International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

SM-DBERT: A Novel Symptom-based Technique for Chronic Disease Classification using DISTILBERT

Swati Saigaonkar[1], Vaibhav Narawade[2]
[1]Research Scholar, Dept of Computer Engineering
Ramrao Adik Institute of Technology, D Y Patil Deemed to be University
Nerul, India
swatiavarma@gmail.com
[2]Professor, Dept of Computer Engineering
Ramrao Adik Institute of Technology, D Y Patil Deemed to be University
Nerul, India
vaibhav.narawade@rait.ac.in

Abstract— Machine learning and deep learning models when applied on EHR systems are considerably augmenting the prediction tasks performed on medical data. Humongous amount of information lies in the free form clinical texts. But there exist challenges associated with such kinds of unstructured data. Transformers based models like Bidirectional Encoder Representations from Transformers (BERT) has revolutionized the work. DISTILBERT, a lighter version of BERT, is even promising as the time required is reduced to nearly one-third without losing the performance. In this research work, we present SM-DBERT, Symptom-based Modified DistilBERT architecture designed for Chronic Diseases. The foundation of SM-DBERT is symptomatology, as an optimal model should prioritize symptoms as they are the key indicators. The existing DISTILBERT architecture has been modified by introducing additional layers and extra embeddings of external knowledge and presented along with input ids and attention masks. These extra knowledge helps the model to learn more relevant information. SM-DBERT has demonstrated notable improvement in the results. The accuracy obtained with this novel approach is 0.98 as against the basic DISTILBERT model.

Keywords- Clinical notes, DistilBERT, Electronic Health Record, Embedding Layer, Medical Information Mart for Intensive Care, Natural Language Processing.

I. INTRODUCTION

There has been a good adoption rate of EHR systems in many developed countries. But in many developing countries like India, a substantial amount of medical information is documented in clinical notes like prescriptions, discharge reports, laboratory reports, etc. These unstructured data contain massive amount of information and can provide meaningful insights to medical practitioners.

According to a recent survey, the utilization of Electronic Health Record (EHR) systems has increased by nine times compared to the survey conducted in 2008[1]. EHR systems are capable of managing both structured and unstructured data, such as patient information, admission records, diagnosis and procedure data, vitals of patients, laboratory results, discharge summaries, data from various sensors etc. The unstructured data, though a good source of information, are high-dimensional and heterogeneous leading to increased complexity [2], [3].

Symptoms and diseases, one disease with another disease is often correlated and also follows a sequential pattern or a chronological order. Some examples could be:

- Frequent coughing, cold symptoms, absence of fever, and wheezing are indicative of asthma.
- Recurrent cold symptoms without fever may be a sign of allergic rhinitis.
- Glycosylated Hemoglobin levels can indicate the presence of pre-diabetes or diabetes.
- Diabetes leading to coronary artery disease

Chronic diseases are diseases which last for more than three months[4]. If detected early can lead to better assessment and treatment.

In working with unstructured data such as clinical notes, several challenges arise. These include the heterogeneous nature of the data, as well as its unstructured format. Clinical notes often lack proper grammatical structures and are composed primarily of phrases, which can make it difficult to extract meaningful information. Additionally, the use of abbreviations further complicates the task of processing clinical notes. These challenges highlight the need for specialized techniques and tools to effectively work with unstructured data in the clinical domain. By addressing these obstacles, researchers and practitioners can develop more

accurate and efficient methods for analyzing clinical notes, ultimately leading to improved patient care and outcomes.

Machine learning models replaced the rule-based techniques which were used prominently earlier, but with the advent of deep learning models and specifically after the adoption of recurrent neural networks and its variants, there has been a significant improvements in the field. Transformer based language models have further revolutionized the work. These methods have the potential to pave the way for the development of dependable medical decision support systems and personalized medicine [5].

The following paragraph describes the anatomy of EHR data followed by relevant research that has been conducted in this area. It mainly highlights the work done in the prediction of diseases and states the research gaps found.

