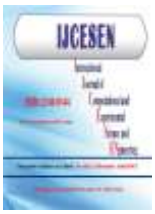
International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 12-No.1 (2026) pp. 1580-1586
http://www.ijcesen.com



ISSN: 2149-9144

Research Article

NLP-Based Predictive Analytics Framework for Early Detection of Medical Coding Errors Using Transformer-Based Clinical Text Analysis

Triveni Kolla*

Independent Researcher, USA
* **Corresponding Author Email:** triveniacheiver@gmail.com- **ORCID:** 0000-0002-5247-7899

Article Info:

DOI: 10.22399/ijcesen.5096
Received : 05 Feruary 2026
Revised : 25 March 2026
Accepted : 27 March 2026

Keywords

Natural Language Processing,
Medical Coding Accuracy,
Transformer-Based Embeddings,
Predictive Analytics,
Clinical Documentation

Abstract:

Accurate medical coding is essential for healthcare reimbursement, regulatory compliance, and accreditation, serving as the foundation for quality clinical documentation. Current post-coding review approaches for error detection are resource-intensive and reactive, failing to alert coders about documentation discrepancies before code assignment occurs. An NLP-based predictive analytics framework has been developed to identify medical coding errors by comparing clinical narratives to assigned diagnosis and procedure codes. The framework employs specialized transformer models to extract clinical details from unstructured documentation and assess alignment between written clinical information and assigned codes. Transformer-based architectures enable semantic understanding of clinical text and generation of code representations for both ICD-10-CM and CPT nomenclature. Structured features—including encounter type, provider specialty, and documentation length—complement neural embeddings for comprehensive encounter characterization. Supervised machine learning classifiers predict coding error risk across diverse medical specialties. Cost-sensitive learning approaches address the inherent class imbalance in medical coding datasets, prioritizing minority class (error) detection. The framework demonstrates substantial performance improvements compared to rule-based validation systems across multiple clinical domains. NLP-based predictive systems for coding error detection offer healthcare organizations the opportunity to shift from retrospective audit models to proactive risk identification, enabling more efficient resource allocation, enhanced compliance outcomes, and improved revenue cycle accuracy.

1. Introduction and Healthcare Coding Challenges

Medical coding is the process of transforming clinical services into standardized codes and serves as the foundation for submitting claims for reimbursement, fulfilling regulatory requirements, and analyzing population health data. Coding accuracy affects the financial viability of health systems, compliance with federal regulatory requirements, and the quality of data used to measure the quality of care provided and for research. Despite the importance of accurate coding to the healthcare enterprise, healthcare organizations continue to encounter difficulties in sustaining coding accuracy for high-volume clinical documentation workflows [1].

The integration of artificial intelligence in medical coding has emerged as a transformative approach to address these systemic challenges. AI-based systems can analyze clinical narratives at scale and flag potential coding discrepancies before submission, enabling proactive rather than reactive validation. Healthcare organizations are increasingly evaluating AI solutions to enhance revenue cycle management and reduce the manual burden of retrospective audits [1].

Recent studies continue to demonstrate variability in ICD-10-CM code accuracy across clinical conditions and healthcare settings. The concordance between coded diagnoses and clinical documentation remains inconsistent, with accuracy varying significantly depending on the type of diagnosis, the clinical specialty, and the healthcare provider's documentation practices [2].

Coding errors in the ICD-10-CM classification system can be categorized as undercoding, overcoding, and miscoding. Undercoding occurs when clinical documentation supports a more precise code, but a less specific code is assigned,

failing to reflect the complexity and resources used on the patient. Overcoding, conversely, can occur when insufficient clinical information exists to substantiate the assigned code, or when more specific codes are used when less specific codes are clinically supported [2]. Given the above multidimensional challenge, it is clear that technology-based approaches to targeted auditing of higher-risk documentation and more clever decision support for coders are preferable to retroactive error correction.

2. Clinical Documentation and Coding Alignment Challenges

All medical coding quality efforts are limited by the intrinsic difficulty of converting unstructured text from clinical documentation into discrete, standardized code sets without losing the semantic meaning and clinical context preserved in provider documentation. Clinical documentation may differ considerably from provider to provider, from specialty to specialty, and from one type of encounter to another. This variability makes it difficult for coding professionals to find clinical information and create the correct derive ICD-10-CM diagnoses and CPT procedural codes from narrative text, adhering to complex coding rules and conventions.

