SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V13I5P108

Volume 13 Issue 5, 77-87, May 2026

Original Article

Augmenting Multi-Disease Indian Clinical Notes via Transformer-Based Symptom Fusion Techniques

Swati Varma[1], Anil Hingmire[2], Megha Trivedi[3], Smita Jawale[4], Rucha C. Samant[5], Neeta Deshpande[6]

[1,2,3,4]*Department of Computer Engineering, Vidyavardhi's College of Engineering and Technology, University of Mumbai, Vasai, Palghar, India.*
[5,6]*Department of Computer Engineering,*
Gokhale Education Society's R H Sapat College of Engineering, Management Studies and Research, Nashik, Maharashtra, India.

[2]*Corresponding Author : anil.hingmire@vcet.edu.in*

Received: 08 February 2026 Revised: 09 March 2026 Accepted: 09 April 2026 Published: 27 May 2026

Abstract *- The success of clinical note research has been experienced all over the world, but without emphasis on real-time Indian data. The study first consisted of the formulation of models based on the clinical notes of the Medical Information Mart Intensive Care (MIMIC) database, with a specific focus on such conditions as Asthma, Myocardial Infarction (MI), and Chronic Kidney Disease (CKD). These trained models were then tested on the actual Indian healthcare clinical data in real-time, received in two hospitals. Due to privacy concerns, the volume of collected notes was limited. To augment the data, two methods were devised. The first used a two-step strategy with weighted symptoms and fuzzy techniques for similarity calculations, followed by a synonym replacement technique. The second method augmented data using symptoms of co-occurring diseases. Augmentation was extended by concatenating these notes with a synonym replacement strategy. Bidirectional Encoder Representations from Transformers (BERT), Distilled BERT (DISTILBERT), and Symptom-Driven BERT (SMDBERT) were used on both strategies and compared with the baseline models: Easy Data Augmentation (EDA) and Synthetic Minority Over-sampling Technique (SMOTE). Method 1 outperformed both SMOTE and EDA for all models, while Method 2 gave superior results with DISTILBERT and SMDBERT, attaining, respectively, accuracy levels of 0.98 and 0.96, compared to 0.92 and 0.94 with EDA.*

Keywords *– DistilBERT, Data Augmentation, Real-time Clinical Notes, Neuro-Fuzzy Approach, Transfer Learning.*

1. Introduction

The use of EHR systems is on the rise [1]. Clinical notes are frequently stored within EHR systems, and in developing countries like India, they are largely used. Clinical notes are often kept as scanned copies, which need to be manually typed into the system to have access to the information stored in them. Many hospitals also do not share clinical notes due to issues of patient confidentiality, hence making the data available always inadequate. There is, therefore, a need for approaches like data augmentation to increase the size of the dataset while observing patient privacy.

MIMIC, i2b2, and eICU Collaborative Research Database are publicly available data sources that are an excellent source of clinical and healthcare analytics-related research. The MIMIC III has more than 50,000 intensive care unit patients [2], whereas the eICU has data from around twenty lakh patients [3]. The author in [4] summarizes the data sets that are publicly available and highlights the lack of Indian datasets. Transformer-based models are finding applications for a wide range of prediction tasks in healthcare, such as mortality prediction [5], readmission of patients [6], and the length of stay in hospitals [7]. Application areas of these models also include entity extraction [8, 9], identification of phenotypic traits [10, 11], establishing relations among various entities in medicine, and imitation of patient trajectories.

Augmentation of data is one of the methods to enhance the quantity of a dataset, as well as introduce variations to it. It is a well-known method when it comes to image datasets, but its application when it is used with text data is not common. Augmentation methods that are usually used in image datasets include image rotation, flipping, scaling, or cropping, or noise addition to images. Text augmentation, however, generally involves the substitution of synonyms, the insertion or removal of words, or the rearrangement of the word sequence. However, these methods pose challenges as they may inadvertently alter sentence meaning. Additionally, in clinical data, where precision is crucial, altering important information could have serious consequences.

Conditions that persist for more than three months are referred to as chronic diseases [12], and early identification leads to enhanced assessment and care. Approximately 26 million individuals worldwide annually suffer from heart failure, which is a challenge for cardiologists, surgeons, and clinicians to predict the onset of heart failure [13]. The author in [14] states that Asthma stands as the most common long-term lung condition globally. Moreover, 800 million individuals globally, or more than 10% of the global population, suffer from Chronic Kidney Disease (CKD), a degenerative illness [15]. Given their substantial prevalence, we have chosen these three diseases as our primary focus.

Objectives of the paper:

1. Demonstrates why symptoms of co-occurring diseases can be included, which can eventually make the models more robust.
2. Develop and evaluate novel data augmentation methods tailored to Indian clinical data to overcome limitations in data availability and privacy concerns.

2. Literature Review

Literary works offer a couple of strategies for augmentation, yet there exists a scarcity of methods integrating domain knowledge into the augmentation process [16]. Further, the utilization of augmentation techniques within medical applications is even more insufficiently addressed in the literature.

There are several techniques used on image data, as the author in [17] demonstrated, the generation of samples with a distribution resembling the original actual data after learning the distribution of the input data samples.

Generation of domain-guided augmented data is presented in [18] for classifying fetal ultrasound views by adding a novel context-preserving cut-paste procedure, resulting in models with performance comparable to those trained using traditional augmentation methods.

Researchers are working even in the field of audio data, too; the author in [19] uses both clinical data and voice recordings, though limited in number. Techniques like pitch shifting, time stretching, and noise injection were used to augment the voice data.

Text data augmentation techniques function across various levels, including character, word, sentence, and document levels. The general and commonly used techniques that are used for data augmentation are EDA [20] with Synonym Replacement (SR), Random Insertion (RI), Random Deletion (RD), and Random Swap (RS), Back translation [21], wherein a sentence a converted into another language first and then again back to its source language, Text synthesis, and use of Large Language Models [22].

