

Explainable zero-shot learning and transfer learning for real time Indian healthcare

Swati Saigaonkar, Vaibhav Narawade

Department of Computer Engineering, Ramrao Adik Institute of Technology, D Y Patil Deemed to be University, Nerul, India

Article Info

Article history:

Received Aug 28, 2024

Revised Oct 15, 2024

Accepted Nov 19, 2024

Keywords:

DistilBERT
Explainability
Indian clinical notes
Transfer learning
Zero-shot learning

ABSTRACT

Clinical note research is globally recognized, but work on real-time data, particularly from India, is still lagging. This study initiated by training models on medical information mart for intensive care (MIMIC) clinical notes, focusing on conditions like chronic kidney disease (CKD), myocardial infarction (MI), and asthma using the structured medical domain bidirectional encoder representations from transformers (SMDBERT) model. Subsequently, these models were applied to an Indian dataset obtained from two hospitals. The key difference between publicly available datasets and real-time data lies in the prevalence of certain diseases. For example, in a real-time setting, tuberculosis may exist, but the MIMIC dataset lacks corresponding clinical notes. Thus, an innovative approach was developed by combining a fine-tuned SMDBERT model with a customized zero-shot learning method to effectively analyze tuberculosis-related clinical notes. Another research gap is the lack of explainability because deep learning (DL) models are inherently black-box. To further strengthen the reliability of the models, local interpretable model-agnostic explanations (LIME) and shapley additive explanations (SHAP) explanations were projected along with narrative explanations which generated explanations in a natural language format. Thus, the research provides a significant contribution with ensemble technique of zero-shot learning and SMDBERT model with an accuracy of 0.92 as against the specialized models like scientific BERT (SCIBERT), biomedical BERT (BIOBERT) and clinical BioBERT.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Swati Saigaonkar

Department of Computer Engineering, Ramrao Adik Institute of Technology

D Y Patil Deemed to be University

Nerul, India

Email: swa.sai.rt21@dypatil.edu

1. INTRODUCTION

The adoption of electronic health record (EHR) systems is increasing significantly [1]. Publicly accessible resources such as the medical information mart for intensive care (MIMIC), and enhanced critical care unit (eICU) collaborative research databases provide valuable information on critically ill ICU patients. Specifically, the MIMIC III database includes data from over 50,000 ICU admissions [2], and the eICU dataset covers nearly two million admissions [3]. According to [4], while these datasets are widely available, there is a notable lack of datasets from Indian patient populations.

Chronic diseases are defined as conditions lasting more than three months [5], and early detection is crucial for improving diagnosis and management. Globally, heart failure affects approximately 26 million people annually, posing significant challenges for surgeons, cardiologists, and clinicians in predicting its

start [6]. Asthma, identified as the most prevalent chronic pulmonary condition worldwide, further underscores the need for effective management [7]. Chronic kidney disease (CKD), a progressive ailment affecting over 10% of the world's population, affecting over 800 million individuals [8]. Given their high prevalence, these three conditions are the primary focus of our research. During our analysis of Indian data, we encountered cases of tuberculosis. However, when applying the structured medical domain bidirectional encoder representations from transformers (SMDBERT) model, which was trained on the MIMIC dataset lacking tuberculosis cases, these instances were misclassified. This highlights the challenge of transferring models across datasets with different distributions. As demonstrated in [9], while models trained on one dataset often perform well on similar datasets, zero-shot learning techniques can be leveraged to address the limitations of datasets lacking sufficient examples.

Handling unstructured data, such as medical notes, presents several unique difficulties. One of the main difficulties arises from the diversity and lack of organization within the data [10], [11]. Clinical notes are often written in a phrase-like manner without standardized grammatical structures, which complicates their analysis. In the case of Indian data, there is an added complexity due to the predominance of non-digital formats, which makes it challenging to integrate this information seamlessly into digital systems. These clinical notes were not originally digital, in contrast to datasets that are accessible to the public, complicating their accessibility and analysis.

The practice of transfer learning, which involves using a model created for one task as the foundation for a model on a different task, has demonstrated particularly effective in healthcare applications. Transformer-based models, such as BERT, have revolutionized natural language processing (NLP) by enabling models to understand context more deeply through attention mechanisms. These models excel in tasks involving unstructured clinical data, such as extracting meaningful insights from EHR and clinical notes, making them invaluable tools for advancing patient care.

The interpretability of predictive models using unstructured clinical data has grown to be an important field of study in recent years. As machine learning (ML) and NLP techniques continue to evolve, their application to clinical notes, a rich source of unstructured data, possesses the ability to transform healthcare. However, the complexity and opacity of these models, often referred to as "black boxes," pose significant challenges in clinical settings where transparency and understanding are paramount.

Transformer-based models are increasingly used in healthcare for various predictive tasks, including forecasting mortality rates [12], predicting patient readmissions [13], and estimating hospital stay durations [14]. These models have also proven effective in tasks such as extracting entities [15], [16], identifying phenotypic characteristics [17], [18], modeling patient trajectories, and elucidating relationships among different medical entities. The integration of transformer models into healthcare research demonstrates their adaptability and efficacy in tackling a wide range of issues, advancing clinical decision-making systems, and enhancing our understanding of complex medical conditions.

In healthcare data analysis, a range of methodologies are employed, from rule-based systems to advanced ML and deep learning (DL) techniques. Rule-based methods depend on established guidelines derived from specialized knowledge, but these systems can be inflexible when faced with novel data, often requiring updates or modifications to handle new concepts. ML methods, on the other hand, learn from data directly. Supervised learning, for example, uses labeled datasets to perform tasks like classification, while unsupervised learning identifies patterns in unlabeled data. ML algorithms such as logistic regression (LR), support vector machines (SVM), and random forests are often challenged by high-dimensional, multimodal datasets.

