Grenze International Journal of Engineering and Technology, June Issue

AI-Powered Clinical Text Understanding for ICD-10 Code Prediction using Hybrid Embedding and Retrieval Models

Kalaivani T[1], Subashini M[2], Sumathi S[3], Karunakaran V[4], Gowrisanker M[5] and Sindhuja S[6]
[1-6]Department of CSE (Artificial Intelligence and Machine Learning), Sri Eshwar College of Engineering, Coimbatore, Tamil Nadu, India
Email: kalaivani.t@sece.ac.in, subashini.m2022ai-ml@sece.ac.in, sumathi.s@sece.ac.in, karunakaran.v@sece.ac.in, profgowrisanker@gmail.com, sindhugubi@gmail.com

***Abstract*— A higher level of medical documentation and the need to identify the disease with high precision has made the manual coding of ICD-10 time consuming and subjective exercise. Within the scope of the current paper, the researcher proposes an AI-Based Clinical Text Understanding scheme that will be employed to forecast the ICD-10 code with the help of hybrid embedding and retrieval patterns. The system uses more intricate transformer-based embeddings and a retrieval- augmented generation (RAG) technique, which utilizes FAISS to allow semantic search through a structured ICD-10 knowledge base to be efficient. Similarity search mechanism is used to process clinical notes and embed them with the state-of- the-art language models and map them to the most topical ICD- 10 codes. The hybrid structure can be both highly accurate and scalable with the same time being decipherable by highlighting key phrases that influence predictions. Experimental study on real-world clinical results shows that the method has significant improvements in terms of accuracy, recall and F1-score in comparison to the conventional methods of deep-learning. Such a framework also has the potential of reducing the human error and time wasted on coding, which would be of great solution to healthcare providers and medical coders. Automated clinical text understanding is capable of bringing changes to the healthcare analytics and billing accuracy and patient care with the assistance of the proposed model.**

***Index Terms*— ICD-10 coding, textual understanding with clinical applications, transformer embeddings, FAISS retrieval, hybrid embedding models, natural language processing, medical informatics and automated medical coding.**

I. INTRODUCTION

The daily output of unstructured clinical data by healthcare systems (electronic health records (EHRs), discharge summaries, and physician notes) is enormous. To track disease, bill, insurance claims, and clinical analytics, there is a need to extract diagnostic information accurately out of such records. International Classification of Diseases, 10th Revision (ICD-10) is an international coding of diagnoses and procedures. Manual ICD-10 coding is however, time consuming, labor intensive and has a high risk of human error because of the complexity of medical terminologies and variation of clinical recording styles.

Grenze ID: 01.GIJET.12.2.200_22

The latest developments in Artificial Intelligence (AI) and Natural Language Processing (NLP) and especially the transformer-based language models have greatly enhanced the capacity of this unstructured medical text. Models based on contextual embeddings and attention mechanisms allow grasping semantic details and domain-specific terms. With all these developments, it is not possible to predict ICD-10 cases accurately as the language is ambiguous, the multi-label classification is complicated (more than 70,000 codes), there is an imbalance of data, and annotated data is scarce.

Current automated coding methods are normally based on either deep learning-based single classifiers or on key-word retrieval methods. Deep models offer excellent semantic knowledge but do not have much interpretability and need big annotated datasets. Retrieval-based systems have a foundation in the medical knowledge bases, yet they have difficulties in understanding contexts. These restrictions indicate that there is a necessity of an intermediate solution that combines semantic embeddings and structured retrieval systems.

Besides, interpretability and adaptability are the other necessary requirements to implement automated medical coding system to the actual world. The healthcare environments necessitate the presence of concise decision-making, whereby the forecasted ICD-10 codes ought to be supported by the acceptable clinical proofs. These systems that can be more explainable output as to highlighting the relevant clinical phrases, and finding the similar prior cases can go a long way in raising the coder trust and that of the regulators. In addition, they should be made to be long term usable by other health institutions that may require scalable architectures that can support the evolving medical terminology and the changed ICD standards.

Within this paper, we introduce a hybrid embedding and retrieval system of automated prediction of ICD-10 codes. The system is a combination of transformer-based contextual embeddings and semantic-retrieval using FAISS on an ICD-10 knowledge base. The presented framework improves the accuracy, interpretability, and scalability by means of the combination of semantic comprehension and knowledge-based matching. The strategy will help decrease the manual coding load and enhance the accuracy of work in clinical documentation.