EHR Anatomy

EHR data may contain following categories of data:

1. Patient Data - Patient data like patient id, date of birth, date of death (if patient has expired), expire flag (if patient has expired), allergies,and also any other demographic data. Patient names are de-identified in order to maintain anonymity. Also the dates are shifted for the same reason
2. Hospitalization Data- Admission id, location in hospital, insurance, in-time, outtime, etc.
3. Medication - drug names (usually stored as drug code), treatment procedures (coded), inputs(fluids administered to the patients), outputs(fluids excreted or extracted from the patient)
4. Vitals - patient's vitals data
5. Lab tests - type of measurement and its value
6. Diagnoses - the diagnoses information(usually coded)
7. Clinical notes - notes filled by clinicians. It can be discharge summaries, lab reports,
8. ECG reports, Nursing notes, physician's notes, General notes, etc

As with many countries investing in EHR systems, and with the advancements in the field of artificial intelligence, there arises a need to make optimum use of technology and aid the medical practitioners, care providers with proper inputs. EHR systems store patient related information. Initial work on such systems started with rule based techniques. Electronic Health Record (EHR) systems consist of both structured and unstructured data. Structured data are organized in a systematic manner and is typically stored in database systems. These type of data are often assigned codes such as ICD9 codes for diagnoses and procedures, LOINC codes for laboratory test orders and results, and CPT codes for procedures, including surgical ones. This information can also be used for statistical analysis, insurance claims, and other related purposes.

The data bases that are publicly available for research purposes are Medical Information Mart for Intensive Care (MIMIC), I2b2 datasets, eICU Collaborative Research Database. They contain information on critically ill ICU patients. The MIMIC III database consists of more than 50,000 ICU admissions [6], while the eICU Collaborative Research Database contains nearly 2,00,000 admissions [7]. I2b2 datasets contain data from annual NLP challenges. The common tasks that are performed on EHR data and have been studied are entity extraction using deep learning models[8, 9], length of ICU stay prediction using machine learning and the publicly available data sets [10], mortality prediction using deep learning and making use of unstructured data like clinical notes[11],hospital readmission prediction[12], Phenotyping, [13] highlights the advancement of electronic health record (EHR) phenotypes to detect individuals who have stage 4 solid-tumor cancer or stage 4-5 chronic kidney disease (CKD), and finally disease prediction.

The work in [14] focuses on developing models for predicting chronic cough. Machine learning models, namely, Logistic Regression, SVM, kNN, Random Forest, BiLSTM, and deep learning model BERT were used along with NLP techniques. With only structured data, sensitivity and specificity was 0.856 and 0.866, respectively, which later increased to 0.952 and 0.930 when both structured and unstructured data were used. This demonstrates that unstructured data improves the model. But the time interval in the study was limited to 120 days and the validation of entire study cohort was not done.

In the research paper[15], the authors have first extracted clinical events related to cervical cancer, and represented them with appropriate techniques in order to predict cervical cancer. Four classifiers namely, Random Forest, SVM, Bernoulli Naive Bayes, the Complement Naive Bayes classifier, were used. The classifiers were evaluated using various metrics. Random Forest gave the highest score for precision while Bernoulli Naive Bayes for recall. The future study involves inclusion of primary health records that precede the hospital admission.

The work in [16] is based on vector space and NMF topic modeling to infer demonstrative feature space and then it was used to make accurate disease prediction using deep neural architectures. But this study does not consider clinical data in real-time. The research work in [17] employed NLP algorithms and supervised machine learning techniques, which included discriminative sequence labeling models like Conditional Random Field and classifiers such as Support Vector Machine (SVM) and Random Forest. These techniques were used to identify positive mentions of cough through entity recognition and to classify contextual qualifiers for each mention and then detect chronic cough.

BERT based models have proved to provide better results. In [18], transformer-based architecture, based on time, is used to predict depression. Their model gave an increase of 0.06 in the PRAUC, or precision-recall area under the curve in

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

comparison to the base line model. The work in [19], BEHRT gave nearly 8% improvement as compared to the models before it. Interpretability is a bigger concern when dealing with deep learning models and this work has focused on the same.

The work specified in research paper[20], focuses on developing a model by using transformer based language models RoBERTa and BERT. Polish data were used for training and then later the model was fine tuned in order to perform classification (multi-label). The performance was improved after adding clinical text data and the future direction is to improve the accuracy further by constructing an ensemble of diagnosis predictors.

The work mentioned in [21] is about predicting cancer metastasis from clinical notes. They fine-tuned BERT-based models namely BERT, BlueBERT, BioBERT, ClinicalBERT, and PubmedBERT and found that PubmedBERT gave better performance, which was named as METBERT. This fine-tuned model was then tested on an independent data-set by employing transfer learning and it demonstrated high performance with an AUC of 0.94.