DL techniques, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformer models like BERT and distilled BERT (DistilBERT), have demonstrated superior performance in extracting features from raw data and understanding context. BERT is particularly effective for generating powerful representations from unlabeled data due to its ability to understand context in both directions, making it highly suitable for various ML tasks, including text classification. DistilBERT, a lighter and more efficient version of BERT, achieves similar performance levels while being computationally less demanding [19]. In specific domains such as healthcare, domain-adapted models like SMDBERT, which incorporate specialized medical knowledge, have shown to outperform more general models like DistilBERT due to their enhanced capability to leverage domain-specific information [20].

Research on using zero-shot learning models specifically for predicting tuberculosis (TB) is relatively limited. Most studies focus on more traditional ML approaches or DL techniques applied to medical imaging data like chest X-rays. While there have been some progress, the application of these models for TB prediction using clinical data or notes remains a relatively unexplored area. Capellan-Martin *et al.* [21] discusses the use of vision transformers (ViT) in a self-supervised learning paradigm for improved TB detection in chest X-rays. Zero-shot pediatric TB detection is a method, which has shown appreciable gains in TB detection efficacy. when compared to fully-supervised models. The work in [22] proposes a

generalized zero-shot learning (GZSL) method using self-supervised learning (SSL) for medical image classification. It focuses on training a feature generator and choosing anchor vectors from various disease classes. The work in [23] introduces a multi-label GZSL network that uses chest X-ray images to simultaneously predict several diseases, both seen and unseen. It uses a visual representation guided by semantics taken from a substantial corpus of medical literature.

There are some works using prompt-based large language models (LLMs). Zhu *et al.* [24] looks into the flexibility of LLMs like GPT-4 to EHR data for zero-shot clinical prediction. In crucial tasks like mortality, length-of-stay, and 30-day readmission, it exhibits enhanced prediction performance. Similarly, HealthPrompt [25], utilizes prompt-based learning, allowing pre-trained language models that don't require more training data to adjust to new tasks. The study demonstrates that HealthPrompt can function effectively and efficiently in capturing the context of clinical texts in various clinical NLP tasks, showcasing the ZSL's potential to improve clinical judgment and reduce dependency on large annotated datasets.

This overview highlights the untapped potential of zero-shot learning for TB prediction using clinical data, suggesting a promising direction for future research beyond the more commonly explored areas of medical imaging and EHR-based predictions. Predictive models using clinical notes face unique challenges due to the nature of the data. Several studies have highlighted the importance of model interpretability in healthcare. The interpretability of ML models is a critical aspect of healthcare, as understanding the rationale behind model predictions is essential for ensuring trust and improving patient care. Two prominent techniques for interpreting complex models are local interpretable model-agnostic explanations (LIME) and shapley additive explanations (SHAP).

The work in [26] highlights the application of LIME and SHAP in detecting alzheimer's disease, demonstrating how these methods can increase the transparency and reliability of artificial intelligence (AI)-based predictions. The review emphasizes that although LIME and SHAP provide significant insights into model decisions, they also have limitations, one being the need for tailoring to specific medical contexts. Another study [27] investigates the use of LIME and SHAP for autonomous disease prediction, noting that while these techniques excel in generating local explanations, they face challenges when applied to more complex models or larger datasets. The research underscores the importance of these interpretability methods in making AI-driven decisions more understandable, but also points out the difficulties in scaling these explanations effectively.

As a result, our research is motivated by the need to simplify explainability by using clear and accessible language. Additionally, we aim to address the gaps caused by varying disease characteristics across different countries, ensuring that the models could be effectively applied using transfer learning. Objectives of the paper:

- 1) Development of ensemble approach using SMDBERT and zero shot model for tuberculosis classification.
- 2) Development of simple yet effective narrative based interpretability model for transformer based models.

2. RESEARCH METHOD

2.1. Data pre-processing

The diseases targeted were Asthma, myocardial infarction (MI) and CKD with ICD-9 codes '49320', '5849', '41001', '41011', and '41021', respectively, where the codes '41001', '41011', and '41021' were grouped together under MI. The data used for generating the base model was MIMIC [2]. Initially, 170,446 samples were collected and then refined, narrowing the dataset down to 5,234 samples specifically focused on discharge summaries for analysis. SMDBERT model was fine-tuned using these clinical notes, extracted from the NOTEVENTS table of the MIMIC dataset, along with other structured data. Basic preprocessing was applied before model training, including converting text to lowercase, removing special characters, URLs, and non-alphanumeric elements.

After the initial phase, the models were used on clinical notes from two nursing homes of Mumbai, India. The notes were gathered during two periods: January to April 2023 and October to November 2022. A total of 145 clinical notes, specifically related to the diseases used for training and TB, were selected for analysis. Figures 1 and 2 show samples of clinical notes before and after basic pre-processing. An example of an anonymous Indian clinical note is shown in Figure 3.

2.2. Architecture

As the SMDBERT model gave better performance than DistilBERT, it was chosen as a base model. After training the model, the model was then applied on real time Indian data collected from 2 hospitals of Mumbai, India using transfer learning. Overall, more than 1,500 clinical notes, especially discharge summaries were collected, out of which 145 notes were used. The diseases targeted were the same as those

used for training, with the addition of TB, which is prevalent in India but notably absent from the publicly available dataset, which predominantly represents the United States. This highlights the regional variation in disease prevalence across different countries. Figure 4 depicts the architecture of application of SMDBERT model on the Indian dataset.

