

Enhancing EHR Text Classification: A Hybrid BERT Hashing Framework for Interpretability, Robustness, and Data-Efficient Medical Text Classification

A. Sandhya^{1,*}, G. Deena², Mohamed Ahmed Hassan Youssef³, Shivani Arora⁴, Yahya Ali Saleh Al-Quhali⁵, Ara Kiarash⁶, Cao Yu⁷

^{1,2}Department of Computer Science and Engineering in Artificial Intelligence and Machine Learning, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamil Nadu, India.

³College of Medicine, Zagazig University, Zagazig, Sharqia Governorate, Egypt.

⁴Department of Data and Analytics, Infosys Limited, Edmonton, Alberta, Canada.

⁵Department of Basic Sciences, Amran Community College, Amran, Yemen.

⁶Department of Electrical and Chemical Engineering, McMaster University, Hamilton, Ontario, Canada.

⁷Department of Computing and Information Systems, Singapore Management University, 81 Victoria Street, Singapore. sandhyalagar@gmail.com¹, deenag@srmist.edu.in², 20512022101866@medicine.zu.edu.eg³, shivani.ar80@gmail.com⁴, yahyaalquhali2015@gmail.com⁵, arak@mcmaster.ca⁶, yu.cao.2025@phdcs.smu.edu.sg⁷

Abstract: The clinical narratives contained in EHRs are long, unstructured, with dense information content, and difficult to classify due to uneven length, structure, terminology, and information density. This work presents a hybrid machine learning system that uses classifiers to classify medical transcription, focusing on data efficiency, interpretability, and robustness. The pipeline combines natural language preprocessing, Domain-informed feature extraction, Hashing Vectorizer representation and BERT representation for context. Five features identified in the original study document length, bigram type, relationship words, word count, and noun/verb counts are included to capture complementary lexical, syntactic, and domain information. The Medical Transcriptions Dataset consists of 80% training and 20% testing data. The reported evaluation includes 2321 test samples. Test-set accuracy is about 74.2%, weighted precision is about 74.4%, weighted recall is about 74.2%, and weighted F1 score is about 72.7%. The average macro F1 is around 50.0%, which shows significant class-to-class variation. Many low-support categories have low or zero F1 values, while several high-support categories have strong F1 values. After 6 epochs, the training loss drops from 0.67 to 0.46, and the reported validation F1 rises from 66% to 77% and ultimately levels off. These results show that a hybrid representation of heterogeneous medical text is feasible, but also indicate significant class imbalance and limited minority-class examples. The updated framework thus suggests balanced learning, external validation, systematic ablation, robustness testing, and clinically assessed interpretability before real-world deployment.

Keywords: Electronic Health Records; Clinical NLP; Hashing Vectorizer; Medical Text Classification; Machine Learning; Clinical Text Mining; Medical Text Classification.

Received on: 24/08/2025, **Revised on:** 07/11/2025, **Accepted on:** 26/11/2025, **Published on:** 03/06/2026

Journal Homepage: <https://www.fmdbpublish.com/user/journals/details/FTSBE>

DOI: <https://doi.org/10.69888/FTSBE.2026.000793>

Cite as: A. Sandhya, G. Deena, M. A. H. Youssef, S. Arora, Y. A. S. Al-Quhali, A. Kiarash, and C. Yu, “Enhancing EHR Text Classification: A Hybrid BERT Hashing Framework for Interpretability, Robustness, and Data-Efficient Medical Text Classification,” *FMDB Transactions on Sustainable Biomedical Engineering*, vol. 1, no. 2, pp. 130–142, 2026.

Copyright © 2026 A. Sandhya *et al.*, licensed to Fernando Martins De Bulhão (FMDB) Publishing Company. This is an open access article distributed under [CC BY-NC-SA 4.0](https://creativecommons.org/licenses/by-nc-sa/4.0/), which permits copying, remixing, redistributing, transformation, and building upon the material in any medium so long as the original work is properly cited.

*Corresponding author.

1. Introduction

In today's healthcare environment, electronic health records (EHRs) have become a key source of information for clinical documentation, reimbursement, quality improvement, healthcare operations, and healthcare delivery [1]. Much of an EHR consists of free-text data, such as patient admissions, discharge notes, diagnoses, treatments, procedure notes, progress notes, and clinical observations [2]. These stories include symptom information, clinical history, treatment decisions, demographic information, relationships, and hospital course, which can be hard to encode as tabular data, as in more traditional tabular learning approaches [3]. These stories include symptom information, clinical history, treatment decisions, demographic information, relationships, and hospital course, which can be hard to encode as tabular data, as in more traditional tabular learning approaches [4]. Given these challenges, this manuscript addresses text classification with a focus on robustness, interpretability, and data efficiency [5]. The usefulness of automated EHR text classification is driven by the volume and the rate at which clinical data is being produced [6]. Medical professionals use notes to record information about the care provided, share it with other staff, record observations, support diagnosis and treatment, and complete medical coding. The original manuscript also highlights diagnosis-related groups (DRGs), groups of inpatient cases classified by principal and secondary diagnoses, procedures, age, sex, length of stay and discharge status. While DRG information has long played a critical role in reimbursement, it can also inform hospital operations and resource allocation. Inferring useful categories directly from clinical language is difficult because the system must handle heterogeneous text and try to emulate some of the practices expert coders follow. Several features of clinical text create classification challenges. One is that document length varies widely [7].

A hospital course can be a lengthy description covering the entire admission, while a diagnosis or treatment field may be a short paragraph. Second, EHR fields serve different purposes, so they use different language features [8]. Thirdly, medical writing has its own terminology, abbreviations, acronyms and context-dependent expressions. A good classifier for "ordinary language" use may not be appropriate in the context of clinical language. Conventional language-model strategies require substantial data, memory, and computation, which is problematic in low-resource medical applications, according to the original manuscript. Efficient learning is crucial in healthcare, where data is essential [9]. When a new disease, unusual condition or institution-specific clinical practice arises, there may be relatively few examples, even though accurate classification remains desirable. This manuscript motivates meta-learning as a potential solution. Meta-learning learns to represent multiple tasks in one model, then adapts to a new task with a small support set. This concept is particularly appealing in clinical NLP because labelled notes often vary widely in availability across diseases, institutions, specialties and coding categories [10]. Transformer-based language models, such as BERT, have revolutionized NLP. BERT models bidirectional contextual representations to capture relationships between words in context. This is useful in biomedical text because surrounding words, negation, abbreviations, symptoms, medications, and other medical concepts can influence a term's meaning. The source manuscript describes BERT as useful for text classification tasks and for extracting clinically relevant information from notes. However, BERT models can remain resource-heavy and require large amounts of task-specific training data. These restrictions lead us to consider lighter representations and complementary feature engineering. The proposed direction in the manuscript combines language-model processing with a hashing-vectorizer strategy and domain-informed feature selection [11].

