Enhanced NLP for Medical Text Classification: A Deep Active Learning Approach

Palaparthi Seethalakshmi
Centurion University of Technology and Management, Odisha, India
seethalakshmi.palaparthi1983@gmail.com (corresponding author)

Dhawaleshwar Rao CH
Department of Computer Science and Engineering, Centurion University of Technology and Management, Odisha, India
dhawaleswarrao@gmail.com

K. Swaroopa
Department of Computer Science and Engineering (Data Science), Aditya Institute of Technology and Management, Tekkali, Srikakulam Andhra Pradesh, India
swaroopachalam@gmail.com

Received: 13 May 2025 | Revised: 17 June 2025, 8 July 2025, and 22 July 2025 | Accepted: 27 July 2025

 DOI: https://doi.org/10.48084/etasr.12114

ABSTRACT

This paper presents an enhanced approach for classifying medical texts, combining Deep Active Incremental Learning (AIL) with state-of-the-art techniques to optimize healthcare authorization decisions. Using a Bi-LSTM architecture that is enhanced with contextual embedding and attention mechanisms, the model can dynamically learn from a few labeled data and update its predictions in real-time via entropy-based uncertainty sampling. The proposed framework adopted SMOTE and undersampling strategies. 117,000 actual medical authorization submissions were semantically processed through BioBERT embeddings and Named Entity Recognition (NER). The experimental results show that after 100 active phases of learning, the model achieved a gain of 4% balanced accuracy, indicating its ability to iteratively optimize predictions with minimal guidance. Through the optimization of performance in a constrained resource environment, this approach also enables faster and more efficient processing of medical claims, which can help build scalable and adaptive decision-making capacities.

Keywords-NLP; text classification; active-learning; machine learning; deep learning

I. INTRODUCTION

The healthcare domain is marked by the creation of a lot of data that ranges from electronic health records, clinical notes, medical claims, and paperwork. The ability to uncover meaningful insights from such datasets is critical to improving patient outcomes, determining clinical choices, and reducing expenses. The richness and complexity of healthcare data challenge the task of effective analysis and interpretation. In response to these challenges, attention has been drawn to medical text classification using Natural Language Processing (NLP) techniques. This study focuses on automating the classification of healthcare authorization requests to distinguish between legitimate and potentially fraudulent medical claims, reducing processing time and improving accuracy in insurance decision-making processes [1].

Several studies have explored medical text classification using deep learning approaches. In [2], an active learning model was used for the classification of clinical text, achieving better performance than random sampling with 78.5% accuracy. In [3], a deep active learning model was used to classify cancer pathology reports, reaching 85% accuracy but requiring full dataset retraining for each update cycle. In [4], a supervised machine learning model was developed with active learning for radiology reports classification, showing 82.3% accuracy but limited to single-domain applications without real-time adaptation capabilities. In [5], a Convolutional Neural Network (CNN) model for medical text classification achieved 79.1% accuracy on clinical notes but lacked active learning integration. In [6], a neural network-based model for medical text classification achieved 76.8% balanced accuracy, but without addressing class imbalance issues inherent in medical datasets.

The use of medical text classification in processing healthcare authorization requirements is increasing. Such operations require the prior approval of the healthcare insurers before clinically indicated services or prescriptions are delivered to patients. Solving such tasks usually takes a lot of

time, but can be streamlined and automated through medical text request classification. Medical text classification is promising, but faces complex and dynamic data issues that make it considerably difficult to develop effective classification models [7, 8]. Previous efforts to apply conventional machine learning algorithms to these classification problems faced limitations, especially when processing large datasets [8]. Therefore, there is a need to introduce new solutions to improve traditional approaches.

Deep Learning (DL) can process complex, high-dimensional, and unstructured data to automatically identify features [9]. In contrast, Active Learning (AL) is famous for the acceleration and cost-efficiency of data labeling systems [10]. Since DL and AL support each other, their combination, Deep Active Learning (DAL), brings the positives of both to the maximum and reduces the negatives that come with them. Consequently, the expected performance is enhanced, and the applicability of both methods can be broadened.

