method show indeed better than the other methods in understanding which parts of the discharge summary contribute to code predication. Thus, the study's conclusions are summarised as follows in order to respond to the research questions given in the introduction.

We have built a novel model architecture along with mathematical proof description as theory building on concept of hyperbolic embedding that help on capturing the ICD code hierarchy as external explicit knowledge in the prediction of ICD code from discharge summaries combining with BERT approach. The BERT approach aimed to embed discharge summary text while the hyperbolic embedding to utilises code hierarchy in an end-to-end manner in addition to exploration of which method of interpretability is relevant to the new model on case study of NCoD data and publicly open MIMIC dataset to answer the first research questions.

The analysis of our new model effectiveness discussed in the previous section along with the performance result gained directly relating from incorporating hyperbolic embeddings with BERT-based representations learning, and optimising them using Riemannian adaptive optimisation methods is indeed the cause of the effectiveness of our new model. By comparing the state of the art models performance with the new models with our new model's performance improvement are gained in this study thus, we have worked towards increasing the AUC micro metric of 0.003 and the F1 macro average of 0.02 on HyperBERT model to be able to confirm as the experiment results analysis. To answer the second research question as shown from the case study of NCoD data, the performance results gained as shown and discussed from Section 4.1 in Table 1 indicate that our new proposed model HyperBERT achieved best result from NCoD data in AUC macro and micro 0.92 and 0.95, respectively. Moreover, the F1-score mac 0.542 and micro 0.822 respectively indicate that there is improvement in AUC micro of 0.02. The demonstrations of using BERT visualisation tools our model as practice oriented and interpretable AI using code wise label attention approach in (Figures 2,3,4) of model view and head view, respectively in the context of Ethiopia healthcare system.

4.5 *Findings*

Generally, the main findings of this study in integrating BERT with hyperbolic embeddings for ICD code hierarchy prediction leverages both the contextual power of BERT and the hierarchical representation of hyperbolic space. Here is an analysis of its effectiveness in the context of automatic ICD code prediction from clinical notes in the MIMIC III dataset:

4.5.1 *Contextual representation learning with BERT*

- *Strengths:* BERT excels at capturing contextual meaning in text. Since ICD codes are linked to clinical terms, symptoms, and diagnoses, BERT helps generate rich, contextually-aware embeddings for each note.
- *Challenges:* ICD codes often have complex hierarchical structures that are not directly encoded by BERT. BERT captures semantic similarities in a Euclidean space, which can struggle to reflect the complex ICD relationships without additional structures.

4.5.2 Hyperbolic embeddings for hierarchical representation

- *Strengths*: hyperbolic space effectively represents hierarchical data like ICD codes, where some categories are broad while others are highly specialised. Poincaré embeddings map ICD codes into a hyperbolic space, capturing these hierarchical relationships more naturally than Euclidean embeddings.
- *Effectiveness*: embeddings in hyperbolic space tend to more accurately reflect the distance and containment relationships that naturally exist between ICD codes. For instance, a code under ‘cardiovascular diseases’ should be close to ‘heart diseases’ and much farther from ‘dermatological diseases’.

4.5.3 Combined power of BERT and hyperbolic embeddings

- *Integration effectiveness*: by combining BERT’s contextual embeddings with hyperbolic embeddings, the model can learn both the textual context and the hierarchical ICD structure. This combination allows for effective prediction, even in cases where notes might describe symptoms or conditions indirectly related to the final ICD classification.
- *Document-code similarity*: the integration enables BERT’s contextual information to be ‘anchored’ in hyperbolic space, where similar codes are closer in hierarchy. This reduces ambiguity in hierarchical classification by ensuring context is interpreted within the structured ICD code relationships.
- *Attention mechanism*: implementing code-wise label attention allows the model to prioritise specific portions of the document relevant to the target code. This makes the prediction more precise and relevant to the context in clinical notes.

5 Conclusions

As one of the upcoming issues, this paper addresses the limitations of current methods for intelligent ICD coding assignment in clinical text using deep neural networks, highlighting challenges such as large label space, unbalanced label distribution, large text input size, and the lack of code hierarchy usage as external knowledge. A review of literature shows that BERT-based models are the most effective for automatically assigning ICD codes to clinical narratives, although their performance varies with datasets and fine-tuning procedures. To improve upon existing models, we propose the HyperBERT, which utilises hyperbolic embeddings to better represent the hierarchical relationships in the ICD code system. Our approach