The feature-selection stage considers document length, bigram information, relationship words, total word count, and noun and verb counts. NLTK handles tokenization and part-of-speech tagging, and a bigram probability model captures local word relationships. Relationship words are addressed specifically because they can, in some cases, contain clinically relevant contextual information [12]. The resulting framework aims to retain useful information without relying on an excessively large representation. Interpretability is also a key consideration. When a clinical classifier's results can affect patient care, coding, or resource allocation, it's more than a text-mining tool. Clinicians and healthcare administrators should understand the rationale behind a system's prediction [13]. The manuscript then talks about interpretable deep learning models with attention mechanisms, medical concept embeddings, and feature importance. Attention can identify which parts of a clinical note matter for the prediction, and medical concept embeddings let clinicians inspect key concepts in a clinical note, such as diagnoses, drugs, and procedures. The second important requirement is robustness. Noisy documentation, incomplete notes, spelling variations, abbreviations, distribution shifts and deliberately manipulated inputs can all affect clinical NLP systems. It discusses adversarial learning to improve resilience and interpretability. A good classifier should not change its result drastically due to minor, unimportant or even incorrect input modifications. In healthcare, this can be especially critical, as mispredictions can propagate to downstream clinical or administrative processes [14]. Accordingly, this work addresses the following question: How can researchers design a classification pipeline that efficiently learns discriminative representations from heterogeneous medical transcription text, incorporates domain information, provides interpretable signals, and maintains stable prediction performance? The study utilizes the Medical Transcriptions Dataset from the source manuscript [15].

Because privacy or regulatory considerations limit access to real clinical data, publicly available transcription data serves as a practical proxy for methodological development [16]. This experimental design uses 80% for training and 20% for testing. The principal evaluation measures are accuracy and F1 score; additional measures provide a more detailed examination of

classification performance: precision, recall, class-level support, macro averages, weighted averages, and sample averages [17]—the classification summary above shows 2,321 evaluated samples for the source results. Weighted precision is around 0.744, weighted recall is around 0.742, and weighted F1 is around 0.727. The macro F1 is around 0.500, indicating large differences between classes. This distinction matters because even a seemingly high score may not mean minority classes are easy to identify. The manuscript also includes training/validation behaviour for 6 epochs. The training loss drops from around 0.67 in epoch 1 to around 0.46 in epoch 6. As shown in Figure 1, validation accuracy rises from ~66% to ~77% over the same period. Still, compare the plotted metric to the exact experimental setup before publication, as some validation measures use different terminology for the same metric [18]. A separate plot named "flat validation accuracy" displays values around the lower end of the scale. This implies that any final manuscript should explicitly distinguish between the validation measure based on F1 and flat validation accuracy, rather than using them synonymously. The main finding of the revised study is therefore not that one representation is generally superior to all modern clinical language models, but that a unified, domain-aware classification pipeline for heterogeneous medical text is warranted [19].

This approach involves preprocessing, feature extraction, hashing-based representation, BERT-based contextual learning, classification and evaluation. The manuscript also stresses the importance of interpretability and robustness in deployment [20]. This approach makes the work more scientifically defensible and more congruent with the evidence in the original study. The study also highlights the distinction between clinical usefulness and predictive performance. An algorithm may have a good overall performance score but perform poorly in a specific clinical setting. Systematic under-recognition of a rare category can lead to coding or resource-allocation errors for administrative grouping, for instance. The threshold for acceptable performance would be even higher for decision support applications [1]. The updated manuscript thus does not consider accuracy and F1 as direct evidence of clinical safety; instead, they serve as research evaluation measures. The manuscript shows how representation matters for the input structure. The hospital record is not a single composed whole. It can include short diagnostic statements and long narratives, procedure lists, drug information, and copied or pre-formatted text. It's unlikely that one feature family will represent all these forms equally well. The length and count features represent document structure, the bigrams represent local lexical relationships, hashing provides scalable sparse features, and BERT captures context [3]. This complementary design is a main methodological driver of the proposed framework. From a systems perspective, the pipeline is a series of increasingly complex representations. The pipeline first converts raw transcriptions into standardized tokens. It then converts these tokens into lexical and syntactic properties, and then into contextual representations [2]. The classifier is not merely text-based, but works on the combined representation. This separation makes debugging easier, as the system has four error stages to examine: preprocessing, feature extraction, representation and classification [4].

The study is significant for situations in which access to large, well-labeled clinical data is limited. Constructing a large dataset for each classification can be challenging because of privacy laws or institutional governance [6]. While public medical transcription resources may serve as an initial point for methodological experimentation, privacy-preserving assessment and institutional approval are still necessary. Other complementary approaches that remain promising for reducing the need for annotations include meta-learning, transfer learning and active learning. Last but not least, the manuscript contributes to a broader conversation on responsible AI in healthcare [7]. Accuracy is required, but not enough. Systems should be comprehensible to the people who use them, resilient to unexpected inputs and tested in relevant subgroups and data sources. The study explicitly introduces interpretability and robustness to text classification, moving beyond a narrow benchmark perspective. This proposed framework is therefore placed as a stepping stone for future prediction-based EHR analytic systems that can offer meaningful predictions while providing an appropriate amount of evidence for human review. The remainder of the paper reviews related clinical NLP and text-classification approaches, describes the proposed pipeline and algorithm, presents the experimental findings, interprets the class-level and epoch-level behaviour, discusses practical implications and limitations, and concludes with future directions for integrating predictive models with hospital resource-management systems [8]. The ultimate objective is to support more efficient and trustworthy extraction of information from EHR text while acknowledging the limitations imposed by dataset size, class imbalance, public transcription data, and the currently reported evaluation design [12].

2. Literature Review

The evolution of Clinical NLP from manually engineered lexical representations to distributed representations and transformer models. The progress of Clinical NLP from handcrafted lexical features to distributed representations and transformer models. Early text classification systems largely relied on bag-of-words, n-grams, TF-IDF, part-of-speech patterns, and traditional classifiers. These methods still apply to small datasets because they are relatively simple and efficient. They tend to struggle with ambiguous words, long-distance dependencies, negations, abbreviated phrases, and domain-specific terminology. EHR text exacerbates these problems, as the same clinical concept can be expressed in several ways, and standardized language may be mixed with highly variable language in the notes. The manuscript provides an overview of research on medical text classification. López-García et al. [1] studied transformer-based models for Spanish clinical coding on cancer cases and applied transfer learning to adapt models to a private collection of cancer clinical cases. Their work shows that domain adaptation (i.e.,

introducing more exposure to the terminology and style of the target medical domain) is important for a general-purpose transformer to be useful. This observation is also relevant to this research because it highlights the challenges of interlingual and interinstitutional transfer and supports using contextual language representations. Flores et al. [2] introduced an active-learning approach for biomedical text classification that uses automatically generated regular expressions, called CREGEX. When expert annotation is costly, active learning lets the classifier focus on examples expected to be informative. The study is particularly significant in medical applications where unlimited data may be available, but the domain expert cannot annotate it all. The current manuscript also acknowledges data limitations and does not rely on the availability of very large labelled data sets.