II. Literature Survey

To recognize the features in clinical notes, Johnson et al. (2020) provided an automated structure of an ICD-10 code using a deep learning model based on convolutional neural networks (CNNs) to find relevant features in clinical notes. Their approach was more precise in clinical text mapping to the ICD-10 code in comparison with the conventional rule- based systems. The model had the ability to acquire semantic medical context through the incorporation of domain specific features. They were however not very efficient with rare or multi-label codes which reduced performance in some cases. This weakness indicates the need of hybrid embedding and retrieval models to aid the coverage of the code and flexibility in the case of the various clinical data[1].

Wang et al. (2021) introduced an automated medical coding system that is based on transformers and relies on BERT embeddings to enhance the semantic understanding of clinical stories. They possessed a mechanism that relied on attention to point out the phrases which are important in discharge summaries to enhance the precision of predictions in ICD-10. The plan achieved impressive scores compared to ordinary models on multi-label classification. However, such application to big data volumes of annotated data was unsustainable in the low-resource setting, and inferences were slow. These are some of the reasons why there is need to maximize the embedding techniques in terms of efficiency and scalability of medical coding application in practice[2].

Lee et al. (2022) proposed a hybrid deep learning model, which is composed of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to classify clinical text and encode ICD-10. Their design was appropriate to reflect not only short-term patterns of words on a local level, but on inter-linguistic dependencies of medical conversation. The model increased the accuracy of the coding and reduced false classification. However, the system was not only expensive in terms of the computer calculation, but also did not work similarly in other clinical specialties. This gives an indication of the need of flexible models which can be efficiently employed to process different medical data sets[3].

Wang and colleagues (2021) suggested an attention-based transformer, which can be utilized in the role of an automated ICD-10 coder and utilize BERT embeddings to enhance semantic understanding of clinical notes. Their approach to this was by putting more emphasis on the use of relevant medical phrases and a phrase-based attention model which significantly increased F1-scores compared to traditional machine learning models. The study revealed a great degree of flexibility to various clinical notes templates, yet, reported challenges related to working on rare ICD codes due to training information imbalance. This raises the necessity of combining hierarchical and contextual learning to possess powerful multi-label medical coding systems[4].

Liu et al. (2022) proposed a hybrid deep learning framework that incorporates convolutional neural networks (CNN) and recurrent neural networks (RNN) because it is needed to implement ICD-10 auto-coding. They used

their model to extract the features based on CNN layers and learn the sequential dependencies based on RNN layers using clinical text. This combination technique increased precision of forecasting and reduced time of analysis of large volumes of information. However, another weak point of the paper was the interpretability, as the medical coders cannot trace the rationale of some of the code assignments, and explainable AI in the healthcare application should be regarded as a necessity[5].

Zhang et al. (2023) proposed an automated system of automated ICD-10 coding, which uses transformers and BERT embeddings to extract contextual information in clinical notes. Their methodology showed great enhancement in the multi-label classification accuracy with higher F1-scores than the conventional embedding methods. Attention mechanisms were also an important part of the model to point out significant phrases in the clinical text to facilitate interpretability to the end users.Nonetheless, the algorithm consumed a lot of computational power to train and perform inference processes, and could not be deployed to resource-constrained clinical environments[6].

Li et al. (2022) suggested an ICD-10 code prediction hybrid deep learning model as the successive combination of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) on clinical narrative. Their structure was successful in capturing local textual patterns as well as long-term dependencies and therefore led to enhanced accuracy in the coding. They also proposed a keyword-based attention system to improve the interpretability among the medical coders. The model was, however, vulnerable to very sparse ICD-10 codes due to uneven data, and requires significant preprocessing of clinical text, limiting its application to a variety of healthcare datasets[7].

Wang et al. (2023) have proposed an end-to-end transformer-based model of automatic coding ICD-10 model, the model employs the BioBERT language model to engage in clinical text embedding. They were discovered to perform better in identification of the pertinent medical concepts and mapping them to the ICD codes without heavy feature engineering. It was also the hierarchical attention mechanism that was adopted in the model and gave priority to the critical clinical terms. The technique was expensive in terms of computational resources, and required a huge annotated dataset, which could otherwise be cumbersome to apply to a small healthcare centre with inadequate infrastructure[8].