This study attempted to surmount this critical research barrier by studying the use of DAL to enhance medical text classification. This work presents a novel method for medical text classification, combining deep neural networks, Active Incremental Learning (AIL), sophisticated NLP tools, such as contextual embedding and semantic text representation, along with an active sampling procedure to send uncertain or informative samples for annotation in the AL process, thus extending the model with new data. In addition, a set of real medical requests was used for the analysis during the investigation. The results of this study have a wide reach and could help strengthen the healthcare authorization process by speeding up decision-making, minimizing cost, and providing better care for patients [11].

II. METHODOLOGY

A. Data Collection and Preprocessing

This task involved assembling and preparing the data for medical text classification. The primary dataset comprised 117,342 clinical notes. This study extracted 85,000 clinical notes from the NOTEEVENTS table and 25,000 records from the ADMISSIONS table of MIMIC-III clinical database [12], selecting ICU patients with complete diagnostic codes (ICD-10). From MIMIC-IV [13], 5,000 updated discharge summaries were used from the DISCHARGE table. From the Clinical Practice Research Datalink database (CPRD) [14], 2,342 primary care authorization requests were incorporated, focusing on routine medical procedures and pharmaceutical requests, supplemented by additional medical records. The MIMIC-III database provides de-identified health data from ICU patients, including clinical notes, discharge summaries, and medical authorization records, while MIMIC-IV offers comprehensive medical records with enhanced data quality. CPRD contributes longitudinal health data from UK primary care settings, providing diverse healthcare authorization scenarios. Legitimate requests (117,068) and fraudulent requests (279) were classified based on ICD-10 code consistency and cost estimate validation rules. The features used are clinical notes, diagnostic codes, procedure descriptions, and cost estimates from each dataset.

The preprocessing procedures used NLP tools, including [15]:

- Named Entity Recognition (NER) to detect and classify diseases, medicines, and procedures, expanding the model's ability to understand the clinical background.
- Text normalization, utilizing special medical dictionaries to maintain consistent term descriptions across the dataset.
- Semantic analysis, including domain-specific semantic analysis to determine how medical vocabulary relates to each other.

Appropriate encoding is essential in properly handling categorical features within a neural network. One-hot coding is one of the standard ways to convert categorical variables to vectors, but it can lead to significant problems. This study exploited state-of-the-art word embedding methods, based on context information, to vectorize medical text data:

- Contextual embeddings were extracted using BioBERT, a clinical embedding based on BERT, to improve the accuracy of contextual word representation and the ability to represent medical jargon.
- Domain-adapted word vectors: Pre-trained embeddings were adapted to medical data to improve the medical-specific language representation. Pre-trained Word2Vec embeddings were fine-tuned on a medical corpus of 500,000 clinical notes, adapting general language vectors to better capture medical terminology relationships and semantic similarities specific to healthcare authorization contexts.

One-hot encoding was combined with these embedding-based approaches to treat categorical data. Categorical variables requiring one-hot encoding included: provider specialty codes (23 categories), procedure type classifications (15 categories), insurance plan types (8 categories), and geographic regions (12 categories).

Legitimate requests were characterized by: (i) ICD-10 codes that match documented symptoms and procedures, (ii) cost estimates within 50% of standard medical fee schedules, and (iii) appropriate provider credentials verification. Fraudulent requests exhibited: (i) ICD-10 code inconsistencies with reported symptoms, (iii) cost inflation exceeding 50% of standard rates, and (iii) missing or invalid provider information. This classification was performed by three medical coding specialists with 5+ years of experience, achieving 94% inter-rater agreement.

These data were imbalanced, a situation in which the distribution of classes in the training set is substantially uneven [16]. There was a very skewed class distribution with 117,068 legitimate requests and only 279 fraudulent requests. Most machine learning algorithms struggle to create models that accurately classify instances of the minority class, as the dataset is biased toward the majority class due to its predominance. Random sampling was used to reduce redundancy in the data. If a model sees redundant samples repeatedly for training, it can become biased and overfit, resulting in inaccurate predictions, and learning from new data

might be limited. The SMOTE algorithm was used to oversample minority class instances and balance the dataset.