'Admission Date: [**2182-3-28**] Discharge Date: [**2182-4-1**]\n\nDate of Birth: [**2125-1-8**] Sex: F\n\nService: CCU\n\nHISTORY OF THE PRESENT ILLNESS: The patient is a 57-year-old\nfemale with a history of hypertension and\nhypercholesterolemia but no prior coronary artery disease\nhistory who presented with chest pain. The patient reports\nthe pain began at 4:00 p.m. on the day of admission. She\nreports that it had essentially started earlier in the day\nwith some GERD-like symptoms. She had taken Maalox but had\nno relief with this. The pain was intermittent through the\nday and then at 4:00 p.m. increased and was also associated\nwith nausea and diaphoresis. She reports that the pain shot\nthrough her back. The patient has never had these symptoms\nin the past. She arrived in the Emergency Room at 8:00 p.m.\nAt that time, the pain was unremitting.\n\nIn the Emergency Room, the patient was given Lopressor,\nnitroglycerin, and aspirin. A TP3A inhibitor was not given\nfor concern for dissection. The patient was found to have [**Street Address(2) 12501**] elevations in I, aVL, and V1 through V2 and ST\ndepressions in II, III, and aVF.\n\nThe patient was taken emergently to the Cardiac\nCatheterization Laboratory.\n\nPAST MEDICAL HISTORY:\n1. Hypertension.\n2. Hypercholesterolemia.\n3. Chondrocalcinosis.\n4. Carpal tunnel syndrome.\n5. GERD.\n\nADMISSION MEDICATIONS:\n1. Atenolol 25.\n2. Crestor 10 q.d.\n3. Uniretic 15/25.\n\nALLERGIES: The patient has no known drug allergies.\n\nSOCIAL HISTORY: The patient denied tobacco. The patient\ndenied alcohol or other drug use.\n\nPHYSICAL EXAMINATION: Vital signs: On admission, the\npatient was afebrile, blood pressure 146/88, pulse 71,\nrespirations 18. General: The patient was a well appearing\nfemale in no acute distress. Cardiac: Regular rate and\nrhythm, normal S1, S2. No murmurs were appreciated. Lungs:\nClear to auscultation bilaterally. Abdomen: Soft,\nnontender, nondistended, good bowel sounds. Extremities:\nThere was no edema in the extremities. The dorsalis pedis\npulses were 2+ bilaterally.\n\nLABORATORY AND RADIOLOGIC DATA: On admission, white count\n10.8, hematocrit 38.2, platelets 324,000.\n\nHOSPITAL COURSE: 1. CORONARY ARTERY DISEASE: The patient\nwas taken to Cardiac Catheterization where she was found to\nhave two vessel disease. The proximal LAD was found to have\na 90% lesion with a septal branch with a 95% stenosis and a\n\ndiagonal branch with 99% stenosis. The patient had twin LAD\nstents to the LAD diagonal and LAD septal branches with Taxus\nstents. This restored flow appropriately. The patient also\nhad proximal RCA lesion of 80% and a mid RCA lesion of 80-90%\nwhich were not intervened upon. The patient was given an\naspirin and Plavix and in addition started on a beta [**Last Name (LF) 7005**],\n[**First Name3 (LF) **] ACE inhibitor, and these were titrated up as her blood\npressure tolerated. Given the fact that the patient had\nissues with muscle cramping with Atorvastatin in the past,\nshe was started on Pravastatin which should give less of\nthese side effects. The patient will follow-up with\nCardiology in one month's time for further management of her\ncoronary artery disease and further evaluation of her\nremaining RCA lesions.\n\n2. PUMP: The patient had an echocardiogram to assess her LV\nfunction. She was found to have an ejection fraction of\n30-35% as well as an akinetic apex and apical mid and septal\nakinesis. Given the fact that the patient had significant\napical akinesis, she was started on heparin with a transition\nto Coumadin for anticoagulation for stroke prevention in the\nsetting of apical akinesis.\n\nOn discharge, the patient was given Lovenox injections which\nshould be continued until the patient reaches a therapeutic\ndose of Coumadin.\n\n3. RHYTHM: The patient was monitored on telemetry\nthroughout her hospitalization with no significant events.\n\nThe patient had an EP evaluation and will be followed-up by\nDr. [**Last Name (STitle) **]. In-house, the patient had a signal-averaged\nEKG. She will follow-up with Dr. [**Last Name (STitle) **] with an\n\nechocardiogram on [**2182-5-17**] in a meeting to discuss risk\nstratification for sudden cardiac death and possible ICD\nplacement.\n\n4. NEUROLOGY: The patient complained of left lower\n\nextremity weakness with ambulation two days after her cardiac\ncatheterization. The patient had no evidence for weakness on\nexamination with good proximal and distal strength in the\nlower extremities as well as intact sensation. The patient\nworked with Physical Therapy and was able to ambulate without\ndifficulty. She was also able to ascend stairs without\ndifficulty. The patient was already on an aspirin and Plavix\nshould this represent a small stroke. However, there was no\nevidence for neurologic deficit on examination and this will\nbe followed-up by her primary care physician.\n\n5. HEMATOLOGY: The patient was discharged on Coumadin for\napical akinesis and stroke risk. This will be further\nmonitored by her primary care physician, [**Last Name (NamePattern4) **]. [**Last Name (STitle) 8499**] who will adjust her Coumadin dose.\n\nCONDITION ON DISCHARGE:

Figure 1. Clinical note from MIMIC

2.3. Method

The SMDBERT model was selected for its superior performance compared to other transformer-based models. SMDBERT incorporates additional knowledge by integrating symptom and disease information. The model was fine-tuned on data from the MIMIC database and then applied to real-time Indian data through transfer learning. Since TB was not included in the training set due to a lack of available data, a zero-shot model was employed for its classification. The architecture of ensemble model of SMDBERT and customised Zero shot model was constructed as given in Figure 5.