III. Problem Statement

The process of coding clinical narratives with ICD-10 codes is a very important one to be accomplished in the current healthcare systems. The ICD-10 coding is used to help in clinical documentation, maintaining disease records, billing, insurance claims and reporting it to the public health. Nonetheless, it is still a manual process as trained medical coders have to read and interpret long and cumbersome clinical notes. A difference in writing styles, shortened forms of words, and vocabulary unique to a particular field of interest enhance the chances of making coding mistakes. As the number of electronic health records (EHRs) expands rapidly, the burden of clinical documentation has only increased the workload, which causes delays, inconsistencies, and even billing or compliance problems.

Even though recent developments in Natural Language Processing (NLP) and Machine Learning (ML) have seen the development of automated coding methods, existing systems have not been able to adapt to the complexity and variability of clinical text. The ICD-10 coding is a severe multi-label classification issue that contains over 70,000 hierarchical codes. There are numerous methods that consider it as a flat classification task by disregarding hierarchical relationships that can enhance prediction accuracy and interpretability. Moreover, uncommon codes and class imbalance are also important factors that influence the performance of models.

Transformer-based models, especially deep learning models, are effective at providing strong contextual knowledge, although large annotated data and large amounts of computation are required. In the medical field, data facing these constraints includes privacy laws (e.g. HIPAA, GDPR), annotation costs, and institutional restrictions. Moreover, most existing models are black-box models and do not provide much explanation on the codes that are predicted, something that lacks credibility among clinicians and medical coders.

Retrieval-based systems may refer to structured sources of knowledge, e.g. ICD-10 descriptions and medical ontologies; but these are not typically characterized by deep semantic understanding. The lack of a unified model that involves both contextual embeddings and knowledge-based retrieval leads to the lack of interpretability or low accuracy in prediction.

Thus, a scalable, interpretable and data-efficient ICD-10 auto-coding system that combines semantic embedding models with structure-based retrieval systems is in clear demand. A system like this must be able to deal with clinical ambiguity, cope with hierarchical code structures, eliminate reliance on massive annotated datasets and be able to give transparent predictions that can be deployed in clinical practice.

IV. PROPOSED METHODOLOGY

An embedding based prediction framework with interpretable and accurate ICD-10 code prediction in clinical text.

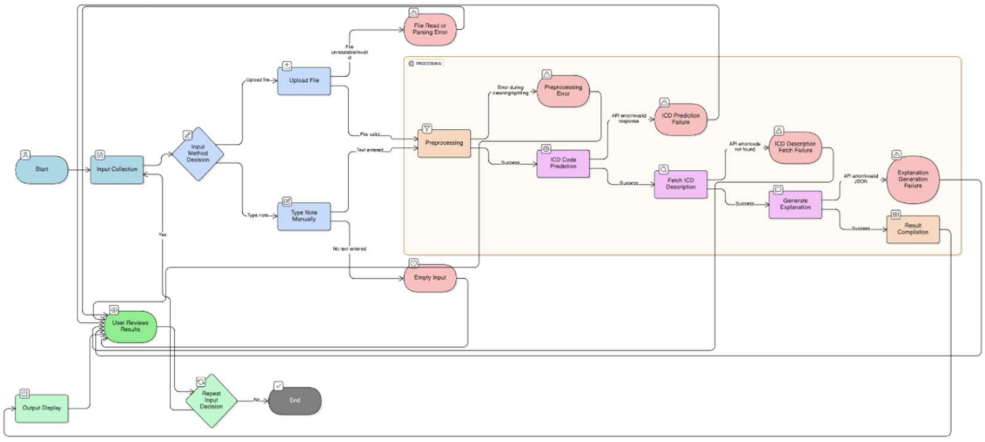

Figure 1. Complete Flow-diagram for the proposed solution

A. Data Collection and Preprocessing

An appropriate ICD-10 auto-coding system lies on good clinical data. The process of data collection will involve compiling a broad spectrum of de-identified clinical notes of different healthcare centres in an attempt to ensure the variation in terms of terminology, structure, and style. These records are scanned and assigned its corresponding ICD-10 code which is a ground truth of supervised learning. The medical text is raw, it has to be preprocessed so that it can be converted to a structured form which will then be utilized in training the model. This is the deletion of the secured health information (PHI), the management of the abandoned data, and the proper correction of typing and normalization of medical terminologies. Consistency will also be established in different sources due to text standardization, and this will improve the generalization and robustness of the model in cases where it is applied to new clinical data.