Roitero et al. [3] presented a domain-specific BERT model trained and fine-tuned on a relatively small domain-specific corpus. The researchers conclude that domain-specific contextual representations can match the performance of very large general-purpose models, but without the data volume. This motivates the use of both BERT-like contextual understanding and a lightweight feature representation in the present work. Gao et al. [4] studied the challenges transformers face in clinical text classification. Their findings have particularly practical implications: They warn against taking the transformer architecture for granted when tackling all clinical NLP challenges. A clinical note may be lengthy, verbose, varied, and formal, unlike normal written communication. In the present manuscript, researchers explicitly incorporate document length, word-level relationships, part-of-speech information, and domain-specific features into contextual modelling to address these challenges. Berge et al. [8] proposed CGMVQA, a method for medical visual question answering that combines classification and generation for complex medical questions. Although the task differs from EHR classification, the work highlights a broader trend toward models that can accommodate heterogeneous information and complex biomedical semantics. Jin et al. [5] presented TBLC-rAttention, an attention mechanism for recognizing emotional tendency in Chinese medical comments [6], to demonstrate the application of attention in medical-language representation learning. The relevance of the present study is that learned attention or feature weighting selects information relevant to predictions. Zhou et al. [6] suggested using a neural network architecture to extract information from Chinese drug package inserts, where they classified the sentences and also added classification and error correction. In this work, the authors highlight the need to access noisy/imperfect biomedical text. Ruan et al. [7] first proposed a multi-view graph learning approach to relation extraction from Chinese clinical records. Their strategy emphasizes local dependencies and global information. Taken together, these studies demonstrate the value of medical NLP representations that capture not just token evidence but relationships.

Berge et al. [8] studied the Tsetlin Machine for human-interpretable rules in text categorization with medical applications. From the perspective of the current work, interpretability is the key contribution. Human users can examine this explicit evidence for classification decisions more easily than a high-dimensional neural representation. In this context, the latter approach combines domain features and contextual modelling to make predictions more comprehensible. The literature also includes substantial research on pretrained language models. Later, Vaswani et al. [15] adopted the transformer architecture, laying the foundation for attention-based models. Devlin et al. [20] proposed BERT, a bidirectional, transformer-based model with contextual representations. For clinical text, BERT's ability to account for context before and after a word is very helpful, as a word's meaning can be influenced by nearby symptoms, medications, negation words or procedural descriptions. Domain-specific embeddings and pretrained models have been the focus of subsequent biomedical NLP research. BioWordVec combines subword information and MeSH knowledge to enhance biomedical word representations. These domain-aware representations are useful because clinical terms may have domain-specific morphology and abbreviations. This shift from task-specific feature engineering to reusable, pretrained representations is also well documented in surveys of pretrained NLP models. But transfer learning does not eliminate data imbalance, domain shift or interpretability issues. The challenge of a long clinical note is yet another one. LongFormer was originally proposed as a long-document transformer designed to address the computational complexity of long sequences. Long-document methods are directly related to the problem at hand, since a hospital-course narrative can be much longer than a diagnosis or treatment description. So, a single uniform representation may not suit all EHR fields. To tackle this heterogeneity, this work introduces a preprocessing and feature extraction module along with a contextual BERT module. Researchers also explore several EHR deep learning architectures.

Comparative studies like Solares et al. [13] illustrate the wide range of neural techniques used with EHR data. Other studies used deep learning to extract primary cancer sites from pathology reports and accompanying clinical narratives. These studies confirm the importance of automated information extraction and indicate that task, data, and clinical context play significant roles in its performance. Clinical coding is one of the most significant applications. Flores et al. [2] compared deep-learning approaches for ICD coding of clinical records. While automated coding can reduce manual effort, models still need to map free-text descriptions to standardized categories. This manuscript extends that idea to DRG motivation. A useful classifier must recognize clinical concepts and their relationships while also coping with coding conventions. To make BERT more suitable for classification, researchers fine-tuned it using standard techniques. The fine-tuned system may perform well on the specific task, but it may require substantial computation and data. In this work, researchers explore whether hashing-based representation and domain-specific feature selection could complement contextual modelling. The key features of the Hashing Vectorizer are that it does not rely on a large vocabulary, and it transforms tokens into a fixed-dimensional vector space. This

can help reduce usage and make processing more straightforward in the vector space and/or a large rapidly evolving feature space. A crucial way to reduce annotation cost is active learning. Delvin et al. [20] used support-vector-machine active learning, and Zhu et al. [18] suggested the condensed stopping criteria for active-learning annotation. These works align with the broader idea that intelligent sampling can facilitate data collection. Active learning is not the main focus of the present manuscript, but its concept fits well with the study's focus on data efficiency. It could be a logical direction for future research. The evaluation method is also crucial. Sun et al. [17] discuss repeated cross-validation as an approach to estimating model accuracy. However, in clinical NLP, a single train-test split can yield an unstable estimate with limited data or a skewed distribution. The 80:20 split in the source experiment should therefore be viewed as a starting point, not a definitive point of generalization. Future studies should use repeated stratified splits or cross-validation, confidence intervals, and external validation. Representation learning approaches such as GloVe and multimodal or vision-language models like VL-BERT have also been used.

These methods show that information can be captured even by sparse lexical methods that cannot capture contextual and distributed representations. However, rich representations can be computationally complex and less transparent. In the current work, researchers balance these by applying a hashing-based feature representation alongside BERT and explicit domain features. Data augmentation can affect deep learning performance. With augmentation, caution is needed in clinical text; synthetic changes may alter medical meaning. Similarly, in medical domain benchmark tasks, Image CLEF-Med illustrates the broader challenge of visual question answering. These observations highlight the need for a task-specific approach to evaluating medical AI and the importance of avoiding a single overall metric. Overall, the literature in this area shows four persistent gaps. First, clinical text is heterogeneous in length and structure. Second, labeled data may be limited, particularly for unusual cases or specific tasks [8]. Third, transformer models are complex and hard to interpret, yet they offer high-level representations. Fourth, clinical deployment requires robustness to noise, distribution shifts and potential adversarial manipulation. The proposed work addresses all of these by implementing a hybrid pipeline that includes preprocessing, domain-aware feature extraction, explicit attention to interpretability and robustness, hashing-based representation, and BERT contextual learning. The literature thus justifies the need for a system that lies between a classical bag-of-words classifier and an unconstrained large transformer. Instead, the correct design supplements contextually represented information with complementary information from inexpensive lexical and syntactic representations and implements it as a layered structure. This design is especially appropriate for public medical transcription datasets with documents that vary widely in length and content, where class imbalance may be significant. The results also showed a low macro F1 of around 0.500 and a higher weighted F1 of around 0.727, further emphasizing the importance of evaluating not only overall performance but also minority-class behaviour.