Admission Date: [2182-3-28] Discharge Date: [2182-4-1] Date of Birth: [2125-1-8] Sex: F Service: CCU HISTORY OF THE PRESENT ILLNESS: The patient is a 57-year-old female with a history of hypertension and hypercholesterolemia but no prior coronary artery disease history who presented with chest pain. The patient reports the pain began at 4:00 p.m. on the day of admission. She reports that it had essentially started earlier in the day with some GERD-like symptoms. She had taken Maalox but had no relief with this. The pain was intermittent through the day and then at 4:00 p.m. increased and was also associated with nausea and diaphoresis. She reports that the pain shot through her back. The patient has never had these symptoms in the past. She arrived in the Emergency Room at 8:00 p.m. At that time, the pain was unremitting. In the Emergency Room, the patient was given Lopressor, nitroglycerin, and aspirin. A TP3A inhibitor was not given for concern for dissection. The patient was found to have [Street Address(2) 12501] elevations in I, aVL, and V1 through V2 and ST depressions in II, III, and aVF. The patient was taken emergently to the Cardiac Catheterization Laboratory. PAST MEDICAL HISTORY: 1. Hypertension. 2. Hypercholesterolemia. 3. Chondrocalcinosis. 4. Carpal tunnel syndrome. 5. GERD. ADMISSION MEDICATIONS: 1. Atenolol 25. 2. Crestor 10 q.d. 3. Uniretic 15/25. ALLERGIES: The patient has no known drug allergies. SOCIAL HISTORY: The patient denied tobacco. The patient denied alcohol or other drug use. PHYSICAL EXAMINATION: Vital signs: On admission, the patient was afebrile, blood pressure 146/88, pulse 71, respirations 18. General: The patient was a well appearing female in no acute distress. Cardiac: Regular rate and rhythm, normal S1, S2. No murmurs were appreciated. Lungs: Clear to auscultation bilaterally. Abdomen: Soft, nontender, nondistended, good bowel sounds. Extremities: There was no edema in the extremities. The dorsalis pedis pulses were 2+ bilaterally. LABORATORY AND RADIOLOGIC DATA: On admission, white count 10.8, hematocrit 38.2, platelets 324,000. HOSPITAL COURSE: 1. CORONARY ARTERY DISEASE: The patient was taken to Cardiac Catheterization where she was found to have two vessel disease. The proximal LAD was found to have a 90% lesion with a septal branch with a 95% stenosis and a diagonal branch with 99% stenosis. The patient had twin LAD stents to the LAD diagonal and LAD septal branches with Taxus stents. This restored flow appropriately. The patient also had proximal RCA lesion of 80% and a mid RCA lesion of 80-90% which were not intervened upon. The patient was given aspirin and Plavix and in addition started on a beta [Last Name (LF) 7005], [First Name 3 (LF)] ACE inhibitor, and these were titrated up as her blood pressure tolerated. Given the fact that the patient had issues with muscle cramping with Atorvastatin in the past, she was started on Pravastatin which should give less of these side effects. The patient will follow-up with Cardiology in one month's time for further management of her coronary artery disease and further evaluation of her remaining RCA lesions. 2. PUMP: The patient had an echocardiogram to assess her LV function. She was found to have an ejection fraction of 30-35% as well as an akinetic apex and apical mid and septal akinesis. Given the fact that the patient had significant apical akinesis, she was started on heparin with a transition to Coumadin for anticoagulation for stroke prevention in the setting of apical akinesis. On discharge, the patient was given Lovenox injections which should be continued until the patient reaches a therapeutic dose of Coumadin. 3. RHYTHM: The patient was monitored on telemetry throughout her hospitalization with no significant events. The patient had an EP evaluation and will be followed-up by Dr. [Last Name (STitle)]. In-house, the patient had a signal-averaged EKG. She will follow-up with Dr. [Last Name (STitle)] with an echocardiogram on [2182-5-17] in a meeting to discuss risk stratification for sudden cardiac death and possible ICD placement. 4. NEUROLOGY: The patient complained of left lower extremity weakness with ambulation two days after her cardiac catheterization. The patient had no evidence for weakness on examination with good proximal and distal strength in the lower extremities as well as intact sensation. The patient worked with Physical Therapy and was able to ambulate without difficulty. She was also able to ascend stairs without difficulty. The patient was already on an aspirin and Plavix should this represent a small stroke. However, there was no evidence for neurologic deficit on examination and this will be followed-up by her primary care physician. 5. HEMATOLOGY: The patient was discharged on Coumadin for apical akinesis and stroke risk. This will be further monitored by her primary care physician, [Last Name (NamePattern4)]. [Last Name (STitle) 8499], who will adjust her Coumadin dose. CONDITION ON DISCHARGE: Stable. DISCHARGE STATUS: To home. DISCHARGE DIAGNOSIS: ST elevation myocardial infarction, status post left anterior descending artery stent. DISCHARGE MEDICATIONS: 1. Aspirin 325 q.d. 2. Lisinopril 5 q.d. 3. Toprol XL 100 q.d. 4. Coumadin 5 q.d. 5. Plavix 75 q.d. 6. Lovenox 60 mg b.i.d. until therapeutic on Coumadin. 7. Pravastatin 80 q.d. FOLLOW-UP PLANS: The patient will follow-up with her primary care physician in the week following discharge for

Figure 2. Pre-processed clinical note from MIMIC

Age: 45
 Gender: Female
 Final Diagnosis: Tuberculosis

Symptoms:
 Paper was brought by relative, was conscious and oriented. Chest pain, hemoptysis, night sweats, fatigue

H/o past illness: k/c/o:

Temp: 100.5
 Pulse: 96 /min
 Respiration: 19 /min
 BP: 110/68 mmHg
 SPO2: 93%
 Creatinine: 0.7 mg/dL
 Hb: 12.4 g/dL
 WBC: 13,800 /μL