The work [23] examined the performance of a model on sixteen different categorization tasks using twenty different augmentation methods. The assignments were divided into four groups based on the prevalence of diseases. Every group was subjected to a systematic application of 20 augmentation methods. Every augmentation approach was assessed using the Transformer Encoder model. The baseline model without augmentation and various augmentation techniques were compared. The results showed that, across a range of strategies and tasks, the model performance was consistently improved by the Splitting Augmenter, with statistically significant improvements for several key metrics like AUC-ROC and F1 Score. For instance, there were improvements of 0.13 in F1 score and 0.34 in AUC-ROC.

Authors of the research paper [24] have applied EDA tasks to Named Entity Recognition datasets. The authors have taken the benchmark datasets that are generally used to evaluate NER tasks in the medical field. Bio BERT model was fine-tuned with and without the augmented data, and found that EDA methods, though simple, yield good results, especially for small datasets.

The paper [25] focuses on using neural networks, more specifically recurrent networks, to predict clinical conditions based on the history of a patient's hospital admissions. Given the sparsity of clinical data available for machine learning (ML), the study goes on to suggest ways of augmenting existing data to increase the efficiency of predictive algorithms. They have introduced two methods, one concentrates on portions of the patient journey to emphasize changes between different visits to hospitals, and the second one relies on the structure of diagnostic codes in order to form trajectories resembling the real world.

These results provide significant improvements on experiments conducted on two datasets and indicate the feasibility of data augmentation in clinical scenarios. SMOTE addresses class imbalance by artificially generating instances of the minority class to improve the representation of the dataset for better model performance [26].

There are challenges when applying augmentation techniques to medical data, as it may alter the meaning altogether, and the patient’s life could be put at stake. So, careful strategies need to be devised while dealing with clinical text. Moreover, domain knowledge is rarely integrated into the existing techniques. The recent advances, like the study on doctor-patient conversation summarization, introduced the MTS-Dialog dataset, demonstrating how guided summarization techniques improve factual consistency in generated clinical notes [27]. Another approach leverages synthetic transcript generation using generative models, supplementing real-world clinical documentation to refine NLP-based transcription accuracy. For medical document classification, ontology-guided augmentation has also been

proposed [28], which uses domain-specific knowledge to improve feature extraction and thus classification performance. Taking all together, these methods indicate a greater focus on more structured augmentation techniques tailored to medical NLP applications.

Methodologies in healthcare data analysis range from rule-based approaches to machine learning and Deep Learning (DL) techniques. Rule-based techniques rely on predefined rules, based on the expertise of experts, but may struggle with novel concepts and necessitate rule modifications. ML techniques derive patterns directly from data. In supervised learning, labeled datasets are used for tasks such as classification, while unsupervised learning groups similar data points through clustering.

While common ML techniques like LR, SVM, and Random Forests may face challenges in dealing with high-dimensional and heterogeneous multimodal data, DL methods like CNNs, RNNs (including LSTM and GRU), and transformers such as BERT and DISTILBERT are promising, extracting relevant features from raw data and showing improved contextual understanding.

Pretrained for bidirectional contextualized understanding, BERT learns effective representations from unlabeled data and performs excellently in many ML tasks. The fine-tuned versions of the model require minimal additional layers for downstream tasks [29]. Whenever classifying texts, BERT is preferred. A lighter form of BERT is called DISTILBERT, which gives as promising results as BERT [30]. SMDBERT [31], which incorporates domain knowledge, has been shown to give better performances as compared to DISTILBERT. SMDBERT could be expanded further to perform domain adaptation by incorporating extra embeddings [32] and could also be converted into a hybrid model by combining with a zero-shot model, as demonstrated in [33].