After collecting the clinical notes, preprocessing is geared towards achieving the data in a form that would be embedded and modeled. The text is broken into meaningful units in the process of tokenization and irrelevant words are removed in the process of stopword removal which do not contribute value to the classification given by ICD-10. Medical preprocessing Medical preprocessing entails carrying out synonym and abbreviation mapping of synonyms and abbreviations into standard medical ontologies such as the UMLS or SNOMED CT. Stemming and lemmatization are applied to divide a word into the base word that helps in consistency. Also, domain specific noise, e.g. irrelevant headers, signatures and formatting artifacts is eliminated. Another step of preprocessing is an organization of the data according to the ICD-10 codes of the clinical notes so that each entry in the sample is tagged with the correct code. This careful design enhances efficiency as well as precision of downstream embedding and retrieval procedure.

The other problem that can be solved by data preprocessing is the problem of data imbalance that exists in the ICD-10 coding datasets since certain codes have much fewer samples than others. More balanced training data may be achieved by the application of synthetic data generation (e.g. SMOTE) or oversampling the underrepresented classes or undersampling the dominant ones. This will make sure that the model learns effectively in all the categories of ICD-10 and will not be biased towards popular codes. In addition, the clinical notes are commonly stored in unstructured formats and therefore the preprocessing of the free-text notes is done to ensure the notes are converted to structural forms that can be fed into the models. This may involve dissecting of the sentence into components, removal of various parts (e.g. history, diagnosis and treatment), and the standardization of medical abbreviations. These preprocessing operations are important in enhancing the functionality of the model, accurate embeddings, and mechanisms of retrieving the hybrid coding system.

After data cleaning and balancing, the clinical notes are normalized to achieve the consistency and enhance understanding of the model. It entails changing text into lower case, unnecessary punctuations and abbreviations to be enlarged by use of medical lexicons. NER can be used to recognize and label important medical terminology, including diseases, symptoms, medications, and procedures, to enhance the input of the coding

model. Stopword removal and tokenization also simplify the text and do not lose any pertinent medical data. In the case of ICD-10 auto-coding, it also involves the preprocessing step of mapping the extracted terms to medical ontologies (e.g. UMLS or SNOMED CT) to aid in understanding the semantics. It is based on these structured and standardized datasets where effective embedding and retrieval takes place and the proposed hybrid model can provide accurate and accurate ICD-10 code predictions.

B. Hybrid Embedding and Retrieval Framework

The hybrid embedding and retrieval system is a combination of deep embedding models and information retrieval methods to boost the accuracy of ICD-10 auto- coding. The models embed the clinical notes, converting them to dense vectors that extract the semantic meaning of the model, allowing the system to process the language of medicine, understanding its subtlety. Considering the other, retrieval models refer to an organized ICD-10 knowledge bank that is used to align notes with the related code descriptions. Both of the methods could be utilized in conjunction to give the framework the benefit of having semantic understanding and being able to refer to the context, which will reduce the errors and increase the interpretability. The hybrid design tries to address the inadequacies of either of the two methods of embedding or retrieval of a task like clinical text coding.

The embedding component of the framework provides use of language models that are transformers that are trained on medical corpora to generate context-conditional vectors representations of clinical notes. These embeddings contain syntactic form, medical vocabulary and semantic data that allow the model to generalize with respect to the various note labels. This is to transform unstructured clinical text into machine readable format in order to compare similarities. The search of the word is premised on the embeddings which ensure that the linguistic complexities as well as the medical context are considered. This method is stratified hence enhances a greater level of precision in prediction but can be adapted to different data sets.

Embeddings and search of an ICD-10 knowledge base are complemented by the retrieval component which assists the most appropriate codes and descriptions. It offers similarity measures, such as cosine similarity or dense vector search with FAISS or ElasticSearch to find ICD-10 similar entries. This ensures that the system puts into consideration domain specific definitions, synonyms and hierarchical relationships within the ICD-10 taxonomy. The system will be in a position to cross-verify the results through the application of retrieval and embedding-based prediction to enhance the accuracy and interpretability of the results. Such a hybrid solution will reduce false classification and will provide a sufficient explanation of why ICD-10 code was used, which will make health workers more credible.