Figure 3. Symptom extracted clinical note of Indian dataset

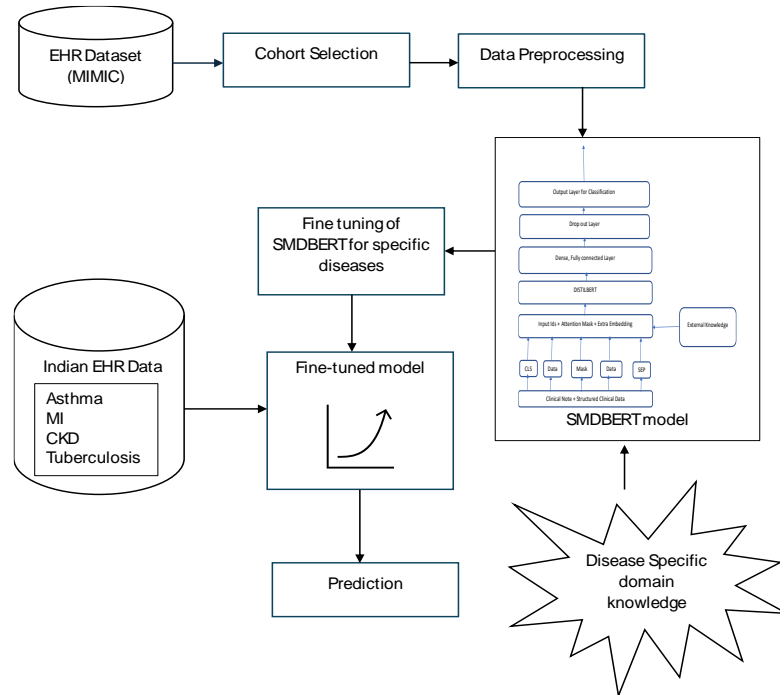


Figure 4. Architecture

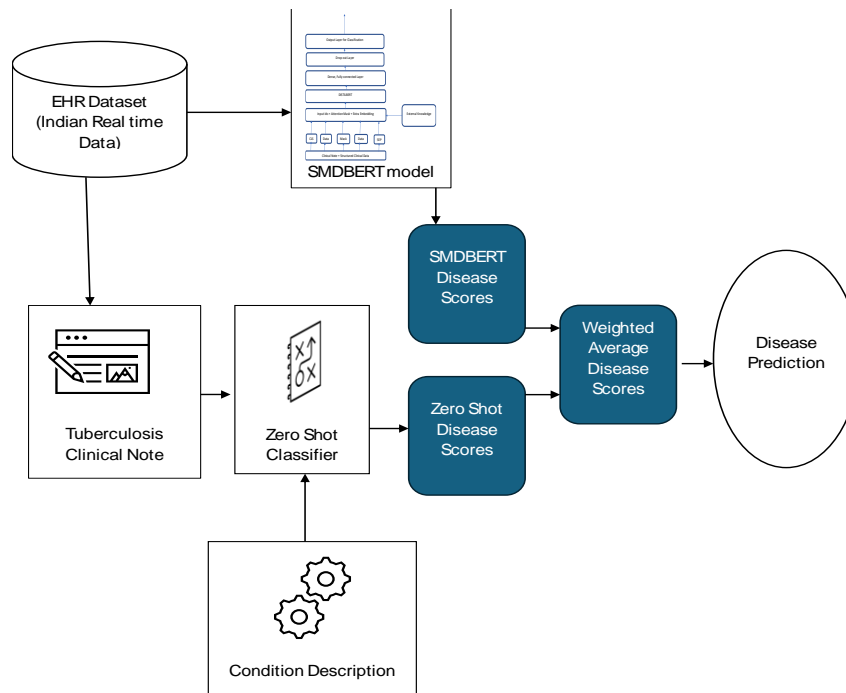


Figure 5. Ensemble architecture

The model can be expressed mathematically as:

Let S : set of input sentences (clinical notes).

D : set of disease labels, where $D=\{0,1,2,3\}$

$\text{SMD}_{\text{pred}}(s)$: probability distribution over the disease labels predicted by the SM-DBERT model for input sentence s .

$ZS_{pred}(s)$: probability distribution over the disease labels predicted by the zero-shot model for input sentence s .

α, β : weights assigned to SM-DBERT and zero-shot predictions, respectively, where $\alpha+\beta=1$.

$Combined_{pred}(s)$: final combined prediction score for each disease label $d \in D$ for sentence s

$FinalPred(s)$: the final predicted label for sentence s .

Steps:

SMDBERT prediction:

$$SMD_{pred}(s) = \left\{ P_{SMD} \left(\frac{d_i}{s} \right) \mid d_i \in D \right\}$$

where $P_{SMD} \left(\frac{d_i}{s} \right)$ is the probability assigned by the SMDBERT model to disease d_i for sentence s .

Zero-shot prediction:

$$ZS_{pred}(s) = \left\{ P_{ZS} \left(\frac{d_i}{s} \right) \mid d_i \in D \right\}$$

where $P_{ZS} \left(\frac{d_i}{s} \right)$ is the probability assigned by the zero shot model to disease d_i for sentence s .

Combined prediction:

$$Combined_{pred}(s) = \left\{ \alpha \cdot P_{SMD} \left(\frac{d_i}{s} \right) + \beta \cdot P_{ZS} \left(\frac{d_i}{s} \right) \mid d_i \in D \right\}$$

Here, α and β represent the relative importance or weight of the SMDBERT and zero-shot predictions, respectively.

Final prediction:

$$Final_{pred}(s) = \arg \max Combined_{pred}(s)$$

The final predicted label is the one with the highest combined prediction score.

3. RESULTS AND DISCUSSION

Table 1 shows the performance of different models on the Indian data. Models, namely BERT, DistilBERT and SMDBERT were initially fine tuned on MIMIC notes. Following the training phase, the models underwent practical application using Indian clinical records obtained from hospitals situated in Mumbai, India. A total of 145 clinical notes were meticulously selected, concentrating exclusively on the targeted diseases. The transcription of clinical notes was conducted manually due to the provision of scanned copies by the hospitals. It is worth noting that treatment specifics were intentionally excluded from the dataset to align with the research's focus on capturing symptoms. Specialized models were also tested on the Indian data collected. SMDBERT gave better results as compared to other models as it takes, symptom disease information as an additional embedding. As it can be seen our method with an ensemble of SMDBERT and zero shot learning gave better results with an accuracy of 0.924.