Undercoding, overcoding, and miscoding are forms of documentation-coding misalignment. Undercoding occurs when clinical documentation supports a more precise code, but the code does not represent that level of specificity, so it fails to reflect the complexity and resources used on the patient. Data harmonization across healthcare systems remains a critical challenge in standardizing clinical coding practices. The Common Data Model (CDM) and similar data harmonization frameworks aim to map diverse clinical documentation formats and coding conventions to unified standards, facilitating consistent interpretation of clinical information across databases [3].Implementing a CDM architecture requires systematic processes to transform heterogeneous clinical data into standardized representations. This includes mapping codes, normalizing clinical concepts, and aligning documentation practices across different healthcare providers and systems. Such harmonization efforts are essential for establishing reliable coding practices and enabling meaningful clinical data exchange [3].Another related coding error is overcoding, which can occur when there is not enough clinical information in the medical record to substantiate the code assignment or when more specific codes are used when less specific codes are clinically supported. Errors caused by semantic complexity and context-dependent meanings are common in clinical text. Automated extraction of relationships among findings had 18-34% agreement for synonyms indicating a relationship. Hierarchical relationships showed only 4–9% agreement among findings, such as sudden changes in mental status, unexplained medical conditions, and echocardiogram results [4]. These may be due to coding professionals misunderstanding general clinical terms or mistakenly coding ruled-out diagnoses as confirmed diagnoses.

Care visits encounter documentation gaps when the provided documentation lacks sufficient detail for the clinical coder to assign the correct codes. Research using automated methods to extract medical terms has revealed significant differences in clinical synonyms and abbreviations, and experts do not always agree on the importance of certain medical terms across different fields (Fleiss' kappa 0.29-0.38). For instance, the suggested use of NLP to find synonyms in clinical text received approval ratings ranging from 25% to 43%, while extracting hierarchical relationships was rated between 31% and 67%, depending on the specific text used. A further problem is semantic ambiguity. Howard et al. note that the semantics of clinical text depend on the exact context, which poses a challenge for rule-based text processing systems.

3. NLP and Transformer-Based Approaches for Clinical Text Analysis

In the past 10 years, NLP methods have moved from rule-based systems and bag-of-words representations to more complex neural network architectures capable of more complex semantic and contextual analysis of clinical text. The application of NLP to clinical notes and other unstructured text has shown promise as a means of extracting structured information, identifying clinical concepts, and assessing relationships between narrative text and standardized medical vocabularies. These abilities provide the technical basis for automated coding help and prediction systems, which can review large amounts of documents and perform certain clinical tasks nearly as well as humans.Transformer architectures are used in clinical NLP to fix the problems of older RNN sequence models by speeding up the processing of long documents and enhancing the understanding of text relationships through self-attention mechanisms. BERT (Bidirectional Encoder Representations from Transformers) has been widely used in clinical NLP after further domain-specific pre-training on large clinical note

corpora. Clinical BERT models, which were trained on around 2 million anonymous clinical notes from electronic health records, have been found to perform better than general language models in tasks like extracting medical concepts, recognizing named entities, and identifying relationships in clinical data. They were reported to improve performance over general BERT and BioBERT on medical named entity recognition tasks. Clinical BERT achieved F1 scores between 86.4% and 87.8% at the i2b2 2010 concept extraction task, compared to 83.5% when using general BERT models. Clinical BERT models have achieved an F1 score of 78.9% at the i2b2 2012 temporal relation extraction task [6].The self-attention mechanism of the transformer architecture has special advantages for clinical coding tasks, because transformer architectures do not process text sequentially from left to right as earlier deep learning architectures do but can instead simultaneously process all of the relationships between all of the words in a document. This is particularly useful for clinical notes, where relevant information for a diagnosis may be located in different places: initial complaints in the chief concern, supporting findings in the physical examination, and diagnostic conclusions in the assessment.Recent advances in transformer-based architectures have demonstrated significant improvements in analyzing clinical text and medical signals. Survey research on transformers and large language models applied to medical diagnosis indicates that domain-specific transformer models, when trained on large clinical datasets, outperform general-purpose language models in clinical tasks [5].Transformer architectures employ self-attention mechanisms that process multiple relationships simultaneously across documents, rather than sequentially. This capability is particularly valuable for clinical notes, where relevant diagnostic information may be distributed across different sections: initial clinical findings in the chief complaint, supporting observations in the physical examination, and diagnostic conclusions in the assessment [5].The application of transformer-based models to clinical text analysis has shown promise in extracting structured information from unstructured clinical narratives, identifying clinical concepts, and assessing relationships between narrative text and standardized medical vocabularies. These capabilities provide the technical foundation for automated coding assistance and prediction systems that can review large volumes of documents [5].