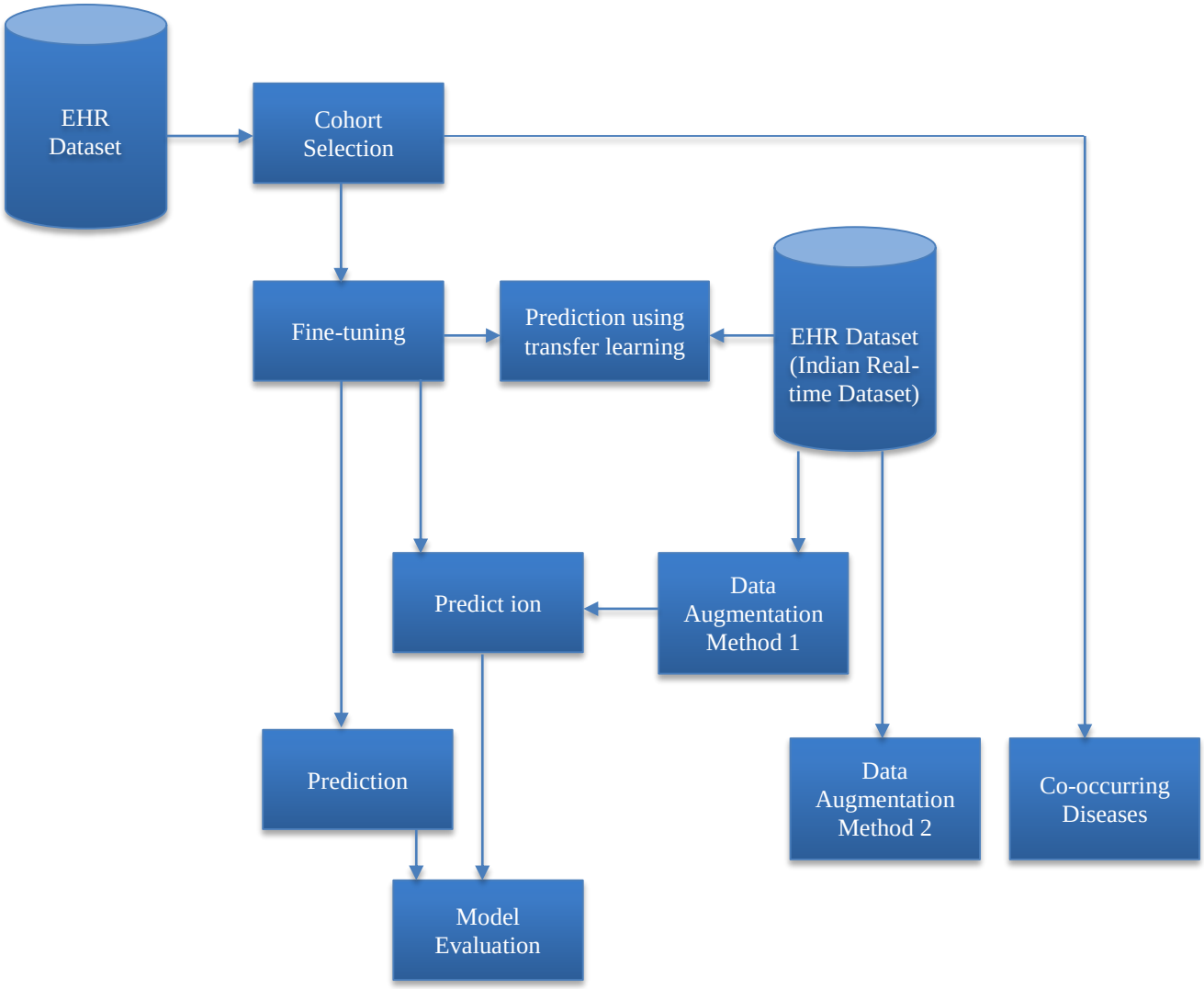

Fig. 1 Overall architecture

3. Materials and Methods

3.1. Overall Architecture

The entire architecture of the system is illustrated in Figure 1. Input data for the system is sourced from the MIMIC dataset. Next, cohort selection is carried out to focus on particular diseases. The models are fine-tuned for these specific diseases and subsequently applied to real-time Indian datasets using a transfer learning approach. Because of the restricted availability of data, primarily due to privacy considerations, techniques for data augmentation are employed to boost the dataset's volume. The augmented data undergoes model application and subsequent evaluation and comparison against the baseline model.

3.2. Data Selection

The MIMIC dataset is used to create the base models. The targeted diseases were Asthma, kidney disease, and Myocardial Infarction with ICD9 codes of 49320,5849,41001,41011,41021 respectively. MI is further broken down into codes 41001, 41011, and 41021. First of all, 170446 samples were retrieved, and according to these, a refining process was carried out, thus leaving 5234 samples to be analyzed, limited to discharge summaries. BERT and DistilBERT models and our own model, SMDBERT, were adapted using clinical records obtained on the MIMIC dataset, as well as other structured clinical data. Before fine-tuning, there was initial text preprocessing, which involved lowercasing of text and removing special characters, URLs, and non-alphanumeric characters. These models were trained on clinical documents that were gathered in two hospitals in Mumbai, India. The records were gathered during two timeframes: October to November 2022 and January to April 2023. In total, 132 clinical notes centered on the diseases of interest were chosen for evaluation.

3.3. Methodology

The decision to incorporate symptoms of co-occurring diseases is formed through empirical evidence by analyzing the MIMIC dataset. The dataset consists of about 46,000 admissions or unique SUBJECT_IDS. Many of the clinical notes are multi-labeled in nature. A query targeting unique SUBJECT_IDS with a single label yielded a count of 190. This means that only 190 were given a single ICD9 code, while the remaining patients had multiple ICD9 codes assigned. An important observation is that comorbid patients, for instance, Asthma and diabetes, will not have just symptoms of Asthma alone, but also symptoms of diabetes. Moreover, it is likely that symptoms of one condition, say diabetes, will overshadow those of another, say Asthma.

As a result, when selecting a clinical note labeled with ‘asthma’ for model training, it may predominantly contain symptoms related to diabetes. To account for such scenarios, symptoms of co-occurring diseases were incorporated to augment the data. Figure 2 displays a heat map showing the main diseases and their co-occurring diseases.

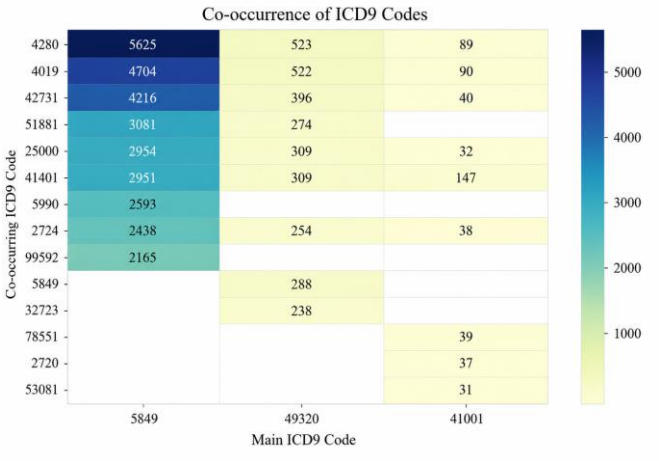

Fig. 2 Co-occurrence of diseases

As can be seen from the heat map, the top five common diseases are the ones with codes 4280, 4019, 42731, 51881, and 25000. Out of these 4280, 42731, and 51881 are related to heart failure and respiratory failure, so we avoided them, as these are also our primary diseases. We narrowed our focus to the remaining two codes, 4019 and 25000, corresponding to hypertension and diabetes, respectively. Also, these conditions are frequently combined with other diseases.

The initial process of our methods is described below. Both methods first calculate the frequency of each unique symptom, and then a weighted symptom list is created. The rationale behind this is that the symptoms that occur more frequently will appear more times in the list. Similarity of these symptoms and domain-specific symptoms is calculated using a fuzzy string-matching technique.

The commonly used metric for fuzzy string matching is the Levenshtein distance. A combination of existing symptoms and domain-specific symptoms is generated, which satisfies the minimum threshold value. A randomizer function is then defined, which uses existing data and randomly selects demographics data from it, and then appends it to every record generated.

As illustrated in Figure 7, the above strategy constitutes the initial layer of our data augmentation strategy. Following this, the synonyms are generated for each record of the augmented data layer 1. This constitutes the second layer of the data augmentation process. The final layers are the combination of AD Layer 1 and AD Layer 2, which represent domain-enriched data and synonym-replaced data, respectively.