Table 1. Performance metrics of fine-tuned models on Indian data

Type of Model	Models	Accuracy	Precision	Recall	F1-score
Fine-tuned	BERT	0.55	0.5	0.55	0.52
	DistilBERT	0.64	0.65	0.64	0.62
	SMDBERT	0.75	0.69	0.75	0.72
Specialised	SCIBERT	0.76	0.83	0.76	0.72
	BIOBERT	0.62	0.52	0.62	0.54
	Clinical BioBERT	0.61	0.51	0.61	0.53
	Proposed Ensemble approach	0.924	0.927	0.924	0.925

In (1), (2), and (3) describes the computational formulations for precision, recall, and the F1-score, respectively.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (1)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (2)$$

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (3)$$

Figure 6 displays the models' performance graphically with Figure 6(a) showing model performance with respect to the specialized models and Figure 6(b) showing the proposed model performance with respect to fine-tuned models. As shown in the diagrams the ensemble approach of SMDBERT and zero shot has given better results than other fine-tuned models or specialized models. The specialized models considered are scientific BERT (SCIBERT), biomedical BERT (BIOBERT) and Bio_ClinicalBERT, while fine-tuned models are BERT and DistilBERT.

The confusion matrix demonstrated in Figure 7, shows that the model effectively classifies a majority of instances correctly, indicating strong performance and reliable predictive accuracy across multiple classes. Figure 8 shows the output of LIME, which explains predictions made by the ensemble approach by highlighting the contributions of different input features. Prediction probabilities section shows the predicted probabilities for each class. The model predicts this class with a probability of 0.80. The middle section shows the features that contribute to the prediction of the class. The right side of the diagram show the specific words or phrases, which are highlighted, reflecting their importance in the model's decision-making process.

Figure 9 shows the clinical note given as input, it includes various attributes like age, gender, final diagnosis, complaints on admission past medical history, vital signs, and other relevant clinical information. The horizontal bar at the top represents the SHAP value for the highlighted output which is Output 0. The values on the bar range from 0.0 to 1.0, indicating the probability or confidence of the model's prediction for each output.

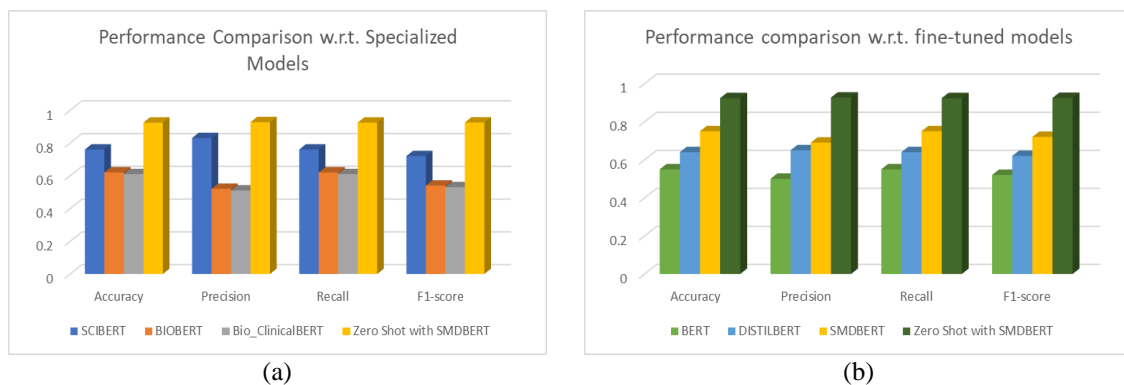


Figure 6. Graph displaying performance comparison of proposed approach (a) with specialized models (b) with fine-tuned models

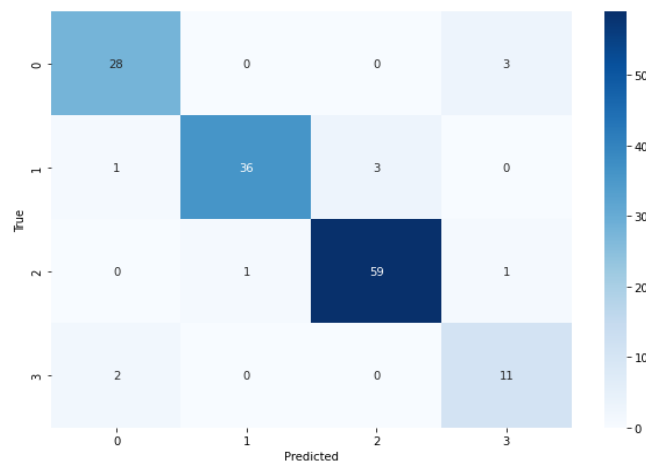


Figure 7. Confusion matrix of the proposed approach

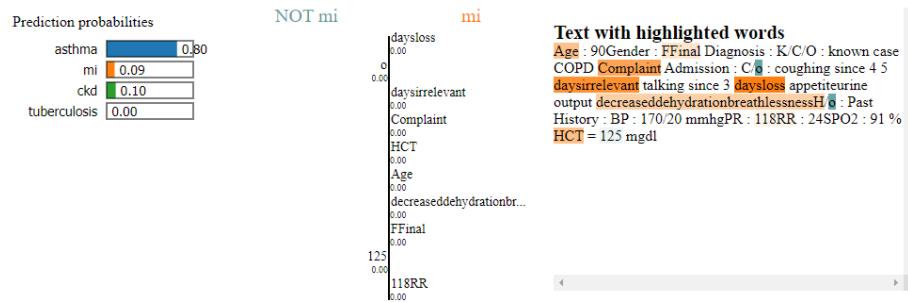


Figure 8. LIME explanation

Figure 10 shows the combined interpretability technique wherein Figure 10(a) presents narrative explanations, while Figure 10(b) provides corresponding LIME explanations. These LIME justifications are produced alongside the narratives to facilitate a clearer understanding of the model's predictions. The narratives offer a written summary, detailing how specific words or features have a major impact on the model's decisions. By presenting both narratives and visual explanations together, users can more easily interpret the main elements that the model considers when making its predictions.