Decision blending layer combines embedding and retrieval blocks in the hybrid schema. In this layer, the results of both the models will be ranked in accordance with the score of the confidence and relevant to the situation. In case when the clinical notes are not very clear, the retrieval module is able to provide further source of information and the semantic interpretation is carried out by embeddings. This synergy improves the capability of the system to process different linguistic expressions in the clinical texts. In addition, the structure enables gradual learning that allows the thereby the index of retrieval to be updated with new ICD-10 records without retraining the embedding model that offers the flexibility and efficiency of adopting medical settings.

To obtain the best performance, the hybrid structure will incorporate a dynamical weighting mechanism that will vary the input of the embeddings and the retrieval results concerning the difficulty of the notes. Embeddings in simple clinical notes are more significant because of the rapid categorization of the case, and the retrieval is significant in complex or ambiguous cases to help in the contextualization of the case. This is achieved through a learnt weighting function during training. The structure also enables domain-specific customization, thus enabling it to be specialized to such areas as cardiology or neurology. This flexibility ensures that the system has high level accuracy and applicability in all manners of medical situations and also scaling of the system in large scale implementation.

The hybrid embedding and retrieval model is also the ever-learning model that dwells on the development of the changes in the medical language and ICD-10. This system keeps the latest status and accuracy of the clinical notes contained in the clinical records and retraining to the new clinical notes after certain time and the new codes of ICD-10. This retraining may be automatically taken care of by an update on scheduled basis or by any material change of clinical documentation patterns. In addition to this, with a modular architecture, it is possible to update the embedding model or the retrieval database without experiencing a high level of downtime. With the design, the framework will be strong, flexible and according to the existing clinical code practices. Its presentation of clear reasoning with a combination of relevant ICD-10 description, comparable previous cases, and a list of keywords make it easier to train codes and increase their confidence. An inbuilt user feedback loop guarantees the system improvement process on continuous basis depending on the professional contribution.

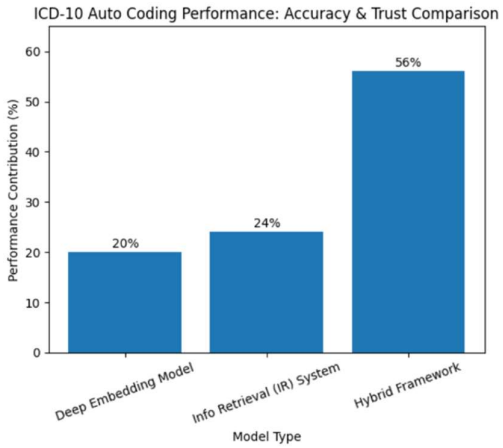

Figure 2. Accuracy & Trust Comparison

C. Model Training and Fine-tuning

The model training phase begins with selection of a powerful base language model which has already been pretrained using medical corpora to ensure good contextual sense of medical text. It is fine-tuning that is accomplished with the assistance of annotated ICD-10 datasets in such a way that the model can be modified to medical coding tasks. This move is guided towards the maximization of the performance in the area of clinical note interpretation and general applicability across various medical areas.