4. Predictive Framework Architecture and Feature Engineering

An NLP-based coding error prediction system takes multiple input sources and feature extractors and builds a model of the complex relationships between clinical documentation, coding assignments, and coding performance. The pipeline consists of a series of stages, including the processing of clinical text, the use of neural embeddings and structured clinical metadata to extract features from the clinical notes, and supervised classifiers to predict specific coding errors.

Preprocessing is necessary due to unique elements of clinical notes, such as heterogeneous styles, frequent use of abbreviations, and specialized languages. Text normalization includes tasks such as standardizing headings, expanding clinical abbreviations and splitting notes into standard documentation structures. Yao et al. showed that using a model with domain knowledge, 200-dimensional pre-trained word embeddings (trained on MIMIC-III clinical notes), and UMLS CUI entity embeddings achieved an overall Macro F1 score of 0.8016 on the textual task and 0.6768 on the intuitive task, better than rule-based baseline systems ($p < 0.05$) [7].

The main technical step of the predictive model is feature extraction. The first pathway employs pre-trained clinical transformer models (kernel size of 5, 256 convolution filters, and hidden layer size of 128), which make dense vector representations of clinical notes [7]. These embeddings are useful in understanding clinical information, as they group similar medical ideas together, helping to identify connections between different terms used for the same concepts.

Another related approach is to build code-based embeddings by jointly embedding diagnosis and procedure codes into the same semantic space. Choi et al. trained medical concept embeddings using a large dataset of over 4 million patients' insurance claims from 2005 to 2013, which included ICD-9 diagnosis codes, CPT procedure codes, lab test results, and drug prescriptions. Using d=200 the r=5, MCEMC's Medical Relatedness Measure achieved scores of 0.4536 (fine-grained) and 0.4804 (coarse-grained) with the CCS hierarchy [8]. For the May-Treat relationship in NDF-RT, MCEMC's MCEMCmonth model achieved a hit rate of 19.24% in the neighbor hit rate test. For analogical reasoning, the hit rates were 37.68%/60.57% (average/max seed). The hit rates of MCEMCmonth,hs for the May-Prevent relation are 8.82%/30.20%/57.35% (average/max seed). The MCECN-SVD model achieved the highest hit rates: for the May-Treat neighbor, 52.26%, and for the May-Prevent neighbor, 39.71%. These

experiments indicate that the embeddings created from large clinical data effectively captured important medical relationships, with diagnosis codes grouping together with their related treatments in the same space.

The structured metadata features add to the neural embeddings by including various details from each patient visit that can influence how accurately codes are assigned, like the day of the week, the time of day, and the workload differences between coding shifts throughout the week.

Neural and code embeddings, along with structured metadata and population-level reference features, create a complete picture of each encounter that helps predict the chances of coding errors in different clinical situations.

5. Model Development and Training Methodology

When developing supervised machine learning models for predicting coding errors, researchers should take into account the characteristics of the training data, the model architecture, and the strategies to reduce class imbalance. They should also consider validation approaches for healthcare decision support. The modeling strategy relies on retrospective audit data to produce ground-truth labels for coding errors through expert review of encounters.

Class imbalance is a significant problem in medical coding error prediction models, as audit data typically contains substantially more correctly coded encounters than incorrectly coded ones. Without addressing this imbalance, classifiers could achieve high accuracy simply by predicting the majority (correct coding) class while failing to identify actual coding errors.

Modern approaches to class imbalance in medical classification include cost-sensitive learning methods, which assign differential penalties to misclassifications of the minority class versus the majority class. Cost-sensitive algorithms force classifiers to focus more attention on the minority class by increasing the weight assigned to minority class errors during training. This approach has demonstrated superior performance compared to traditional methods when applied to imbalanced medical datasets [9].

Comparative analysis of cost-sensitive methods shows that cost-sensitive random forest and cost-sensitive XGBoost achieve high precision, recall, and F-measure scores on imbalanced medical datasets. For example, on the cervical cancer dataset (93.59% imbalance), cost-sensitive random forest achieved precision of 1.000, recall of 1.000, and F-measure of 1.000, demonstrating the effectiveness of cost-weighted methods on highly imbalanced medical data [9].