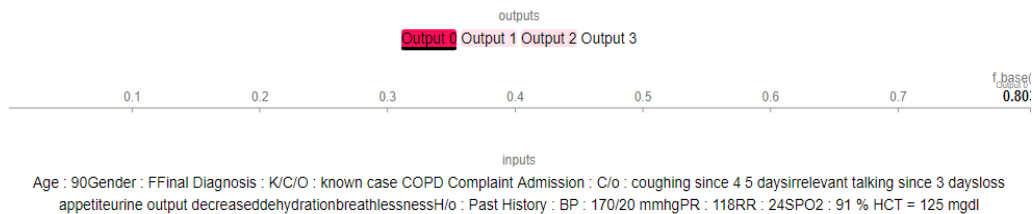


Figure 9. SHAP explanation

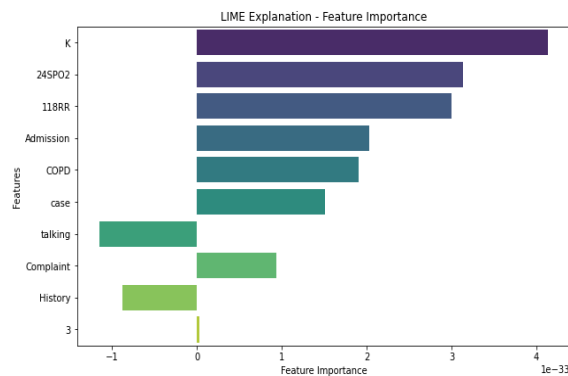
The model identified 'Asthma' due to the following features:

- The presence of 'K' contributes significantly to the prediction.
- The presence of '24SPO2' contributes significantly to the prediction.
- The presence of '118RR' contributes significantly to the prediction.
- The presence of 'Admission' contributes significantly to the prediction.
- The presence of 'COPD' contributes significantly to the prediction.
- The presence of 'case' contributes significantly to the prediction.
- The presence of 'talking' contributes significantly to the prediction.
- The presence of 'Complaint' contributes significantly to the prediction.
- The presence of 'History' contributes significantly to the prediction.
- The presence of '3' contributes significantly to the prediction.

(a)

Highlighted Text:

Age: 90Gender: FFFinal Diagnosis: K/C/O : known case of COPD Complaint on Admission: C/o: coughing since 4 to 5 daysirrelevant talking since 3 daysloss of appetiteurine output decreaseddehydrationbreathlessnessH/o: Past History: BP: 170/20 mmhgPR: 118RR: 24SPO2: 91%HCT = 125 mgdl



(b)

Figure 10. Combined interpretability technique (a) narratives for important words (b) LIME explanation generated along with narratives

4. CONCLUSION

In this study, we addressed the challenges of applying ML models to real-time clinical data, particularly in the Indian context. By utilizing the SMDBERT model fine-tuned on publicly available datasets, MIMIC, the potential of transformer-based models for the classification of chronic diseases such as Asthma, MI, and CKD was demonstrated. However, the limitations of these models were evident when they were applied to Indian clinical data, where diseases like tuberculosis are more prevalent but not represented in the training datasets. To overcome this, an innovative ensemble approach that combines the strengths of the SMDBERT model and a customized zero-shot learning method was developed, enabling effective classification of tuberculosis cases despite the lack of corresponding examples in the training data.

Additionally, the issue of model interpretability was tackled, which is critical for clinical adoption. The research introduced narrative-based interpretability technique, along with LIME interpretability to provide natural language explanations of the model's decision-making process. This approach not only enhances the transparency of predictions but also helps healthcare professionals in understanding the rationale behind the model's outputs, making these tools more reliable and trustworthy in clinical practice.

Our research's findings highlight the significance of customizing machine learning models to the unique requirements of various demographics and the need of creating strong interpretability frameworks for intricate models. By integrating an ensemble of zero-shot learning with SMD-BERT, our approach achieved an accuracy of 0.92, outperforming specialized models like SCIBERT, BIOBERT, and clinical BioBERT. This research contributes significantly to the field by demonstrating how advanced NLP techniques can be adapted for diverse healthcare environments and by proposing a new direction for the interpretability of AI models in clinical surroundings. Moving forward, expanding the scope of real-time datasets and further refining interpretability methods will be crucial for advancing AI's role in healthcare.

ACKNOWLEDGEMENTS

The authors are thankful to the Department of Computer Engineering, Ramrao Adik Institute, D Y Patil Deemed to be University, to provide facilities and support to carry out the research work.




REFERENCES

- [1] J. Henry, Y. Pylypchuk, T. Searcy, and V. Patel, "Adoption of electronic health record systems among U.S. non-federal acute care hospitals: 2008-2015," *ONC Data Brief*, vol. 5, no. 35, 2016.
- [2] A. E. W. Johnson *et al.*, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, no. 1, p. 160035, May 2016, doi: 10.1038/sdata.2016.35.
- [3] T. J. Pollard, A. E. W. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi, "The eICU collaborative research database, a freely available multi-center database for critical care research," *Scientific Data*, vol. 5, no. 1, p. 180178, Sep. 2018, doi: 10.1038/sdata.2018.178.
- [4] J. Yang *et al.*, "Brief introduction of medical database and data mining technology in big data era," *Journal of Evidence-Based Medicine*, vol. 13, no. 1, pp. 57–69, Feb. 2020, doi: 10.1111/jebm.12373.
- [5] R. Alanazi, "Identification and prediction of chronic diseases using machine learning approach," *Journal of Healthcare Engineering*, vol. 2022, pp. 1–9, Feb. 2022, doi: 10.1155/2022/2826127.
- [6] S. Ahmed *et al.*, "Prediction of cardiovascular disease on self-augmented datasets of heart patients using multiple machine learning models," *Journal of Sensors*, vol. 2022, pp. 1–21, Dec. 2022, doi: 10.1155/2022/3730303.
- [7] L. J. Spencer, A. Dedu, H. A. Kalkidan, M. A. Solomon, A. Cristiana, and Nooshin Abbasi, "Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017," *The Lancet*, vol. 392, pp. 1789–1858, 2018.
- [8] C. P. Kovesdy, "Epidemiology of chronic kidney disease: an update 2022," *Kidney International Supplements*, vol. 12, no. 1, pp. 7–11, Apr. 2022, doi: 10.1016/j.kisu.2021.11.003.
- [9] S. Saigaonkar and V. Narawade, "Domain adaptation of transformer-based neural network model for clinical note classification in Indian healthcare," *International Journal of Information Technology (Singapore)*, Aug. 2024, doi: 10.1007/s41870-024-02053-z.
- [10] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 5, pp. 1589–1604, 2018, doi: 10.1109/JBHI.2017.2767063.
- [11] Y. Meng, W. Speier, M. Ong, and C. W. Arnold, "HCET: hierarchical clinical embedding with topic modeling on electronic health records for predicting future depression," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 4, pp. 1265–1272, Apr. 2021, doi: 10.1109/JBHI.2020.3004072.
- [12] J. Ye, L. Yao, J. Shen, R. Janarthnam, and Y. Luo, "Predicting mortality in critically ill patients with diabetes using machine learning and clinical notes," *BMC Medical Informatics and Decision Making*, vol. 20, no. S11, p. 295, Dec. 2020, doi: 10.1186/s12911-020-01318-4.
- [13] K. Huang, J. Altaosaar, and R. Ranganath, "ClinicalBERT: modeling clinical notes and predicting hospital readmission," *arXiv preprint arXiv:1904.05342*, 2019, [Online]. Available: <http://arxiv.org/abs/1904.05342>.
- [14] J. Wu, Y. Lin, P. Li, Y. Hu, L. Zhang, and G. Kong, "Predicting prolonged length of ICU stay through machine learning," *Diagnostics*, vol. 11, no. 12, p. 2242, Nov. 2021, doi: 10.3390/diagnostics11122242.
- [15] S. A. Moqurrab, U. Ayub, A. Anjum, S. Asghar, and G. Srivastava, "An accurate deep learning model for clinical entity recognition from clinical notes," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 10, pp. 3804–3811, 2021, doi: 10.1109/JBHI.2021.3099755.