employs the Poincaré ball model to learn hyperbolic representations, enhancing document-code similarity and capturing code hierarchy. Our experimental results demonstrate that HyperBERT outperformed state-of-the-art models, achieving improvements in AUC micro metrics and F1-scores on the MIMIC III and NCoD datasets. Key strategies included applying hyperbolic embeddings, optimising for semantic similarity, and developing code-aware document representations that dynamically attend to relevant ICD codes. We also implemented a code-wise label attention mechanism to focus on specific codes during predictions and used visualisation techniques to analyse attention weights and embeddings. Future improvements may entail addressing the computational complexity associated with integrating BERT and hyperbolic space and expanding the dataset beyond the initial 1,200 records to enhance model performances of words taken from discharge summaries data lung disease class of NCoD data due to resources constraints during execution of the experiments. These authors in future would like to suggest improvements in utilising domain-specific pre-training models like clinical BERT other than generalised BERT model since pre-training BERT on a more extensive corpus of medical literature specific to ICD codes could enhance the contextual understanding of the embeddings and improve the overall performance. The model integration with the healthcare system of Ethiopia challenges and strategies includes ensuring compatibility with existing standards and mapping hyperbolic embeddings back to traditional code system. Beside addressing privacy and ethical considerations when working with healthcare data and ensure compliance with relevant regulations and guidelines of national NCoD.

References

- Abate, A. (2017) *National Classification of Diseases (NCoD) Instructions Manual*, July, EMOH, Addis Ababa.
- Arkey, L.S. (1996) ‘Combining classifiers in text categorization’, *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.289–297.
- Bhavani Singh, A.K., Guntu, M., Bhimireddy, A.R., Gichoya, J.W. and Purkayastha, S. (2020) *Multi-Label Natural Language Processing to Identify Diagnosis and Procedure Codes from MIMIC-III Inpatient Notes*, ArXiv, <https://arxiv.org/abs/2003.07507>.
- Bonnabel, S. (2013) ‘Stochastic gradient descent on Riemannian manifolds’, *IEEE Transactions on Automatic Control*, Vol. 58, No. 9, pp.2217–2229, <https://doi.org/10.1109/TAC.2013.2254619>.
- Cao, P., Chen, Y., Liu, K., Zhao, J., Liu, S. and Chong, W. (2020) ‘HyperCore: Hyperbolic and co-graph representation for automatic ICD coding’, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, pp.3105–3114.
- Creswell, J.W. (2009) ‘Research design’, *Qualitative, Quantitative and Mixed Methods Approach*, p.270, SAGE Publication, Los Angeles.