It is evident from the literature that transformer-based models are outperforming other models in various tasks, indicating the need for further investigation of their potential. This highlights the significance of exploring transformer-based models in more detail, as they have shown promising results in numerous applications. As such, it is crucial to delve deeper into their capabilities to better understand how they can be optimized for disease prediction task.

II. MATERIALS AND METHODS

A. Architecture

The overall architecture of SM-DBERT is presented in Figure 1. The inputs to the system are structured as well as unstructured data of patients. Cohort selection is done next to extract the data related to our study of chronic diseases. The extracted data are then pre-processed as per the need of the study and then the modified DISTILBERT architecture is fine-tuned with the processed data, which is further applied to the classification tasks to predict the diseases.

B. Study Cohort

The data that have been used is clinical notes taken from the MIMIC III database. It is a comprehensive collection of data related to intensive care unit patients with more than 50,000 admissions. Structured data like diagnosis code, subject id, admission id, and unstructured data mainly consisting of clinical notes from the NOTEVENTS table are used.

The study cohort consists of patients with diseases namely ‘Chronic obstructive asthma’, ‘Rheumatic heart failure’, ‘Chronic kidney disease’, ‘Rheumatoid arthritis’ having ICD9 codes ‘49320’, ‘39891’, ‘5853’, and ‘7140’ respectively. To conduct the research, we utilized three tables from the MIMIC database: NOTEEVENTS, D_ICD_DIAGNOSES, and DIAGNOSES_ICD. While the other two tables include organized data, the NOTEEVENTS table contains unstructured clinical notes. We established a relationship between the tables as follows:

DIAGNOSES_ICD ⋈ DIAGNOSES_ICD.SUBJECT_ID = NOTEEVENTS.SUBJECT_ID (NOTEEVENTS)

200 patients of every disease were considered resulting in total of 55,794 clinical notes.

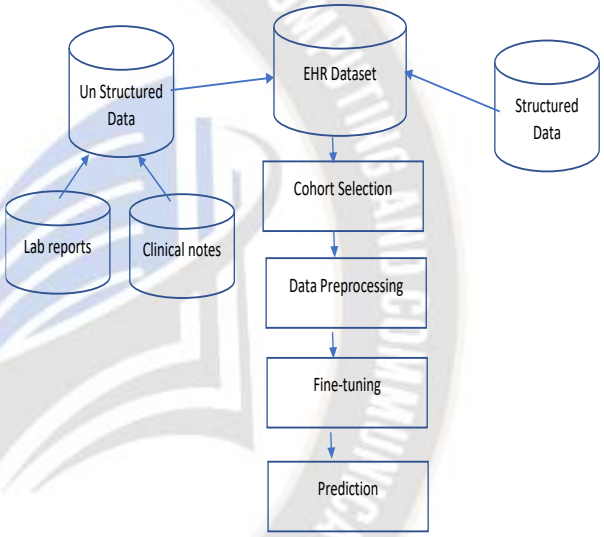

Fig 1 : Overall Process

C. Data Pre-processing

To maintain the quality and precision of the model, several steps were taken to pre-process the dataset. Specifically, all duplicate values were eliminated, as well as patients with no clinical notes were eliminated too. Furthermore, any special characters and URLs present in the data were removed in order to simplify the text and make it easier to analyze. To standardize the data further, regular expressions were employed to substitute any non-alphanumeric characters with a space character. These steps were crucial in preparing the dataset for analysis, allowing for more accurate and effective modelling of the clinical notes data. Figure 2 highlights the pre-processing steps. Finally, the data were pre-processed such that it could be served to a BERT-based model.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

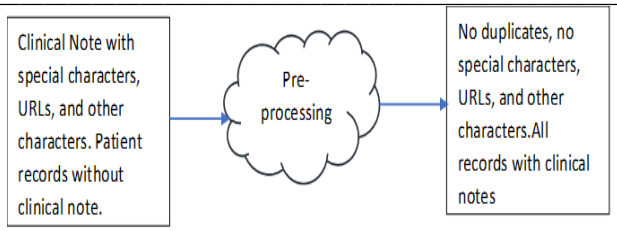

Fig 2: Pre-processing process

PATIENT/TEST INFORMATION:\nIndication: Chest pain. Congestive heart failure. Coronary artery disease. Left ventricular function. Shortness of breath.\nHeight: (in) 64\nWeight (lb): 190\nBSA (m2): 1.92 m2\nBP (mm Hg): 150/97\nHR (bpm): 103\nStatus: Inpatient\nDate/Time: [**2145-6-21**] at 09:05\nTest: Portable TTE (Complete)\nDoppler: Full doppler and color doppler\nContrast: None\nTechnical Quality: Adequate\n\n\nINTERPRETATION:\n\nFindings:\n\nLEFT ATRIUM: Moderate LA enlargement. Elongated LA.\n\nRIGHT ATRIUM/INTERATRIAL SEPTUM: Normal RA size.\n\nLEFT VENTRICLE: Mild symmetric LVH. Severely dilated LV cavity.