The algorithm of the model is:

1. Calculate Symptom Frequencies:
   - For each symptom in the dataset:
     - Count the frequency of occurrence.

2. Create Weighted Symptom List:
   - Generate a list of symptoms weighted by their frequency.
3. Define Function for Symptom Similarity:
   - Define a function to calculate the similarity between a given symptom and domain-specific symptoms using fuzzy matching.
4. Define Function for Combination Generation:
   - Define a function to generate new combinations based on a given label:
     - Include existing symptoms.
     - Add additional domain-specific symptoms based on similarity.
   - Randomizer:
     - Use existing data to select demographic information randomly.
     - Append selected demographic information to every generated combination.
5. Apply Synonym Replacement:
   - For each record:
     - Replace synonyms of existing symptoms.
     - Augment the data with synonym replacements.

As previously mentioned, the second method also builds upon the foundation laid by the first layer. Symptoms of occurring diseases, which are identified as mentioned previously, are incorporated. This combined augmentation forms the second layer of the process, enhancing the dataset's richness and diversity. Subsequently, synonym replacement is applied, constituting the third layer of augmentation. The entire process is illustrated in Figure 8.

Mathematically, our augmentation process of Method 2 can be represented as:

Let fs represent the frequency of symptoms in the dataset D. The frequency can be calculated as given in equation (1):

$$fs = \frac{Number\ of\ occurences\ of\ symptom\ s}{Total\ number\ of\ records\ in\ D} \quad (1)$$

Let W be the weighted list of symptoms, where each symptom si is associated with its frequency weight wi.

This can be represented as:
$W = \{(s1, w1), (s2, w2),...,(sn,wn)\}$ where $wi=fsi$ for each i.

Let $sim\ (s1,\ s2)$ represent the similarity between symptoms $s1$ and $s2$. This is calculated using fuzzy matching techniques, which internally uses Levenshtein distance.

Let $G(label)$ represent the function to generate new combinations based on a given label. This function takes an existing set of symptoms S and returns augmented combinations that include additional domain-specific symptoms.

It can be represented as:
$G(label) = \{S \cup \text{domain-specific symptoms}\}$

Randomizer Demographics represent the randomizer function, which takes existing demographics data and randomly selects demographic information from it. This function then appends the selected demographic information to every record generated during the data augmentation process. It can be represented as:

Randomizer (Demographics)=selected demographic information

Let R represent a record in the dataset D, and SR represent the set of symptoms associated with record R. Augmenting more records by adding symptoms of co-occurring diseases can be represented as:

New record R′= SR ∪ symptoms of co-occurring diseases

Let SR (record) represent the function to apply synonym replacement to a given record. This function replaces words or phrases in the record with their synonyms. It can be represented as:

SR (record)=record with synonym replacements

Figures 3, 4, and 5 show different samples of clinical notes after the augmentation process. Figure 6 illustrates a deidentified Indian clinical note.

'Age 90 Gender: female Symptoms loose motion since 3-4 days, poor oral intake, severe weakness, altered sensorium, mild b/l pedal edema, facial puffiness, Coughing or wheezing attacks that are worsened by a respiratory virus, Chest tightness or pain, Shortness of breath, mucoid sputum'

Fig. 3 Sample augmented note with extracted symptoms and domain-specific symptom

years : 65
gender : Male
Final diagnosing : urosepsis and ckd Symptoms : axerophthol lxv year old b/b relative in our hospital,decently side chest pain , febrility , b/l wheel oedema , gen helplessness H/o past_tense illness : k/c/o : DM , CKD temporary_worker : ninety-nine Pulse:140 /min Respiration:24 /min BP : 110/80 mmhg SPO2 : 98 % Investigation : HB 11.3 WBC 20300 PLT 407000 RBS FBS - 239 CREAT 4.42

Fig. 4 Sample augmented note with synonym replacement

Age 88
Gender: female
Symptoms breathlessness, tachypnea, desaturation, DOE, Slurred speech, left side weakness, gen weakness, bilateral pedal oedema since 4-5 day increased since today morning,tiredness, blood pressure : 140/90, more thirst, fasting sugar above 100

Fig. 5 Sample co-occurring disease symptom augmented note

Age: 67
Gender: Male
Final Diagnosis: Atypical chest pain

Symptoms: pain in epigastric region, nausea, uneasin ess, restlessness, general weakness

H/o past illness: k/c/o: DM, HTN

Temp: 98
Pulse: 77/min
Respiration: 24/min
BP: 160/100 mmhg
SPO2: 98 %
Investigation:
HB11.8
WBC11000
PLT274
ELECTROLYTE 136.1 / 4.4 / 102.2

CREAT1.4
SR. CALCIUM
SR. BILI 0.40.20.6
SGOT/SGPT11.5 / 5.35
CSF ROUTINE 173.20

Fig. 6 Sample indian clinical note

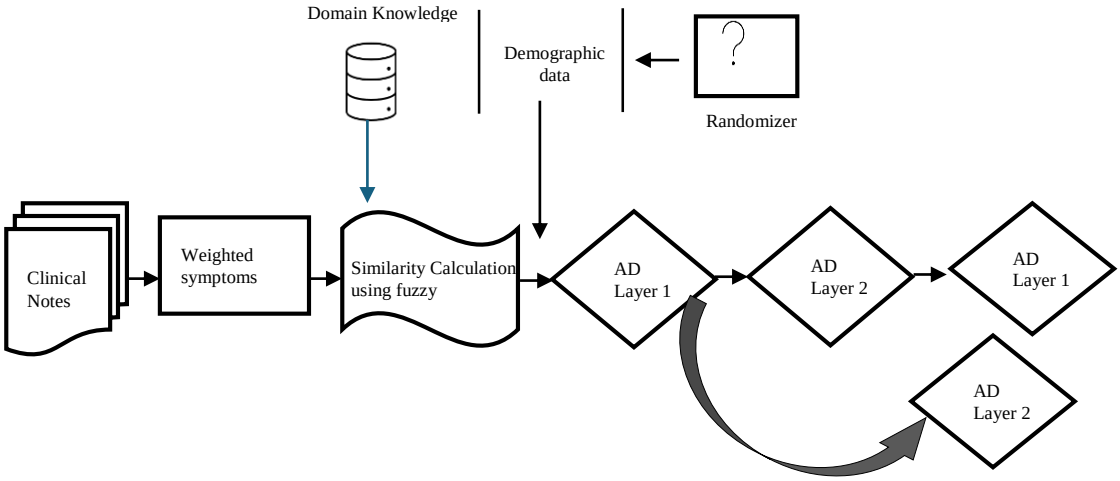

Fig. 7 Augmentation method 1 with weighted symptom integration, domain knowledge, and synonym replacement