3. Proposed System

The proposed system is a hybrid EHR text-classification pipeline that includes preprocessing, domain-informed feature extraction, hashing-based vectorization, contextual BERT representations, classification and interpretability-oriented analysis. Each step follows the source manuscript workflow: Load Dataset, Data Pre-processing, Feature Extraction, Data Transformation, Data Analysis, Model Discovery, Conformance Checking, Performance Analysis, Discover Insights. The new formulation explicitly represents data flow and decouples representation learning from evaluation. The input is the Medical Transcriptions Dataset, which contains medical transcription samples from various specialties. According to the source manuscript, the data was scraped from mtsamples.com and provided by data. The dataset consists of labeled text instances. Clinical notes vary widely in content and length, so preprocessing is crucial before feature extraction. In the first step, researchers “normalized” each transcript while preserving clinically useful words. Using NLTK, researchers tokenize the text. The text is tokenized with NLTK. Tokenization breaks a document into a sequence of lexical units. Part-of-speech tagging is then performed to enable counting nouns, verbs, etc.

These counts are explicitly added as features. The goal is not to replace contextual representations, but to give concise, easy-to-understand indications of the linguistic makeup of a document. The second stage identifies five feature groups: length count, bigram type, relationship words, word count and noun/verb count, which are the main feature groups located in the source manuscript. Length count reflects the text that has been typed into the transcription. One similar but different measure is word count, which is based on tokenized text. Pairs of adjacent words in a bigram can be used to represent local relationships (e.g., medically meaningful phrases). Relationship words are selected from a pre-defined list and used as domain-informed indicators. Noun and verb counts provide a rough description of clinical content and action-oriented language. A bigram probability model estimates the likelihood of a subsequent word based on preceding context. The source manuscript describes this as a Markov-model-style representation. In a first-order formulation, the probability of a sequence can be approximated using conditional probabilities of adjacent tokens as in Equation 1:

$$P(w_i/w_{i-1}) \quad (1)$$

Where, P = probability, w_i = current word in the sentence, w_{i-1} = the previous word of w_i

These bigram statistics serve as additional features, not a full language model. This keeps feature extraction lightweight and lets well-known local word pairs guide classification. The HashingVectorizer provides the primary sparse representation. In hashing maps, instead of explicitly storing a vocabulary of all observed tokens, the system maps tokens to a fixed-dimensional vector space. The advantages are bounded memory usage and the ability to include new tokens without rebuilding the vocabulary. Occasionally, hash collisions occur; thus, the hash dimension must be chosen carefully. The proposed design then follows manual extraction of domain features with hashed lexical information. BERT is used as the contextual language component. BERT generates representations in which the meaning of a word is understood as part of the context. This matters for EHR text because clinical meaning often depends on neighboring terms. The BERT component therefore supplies semantic information that complements the sparse hashing representation. The combined representation can be expressed conceptually as in Equation 2:

$$H = [H_BERT \parallel H_hash \parallel H_domain] \quad (2)$$

where H_BERT are contextual features, H_hash are hashed lexical features, H_domain are explicitly extracted domain features, and $\parallel$ denotes concatenation. The output feeds into the classification layer. Researchers incorporate interpretability during model analysis. If representations of the text are attention-based, the weights of attention can be examined to determine what spans of text are related to predictions. Explicit domain features also include human-readable evidence, since document length, word counts, relationship terms, or grammatical counts can be directly inspected. This is better than using a single "black box score". Robustness is addressed in adversarial-learning terms. A powerful implementation can create noisy versions of training examples, helping the classifier produce consistent predictions. The inner maximization is a worst-case perturbation in a specified constraint, and the outer minimization is to train the model to be less sensitive to this perturbation. In clinical text, perturbations must be carefully devised to avoid clinically unrealistic examples and to avoid changing the target meaning. The training procedure follows the 80:20 split, using 80% of the data for training and 20% for testing. Evaluation is reported in the source manuscript as both accuracy and F1 scores, and precision, recall, and class support are also reported.

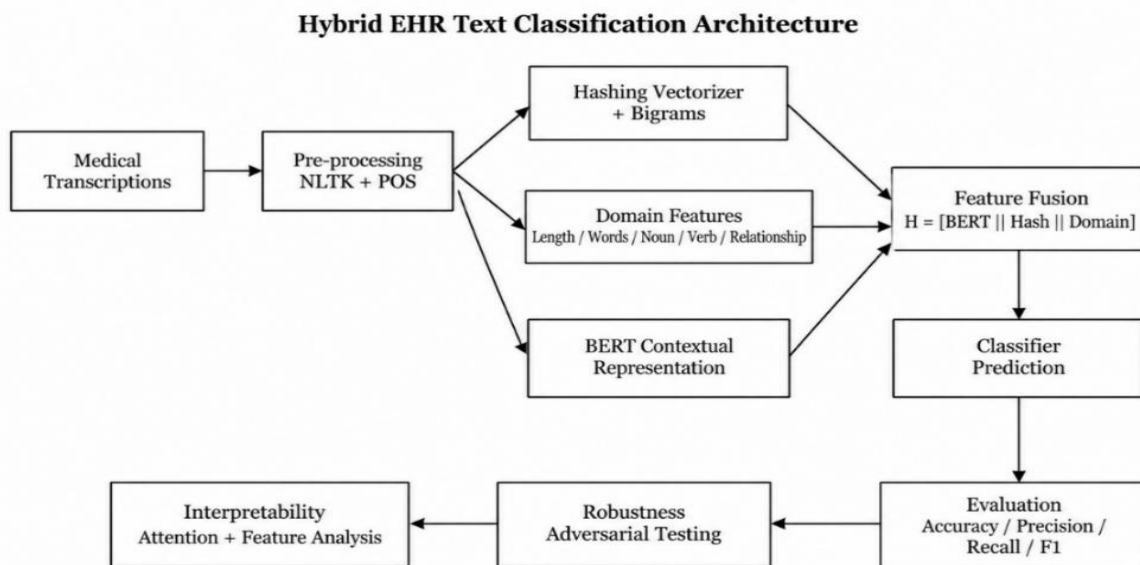


Figure 1: Proposed hybrid architecture for EHR medical transcription classification

During the reported experiment, the classifier is trained for 6 epochs. Researchers track loss and validation performance after each epoch to confirm optimization. The proposed architecture thus comprises 3 complementary information streams as shown in Figure 1. BERT captures contextual semantics, hashing captures scalable lexical patterns, and domain features capture explicit characteristics of medical transcriptions. The final evaluation stage depends not only on aggregate classification quality but also on class behaviour. Importantly, the reported dataset is imbalanced; samples per class range from fewer than 50 to several hundred. The preprocessing step transforms raw medical records into a representation that is suitable for both sparse and contextual representations. The reported workflow utilizes NLTK tokenization and NLTK part-of-speech tagging. Any record that is empty or not usable should be discarded; any record that contains clinically significant information should be retained. Use the same preprocessing pipeline for both training and test sets to prevent leakage. The reported feature-selection strategy includes document length, bigram type, relationship words, word count, and noun/verb counts. These features are then

combined into a representation by combining hashed lexical features with BERT contextual representations. The resulting feature vector is then passed to the classifier.