Fig 3: Sample Clinical Note

patient test information indication chest pain congestive heart failure coronary artery disease left ventricular function shortness of breath height in 64 weight lb 190 bsa m2 1 92 m2 bp mm hg 150 97 hr bpm 103 status inpatient date time 2145-6-21 at 09 05 test portable tte complete doppler full doppler and color doppler contrast none technical quality adequate interpretation findings left atrium moderate la enlargement elongated la right atrium interatrial septum normal ra size left ventricle mild symmetric lvh severely dilated lv cavity

Fig 4: Sample Clinical Note After Pre-processing

Figure 3 shows a sample clinical note taken from MIMIC-III data store and Figure 4 shows the same sample after pre-processing. The case of the data were changed to lower case and all the steps mentioned previously were applied to the sample.

After performing all the pre-processing steps on the clinical notes, it resulted in 54299 clinical notes as records with no clinical notes were eliminated.

D. Methodology

The techniques that have been used on EHR data can be summarized as follows:

Rule based techniques:

It consists of a plethora of rules, which are applied to the antecedents and consequent then can be inferred from it. Rules are generally developed by using expert knowledge, knowledge bases like UMLS (Unified Medical Language System) can also be used. Rules can also be generated automatically from the training dataset. Advantages of rule-based system is that they are simple but they are of less use if new concepts/words/patterns are encountered resulting in low recall.

Machine learning techniques:

Machine learning is a branch of artificial intelligence that learns from the data itself. It can be classified into 2 types namely supervised and unsupervised.

1. Supervised technique- in supervised techniques, labels need to be provided along with the feature set, so that the model learns from the data. Classification, regression are the tasks that follow supervised learning methodology.

2. Unsupervised technique- in unsupervised techniques, labels are not provided and data having similar characteristics are grouped together. Clustering is the task that follows unsupervised methodology

The most common machine learning techniques that are used are: Logistic regression, SVM, Naive Bayes, Random Forests, Conditional Random Fields

Apart from these methods, there are other methods too which are used. Although these statistical methods provide good results, they are incapable of handling high dimensional, heterogeneous multimodal data [22].

Deep Learning techniques:

Deep learning models offer a promising approach as they can automatically extract relevant features from the raw input data, without the need for manual feature engineering like traditional machine learning methods. This means less preprocessing is required, making the modeling process more efficient. Various deep learning techniques are used for different applications, including Convolutional Neural Networks (CNNs) for image recognition and Recurrent Neural Networks (RNNs) for sequence prediction.

CNNs are widely used for image data analysis. However, researchers have also applied CNNs to textual data with some success. Nonetheless, the research community is currently exploring more advanced solutions for text analysis.

RNNs and their variants are widely used in sequence data analysis. RNNs have the ability to handle time dependencies in the input data. However, the traditional RNNs have limitations such as vanishing gradients and exploding gradients issues, which limit their performance. To address these limitations, advanced versions of RNNs such as LSTM (Long Short-Term Memory) and GRU (Gated Recurrent Unit) have been developed. Modern RNNs leverage memory cells and gates that enable them to preserve information for extended periods when contrasted with traditional RNNs. Also, these networks incorporate additive updates instead of multiplicative updates, which effectively address the issue of vanishing gradients..

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

BiLSTM is a type of LSTM, where the prefix "bi" stands for bidirectional. This model has the capability to process input sequences from both the forward and backward directions, which aids in better understanding the context of the input data.

Transformers have become a widely used type of neural network architecture in the field of natural language processing. Unlike RNNs, transformers can handle long-term dependencies and are capable of processing entire sentences or documents at once, making them a promising solution for advanced NLP applications.