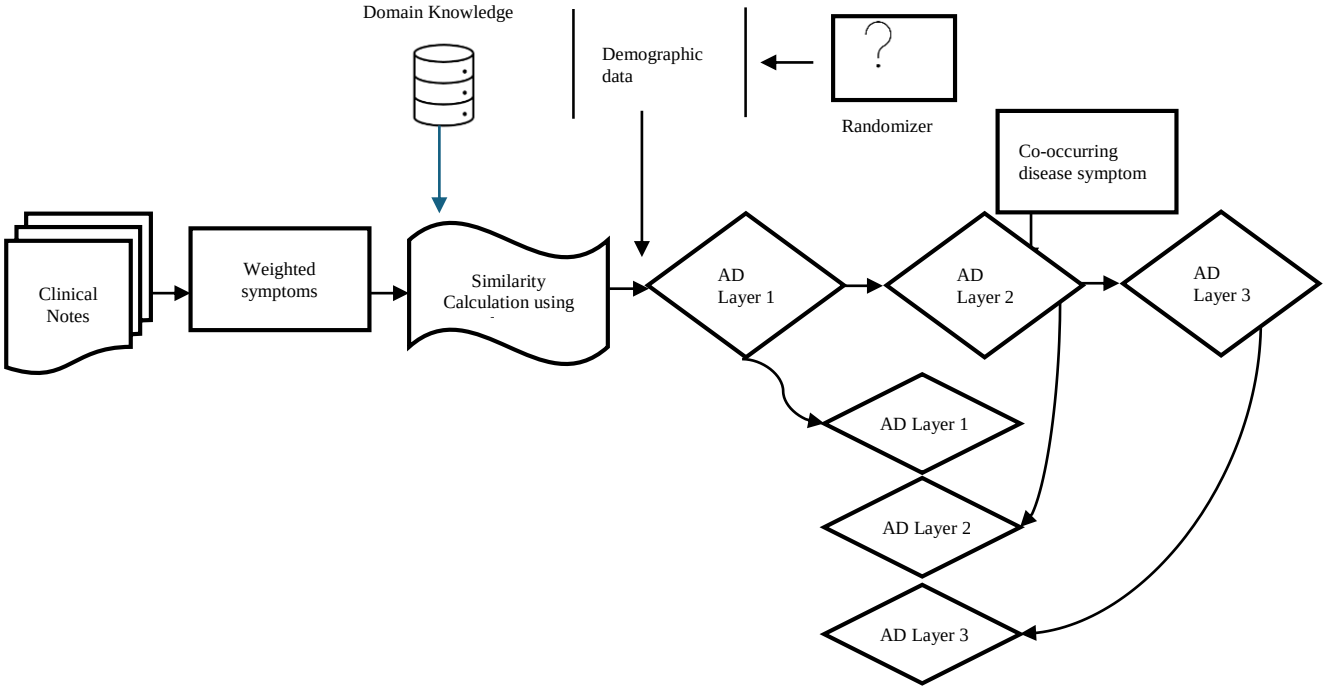

Fig. 8 Holistic augmentation method 2 with weighted symptom integration, domain expertise, synonym replacement, and incorporating co-occurring diseases

4. Results and Discussion

Table 1 presents the performance following the models' application using transfer learning on Indian data. Table 2 displays the performance metrics for various algorithms and data models.

Table 1. Performance metrics of fine-tuned models on indian real-time dataset

Models	Accuracy	Precision	Recall	F1-score
BERT	0.59	0.6	0.59	0.59
DISTILBERT	0.72	0.8	0.72	0.72
SMDBERT	0.84	0.84	0.84	0.84

Table 2. Performance metrics for different models and methods

Method	Models	Accuracy	Precision	Recall	F1 measure
SMOTE	BERT	0.54	0.66	0.54	0.5
	DISTILBERT	0.49	0.49	0.49	0.49
	SMDBERT	0.51	0.51	0.51	0.51
EDA	BERT	0.94	0.94	0.94	0.94
	DISTILBERT	0.92	0.89	0.93	0.9
	SMDBERT	0.94	0.94	0.94	0.94
Method 1	BERT	0.94	0.94	0.94	0.94
	DISTILBERT	0.92	0.94	0.91	0.92
	SMDBERT- Single symptom	0.96	0.96	0.96	0.96
	SMDBERT- Multiple symptoms	0.97	0.98	0.97	0.97
	SMDBERT-All symptoms	0.97	0.98	0.97	0.97
Method 2	BERT	0.87	0.87	0.87	0.87
	DISTILBERT	0.98	0.98	0.98	0.98
	SMDBERT- Single Symptom	0.96	0.96	0.96	0.96
	SMDBERT- Multiple symptoms	0.96	0.97	0.97	0.97
	SMDBERT- All symptoms	0.96	0.96	0.96	0.96

Clinical notes and partial structured data were first used to fine-tune the models. Once this had been done, their performance was assessed on actual Indian clinical data in the hospitals. This was done using a carefully edited collection of clinical notes, which were related to the identified diseases. Since the hospitals provided scanned documents, the notes had to be manually transcribed. Importantly, information regarding treatment was deliberately omitted to maintain the study's emphasis on symptom analysis. As can be seen, our model SMDBERT gave better results.

In employing the EDA method of augmentation, the notes increased from 132 to 528, with 132 records for every task of EDA, i.e., SR, RI, RD, and RS. With the SMOTE method, the number of records increased to 183, as SMOTE generates as many synthetic examples from the minority classes as there are from the majority class. For Method 1, i.e., augmentation with weighted symptom extraction, domain knowledge, and synonym replacement, the notes increased from 132 to 384. 60 records were appended using domain knowledge, and the total of 192 records were added to the previous layer with a synonym replacement strategy.