Gradient boosted decision trees implemented using XGBoost and cost-sensitive random forest algorithms have shown state-of-the-art results on imbalanced classification problems. These ensemble methods handle the sparsity and high-dimensionality of mixed feature spaces effectively, making them suitable for coding error prediction tasks where features include both neural embeddings and structured clinical metadata [9].

Gradient increased decision trees were implemented using eXtreme Gradient Boosting (XGBoost), an efficient, distributed implementation of end-to-end tree boosting that has claimed state-of-the-art results in numerous classification problems. As Chen and Guestrin note, of the 29 solutions winning the Kaggle challenges in 2015, 17 made use of XGBoost. XGBoost produced an AUC of 0.8304 on the Higgs-1M classification benchmark test set and was over 10 times faster than scikit-learn under the same conditions [10]. The system can determine splits while being aware of the data's sparsity in a high-dimensional mixed feature space, which fits the coding error prediction task.

The training method businesses used to get a better idea of how well the model works is called stratified 10-fold cross-validation, which ensures that the classes are evenly distributed in both the test and training groups. The parameters organizations used are the default parameters from Chen and Guestrin [10]: a maximum tree depth of 8 and a shrinkage of 0.1, which were previously validated on several other datasets. Accurate probability estimates from the trained models help set thresholds and make decisions that weigh the sensitivity of detection against the costs of checking more cases.

6. Validation Results and Clinical Decision Support Integration

The assessment of NLP-based coding error prediction systems should consider various performance areas that must be met for them to be used in real-world clinical settings. Quantitatively, common classification metrics (precision, recall, F1, accuracy) can be used, along with operational metrics indicative of the impact of the coding validation process. The predictive method should also be compared against baseline methods to validate the additional value that it adds.

On retrospective clinical datasets, the models may perform better or worse depending on the department or parameter configuration. The model was trained and tested on a dataset of 21,953

clinical records from five departments, with 90% (20,173 records) used for training and 10% (1780 records) for testing. The performance varied by department, with precision ranging from 0.52 to 0.96, recall ranging from 0.85 to 0.99, and F-score ranging from 0.65 to 0.98; all departments achieved clinically acceptable performance. The best performance was achieved when using word2vec embeddings with 128 dimensions, a document length of 200, and convolutional filter windows of sizes 1, 2, 3, 4, and 5 for the cardiology domain with 0.96 precision, 0.99 recall, and a 0.98 F-score. The lowest performance was detected for the nephrology domain, with 0.52 precision and a 0.65 F-score.

Regarding the model itself, testing the parameters shows that they were relevant to the model's final performance. The model was configured with a batch size of 64 and achieved an accuracy score of 0.94, a precision score of 0.66, a recall score of 0.94, an F-score of 0.77, and a loss of 0.03 [11].

Additionally, model performance varies with feature representation and clustering quality. Al-Khamees et al. demonstrated that using a mix of cosine and cityblock distances along with a Z-score adjustment for grouped data significantly improved how well medical data was classified. For example, the Breast Cancer Wisconsin data set (569 samples, 30 attributes) was classified with an accuracy of 0.9825 and ARI of 0.9303, as compared to 0.8752 with the standard Euclidean K-Means. The homogeneity has increased from a value of 0.7721 to 0.8676 when applying a Z-score outlier reassignment with a threshold of 3.0 to eliminate 74 outlier samples [12]. The heart disease dataset, which contains 303 samples and 13 features, has an accuracy of 0.90. Its homogeneity increased from 0.4335 to 0.5352, with an ARI of 0.6334. The running times are 0.0025 seconds and 0.0140 seconds, respectively [12].

This suggests that careful feature engineering and cluster refinement can lead to predictive models with performance far exceeding baseline classifiers, which can benefit the allocation of limited audit and validation resources.