After applying SMOTE to generate synthetic fraudulent samples and random undersampling to reduce the majority class, the final dataset achieved a 50:50 class distribution with equal representation of legitimate and fraudulent requests, resulting in 30,000 legitimate requests and 30,000 fraudulent requests (including 29,721 SMOTE-generated synthetic samples). This balancing approach eliminates class bias and provides an optimal foundation for robust model training and evaluation.

B. Model Design

A Bi-LSTM neural network is the backbone of the proposed model, augmented by attention mechanisms that highlight salient textual attributes. A stream-based AL scenario was used, along with a mix of uncertainty sampling to select samples and online learning to update the model and implement AIL. When samples enter a stream-based AL environment, uncertainty sampling facilitates the selection of informative samples from unlabeled data, submitted to human experts for validation, and updates the model accordingly. The model can be further trained incrementally in a real-time setting to load new data and enable continuous learning without the need to start re-training from scratch. The expert annotation team consisted of two certified medical coders (CCS credentials) and one healthcare fraud investigator with 10+ years of experience. Experts reviewed uncertain samples using standardized criteria: clinical suitability, cost reasonableness, and documentation completeness, with consensus required for final labeling. Figure 1 illustrates the proposed framework.

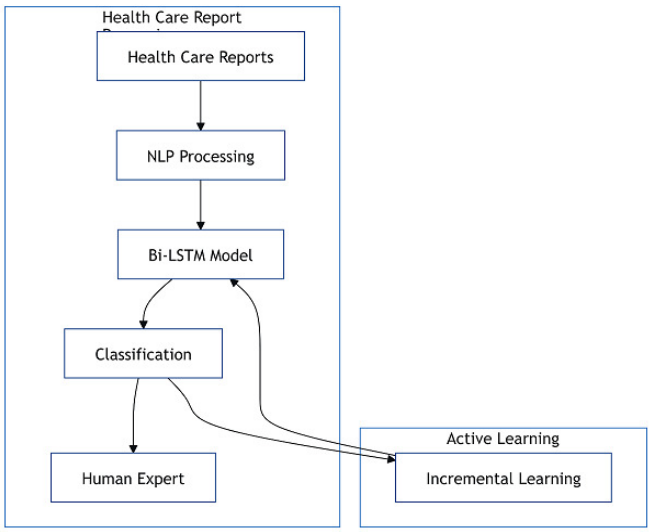

Fig. 1. Proposed framework.

The architecture of the proposed framework involved:

- Two Bi-LSTM layers (256 and 128 hidden units), multi-head attention (8 heads), dropout layers (0.3, 0.2, 0.1), dense layer (64 neurons), sigmoid output.
- BioBERT embeddings (768-dim) with medical NER preprocessing.
- Entropy-based uncertainty sampling with threshold τ=0.8. Samples exceeding this threshold are forwarded for expert annotation and used for immediate model updates.
- Online SGD updates with momentum β =0.9, adaptive learning rate 0.0001, and class-weighted loss with dynamic weight adjustment.

C. Model Training

A set of preprocessed healthcare requests was used to train the model. Stratified k-fold cross-validation, dropout, and L2 regularization were employed to improve model performance and control overfitting issues. The Adam optimizer was used with the following parameters: 50 epochs, batch size 32, learning rate 0.001, L2 regularization λ=0.01, early stopping (*patience*=10), binary cross-entropy loss function, and sigmoid activation. After model training and evaluation, the stage of active incremental learning was started. The model gives predictions for each new batch of data samples, and human experts selected the samples with the greatest model entropy for manual annotation. A threshold of 0.8 was used for entropy to extract the data for annotation.

In turn, the model was constantly updated with the help of newly labeled samples. After 100 iterations, the performance of the model was tested using the test set. The balanced dataset of 60,000 samples was divided into 70% training (42,000 samples), 15% validation (9,000 samples), and 15% test (9,000 samples) using stratified sampling to maintain class distribution.