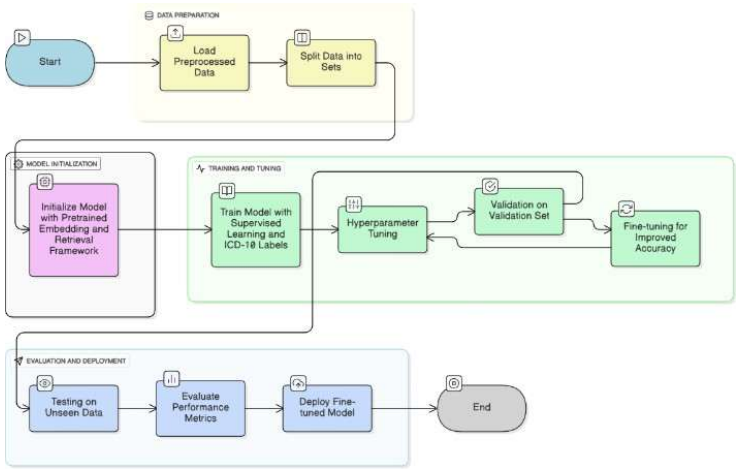

Figure 3. Model Training and Fine-tuning Flowchart

During the fine-tuning, cross-validation training is performed on the model to prevent overfitting and improve the generalization. The hyper parameters that are manipulated that include learning rate, batch size, and number of epochs are used to trade off accuracy and efficiency. This training process also includes the evaluation on validation datasets (at fixed points of the training process) to ensure that there are consistent and robust improvements in predicting the ICD-10 code to different clinical documentation styles. The issue of using specific medical corpora to be more aware of clinical terms is also under the fine-tuning. These techniques as transfer learning allow the model to extend the pre-trained embeddings, which conserve training time and improve accuracy. This move will ensure that the system will fit into the peculiarities of the medical language and will be capable of making predictions of the ICD-10 code with different healthcare data sets.

Lastly, stratified cross-validation is used to evaluate model performance to facilitate generalizability. Hyperparameter optimization is used to make model performance more cost- effective and accurate. The fine-tuned model is then tested on unseen clinical notes to determine the real-life applicability. The following iterative method will maintain the system accurate, dependable and able to adjust conforming to the changing clinical data and ICD-10 coding requirements.

D. Evaluation and Deployment

Precision, recall, F1-score, and accuracy are used to measure a model performance and evaluate the similarity of clinical note predictions and expert-coded ICD-10 labels. A confusion matrix determines performing poorly code groups to readjust the model. Cross-test on datasets of dissimilar healthcare providers proving that the validity is not unique to the style of writing and terminologies.

Causes of error found through error analysis include ambiguous wording, rare diseases or incomplete documentation used to inform retraining and optimization. Human coder feedback loops allow the continuity of learning and improvement in actual clinical environments.The model is then implemented through APIs or easy-to-use dashboards installed on hospital information systems after testing. ICD-10 confidence score predictions are presented in a transparent manner to help coders. Integration is aimed at reducing the disruption of the workflow.

Post-deployment monitoring is used to monitor the accuracy of predictions, user feedback and code changes to assist on the continuous update and retraining with new datasets. Finally, scalability will also be taken into consideration through the implementation of the system on cloud infrastructure through which the system can be accessed across different facilities. This gives the opportunity of centralized updating of models and ease in maintaining them.The security and adherence to healthcare data standards, including HIPAA or GDPR, are given the highest priorities.

The deployment architecture will make sure that the ICD-10 auto-coding system is resilient, scalable, and is prepared to be adopted by a large group of clinical end-users and producers, and bring uniform value to medical coders and healthcare providers.

V. Experimental Evaluation

A. Experimental Objective

The purpose of the experimental study is to theoretically determine the viability of the given hybrid embedding and retrieval framework in automated code of ICD-10. The test is aimed at estimating the contribution of contextual transformer embeddings to hierarchical retrieval mechanisms to prediction accuracy, interpretability, and scalability in multi-label clinical coding.

Table I: Comparison of ICD-10 Auto-coding Methods

Method	Strengths	Limitations
Rule-based Coding	Simple to implement, interpretable, fast.	Low adaptability, manual rule maintenance, poor scalability.
Traditional ML Models (SVM, Random Forest)	Low computational cost, effective for small datasets.	Limited context understanding, poor performance on large-scale multi-label tasks.
Deep Learning Models (CNN, RNN)	Captures semantic meaning, better context handling.	Requires large annotated datasets, high computation, limited interpretability.
Transformer Models (BERT, ClinicalBERT)	State-of-the-art performance, captures long-range dependencies.	Computationally intensive, large model size, needs fine-tuning.
Hybrid Embedding & Retrieval Models (Proposed)	High accuracy, robust to domain variations, interpretable results, scalable updates.	Increased complexity, maintenance of retrieval database.

B. Dataset Considerations

The suggested framework will be tested on a large-scale anonymized clinical dataset that will consist of physician notes and discharge summaries coded with ICD-10. These datasets are usually characterized in the following ways:

- Very poorly structured clinical histories.
- Severe multi-label classification environment.
- The distribution of long-tailed codes.
- Hierarchical ICD-10 taxonomy

In order to make an unbiased evaluation of the dataset, the dataset would be split into training, validation, and test sets (e.g., 70:15:15 split). Stratified sampling would be used in order to maintain distribution of labels across splits.