3.1. Algorithm 1: Hybrid EHR Text Classification

Input: Labelled medical transcription dataset D

Output: Predicted class labels and evaluation metrics

1. Load the medical transcription dataset.
2. Clean and tokenize each transcription using NLTK.
3. Apply part-of-speech tagging.
4. Calculate document length, word count, noun count, and verb count.
5. Identify predefined relationship words.
6. Generate bigram statistics.
7. Transform lexical features using Hashing Vectorizer.
8. Generate contextual BERT representations.
9. Concatenate BERT, hashing, and explicit domain features.
10. Split data into 80% training and 20% testing subsets.
11. Train the classification model and monitor validation performance.
12. Inspect attention and explicit feature signals for interpretability.
13. Evaluate accuracy, precision, recall, F1, and class-wise support.
14. Analyze minority-class errors and convergence behaviour.
15. Return predictions and evaluation results.

4. Results

The experimental results included in this source manuscript demonstrate that the proposed EHR text-classification pipeline can learn useful patterns from heterogeneous medical transcriptions and highlight significant inter-class variation. The evaluation uses the Medical Transcriptions Dataset with an 80:20 training-to-testing split. Researchers use precision, recall, F1 score and class support to assess performance, with training and validation curves provided over six epochs. Researchers list the experimental hyperparameters here, as the source document did not include a complete confusion matrix or a table of all hyperparameters. The model's overall reported test-set accuracy is about 0.742. In a single-label multiclass classification problem, overall accuracy is roughly equal to micro recall, and the reported weighted recall is 0.741922 (or 1,722 correctly classified samples out of 2,321). Weighted precision is 0.744074, and weighted F1 is 0.726999. The weighted values suggest meaningful learning. However, performance varies across categories, as shown by the difference between weighted and macro scores. The macro-averaged precision is 0.583867, macro recall is 0.492141, and macro F1 is 0.500223.

In macro averaging, each class gets the same weight regardless of the number of samples. The macro values are significantly lower than the weighted values, showing that the model scores well on some minority classes but not on the majority classes. This is not a statistical quirk; it is a central finding. A model well suited to a high-frequency class can perform poorly on a rare class or on clinically important cases in an EHR application. Class-level results provide additional insight. Class B is one of the strongest categories, with precision 0.952500, recall 1.000000, and F1 0.975672 across 381 samples. Class E also has recall of 1.000000 and an F1 of 0.876404, supported by 312 samples. Class D reaches precision 0.823105, recall 0.850746, and F1 0.836697 over 268 samples. Class G records precision 0.750779, recall 0.912879, and F1 0.823932 over 264 samples. Class M has precision 0.814815, recall 0.781065, and F1 0.797583 over 169 samples. When there are enough training examples for a category, and there are clear textual patterns in them, the results here suggest that the learned representation functions well. Class C achieves precision 0.767656, recall 0.796089, and F1 0.777032 with 215 samples. Class A achieves precision 0.697143, recall 0.618299, and F1 0.655914 with 197 samples. Class N achieves precision 0.845588, recall 0.595855, and F1 0.690988 with 193 samples.

For N, the difference between precision and recall shows that the predictions made in this category tend to be correct, but many of the true N examples are not retrieved by the model. This behaviour is crucial for error analysis because improving recall will require better intra-class variation or more training examples. Several smaller classes have serious recognition problems. Class F has precision 0.944444 but recall only 0.239437, producing an F1 of 0.382022 across 71 samples. This means the model is conservative in predicting F: if it predicts the class, it is likely to be right, but it is likely to miss most true F examples. In the classification table reported for the classes, precision and recall for classes H, J, L, and Z are zero, as are their respective supports of 49, 39, 62, and 61. Class I has precision 0.800000 but recall 0.100000 and F1 0.177778 over 40 samples. This result is strong evidence that class imbalance and lack of representation are significant reasons for the poor performance of minority

classes. An unequal distribution of support confirms this. The largest reported class is B with 381 samples, followed by E with 312, D with 268, G with 264, C with 215, N with 193, and A with 197. In contrast, I have 40 samples, J has 39, H has 49, Z has 61, L has 62, and F has 71. A classifier trained on this distribution can reduce aggregate loss by focusing on the most prevalent categories. Thus, a weighted F1 of around 0.727 should not be interpreted as evidence that the model reliably recognizes all medical categories.

Training behaviour at the epoch level shows optimization progress. The training-loss plot demonstrates a monotonic loss over the course of the training from about 0.67 at epoch 1 to about 0.46 at epoch 6. The approximate sequence is 0.67, 0.57, 0.53, 0.50, 0.48, and 0.46. As training continues, the continued decrease means that the model is becoming a better fit for the training goal. Decreasing training loss, however, is not sufficient to show generalization and should be interpreted alongside validation behaviour and held-out test results. The validation curve for the F1 score rises from around 66% at the first epoch to around 77% at the fifth and sixth epochs. The curve rises steeply in the early epochs and flattens out later. This trend indicates that most learning occurs in the initial epochs of training, and that later epochs (past 200) yield smaller gains. The 6th epoch result is very close to the highest. This behaviour aligns with the aggregate test F1 (0.727), but it should not be confused with test metrics, as the source document does not specify all splits or evaluation details. In Figure 2, “Flat Validation Accuracy”, there is a much lower curve with scores remaining near the bottom of a 0- 100% scale. According to the manuscript, the suggested model becomes more accurate and shows minimal loss as the number of epochs increases.

Table 1: Class-wise classification performance

Class	Precision	Recall	F1	Support
A	0.697143	0.618299	0.655914	197
B	0.952500	1.000000	0.975672	381
C	0.767656	0.796089	0.777032	215
D	0.823105	0.850746	0.836697	268
E	0.780000	1.000000	0.876404	312
F	0.944444	0.239437	0.382022	71
G	0.750779	0.912879	0.823932	264
H	0.000000	0.000000	0.000000	49
I	0.800000	0.100000	0.177778	40
J	0.000000	0.000000	0.000000	39
L	0.000000	0.000000	0.000000	62
M	0.814815	0.781065	0.797583	169
N	0.845588	0.595855	0.690988	193
Z	0.000000	0.000000	0.000000	61
Macro avg.	0.583867	0.492141	0.500223	2321
Weighted avg.	0.744074	0.741922	0.726999	2321
Sample avg.	0.809881	0.755567	0.767059	2321

However, the source does not fully describe the connection between this plot and the F1 validation curve. The metric name, calculation procedure, and numerical values should be explained in detail to avoid confusion when readers encounter the reported test accuracy of approximately 74%. These results thus give three complementary conclusions. First, the model achieves acceptable overall classification performance, with an accuracy of around 74.2% and a weighted F1 of 72.7%. Second, optimization performance remains consistent across the six reported epochs (training loss decreases while validation F1 increases). Third, the overall performance hides significant class-level weaknesses. The macro F1 is clearly below the weighted F1 (about 0.500), and some minority classes have no F1. This suggests that future work should focus more on balanced recognition than on the overall score. The feature design also offers interpretability. Length, noun and verb counts, relationship-word indicators, and bigram information are directly interpretable. These features link the clinical stories to predictive models. BERT representations add this contextual information, capturing semantic relationships not available from simple counts. A simple lexical representation is provided by hashing. This combination is methodologically reasonable for heterogeneous medical text, but an ablation study with exact numerical results would be necessary to isolate each component's contribution quantitatively.