A labeled healthcare dataset is defined as:

$$Do = \{(a_{1,}\, b_1\,), (a_{2,}\, b_2\,) \ldots . (a_{m,}\, b_m\,) \quad (1)$$

where a_i represents a healthcare request sample and b_i denotes its associated class label. A model, denoted as M_0, is trained using this dataset to learn a mapping from input a to the predicted output $\hat{y}$, i.e., $M_0(a) \rightarrow \hat{y}$. Training is performed by minimizing a loss function ℓ, which measures the difference between predicted and actual labels, yielding the optimal model parameters φ^*. A stream of new, unlabeled healthcare request samples is represented as:

$$A_{str} = \{a_{1,} a_2 \ \ldots \ldots \ldots a_m\} \quad (2)$$

where each a_j is a new input. For every incoming sample, the model predicts the class probability distribution. Prediction uncertainty is calculated using entropy:

$$H(a_j\,) = \sum P\left(b = {}^{c}/_{a_j}\,\right).log\, P\left(b = {}^{c}/_{a_j}\,\right) \quad (3)$$

If the entropy exceeds a predefined threshold τ, the sample is flagged for manual labeling. A domain expert (oracle) provides the correct label b_j^* for the selected sample a_j, resulting in a new labeled pair $(a_j, b_j)^{**}$.

The model is then updated with the new labeled data. To compensate for class imbalance, a weight $\omega_{c,j}$ is used to adjust each sample's loss contribution:

$$l_{update} = \Sigma_j \Sigma_c \omega_c . j . l(Mo(a_j, b_j)) \quad (4)$$

This ensures that minority classes are appropriately emphasized during training. The combined use of active learning and incremental model updating allows continuous performance improvement through selective labeling, without the need to retrain the model from scratch.

D. Dataset Considerations

This study uses only medical datasets (MIMIC-III, MIMIC-IV, CPRD). All are publicly available and de-identified, with no personally identifiable information or human subject data collected during this study, requiring no ethical approval [17]. This ensures both transparency and reproducibility of the research process.

E. Evaluation Metrics

Although accuracy is commonly used as a base measure of classifier performance in the field of machine learning, flaws arise when highly imbalanced datasets are involved. If the classifiers place a higher value on the majority class, they can overestimate the results. Thus, accuracy does not offer useful information on the correct classifications of individual classes. Therefore, balanced accuracy was adopted as the preferred evaluation metric. To take into account the two classes equally, the balanced accuracy is calculated using the mean of sensitivity and specificity. Therefore, it provides a more reliable approach to gauge how well the model works with an imbalanced class distribution.

To improve the standard Bi-LSTM architecture, the implementation incorporated the following:

- Attention mechanism: A hierarchical attention strategy was applied to highlight significant words and sentences, increasing the model's ability to learn relevant diagnostic and procedural information.
- Contextual language representation: Using transformer-based embeddings, the model obtains an interpretative advantage because it collects richer contextual information, leading to a collectively better semantic understanding of medical language.
- Document-level feature extraction: A document-level feature extraction process was employed to derive a clear perspective of the semantic allocation of text in medical documents.

III. RESULTS

The results in Table I reveal a comparison between the model's balanced accuracy in the validation and test sets at the preliminary assessment. The baseline model was a traditional Naïve Bayes classifier using lexical features (domain length, digit ratio, character entropy), which achieved 70.45% balanced accuracy on the validation set and 73.72% on the test set. The base Bi-LSTM model involved Word2Vec embeddings and a single attention layer, achieving 75.32% validation accuracy and 79.32% test accuracy. The Bi-LSTM model with NLP enhancements incorporated BioBERT embeddings, attention mechanisms, and NER, achieving the highest performance with 78.2% validation accuracy and 82.1% test accuracy.

TABLE I. MODEL RESULTS BEFORE AIL

Models	Balanced accuracy	
	Validation set	Test set
Baseline	0.7045	0.7372
Bi-LSTM	0.7532	0.7932
Bi-LSTM with NLP	0.782	0.821

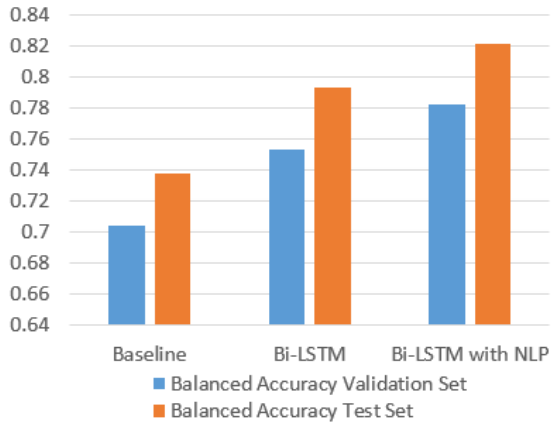

Fig. 2. Model results before AIL.