One widely used transformer model in natural language processing is BERT, created by Google. BERT is pre-trained on a vast amount of text before fine-tuning for a specific natural language processing (NLP) task. This model utilizes stacked encoders to facilitate learning of contextual relationships between words in a sentence. During the pre-training phase, BERT undergoes two tasks: masked language modeling and next sentence prediction. In MLM, random words in a sentence are masked, and BERT is trained to predict them. In NSP, the model checks to see if the second sentence follows the first. BERT's bidirectional approach allows for a better comprehension of sentence context in comparison to other models.

DISTILBERT is a more compact variant of BERT, achieved through a technique called distillation. This method involves compressing the number of layers in the original model, resulting in a lighter and faster model.

In our previous work[23], we compared BERT and DISTILBERT and found that DISTILBERT gave better performance and also the time required to train the model was just one third of the time required to train the BERT model. The next sub section describes our proposed architecture which is based on DISTILBERT.

The modified DISTILBERT architecture is displayed in Figure 5.

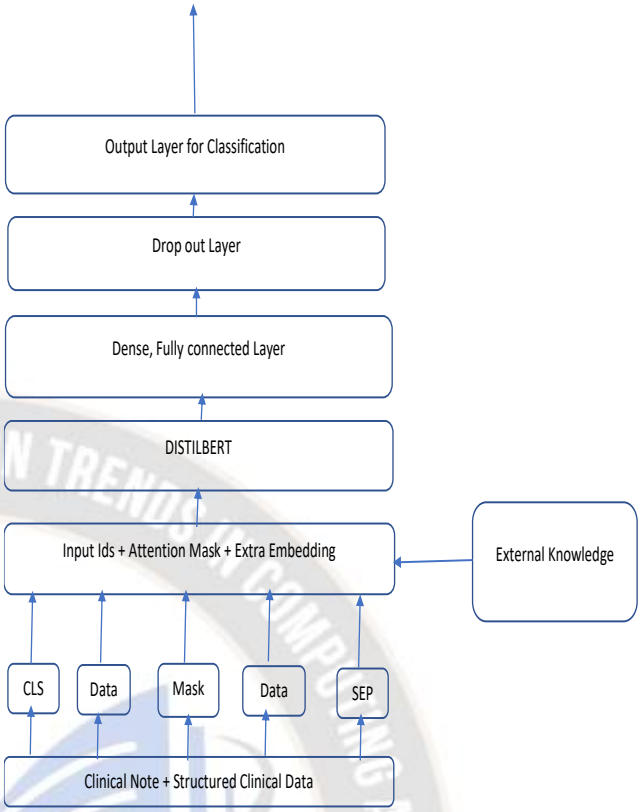

Fig 5 : Architecture of the fine-tuned model

The input to SM-DBERT comprises both structured and unstructured data, which undergo preprocessing, as described in the preceding section, and tailored to meet the specific requirements of the BERT model. The special tokens used in BERT, including [CLS], [SEP], and [MASK], are used to separate and encode the input data. [CLS] is used at the beginning of the input and represents the classification of the entire sequence. [SEP] is used to separate two different sentences or segments within a single sequence. [MASK] is used during training to randomly mask certain tokens in the input and force the model to predict the correct token. Together, these tokens help BERT to effectively encode and process input data for natural language processing tasks. This is then fed to the model.

The model receives three input layers:

The system comprises three input layers: "inps", which accepts integer sequences of length 512; "masks", which accepts integer sequences of length 512 and represents the attention mask for the input sequences; and "extra_inps", which accepts integer sequences of varying length, meaning that the number of tokens may differ for each input sequence. Our novel approach involves the incorporation of external knowledge to enhance the model's performance by providing additional and crucial information. This results in the model paying more attention to the external knowledge, thereby improving its overall accuracy and effectiveness, i.e. the embeddings from “inps” and

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

"extra_inps" are concatenated along the feature dimension (i.e., the output dimension), and passed through a fully connected dense layer with 512 units and a ReLU activation function.

Let the input sequence be denoted by x = {x1, x2, ..., xn}, where xi is the ith token in the sequence. The input sequence is first encoded using the standard DISTILBERT encoder, denoted by E_distilbert, to obtain the hidden states H = {h1, h2, ..., hn}:

H = E_distilbert(x)

The external knowledge is represented as a set of embeddings K = {k1, k2, ..., km}, where ki is the i^{th} external knowledge embedding.