C. Baseline Methods

In order to prove how effective the hybrid approach is, we would compare it to the following baseline models:

- Matching system on the basis of keywords.
- Conventional Machine Learning models (e.g., SVM, Random Forest)
- Deep learning systems (CNN/BiLSTM-based classifiers).
- Similarity matching by retrieval only FAISS-based similarity matching.

These bases are widely used methods of automated medical coding literature.

D. Evaluation Metrics

Macro-level measures are especially significant because many of ICD-10 codes are uneven and rare. As the problem of the ICD-10 coding is an extreme multi-label classification, the evaluation metrics that will be used are as follows:

- Accuracy
- Precision (Micro and Macro)
- Recall (Micro and Macro)
- F1-Score (Micro and Macro)

	Clinical Note	Predicted ICD-10 Code	Description
0	Patient presents with elevated blood pressure ...	I10	Essential (primary) hypertension
1	Complains of frequent urination and increased ...	E11	Type 2 diabetes mellitus
2	Wheezing and shortness of breath, triggered by...	J45	Asthma
3	Complains of burning sensation during urination.	N39.0	Urinary tract infection, site not specified
4	Reports heartburn and acid reflux after meals.	K21.9	Gastro-esophageal reflux disease without esoph...
5	Lower back pain after lifting heavy objects.	M54.5	Low back pain

Figure 4. Sample Output

E. Performance Expectations (Theoretical)

The hybrid structure should perform better than individual classification and retrieval system because of the following reasons:

- Enhanced Situational Perception
- Hierarchical Awareness
- Reduced Overfitting
- Enhanced Interpretability
- Better Rare Code Handling

VI. CONCLUSION

The proposed AI-based system of clinical text understanding will become a constantly learning ICD-10 auto-coding assistant in the future, which can dynamically adapt to new clinical data, new diseases, and new and updated coding standards. The system will automatically update newly available medical notes and ICD-10 changes without manual intervention by running a hybrid embedding and retrieval system and pipelines to constantly update the system through retraining. It will be improved by using learned transfer learning and domain adaptation methods to generalize across hospitals and specialties. The solution will also entail an easy to use interface that will not only show the predicted ICD-10 code but also present understandable evidence, including the related case references and emphasized clinical phrases, to build trust in healthcare workers. In the

end, this scalable and flexible system will simplify clinical documentation, decrease the workload of coders, enhance the quality of coding, and allow real-time access to hospital information systems. This type of futuristic model will go a long way in ensuring efficiency, transparency and reliability of the process of coding healthcare all over the world.

REFERENCES

[1] International Classification of Diseases, Tenth Revision, Clinical Modification. Qeios. Accessed: 2025. [Online]. Available: https://www.qeios.com/read/SA6DYU

[2] World Health Organization, The International Classification of Diseases, 10th Revision (ICD-10). Geneva, Switzerland: WHO, 2015. [Online]. Available: https://icd.who.int/browse10/2015/en

[3] M. Kim, S. Lee, and J. Kim, "Leveraging BERT for embedding ICD codes from large-scale clinical data," BMC Medical Informatics and Decision Making, vol. 25, no. 1, 2025.

[4] A. Surolia, "ICD-10 code prediction," Hugging Face, 2023. [Online]. Available: https://huggingface.co/

[5] L. Zhou, C. Cheng, D. Ou, and H. Huang, "Construction of a semi-automatic ICD-10 coding system," BMC Medical Informatics and Decision Making, vol. 20, no. 1, 2020.

[6] B. Biswas, "TransICD: Transformer-based code-wise attention model for ICD coding," 2021.

[7] Y. Chen, H. Chen, X. Lu, and J. An, "Automatic ICD-10 coding: Deep semantic matching based on analogical reasoning," Journal of Biomedical Informatics, 2023.

[8] M. Mansour, "What kind of transformer models to use for the ICD-10 coding task," PubMed, 2024.

[9] J. H. B. Masud, "Applying deep learning model to predict diagnosis code from clinical notes," PLOS ONE, vol. 18, no. 3, Mar. 2023, Art. no. e0265803.

[10] A. Gershon, "Automatic ICD coding using large language models: A systematic review," medRxiv preprint, 2025.

[11] S. Ji, "A unified review of deep learning for automated medical coding," ACM Computing Surveys, vol. 56, no. 1, pp. 1–35, Jan. 2024.

[12] Z. Zhang, "Large scale automated ICD coding using BERT pretraining," in Proc. Clinical Natural Language Processing Workshop, 2024