The validation set was used for hyperparameter tuning and model selection during training. The test set was an independent evaluation set, never used during training, providing an unbiased performance assessment.

The capacity of the augmented Bi-LSTM model to adapt was tested on more than 100 sample queries in successive steps. The iterative learning approach allows for substantive prediction refinement in every AL cycle. Table II shows the results, where the performance of the proposed model was improved with remarkably fewer labeled samples, highlighting its effectiveness with scarce data and offering a good approach for the medical domain, where resources may be limited. Figure 3 displays how the performance of the proposed model was progressively improved in successive rounds of query refinement/model adjustments, highlighting its potential for growth in multiple training rounds. Through the iterative querying and updating of the model, a consistent 4% improvement in performance was achieved.

TABLE II. MODEL RESULTS AFTER AIL

	Number of query iterations			
Models	20 queries	40 queries	60 queries	100 queries
Baseline+AIL	0.7391	0.7021	0.731	0.764
Bi-LSTM+AIL	0.7521	0.7562	0.7692	0.7952
Bi-LSTM+NLP+AIL	0.7851	0.7912	0.8081	0.8342

The values in Table II were obtained through the following AIL process:

- Initial training: Models were trained on 80% of the labeled training dataset using stratified sampling to maintain class distribution.
- Stream simulation: The remaining 20% was used to simulate a stream of unlabeled healthcare requests, processed sequentially.

- Uncertainty sampling: For each incoming sample, entropy-based uncertainty was calculated using (3).
- Query selection: Samples with an entropy of more than the 0.8 threshold were selected for manual annotation by domain experts.
- Incremental updates: The model was updated using selected samples with class-weighted loss (4) to handle imbalance.
- In performance evaluation, 20 queries refer to 20 samples with an entropy of more than the 0.8 threshold, submitted for expert labeling, then used for incremental model updates. The batch size was chosen for practical annotation constraints and computational efficiency. After each batch, the model was evaluated on the held-out test set to measure balanced accuracy improvement.
- Iterative process: This process continued for 100 iterations, with performance measured at 20, 40, 60, and 100 query intervals.

IV. CONCLUSION

This study presented a novel deep AIL technique in the classification of medical texts that involved improved NLP methods to address the main issues related to the field and reduce the necessity of human annotation by prioritizing informative samples. The ability of the proposed model to keep itself updated with new information ensures an ever-increasing performance gain, which lessens the need for massive labeled training data. Experimental results support the conclusion that the model is capable of learning from a small dataset while improving itself as new labeled samples are present. The flexibility of the model during its creation, starting from the initial results up to the continuous growth, emphasizes its utility for effective and continuous medical analysis to help insurers guide medical decisions more quickly and, as a result, improve patient outcomes.

The primary contributions of this work are fourfold:

1. A novel integration of deep AIL with state-of-the-art NLP techniques, specifically combining Bi-LSTM with BioBERT embeddings and attention mechanisms for medical text classification.
2. An innovative entropy-based uncertainty sampling strategy that enables dynamic model updates with minimal human annotation, achieving a 4% improvement in balanced accuracy over 100 active learning iterations.
3. A comprehensive framework that addresses class imbalance in medical authorization data through SMOTE and undersampling, demonstrating scalability on a real-world dataset of healthcare authorization requests.
4. A practical solution for healthcare authorization processing that enables real-time decision-making while continuously improving model performance through selective learning, thus reducing costs and improving patient care outcomes in resource-constrained medical environments.

REFERENCES

[1] X. Chen and Y. Du, "Enhancing medical text classification with GAN-based data augmentation and multi-task learning in BERT," *Scientific Reports*, vol. 15, no. 1, Apr. 2025, Art. no. 13854, https://doi.org/10.1038/s41598-025-98281-9.