In Method 2 of data augmentation, the number of records increased from 132 to 444. Initially, 192 records were incorporated in the first layer, which was a blend of weighted symptom extraction with domain knowledge. Subsequently, in the second layer, 30 records featuring symptoms of co-occurring diseases were integrated. Finally, all 222 records underwent synonym replacement in the third layer.

The dataset was divided into a twenty percent test set and an eighty percent train set. Adam optimizer was used, with a learning rate of 1e-4, with 6 epochs to train the models.

The computational formulae of the precision, recall, and the F1-score are described in equations (2), (3), and (4), respectively.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1 - score = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (4)$$

While the EDA method showcased satisfactory performance, our augmentation techniques exhibited superior performance except for the majority of the runs. Our model, SMDBERT, showcased a better performance compared to BERT and DISTILBERT models with both of our augmentation strategies. Also, our results show that adding multiple symptoms gives better performance compared to just adding a single symptom to the notes.

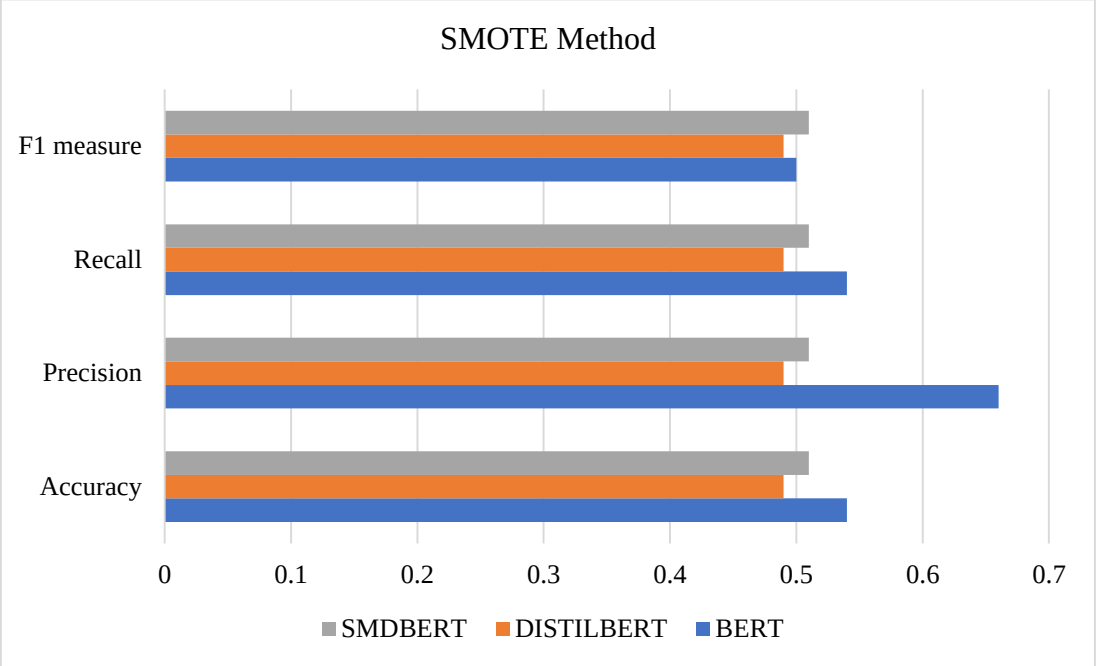

Fig. 9 Graph displaying performance metrics using SMOTE

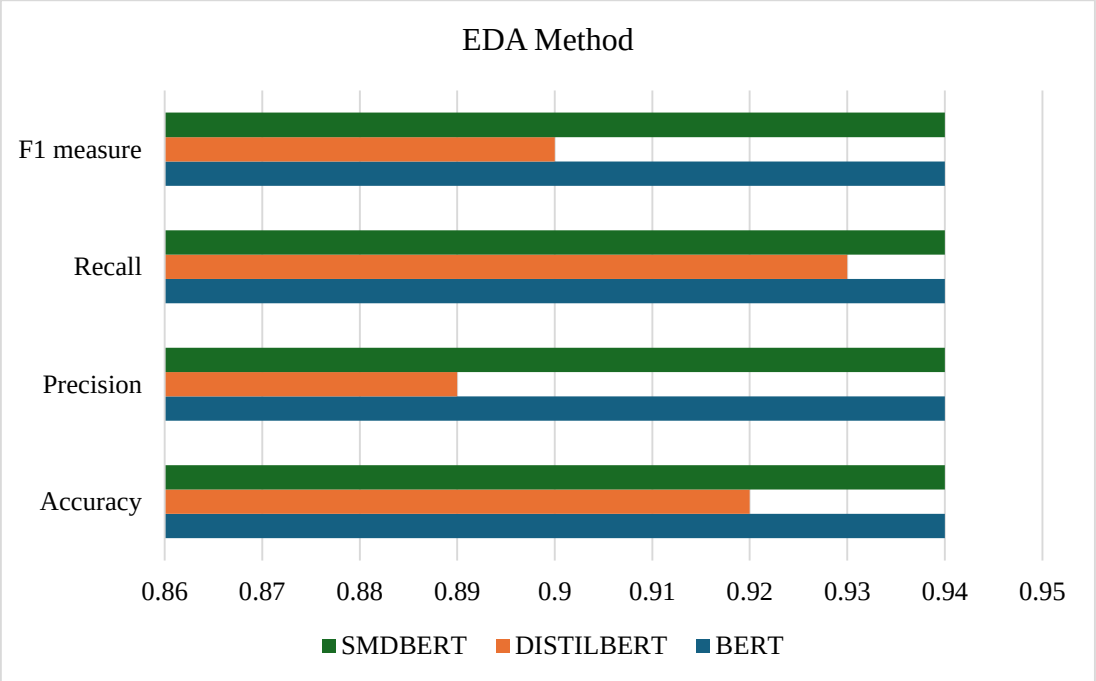

Fig. 10 Graph displaying performance metrics using the EDA Method of augmentation