The external knowledge embeddings are concatenated with the input sequence embeddings to obtain the modified input embeddings E = {e1, e2, ..., en}:

ei = concat(xi, K)

The modified input embeddings are then passed through a set of extra layers, denoted by L_extra, to obtain the final output sequence H_extra = {h'_1, h'_2, ..., h'_n}:

H_extra = L_extra(E)

To mitigate overfitting, the output of the dense layer is passed through a dropout layer with a rate of 0.5. Prior to model training, the dataset was partitioned into separate training and testing sets as per the standard practice. To mitigate the issue of class imbalance that is often encountered, class weights were introduced. These weights were incorporated into the model during training. Finally, the model outputs a probability distribution over the number of output classes using a SoftMax activation function.

The hyperparameter for learning rate in our model was set to 1e-4, and we trained the models throughout the course of four epochs.

III. RESULTS AND DISCUSSION

TABLE I presents the DISTILBERT model's performance metrics, while TABLE II shows the performance metrics for our proposed model.

Table 1. Metrics when DISTILBERT was used

	Precision	Recall	F1-score	Support
Asthma	0.8	0.72	0.76	2463
Arthritis	0.74	0.78	0.76	3043
Renal disease	0.75	0.74	0.75	2925
Heart disease	0.70	0.75	0.72	2429
Accuracy			0.75	10860
Macro avg	0.75	0.75	0.75	10860
Weighted avg	0.75	0.75	0.75	10860

Table 2. Metrics when SM-DBERT was used

	Precision	Recall	F1-score	Support
Asthma	0.98	0.98	0.98	2973
Arthritis	0.98	0.99	0.99	2490
Renal disease	0.99	0.97	0.98	2849
Heart disease	0.97	0.98	0.98	2548
Accuracy			0.98	10860
Macro avg	0.98	0.98	0.98	10860
Weighted avg	0.98	0.98	0.98	10860

Equations (1), (2), and (3) provide the formulae for calculating precision, recall, and f1-score, respectively:

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

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

Table I and Table II display the precision, recall, F1 score, and support for each class, which are standard metrics for evaluating the performance of a multi-class classification model. Our study employed these metrics to assess the performance of our suggested model on a given dataset comprising 10860 samples across four distinct classes.

The precision of our model for classes Asthma, Arthritis, Renal disease, and Heart disease was 0.98, 0.98, 0.99, and 0.97, respectively. This means that out of all the samples our model predicted as belonging to a particular class, the percentage that was correctly classified was 98%, 98%, 99%, and 97%, respectively as against 80%, 74%, 75%, 70% when DISTILBERT was used.

The recall of our model for classes Asthma, Arthritis, Renal disease, and Heart disease was 0.98, 0.99, 0.97, and 0.98, respectively. This means that out of all the samples that actually belonged to a particular class, the percentage that our model correctly classified was 98%, 99%, 97%, and 98%, respectively as against 72%, 78%, 74%, and 75% when DISTILBERT was used.

In addition, our proposed model exhibits better macro-average value, as well as weighted average value, compared to the DISTILBERT model. During the initial training phase of 4 epochs, the DISTILBERT model achieved a training accuracy

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

of 0.7808. However, our model outperformed DISTILBERT, achieving a significantly higher training accuracy of 0.9781. In terms of validation accuracy, our model again outperformed DISTILBERT with a score of 0.9808 as compared to 0.7464 obtained by the DISTILBERT model. These results suggest that our proposed model is more accurate and effective than the DISTILBERT model for predicting the given disease. This shows that our model has outperformed the DISTILBERT model for same number of epochs. Figure 6 and Figure 7 shows the graphs for accuracy vs epochs and loss vs epochs respectively. The accuracy and validation accuracy both show an increasing trend over the epochs. This indicates that the model's performance improved with each epoch, as the accuracy values increased. The validation accuracy closely follows the accuracy, suggesting that the model generalizes well to unseen data. Both the accuracy and validation accuracy curves seem to converge as the number of epochs increases. Similarly, the loss and validation loss curves also appear to converge. This suggests that the model has reached a stable point and further training may not have resulted in significant improvements.