[2] R. L. Figueroa, Q. Zeng-Treitler, L. H. Ngo, S. Goryachev, and E. P. Wiechmann, "Active learning for clinical text classification: is it better than random sampling?," *Journal of the American Medical Informatics Association*, vol. 19, no. 5, pp. 809–816, Sep. 2012, https://doi.org/10.1136/amiajnl-2011-000648.

[3] K. De Angeli *et al.*, "Deep active learning for classifying cancer pathology reports," *BMC Bioinformatics*, vol. 22, no. 1, Dec. 2021, Art. no. 113, https://doi.org/10.1186/s12859-021-04047-1.

[4] D. H. M. Nguyen and J. D. Patrick, "Supervised machine learning and active learning in classification of radiology reports," *Journal of the American Medical Informatics Association*, vol. 21, no. 5, pp. 893–901, Sep. 2014, https://doi.org/10.1136/amiajnl-2013-002516.

[5] M. Hughes, I. Li, S. Kotoulas, and S. Toyotaro, "Medical Text Classification Using Convolutional Neural Networks," in *Studies in Health Technology and Informatics*, IOS Press, 2017, pp. 246–250.

[6] L. Qing, W. Linhong, and D. Xuehai, "A Novel Neural Network-Based Method for Medical Text Classification," *Future Internet*, vol. 11, no. 12, Dec. 2019, Art. no. 255, https://doi.org/10.3390/fi11120255.

[7] A. Salau, N. A. Nwojo, M. M. Boukar, and O. Usen, "Advancing Preauthorization Task in Healthcare: An Application of Deep Active Incremental Learning for Medical Text Classification," *Engineering, Technology & Applied Science Research*, vol. 13, no. 6, pp. 12205–12210, Dec. 2023, https://doi.org/10.48084/etasr.6332.

[8] Y. Wang *et al.*, "A clinical text classification paradigm using weak supervision and deep representation," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, Dec. 2019, https://doi.org/10.1186/s12911-018-0723-6.

[9] Z. Shen and S. Zhang, "A Novel Deep-Learning-Based Model for Medical Text Classification," in *Proceedings of the 2020 9th International Conference on Computing and Pattern Recognition*, Xiamen, China, Oct. 2020, pp. 267–273, https://doi.org/10.1145/3436369.3436469.

[10] B. Settles, "Active Learning Literature Survey," University of Wisconsin-Madison Department of Computer Sciences, Technical Report, 2009.

[11] N. Nissim *et al.*, "An Active Learning Framework for Efficient Condition Severity Classification," in *Artificial Intelligence in Medicine*, vol. 9105, J. H. Holmes, R. Bellazzi, L. Sacchi, and N. Peek, Eds. Springer International Publishing, 2015, pp. 13–24.

[12] A. E. W. Johnson *et al.*, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, no. 1, May 2016, Art. no. 160035, https://doi.org/10.1038/sdata.2016.35.

[13] A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. A. Celi, and R. Mark, "MIMIC-IV." PhysioNet, https://doi.org/10.13026/S6N6-XD98.

[14] E. Herrett *et al.*, "Data Resource Profile: Clinical Practice Research Datalink (CPRD)," *International Journal of Epidemiology*, vol. 44, no. 3, pp. 827–836, Jun. 2015, https://doi.org/10.1093/ije/dyv098.

[15] F. K. Khattak, S. Jeblee, C. Pou-Prom, M. Abdalla, C. Meaney, and F. Rudzicz, "A survey of word embeddings for clinical text," *Journal of Biomedical Informatics*, vol. 100, 2019, Art. no. 100057, https://doi.org/10.1016/j.yjbinx.2019.100057.

[16] M. Badawy, N. Ramadan, and H. A. Hefny, "Big data analytics in healthcare: data sources, tools, challenges, and opportunities," *Journal of Electrical Systems and Information Technology*, vol. 11, no. 1, Dec. 2024, Art. no. 63, https://doi.org/10.1186/s43067-024-00190-w.

[17] F. Dernoncourt, J. Y. Lee, O. Uzuner, and P. Szolovits, "De-identification of patient notes with recurrent neural networks," *Journal of the American Medical Informatics Association*, vol. 24, no. 3, pp. 596–606, May 2017, https://doi.org/10.1093/jamia/ocw156.