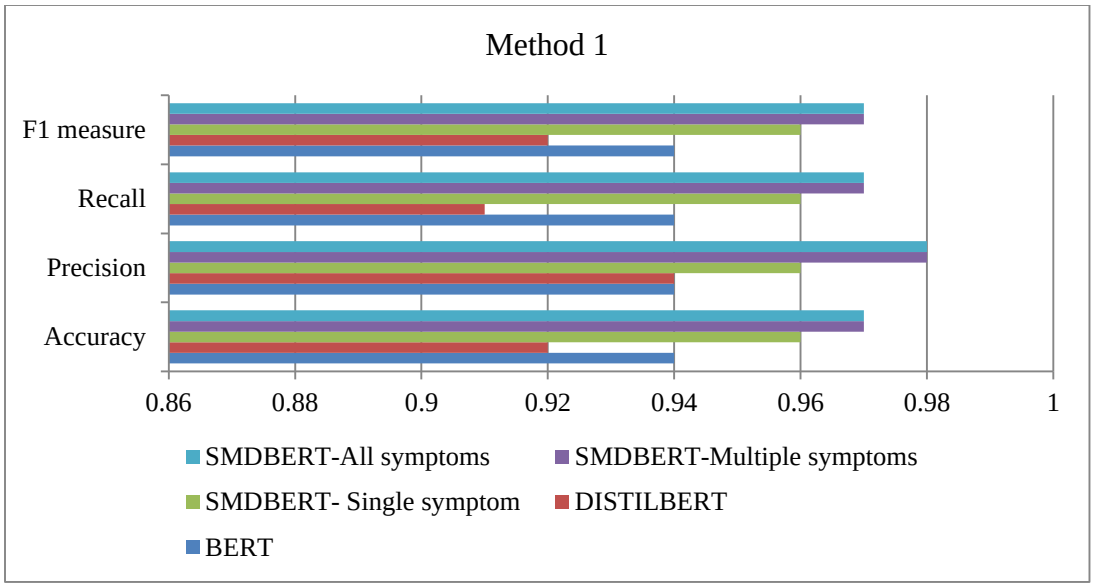

Fig. 11 Graph displaying performance metrics using Method 1

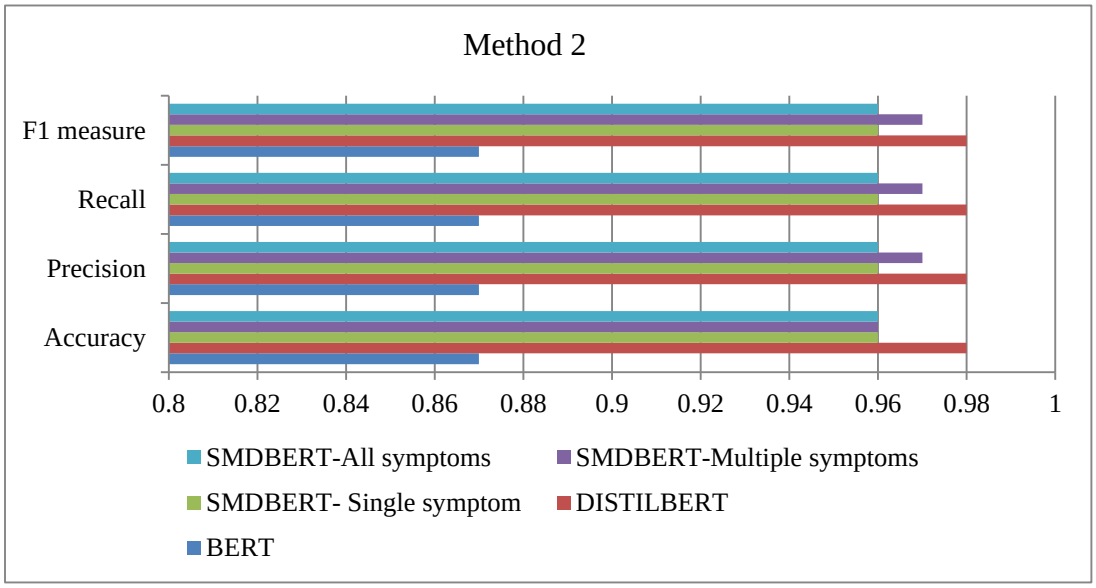

Fig. 12 Graph displaying performance metrics using method 2

The conclusion is that Method 1 showed the best results. Method 2, although it gives strong performance with DISTILBERT, fell short in comparison to Method 1. The EDA method, while effective, was outperformed by the approaches of Method 1 and Method 2, emphasizing the significance of domain-specific augmentation techniques for enhancing model performance in clinical text analysis. (Figure 9-12).

5. Conclusion

A notable gap exists in publicly accessible Indian clinical datasets, which is even worsened by the limited adoption of EHR in smaller healthcare facilities, where clinical notes are often stored in hard copy format. This situation poses a challenge in acquiring relevant clinical data. Also, the increasing use of transformer-based models does offer a potential for possible augmentation. It extends not only the quantity but also the diversity in the data for better robustness in the model. Transfer learning methods reduce the hassle of training from scratch and thus allow creating efficient models. Our initial approach involved integrating domain expertise with extracted symptoms to enrich the clinical notes. We conducted experiments with varying numbers of symptoms, ranging from individual symptoms to all available symptoms. Finally, synonym replacement was carried out on the combined clinical notes to further increase the robustness of the model. Although there is much research on ICD9 codes in the medical domain, the problem of multiple ICD9 codes is

not much taken into consideration. In our second approach, we added information on co-occurring diseases since secondary diseases mostly appear along with the symptoms of a primary disease. So, we chose the most common co-occurring diseases - Diabetes and Hypertension - and added the symptoms of these two conditions to the combined clinical notes. Though the performance of the BERT model slightly decreased, it proved that adding co-occurring symptoms of diseases can be explored as a data augmentation strategy.

In the future, this model can be expanded to make predictions of more diseases with more data and characteristics just by increasing its size. Moreover, certain focus should be given to clinical variables in the course of augmentation, and this is what we would like to address in our future activity.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Funding Statement

On Behalf of all authors, the corresponding author states that they did not receive any funds for this project.