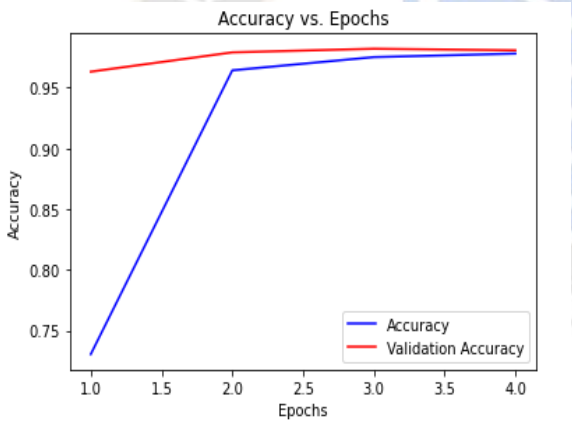

Fig 6 : Graph showing accuracy vs epochs for SM-DBERT

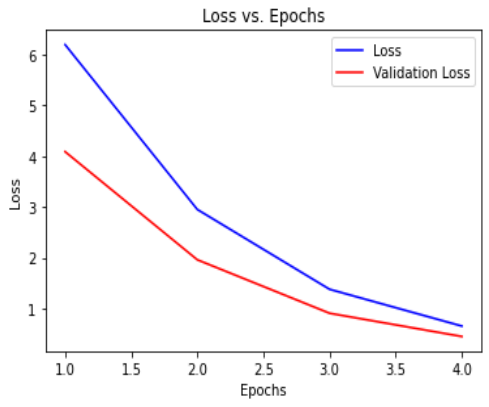

Fig 7 : Graph showing loss vs epochs for SM-DBERT

IV. CONCLUSION AND FUTURE WORK

As the utilization of EHR systems continues to rise, EHR data have been used for various tasks, one of them being disease prediction. Chronic diseases if detected early can lead to better disease management. Clinical notes contain humongous data and with the help of NLP, it can be used in assisting disease prediction. According to recent research, BERT models have shown to perform better than conventional NLP methods. However, the lighter version of BERT known as DISTILBERT has shown to outperform BERT while also taking significantly less time to fine-tune the model. The study utilized DISTILBERT as the foundational architecture, and through the incorporation of supplementary information as an additional embedding layer, our system demonstrated significantly better performance in comparison to the fine-tuned DISTILBERT model. The model demonstrated an accuracy of 0.9808, which is significantly higher than the accuracy achieved by the fine-tuned DISTILBERT model, which was only 0.7464.

The model shows promising results for predicting the four chronic diseases. In the future, this model can be extended to predict other diseases by incorporating relevant data and features. Additionally, transfer learning can be employed to fine-tune the existing model on different datasets with similar features, which can improve the accuracy and efficiency of the model. Overall, the model has significant potential for application in other healthcare domains for predicting various diseases and it also reduces the reliance on EHRs, which is a challenge in developing nations as they have lower rates of structured EHR implementation.

ACKNOWLEDGMENT

We would like to thank our college RAIT, and D Y Patil Deemed to be University for providing us with the state-of-the-art facilities to carry out our research.

REFERENCES

[1] J. Henry, Y. Pylypchuk, T. Searcy, V. Patel, “Adoption of Electronic Health Record Systems among U.S. Non-Federal Acute Care Hospitals 2008–2015”, *ONC Data Brief*, No. 35, 2016.

[2] 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 J. Biomed. Heal. Inform.*, vol. 22, no. 5, pp. 1589–1604, Sep. 2018.

[3] Y. Meng, W. F. Speier, M. Ong, and C. W. Arnold, “HCET: Hierarchical clinical embedding with topic modeling on electronic health record for predicting depression,” *IEEE J. Biomed. Heal. Inform*, doi: 10.1109/JBHI.2020.3004072

[4] Rayan Alanazi, "Identification and Prediction of Chronic Diseases Using Machine Learning Approach", *Journal of Healthcare Engineering*, vol. 2022, Article ID 2826127, 9 pages, 2022. https://doi.org/10.1155/2022/2826127

[5] Sheikhalishahi S, Miotto R, Dudley J, Lavelli A, Rinaldi F, Osmani V, "Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review", *JMIR Med Inform 2019*;7(2):e12239, DOI: 10.2196/12239

[6] Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, et al. "*MIMIC-III, a freely accessible critical care database. Sci Data. 2016*";3: 160035, https://doi.org/10.1038/sdata.2016.35

[7] Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O, "The eICU Collaborative Research Database, a freely available multi-center database for critical care research", *Sci Data.* 2018 Sep 11;5:180178. doi: 10.1038/sdata.2018.178. PMID: 30204154; PMCID: PMC6132188.

[8] 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, Oct. 2021.

[9] N. Liu, Q. Hu, H. Xu, X. Xu and M. Chen, "Med-BERT: A Pre-Training Framework for Medical Records Named Entity Recognition," *IEEE Transactions on Industrial Informatics*, doi: 10.1109/TII.2021.3131180.