Table 1: *Clinical Documentation and Coding Alignment Performance Metrics [3, 4]*

Category	Metric/Measure	Value/Range
Relationship Extraction	Synonym agreement rate	18-34%
Relationship Extraction	Hierarchical relationship agreement	4-9%
Expert Agreement	Fleiss' kappa score	0.29-0.38
NLP Performance	Synonymy extraction approval	25-43%
NLP Performance	Hierarchical extraction approval	31-67%

Table 2: *Overview of NLP and Transformer-Based Approaches in Clinical Text Analysis [5, 6]*

Category	Technique/Model	Key Feature	Outcome/Performance
NLP Evolution	Neural Networks	Context-aware semantic analysis	Improved understanding of clinical text
Clinical NLP Application	Information Extraction	Identifies concepts & relationships	Supports automated coding systems
Transformer Models	Self-Attention Mechanism	Processes all word relationships simultaneously	Handles long clinical documents efficiently
Clinical BERT	Domain-Specific Pretraining	Trained on ~2M clinical notes	Higher accuracy than general BERT/BioBERT
Performance Metrics	NER & Relation Extraction	F1 Scores (86.4%–87.8%, 78.9%)	Enhanced clinical task performance

Table 3: *Model Development and Training Methodology Metrics [9, 10]*

Technique	Dataset	Precision	Recall	F-Measure	AUC	Key Finding
Cost-Sensitive XGBoost	Pima Indians Diabetes	0.767	0.855	0.81	0.83	Best minority class performance on diabetes data
Cost-Sensitive Random Forest	Breast Cancer	0.878	0.9	0.889	0.803	Superior on breast cancer minority class detection
Cost-Sensitive Random Forest	Cervical Cancer	1	1	1	0.988	Near-perfect minority class identification
Cost-Sensitive Random Forest	CKD	0.99	1	0.995	0.986	Excellent performance on imbalanced medical

						data
SMOTE + Random Forest (Prior Work)	Pima Indians Diabetes	0.68	0.86	0.68	0.76	Comparison baseline

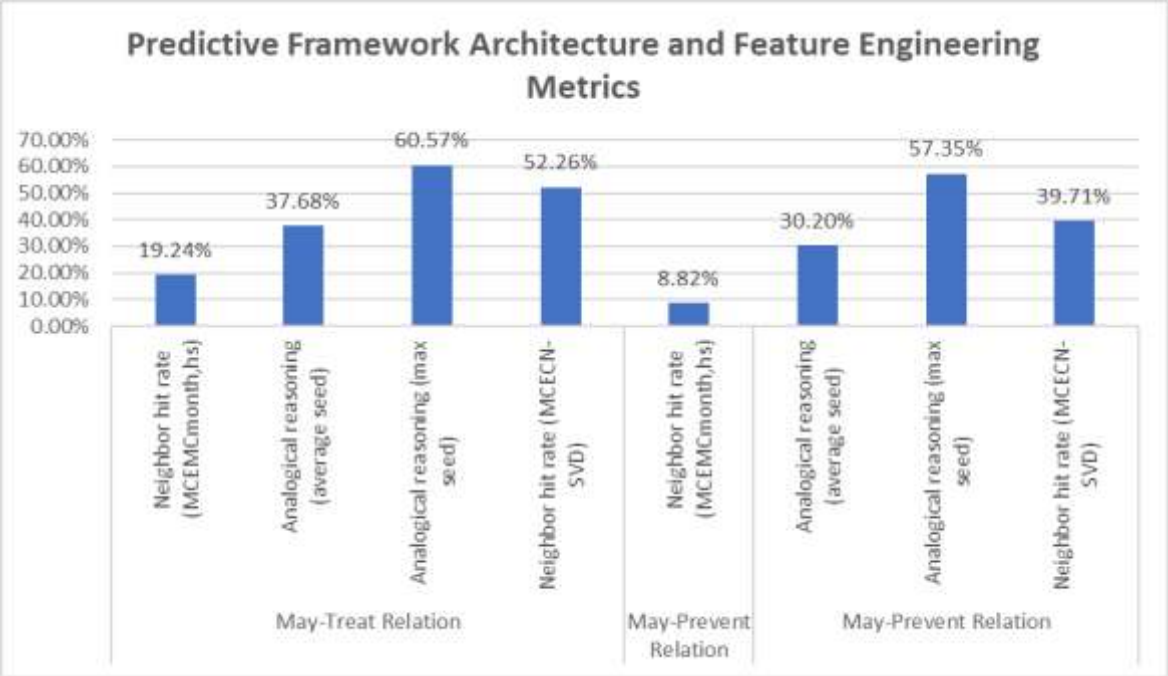


Figure 1: *Medical Concept Embedding Performance: May-Treat and May-Prevent Relationship Identification Across Different Models [7, 8]*

7. Conclusions

NLP-powered predictive models for medical coding errors represent a significant advancement in healthcare revenue cycle management and clinical documentation quality assurance. Domain-specific transformer models, when trained on large clinical text corpora, outperform customary rule-based validation systems by effectively modeling semantic relationships between unstructured clinical narratives and standardized code assignments.