References

[1] JaWanna Henry et al., "Adoption of Electronic Health Record Systems among U.S. Non-Federal Acute Care Hospitals: 2008-2015," *ONC Data Brief*, no. 35, pp. 1-11, 2016. [Google Scholar] [Publisher Link]

[2] Alistair E.W. Johnson et al., "MIMIC-III, A Freely Accessible Critical Care Database," *Scientific Data*, vol. 3, pp. 1-9, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[3] Tom J. Pollard et al., "The eICU Collaborative Research Database, A Freely Available Multi-Center Database for Critical Care Research," *Scientific Data*, vol. 5, pp. 1-13, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[4] Jin 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, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[5] Jiancheng Ye et al., "Predicting Mortality in Critically Ill Patients with Diabetes using Machine Learning and Clinical Notes," *BMC Medical Informatics and Decision Making*, vol. 20, pp. 1-7, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[6] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath, "ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission," *ArXiv Preprint*, pp. 1-9, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[7] Jingyi Wu et al., "Predicting Prolonged Length of ICU Stay through Machine Learning," *Diagnostics*, vol. 11, no. 18, pp. 1-18, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[8] Syed Atif Moqurrab et al., "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. [CrossRef] [Google Scholar] [Publisher Link]

[9] Ning Liu et al., "Med-BERT: A Pre-Training Framework for Medical Records Named Entity Recognition," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 8, pp. 5600-5608, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[10] Natalie 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, pp. 2818-2823, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[11] Pratheeba Jeyananthan, "Machine Learning in the Identification of Phenotypes of Multiple Sclerosis Patients," *International Journal of Information Technology*, vol. 16, pp. 2307-2313, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[12] Rayan Alanazi, "Identification and Prediction of Chronic Diseases Using Machine Learning Approach," *Journal of Healthcare Engineering*, vol. 2022, no. 1, pp. 1-9, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[13] Sumaira Ahmed et al., "Prediction of Cardiovascular Disease on Self-Augmented Datasets of Heart Patients Using Multiple Machine Learning Models," *Journal of Sensors*, vol. 2022, no. 1, pp. 1-21, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[14] Spencer L. James et al., "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," *Lancet*, vol. 392, pp. 1789-1858, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[15] Csaba P. Kovesdy, "Epidemiology of Chronic Kidney Disease: An Update 2022," *Kidney International Supplements*, vol. 12, no.1, pp. 7-11, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[16] Dharini Ramachandran, and R. Parvathi, "A Novel Domain and Event Adaptive Tweet Augmentation Approach for Enhancing the Classification of Crisis Related Tweets," *Data & Knowledge Engineering*, vol. 135, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[17] Ye Zhang et al., "GAN-based One-Dimensional Medical Data Augmentation," *Soft Computing*, vol. 27, pp. 10481-10491, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[18] Chinmayee Athalye, and Rima Arnaout, "Domain-Guided Data Augmentation for Deep Learning on Medical Imaging," *PLoS One*, vol. 18, no. 3, pp. 1-12, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[19] Mohammed Muzaffar Hussain et al., "Enhancing Parkinson's Disease Identification using Ensemble Classifier and Data Augmentation Techniques in Machine Learning," *Clinical eHealth*, vol. 6, pp. 150-158, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[20] Yonatan Belinkov, and Yonatan Bisk, "Synthetic and Natural Noise Both Break Neural Machine Translation," *arXiv preprint*, pp. 1-13, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[21] Djamila Romaissa Beddiar, Md Saroar Jahan, and Mourad Oussalah, "Data Expansion Using Back Translation and Paraphrasing for Hate Speech Detection," *Online Social Networks and Media*, vol. 24, pp. 1-13, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[22] Chenxi Whitehouse, Monojit Choudhury, and Alham Fikri Aji, "LLM-Powered Data Augmentation for Enhanced Cross-Lingual Performance," *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 671-686, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[23] Hongxia Lu, and Cyril Rakovski, "The Effect of Text Data Augmentation Methods and Strategies in Classification Tasks of Unstructured Medical Notes," *Research Square*, pp. 1-29, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[24] Abdul Majeed Issifu, and Murat Can Ganiz, "A Simple Data Augmentation Method to Improve the Performance of Named Entity Recognition Models in Medical Domain," *2021 6th International Conference on Computer Science and Engineering (UBMK)*, Ankara, Turkey, pp. 763-768, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[25] Alexander Ylnner Choquenaira Florez et al., "Augmentation Techniques for Sequential Clinical Data to Improve Deep Learning Prediction Techniques," *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), Rochester*, MN, USA, pp. 597-602, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[26] Nitesh V. Chawla et al., "Smote: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002. [CrossRef] [Google Scholar] [Publisher Link]

[27] Asma Ben Abacha et al., "An Empirical Study of Clinical Note Generation from Doctor-Patient Encounters," *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Dubrovnik, Croatia, pp. 2291-2302, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[28] Mahdi Abdollahi et al., "Ontology-Guided Data Augmentation for Medical Document Classification," *International Conference on Artificial Intelligence in Medicine*, vol. 12299. pp. 78-88, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[29] Rukhma Qasim et al., "A Fine-Tuned BERT-Based Transfer Learning Approach for Text Classification," *Journal of Healthcare Engineering*, vol. 2022, no. 1, pp. 1-17, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[30] Swati Saigaonkar, and Vaibhav Narawade, "Predicting Chronic Diseases Using Clinical Notes and Fine-Tuned Transformers," *2022 IEEE Bombay Section Signature Conference (IBSSC)*, Mumbai, India, pp. 1-6, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[31] Swati Saigaonkar, and Vaibhav 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, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[32] Swati Saigaonkar, and Vaibhav Narawade, "Domain Adaptation of Transformer-Based Neural Network Model for Clinical Note Classification in Indian Healthcare," *International Journal of Information Technology*, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[33] Swati Saigaonkar, and Vaibhav Narawade, "Explainable Zero-Shot Learning and Transfer Learning for Real Time Indian Healthcare," *International Journal of Informatics and Communication Technology*, vol. 14, no. 1, pp. 91-101, 2025. [CrossRef] [Google Scholar] [Publisher Link]