[10] Wu J, Lin Y, Li P, Hu Y, Zhang L, Kong G, "Predicting Prolonged Length of ICU Stay through Machine Learning", Diagnostics (Basel). 2021 Nov 30;11(12):2242. doi: 10.3390/diagnostics11122242

[11] Ye, J., Yao, L., Shen, J. et al. Predicting mortality in critically ill patients with diabetes using machine learning and clinical notes. BMC Med Inform Decis Mak 20, 295 (2020). https://doi.org/10.1186/s12911-020-01318-4

[12] Huang, Kexin et al. "ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission." ArXiv abs/1904.05342 (2019)

[13] Ernecoff, N.C., Wessell, K.L., Hanson, L.C. et al,"Electronic Health Record Phenotypes for Identifying Patients with Late-Stage Disease: a Method for Research and Clinical Application", J GEN INTERN MED 34, 2818–2823 (2019). https://doi.org/10.1007/s11606-019-05219-9

[14] Xiao Luo, Priyanka Gandhi, Zuoyi Zhang, Wei Shao, Zhi Han, Vasu Chandrasekaran, Vladimir Turzhitsky, Vishal Bali, Anna R. Roberts, Megan Metzger, Jarod Baker, Carmen La Rosa, Jessica Weaver, Paul Dexter, Kun Huang, "Applying interpretable deep learning models to identify chronic cough patients using EHR data", Computer Methods and Programs in Biomedicine, Volume 210, 2021, 106395, https://doi.org/10.1016/j.cmpb.2021.106395

[15] Weegar R, Sundstr¨om K. "Using machine learning for predicting cervical cancer from Swedish electronic health records by mining hierarchical representations", PLoS One. 2020 Aug 21;15(8):e0237911. doi: 10.1371/journal.pone.0237911. PMID: 32822401; PMCID: PMC7444577

[16] T. Gangavarapu, G. S. Krishnan, S. K. S and J. Jeganathan, "FarSight: Long-Term Disease Prediction Using Unstructured Clinical Nursing Notes," in IEEE Transactions on Emerging Topics in Computing, vol. 9, no. 3, pp. 1151-1169, 1 July-Sept. 2021, doi: 10.1109/TETC.2020.2975251

[17] Bali, V., Weaver, J., Turzhitsky, V. et al. "Development of a natural language processing algorithm to detect chronic cough in electronic health records", BMC Pulm Med 22, 256 (2022). https://doi.org/10.1186/s12890-022-02035-6

[18] Y. Meng, W. Speier, M. K. Ong and C. W. Arnold, "Bidirectional Representation Learning From Transformers Using Multimodal Electronic Health Record Data to Predict Depression," in IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 8, pp. 3121-3129, Aug. 2021, doi: 10.1109/JBHI.2021.3063721

[19] Li, Y., Rao, S., Solares, J.R.A. et al., "BEHRT: Transformer for Electronic Health Records.", Sci Rep 10, 7155 (2020). https://doi.org/10.1038/s41598-020-62922-y

[20] Anetta K, Horak A, Wojakowski W, Wita K, Jadczyk T., "Deep Learning Analysis of Polish Electronic Health Records for Diagnosis Prediction in Patients with Cardiovascular Diseases", J Pers Med. 2022 May 25;12(6):869. doi: 10.3390/jpm12060869. PMID: 35743653; PMCID: PMC9225281.

[21] Liu K, Kulkarni O, Witteveen-Lane M, Chen B, Chesla D, "MetBERT: a generalizable and pre-trained deep learning model for the prediction of metastatic cancer from clinical notes", AMIA Annu Symp Proc. 2022 May 23;2022:331-338. PMID: 35854741; PMCID: PMC9285138

[22] Ernecoff, N.C., Wessell, K.L., Hanson, L.C. et al,"Electronic Health Record Phenotypes for Identifying Patients with Late-Stage Disease: a Method for Research and Clinical Application", J GEN INTERN MED 34, 2818–2823 (2019). https://doi.org/10.1007/s11606-019-05219-9

[23] S. Saigaonkar and V. Narawade, "Predicting chronic diseases using clinical notes and fine-tuned transformers," 2022 IEEE Bombay Section Signature Conference (IBSSC), Mumbai, India, 2022, pp. 1-6, doi: 10.1109/IBSSC56953.2022.10037512