- de Lima, L.R.S., Laender, A.H.F. and Ribeiro-Neto, B.A. (1998) 'A hierarchical approach to the automatic categorization of medical documents', *Proceedings of the seventh international conference on Information and Knowledge Management (CIKM '98)*, pp.132–139, <https://doi.org/10.1145/288627.288649>.
- Devlin, J., Chang, M-W., Lee, K. and Toutanova, K. (2019) 'BERT: pre-training of deep bidirectional transformers for language understanding', *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, pp.4171–4186.
- Dong, H., Suárez-Paniagua, V., Whiteley, W. and Wu, H. (2021) 'Explainable automated coding of clinical notes using hierarchical label-wise attention networks and label embedding initialisation', *Journal of Biomedical Informatics*, Vol. 116, p.103728.
- Ethiopia MOH (2016) *Information Revolution Roadmap National Health Information Enterprise Architecture*, April, FMOH, Addis Ababa.
- Feucht, M., Wu, Z., Althammer, S. and Tresp, V. (2021) 'Description-based label attention classifier for explainable ICD-9 classification', in *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, Online, Association for Computational Linguistics, pp.62–66.
- Huang, K., Garapati, S. and Rich, A.S. (2020) *An Interpretable End-to-end Fine-tuning Approach for Long Clinical Text*, ArXiv [online] <https://arxiv.org/abs/2011.06504> (accessed 25 January 2025).
- Ji, S. (2022) *A Unified Review of Deep Learning for Automated Medical Coding*, Science Direct, University College, London.
- Johnson, A., Pollard, T., Shen, L. et al. (2016) 'MIMIC-III, a freely accessible critical care database', *Scientific Data*, Vol. 3, p.160035, <https://doi.org/10.1038/sdata.2016.35>.
- Kanwal, N. (2020) *Dilated Convolution Networks for Classification of ICD-9 based Clinical Summaries*, March, Thesis for: Master of Science, University of Stavanger Stavanger, Norway, DOI: 10.13140/RG.2.2.33003.26405.
- Lewis, C.C., Boyd, M.R. and Walsh-Bailey, C. (2020) 'A systematic review of empirical studies examining mechanisms of implementation in health', *Implementation Science*, Vol. 15, No. 21, <https://doi.org/10.1186/s13012-020-00983-3>.
- Li, F. and Yu, H. (2020) 'ICD coding from clinical text using multi-filter residual convolutional neural network', *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, No. 5, pp.8180–8187, <https://doi.org/10.1609/aaai.v34i05.6331>.
- Liu, Z. and Xu, Y. (2021) *THG: Transformer with Hyperbolic Geometry*, ArXiv [online] <https://arxiv.org/abs/2106.07350> (accessed 27 January 2025).
- Moons, E., Khanna, A., Akkasi, A. and Moens, M-F. (2020) 'A comparison of deep learning methods for ICD coding of clinical records', *Applied Sciences*, Vol. 10, No. 15, p.5262, <https://doi.org/10.3390/app10155262>.
- Mullenbach, J., Wiegrefe, S., Duke, J., Sun, J. and Eisenstein, J. (2018) 'Explainable prediction of medical codes from clinical text', *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Association for Computational Linguistics, pp.1101–1111.
- Newman-Griffis, D. and Fosler-Lussier, E. (2021) 'Automated coding of under-studied medical concept domains: linking physical activity reports to the International Classification of Functioning, Disability, and Health', *Frontiers in Digital Health*, Vol. 3, p.620828, <https://doi.org/10.3389/fdgth.2021.620828>.
- Nickel, M. and Kiela, D. (2018) *Learning Continuous Hierarchies in the Lorentz Model of Hyperbolic Geometry*, arXiv [online] <https://arxiv.org/abs/1806.03417> (accessed 25 January 2025).
- Pascual, D., Luck, S. and Wattenhofer, R. (2021) 'Towards BERT-based automatic ICD coding: limitations and opportunities', *Proceedings of the 20th Workshop on Biomedical Language Processing*, Association for Computational Linguistics, pp.54–63.
- Peng, W., Varanka, T., Mostafa, A., Shi, H. and Zhao, G. (2021) *Hyperbolic Deep Neural Networks: A Survey*, ArXiv [online] <https://arxiv.org/abs/2101.04562> (accessed 25 January 2025).
- Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N. and Huang, X. (2020) *Pre-trained Models for Natural Language Processing: A Survey*, ArXiv [online] <https://doi.org/10.1007/s11431-020-1647-3> (accessed 25 January 2025).
- Remmer, H.D. (2021) *Automatic Diagnosis Code Assignment with KB_BERT*, Stockholm University, Department of Computer Science, Sweden.
- Rothman, D. (2021) *Transformers for Natural Language Processing*, Packt Publishing, Birmingham UK.
- Sadoughi, N., Finley, G.P., Fone, J., Murali, V., Korenevski, M., Baryshnikov, S., Axtmann, N. and Miller, M. (2018) *Medical Code Prediction with Multi-View Convolution and Description-Regularized Label-Dependent Attention*, ArXiv. <https://arxiv.org/abs/1811.01468>.
- Shang, J., Ma, T., Xiao, C. and Sun, J. (2019) *Pre-training of Graph Augmented Transformers for Medication Recommendation*, ArXiv, <https://arxiv.org/abs/1906.00346>.
- Si, Y. and Roberts, K. (2021) *Three-level Hierarchical Transformer Networks for Long-sequence and Multiple Clinical Documents Classification*, ArXiv. <https://arxiv.org/abs/2104.08444>.
- Tsai, S.C., Huang, C.W. and Chen, Y.N. (2021) 'Modeling diagnostic label correlation for automatic ICD coding', *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, pp.4043–4052.
- Vu, T., Nguyen, D.Q. and Nguyen, A. (2020) *A Label Attention Model for ICD Coding from Clinical Text*, arXiv preprint arXiv:2007.06351.
- Xie, X., Xiong, Y., Yu, P.S. and Zhu, Y. (2019) 'EHR coding with multi-scale feature attention and structured knowledge graph propagation', *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM '19)*, Association for Computing Machinery, pp.649–658, <https://doi.org/10.1145/3357384.3357897>.

- Yan, C., Fu, X., Liu, X., Zhang, Y., Gao, Y., Wu, J. and Li, Q. (2022) 'A survey of automated International Classification of Diseases coding: development, challenges, and applications', *Intelligent Medicine*, Vol. 2, No. 3, pp.161–173, <https://doi.org/10.1016/j.imed.2022.03.003>.
- Zhang, Z., Liu, J. and Razavian, N. (2020) 'BERT-XML: large-scale automated ICD coding using BERT pretraining', *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, Association for Computational Linguistics, pp.24–34.