More broadly, the evidence reported here favors combining the two proposed methods: an aggregate accuracy of about 64% and meaningful contextual learning in classifying medical transcription text. However, it does not provide evidence of uniformly high clinical reliability. A few qualifications are important: failures in minority classes, class imbalance, evaluation on a single dataset, and limited reporting of experimental details. Interpret the findings as indicative and as a basis for further validation studies, but not as evidence for immediate clinical use. The statistics for samples from different sources in the source

classification table are also informative but should be interpreted with caution. For 2,321 evaluated samples, the sample-average precision is nearly 0.809881, the recall is nearly 0.755567, and the F1 is nearly 0.767059. These averages are calculated differently from the macro and weighted averages and should not replace the main Indicators of Average. They are included to aid reproducibility since they are part of the reported classification output; however, the revised paper emphasizes accuracy, weighted F1, macro F1, and class-wise scores for interpretation (Figure 3). The relationship between precision and recall within classes provides a more detailed view of a classifier's decision boundary. High precision and low recall (as seen for F and I) indicate a conservative prediction: the model makes few predictions for these classes, but the ones it does will have high confidence. In contrast, classes like G and E show high recall, meaning the model correctly classifies many instances. The zero-scored classes indicate an even more serious problem: the model fails to retrieve meaningful examples of those classes.

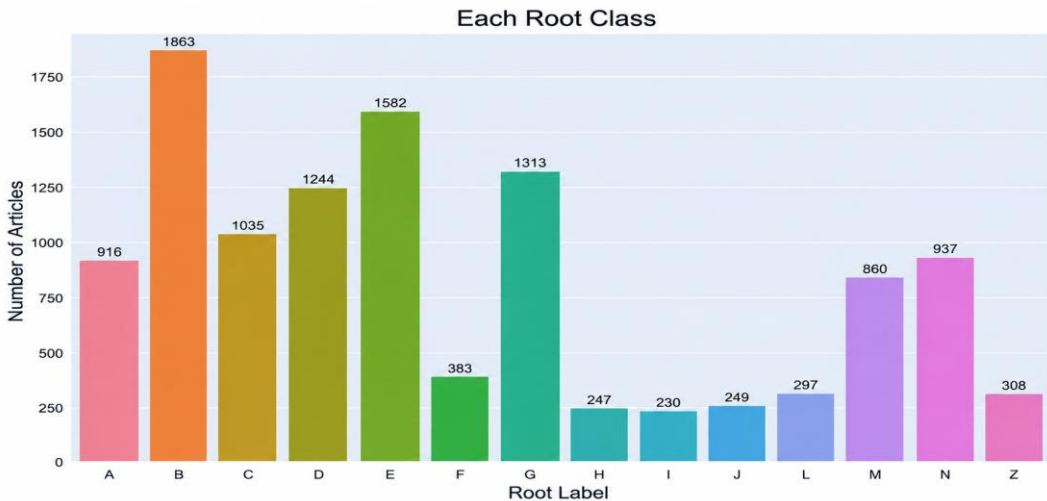


Figure 2: Root-label distribution reported in the source manuscript

These patterns should guide targeted retraining, not blind model enlargement. Figure 5 also reports the distribution of root labels, and the classification task is clearly imbalanced.

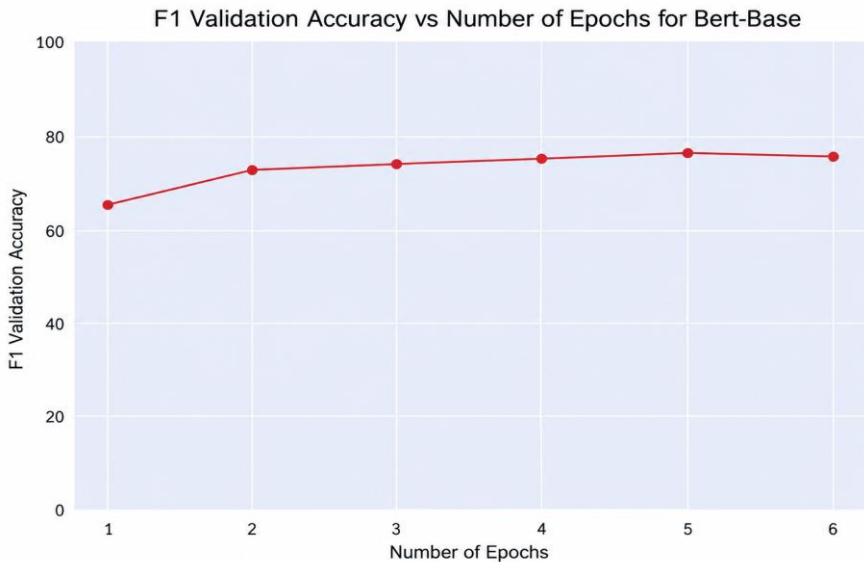


Figure 3: F1 validation accuracy versus number of epochs reported in the source manuscript

The number of articles per root label is highly skewed: some labels have over 1,500 articles, while others have only a few hundred or even fewer. This distribution reasonably explains the disparity between weighted and macro performance (Figure 4).

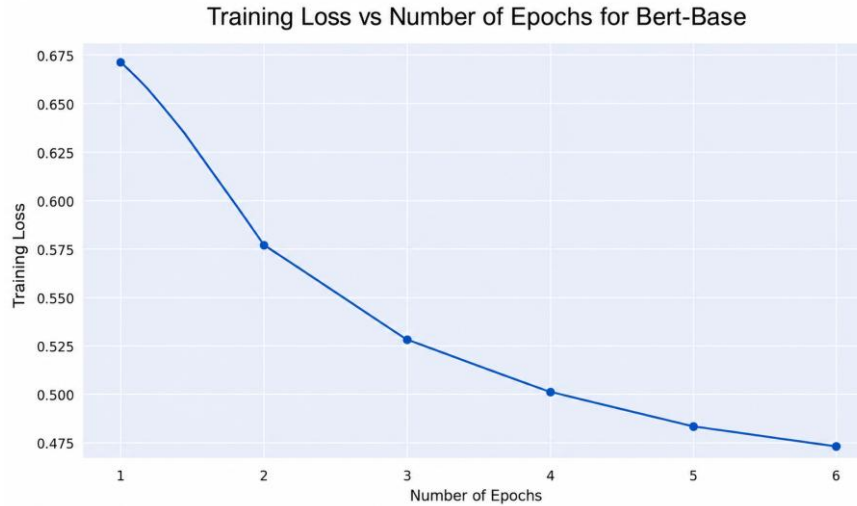


Figure 4: Training loss versus number of epochs reported in the source manuscript

It also recommends that the next iteration of the experiment include a confusion matrix and per-class error counts to identify the sources of minority-class failures directly. Table 1 shows class-wise performance¹ based on class-level classifications.