The integration of multiple feature representation approaches—clinical BERT embeddings, code-specific semantic embeddings, and structured encounter-level metadata—creates comprehensive feature spaces that characterize coding variations across diverse encounters. Advanced class imbalance handling techniques, including cost-sensitive learning methodologies, enable model development that achieves required performance standards across various medical specialties while maintaining focus on minority class (error) detection.

Comprehensive evaluation across diverse clinical datasets demonstrates the framework's effectiveness in improving audit efficiency and accelerating identification of high-risk documentation patterns. The systematic approach to feature engineering, model architecture selection, and validation methodology establishes a reproducible foundation for coding error prediction system development.

Healthcare organizations employing NLP-based predictive frameworks can transition from retrospective, reactive audit processes to prospective risk identification and management strategies. Such transformation enables more efficient allocation of limited audit and validation resources, enhanced regulatory compliance, and improved financial accuracy in revenue cycle operations. As clinical language models expand and training datasets scale, automated coding support systems are anticipated to become increasingly integrated into healthcare information systems supporting both operational and regulatory functions.

The framework addresses critical challenges in medical coding through technology-enabled solutions, positioning healthcare organizations to enhance data quality, strengthen compliance posture, and optimize resource deployment across coding operations.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.
- **Use of AI Tools:** The author(s) declare that no generative AI or AI-assisted technologies were used in the writing process of this manuscript.

References

[1] Saifuddin Shaik Mohammed, "AI in Medical Coding: Transforming the US Healthcare System," International Journal of Innovative Science and Research Technology, 2025. [Online]. Available: https://www.researchgate.net/profile/Saifuddin-Shaik-Mohammed/publication/395893929_AI_in_Medical_Coding_Transforming_the_US_Healthcare_System/links/68d7582ed221a404b2a2e2ca/AI-in-Medical-Coding-Transforming-the-US-Healthcare-System.pdf

[2] Ivan Villar-Balboa et al., "ICD-10-CM coding uncovers the gap between serological and clinically identified coeliac disease prevalence: A population-based study," European Journal of Internal Medicine, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0953620525001414

[3] Elisa Henke et al., "Conceptual design of a generic data harmonization process for OMOP common data model," BMC Medical Informatics and Decision Making, 2024. [Online]. Available: https://link.springer.com/content/pdf/10.1186/s12911-024-02458-7.pdf

[4] Kristina Doing-Harris et al., "Automated concept and relationship extraction for the semi-automated ontology management (SEAM) system," Journal of Biomedical Semantics, 2015. [Online]. Available: https://link.springer.com/content/pdf/10.1186/s13326-015-0011-7.pdf

[5] Mohammed Yusuf Ansari et al., "A survey of transformers and large language models for ECG diagnosis: advances, challenges, and future directions," Artificial Intelligence Review (2025) [Online]. Available: https://link.springer.com/content/pdf/10.1007/s10462-025-11259-x.pdf

[6] Emily Alsentzer et al., "Publicly Available Clinical BERT Embeddings," in Proceedings of the 2nd Clinical Natural Language Processing Workshop, pages 72–78, 2019. [Online]. Available: https://aclanthology.org/W19-1909.pdf

[7] Liang Yao et al., "Clinical text classification with rule-based features and knowledge-guided convolutional neural networks," BMC Medical Informatics and Decision Making, 2019. [Online]. Available: https://link.springer.com/content/pdf/10.1186/s12911-019-0781-4.pdf

[8] Youngduck Choi et al., "Learning Low-Dimensional Representations of Medical Concepts,". [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC5001761/pdf/2381736.pdf

[9] Ibomoiye Domor Mienye and Yanxia Sun, "Performance analysis of cost-sensitive learning methods with application to imbalanced medical data," Informatics in Medicine Unlocked, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S235291482100174X

[10] Tianqi Chen et al., "XGBoost: A Scalable Tree Boosting System," ACM Digital Library, 2016. [Online]. Available: https://dl.acm.org/doi/pdf/10.1145/2939672.2939785

[11] Jakir Hossain Bhuiyan Masud et al., "Applying Deep Learning Model to Predict Diagnosis Code of Medical Records," National Library of Medicine, 2023. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10340491/

[12] Hussein A. A. Al-Khamees et al., "Enhancing classification accuracy in medical datasets using a hybrid distance and cluster refinement-based K-means clustering method," Scientific Reports, 2026. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC12847962/