- [16] N. Liu, Q. Hu, H. Xu, X. Xu, and M. Chen, "Med-BERT: a pretraining framework for medical records named entity recognition," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 8, pp. 5600–5608, Aug. 2022, doi: 10.1109/TII.2021.3131180.
- [17] N. C. Ernecoff *et al.*, "Electronic health record phenotypes for identifying patients with late-stage disease: a method for research and clinical application," *Journal of General Internal Medicine*, vol. 34, no. 12, pp. 2818–2823, Dec. 2019, doi: 10.1007/s11606-019-05219-9.
- [18] P. Jeyanthan, "Machine learning in the identification of phenotypes of multiple sclerosis patients," *International Journal of Information Technology*, vol. 16, no. 4, pp. 2307–2313, Apr. 2024, doi: 10.1007/s41870-024-01735-y.
- [19] S. Saigaonkar and V. Narawade, "Predicting chronic diseases using clinical notes and fine-tuned transformers," in *2022 IEEE Bombay Section Signature Conference (IBSSC)*, Dec. 2022, pp. 1–6, doi: 10.1109/IBSSC56953.2022.10037512.
- [20] S. Saigaonkar and V. Narawade, "SM-DBERT: a novel symptom-based technique for chronic disease classification using DistilBERT," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, no. 9, pp. 2370–2377, Nov. 2023, doi: 10.17762/ijritcc.v11i9.9275.
- [21] D. Capellan-Martin *et al.*, "Zero-shot pediatric tuberculosis detection in chest X-rays using self-supervised learning," *arXiv preprint arXiv:2402.14741*, 2024, doi: 10.1109/ISBI56570.2024.10635520.
- [22] D. Mahapatra, B. Bozorgtabar, and Z. Ge, "Medical image classification using generalized zero shot learning," in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, Oct. 2021, pp. 3337–3346, doi: 10.1109/ICCVW54120.2021.00373.
- [23] N. Hayat, H. Lashen, and F. E. Shamout, "Multi-label generalized zero shot learning for the classification of disease in chest radiographs," *Proceedings of Machine Learning Research*, vol. 149, pp. 461–477, 2021.
- [24] Y. Zhu *et al.*, "Prompting large language models for zero-shot clinical prediction with structured longitudinal electronic health record data," *arXiv preprint arXiv:2402.01713*, 2024, [Online]. Available: <http://arxiv.org/abs/2402.01713>.
- [25] S. Sivarajkumar and Y. Wang, "HealthPrompt: a zero-shot learning paradigm for clinical natural language processing," in *AMIA Annual Symposium Proceedings. American Medical Informatics Association*, 2022, pp. 972–981.
- [26] V. Vimbi, N. Shaffi, and M. Mahmud, "Interpreting artificial intelligence models: a systematic review on the application of LIME and SHAP in Alzheimer's disease detection," *Brain Informatics*, vol. 11, no. 1, p. 10, Dec. 2024, doi: 10.1186/s40708-024-00222-1.
- [27] S. Rao, S. Mehta, S. Kulkarni, H. Dalvi, N. Katre, and M. Narvekar, "A study of LIME and SHAP model explainers for autonomous disease predictions," in *2022 IEEE Bombay Section Signature Conference (IBSSC)*, Dec. 2022, pp. 1–6, doi: 10.1109/IBSSC56953.2022.10037324.

BIOGRAPHIES OF AUTHORS



Swati Saigaonkar    M.E, Assistant Professor, V.C.E.T, is currently pursuing Ph.D. in Computer Engineering, RAIT, D Y Patil Deemed to be University. She received the M.E. degree in the year 2011 and B.E in the year 2005. Her research interest includes neural networks, LLMs, and artificial intelligence. She can be contacted at email: swa.sai.rt21@dyatil.edu.



Vaibhav Narawade    Ph.D., Professor, RAIT, Nerul. He received his Ph.D. and M.E. degrees in 2017 and 2008 respectively, and a B.Tech in the year 1999. He has authored over 50 scholarly works, comprising articles featured in international peer-reviewed journals and conferences. His areas of expertise include wireless sensor networks, image processing, and data science. He can be contacted at email: vaibhav.narawade@rait.ac.in.