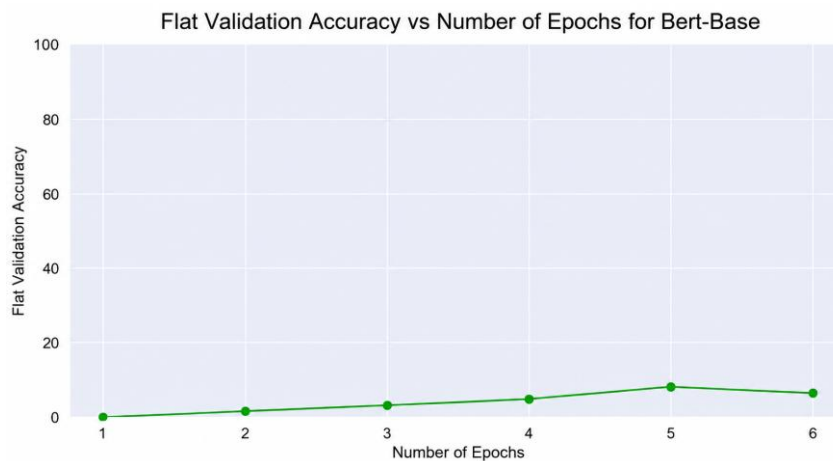


Figure 5: Flat validation accuracy versus number of epochs reported in the source manuscript

5. Discussions and Findings

The results show the potential and the challenges of hybrid machine learning methods for EHR text classification. The biggest difference between weighted and macro is the most salient observation. The weighted F1 is approximately 0.727, a useful aggregate measure of discrimination, while the macro F1 is approximately 0.500 and much lower. This difference stems from unequal class support and the model excluding several low-frequency classes. This suggests that further model development should prioritize sensitivity to minority classes over overall accuracy. Classes B, D, E, G and M performed well, suggesting the model can benefit from repetitive patterns across lexical and contextual levels when sufficient training examples are available. F1 values of B are close to 0.98, and E, D, G, and M are above 0.79. These types are more likely to be better represented in the training data and to have more unique textual evidence. Overall, the results show that fusing contextual BERT representations with explicit lexical and domain features is useful. Meanwhile, the low performance of F, H, I, J, L, and Z reflects the effects of low-class size. In particular, classes H, J, L, and Z have a reported F1 of zero, and I has an F1 below 0.18. F shows another failure mode: extremely high precision and extremely low recall. These patterns matter because they show the errors don't stem from a single cause. Any class that is not very well defined may be hard to predict correctly, as may be any class that contains not enough distinguishing features. One useful objective is interpretability. If the model can reveal attention patterns or explicit feature contributions, researchers can determine whether predictions are driven by medically relevant language or spurious correlations. For instance, if document length is too prominent, the model may correlate long notes with specific labels simply because they are most prevalent in the training data.

Likewise, relationship words or common bigrams might appear throughout the text as surrogates for specific categories rather than as expressions of the conceptual clinical meaning. Thus, interpretability analysis may serve not only as an additional feature but also as an error-detection mechanism. Robustness should be dealt with in the same way. The manuscript correctly recognizes adversarial vulnerability as a worry for clinical text systems. Medical notes may include errors of spelling, abbreviations, sections that are repeated, copied forward text, and noise. The system should retain predictions when words change but those changes are not important to the prediction, and it should respond appropriately when clinically meaningful words change. The current findings do not, however, suggest robustness, as the source manuscript fails to report an adversarial evaluation. This should be expressed as a design goal and an experimental direction, not as an empirical fact. The training curves show positive signs of stable optimization. In Figure 5, the training loss continually decreases over 6 epochs, while validation F1 increases and then levels off. This indicates that the training length is sufficient to achieve most of the reported gain, as shown in the experiment. However, as this difference between training behaviour and class-level test performance illustrates, independent test results should complement early stopping and model selection based on validation results. A model can learn well from the training data, but not benefit from rare classes. The study also has implications for hospital operations. The original manuscript focuses on the relationship between clinical text classification and DRGs and resource management. If clinical reports can be precisely classified, the results would be useful for administrative coding, patient flow analysis, bed planning, staff planning, equipment planning, medication planning and so on in the future.

It is therefore not surprising that researchers propose integrating predictive modelling with resource-management systems. But to be used in operations, predictions would need external verification, validation over time, fairness assessment, monitoring, and human supervision before they could inform high-stakes decisions. Finally, researchers determined the practical benefits of using lightweight and contextual representations. Another advantage of Hashing Vectorizer is that it avoids explicitly storing a large vocabulary, which can help with large-scale text processing. BERT adds contextual semantics, and manually engineered features give explicit signals. This hybrid architecture may be especially useful when computational resources or labelled data are (potentially) limited. It also creates a modular system where each component can be replaced or upgraded without redesigning the entire workflow. However, the existing assessment should be further enhanced. This experiment uses only one public medical-transcription dataset and an 80-20 split. Public transcription data can help with reproducibility, but it may differ from the language, documentation conventions, demographics and coding conventions of a specific hospital. A model that succeeds at one institution may suffer distribution shift at another. External validation with an independent dataset is essential to demonstrate generalizability. Another significant drawback is class imbalance. Several classes have fewer than 70 samples, and the largest class has 381. For most of the categories, Standard accuracy and Weighted F1 naturally give more weight to the dominant categories. Future work should explore class-weighted loss, focal loss, balanced sampling, targeted data augmentation, or active learning. Augmentation should maintain clinical meaning, though. Active learning may be especially useful because it does not require hiring experts for annotation and can be used only for unknown examples or for examples from poorly learned classes. The manuscript must also make its interpretability claims more concrete.

The manuscript must also make its interpretability claims more explicit. Simply stating that attention weights make the model interpretable is not enough. Assess attention using human judgment, perturbation tests, or user consensus with clinically meaningful annotations whenever possible. Likewise, measure feature importance for explicit features across folds and demographic/specialty subgroups to assess stability. This would make interpretability a system property, not just a conceptual feature. The literature reviewed in this manuscript supports this balanced perspective. The evidence for strong contextual modelling comes from transformer studies; active-learning studies show how to handle expensive annotations; domain-specific BERT studies show how to cope with limited data; and interpretable-rule studies show how to create human-readable evidence. The proposed system sits between these directions. The main benefit is not just adopting BERT or hashing, but fusing efficiency, domain knowledge, contextual learning, interpretability, and robustness into the same pipeline. In summary, the findings indicate that the proposed system is a promising research prototype rather than a finished clinical decision-support product. Its aggregate performance is meaningful, its training behaviour is stable, and several well-supported classes are recognized strongly. Yet the low macro performance and zero-F1 minority classes require substantial attention. The most productive next step is a controlled experimental study that compares the hybrid system with BERT-only, hashing-only, classical machine-learning, and domain-specific transformer baselines under identical splits, followed by ablation, cross-validation, external validation, robustness testing, and clinician-centred interpretability evaluation.

6. Conclusion

The authors propose a hybrid approach to medical transcription classification that involves preprocessing, domain-informed feature extraction, Hashing Vectorizer representation, and contextual learning with BERT. Experiments on the Medical Transcriptions Dataset achieved about 74.2% test accuracy, 74.4% weighted precision, 74.2% weighted recall, and 72.7% weighted F1 over 2321 samples. The training loss gradually decreases over 6 epochs, and the validation F1 gradually increases to around 77%. The results show that the proposed approach is feasible for heterogeneous medical text. The macro F1 is around 50.0%, while several low-support classes have an F1 of 0.0%, indicating significant class imbalance and generalization

difficulty. Thus, these findings should be considered a methodological reference, not a basis for clinical use at this time. Future studies should integrate balanced learning, active learning, enhanced domain-specific pretrained models, repeated cross-validation, external datasets, robustness testing and clinically validated interpretability methods. Such integration could support predictive planning for beds, staffing, equipment, meds, and patient flow within hospital resource-management systems. This deployment would continue to involve human oversight and thorough validation to ensure reliable, transparent, and safe automated predictions.

6.1. Future Work

Future development should focus on six areas. To deal with class imbalance, the first approach is class-weighted objectives, balanced sampling, focal loss, or carefully controlled augmentation. Secondly, active learning can lower the number of annotations needed by focusing on uncertain or underrepresented examples. Third, evaluate the hybrid system against the same baselines as BERT-only, Hashing Vectorizer-only, classical machine learning and domain-specific transformer baselines. Fourth, include repeated stratified cross-validation and external validation to evaluate generalization. Fifth, the authors should explicitly assess robustness using clinically realistic perturbations and adversarial tests. Sixth, assess interpretability through clinician review and feature- and/or attention-perturbation experiments. The longer-term vision envisioned in the original manuscript is to integrate predictive models with hospital resource management systems, enabling the flexible deployment of beds, staff, equipment and medications. This integration should remain human-supervised and be implemented only after intensive clinical and operational testing.

Acknowledgment: The authors gratefully acknowledge the support of SRM Institute of Science and Technology at Ramapuram, Zagazig University, Infosys Limited, Amran Community College, McMaster University, and Singapore Management University. The authors sincerely appreciate their valuable support and resources.

Data Availability Statement: The data for this study are available upon request from the corresponding author.

Funding Statement: This manuscript and research paper were prepared without any financial support or funding.

Conflicts of Interest Statement: The authors have no conflicts of interest to declare. This work represents a new contribution by the authors, and all citations and references are appropriately included based on the information utilized.

Ethics and Consent Statement: This research adheres to ethical guidelines and obtains informed consent from all participants. Confidentiality measures were implemented to safeguard participant privacy.

References

1. G. López-García, J. M. Jerez, N. Ribelles, E. Alba, and F. J. Veredas, "Transformers for clinical coding in Spanish," *IEEE Access*, vol. 9, no. 5, pp. 72387–72397, 2021.
2. C. A. Flores, R. L. Figueroa, and J. E. Pezoa, "Active learning for biomedical text classification based on automatically generated regular expressions," *IEEE Access*, vol. 9, no. 3, pp. 38767–38777, 2021.
3. K. Roitero, B. Portelli, M. H. Popescu, and V. D. Mea, "DiLBERT: Cheap embeddings for disease related medical NLP," *IEEE Access*, vol. 9, no. 11, pp. 159714–159723, 2021.
4. S. Gao, M. Alawad, M. T. Young, J. Gounley, N. Schaefferkoetter, H. J. Yoon, X. C. Wu, E. B. Durbin, J. Doherty, A. Stroup, L. Coyle, and G. Tourassi, "Limitations of transformers on clinical text classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 9, pp. 3596–3607, 2021.
5. Q. Jin, X. Xue, W. Peng, W. Cai, Y. Zhang, and L. Zhang, "TBLC-rAttention: A deep neural network model for recognizing the emotional tendency of Chinese medical comment," *IEEE Access*, vol. 8, no. 5, pp. 96811–96828, 2020.
6. R. G. Zhou, S. Chang, and Y. Li, "A neural network architecture for information extraction in Chinese drug package insert," *IEEE Access*, vol. 8, no. 3, pp. 51256–51264, 2020.
7. C. Ruan, Y. Wu, G. Sheng Luo, Y. Yang, and P. Ma, "Relation extraction for Chinese clinical records using multi-view graph learning," *IEEE Access*, vol. 8, no. 11, pp. 215613–215622, 2020.
8. G. T. Berge, O. C. Granmo, T. O. Tveit, M. Goodwin, L. Jiao, and B. V. Matheussen, "Using the Tsetlin machine to learn human-interpretable rules for high-accuracy text categorization with medical applications," *IEEE Access*, vol. 7, no. 8, pp. 115134–115146, 2019.
9. M. Shoenybi, M. Patwary, R. Puri, P. LeGresley, J. Casper, and B. Catanzaro, "Megatron-LM: Training multibillion parameter language models using model parallelism," *arXiv preprint arXiv:1909.08053*, 2019. [Accessed by 17/06/2025].

10. I. Beltagy, M. E. Peters, and A. Cohan, “Longformer: The long-document transformer,” *arXiv preprint arXiv:2004.05150*, 2020. [Accessed by 10/06/2025].
11. F. Petroni, T. Rocktäschel, P. Lewis, A. Bakhtin, Y. Wu, A. H. Miller, and S. Riedel, “Language models as knowledge bases?” *arXiv preprint arXiv:1909.01066*, 2019. [Accessed by 03/06/2025].
12. Y. Liu and M. Lapata, “Text summarization with pretrained encoders,” in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, Hong Kong, China, 2019.
13. J. R. A. Solares, F. E. D. Raimondi, Y. Zhu, F. Rahimian, D. Canoy, J. Tran, A. C. P. Gomes, A. H. Payberah, M. Zottoli, M. Nazarzadeh, N. Conrad, K. Rahimi, and G. Salimi-Khorshidi, “Deep learning for electronic health records: A comparative review of multiple deep neural architectures,” *Journal of Biomedical Informatics*, vol. 101, no. 1, pp. 1–15, 2020.
14. X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, “Pre-trained models for natural language processing: A survey,” *Science China Technological Sciences*, vol. 63, no. 10, pp. 1872–1897, 2020.
15. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Proc. 31st Int. Conf. Neural Information Processing Systems (NeurIPS)*, Long Beach, California, United States of America, 2017.
16. Y. Zhang, Q. Chen, Z. Yang, H. Lin, and Z. Lu, “BioWordVec: Improving biomedical word embeddings with subword information and MeSH,” *Scientific Data*, vol. 6, no. 5, pp. 1–9, 2019.
17. C. Sun, X. Qiu, Y. Xu, and X. Huang, “How to fine-tune BERT for text classification?” in *Chinese Computational Linguistics*, Springer, Cham, Switzerland, 2019.
18. J. Zhu, H. Wang, E. Hovy, and M. Ma, “Condensed-based stopping criteria for active learning for data annotation,” *ACM Transactions on Speech and Language Processing*, vol. 6, no. 3, pp. 1–24, 2010.
19. J. Pennington, R. Socher, and C. Manning, “GloVe: Global vectors for word representation,” in *Proc. 2014 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014.
20. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. 2019 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Minneapolis, Minnesota, 2019.

Publisher’s Note: The publisher remains impartial concerning jurisdictional claims in published maps and institutional affiliations. Responsibility for the content rests entirely with the authors and does not necessarily reflect the publisher’s perspectives.