Predicting chronic diseases using clinical notes and fine-tuned transformers

Ms. Swati Saigaonkar
Research Scholar, Dept of Computer Engineering,
Ramrao Adik Institute of Technology,
D Y Patil Deemed to be University,
Nerul, India
swatiavarma@gmail.com

Dr. Vaibhav Narawade
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*—Electronic health records(EHR) have been used extensively by researchers lately to gain insights and use them as clinical informatics. EHR data contains structured data, as a result of having information systems in-place, and also unstructured data like clinical notes. These unstructured data have a huge scope of exploration and can derive meaningful insights. Challenges exists like the heterogeneous and multi modal nature of such data. This work provides insights into the EHR data, the datasets available for research, the tasks that can be performed on them, the methods that can be applied on them, and then demonstrates how BERT and DistilBERT can be fine-tuned on the medical datasets to predict chronic diseases like asthma, renal diseases, heart diseases and arthritis and how DISTILBERT can be a preferred option over BERT. Both the models BERT and DISTILBERT have been pre-trained and then fine tuned to predict the chronic diseases from the clinical notes.**

Keywords—BERT, DistilBERT, Electronic Health Record, MIMIC, Clinical notes, Natural Language Processing

I. INTRODUCTION

Electronic health record (EHR) recently has gained attention by the researchers, as there has been an extensive use of EHR data especially in developed countries. A tremendous growth can be seen as about 84 percent hospitals are using EHR for recording patient's data. This increase has marked to be nine-fold as compared to 2008[1]. EHR includes data like demographic information of patients, diagnoses and procedures by the care givers, medications or prescriptions, clinical notes by clinicians, images, laboratory results, discharge notes, transfer notes,etc. thus making it high dimensional and heterogenous [2], [3]. Clinical notes written by doctors, family doctors, nurses contain humongous data which if explored can help in early disease identification. With early proof of a disease, it becomes easy for clinicians to manage disease in a better way, and also lead to efficient resource utilization.

Many machine learning as well as deep learning algorithms have been implemented to work on EHR data. This effort may further provide a basis for highly reliable clinical decision support systems or personalized precision medicine[4]. Different modalities of EHR data are being explored for various tasks. There are many developing countries which still have not adopted EHRs, but the entire medical history lies in the form of clinical notes. From Rule based to Machine Learning to Deep Learning, there has been an improvement in methodologies of dealing with EHR data.

Clinical notes pose a challenging problem because of its heterogeneity. The other significant challenge is its un-structuredness, reason being they contain abbreviations which are understood mainly by clinicians only, use of incomplete sentences or phrases, non-grammatical text, etc.

There are various diseases which have sequential symptoms like recurring cold, then fever.

Few examples are as follows:

- Repeated cough, cold, no fever, wheezing - Asthama
- Repeated cold, no fever - Allergic rhinitis
- Glycosylated Haemoglobin - Pre diabetes and Diabetes

Symptoms, diseases are correlated and often clinical events occur in some chronological order. An example of this could be detection of risk factors for coronary artery disease for patients who are suffering from diabetes.

To summarize, the challenges of working with unstructured data like clinical notes are:

- Heterogeneous nature
- Clinical notes – unstructured
- No proper grammatical structures
- Contains mostly phrases
- Use of abbreviations

The rest of the paper is organized as follows. Section II highlights the basics and background needed for the study. Section III highlights the EHR anatomy and the tasks that are performed on it. Section IV presents the literature review. Section V highlights the materials the methods used in the study. Section VI presents the results and discussion and finally section VII presents the conclusion and future work.

II. BASICS AND BACKGROUND

Some terms and techniques to understand that goes while developing systems to work on EHR data:

A. Data Sets

Data can come from EMR, i.e., data that gets generated from one hospital, or from EHR, which combines data from multiple sources. These are recordings that come from practitioners, sensors and instruments used, medical historical data, operational data and recordings of medical events. Some of the data sets that are publicly available and are often used are

978-1-6654-9291-1/22/$31.00 ©2022
IEEE

MIMIC III - MIMIC stands for Medical Information Mart for Intensive Care. It is a database which consists of information of ICU patients admitted between the years 2008 and 2014. The records are deidentified. Records of more than 50,000 ICU admissions are stored. It contains 26 tables which has been recorded during the course of the treatment [5].

I2b2 datasets- the data sets are the result of annual NLP challenges that were conducted as a part of the i2b2 project and are publicly available for use after signing up the agreement.

eICU Collaborative Research Database – this too contain data of critically ill ICU patients and records more than 2,00,000 admissions. It contains 31 tables. The data too is freely available [6].

B. Data characteristics

Structured data

The data can contain some structured data like patient demographics, admission details, procedures conducted, diagnosis details, prescriptions, charted events, services, transfers, laboratory results etc. There is a structure to this as it generally comes from some operational system which the hospital uses.

Unstructured data

Apart from the structured data, these kind of EHR data also contains unstructured data like clinical notes which can be discharge summaries, clinician's notes, nursing notes, laboratory results etc.

The data must be vectorized and then applied to DL models to make some predictions or determine some outcome.

C. Codes

ICD 9 codes are commonly used to assign codes to diagnoses and procedures, LOINC codes are commonly used to assign codes for laboratory test orders and results, CPT codes are used to assign codes to procedures, mainly surgical, insurance claims, and other information which can be used for statistical purposes. There are other coding systems too which are also used. Apart from serving medical purposes, these coding systems also helps in research.

III. EHR ANATOMY AND COMMON TASKS

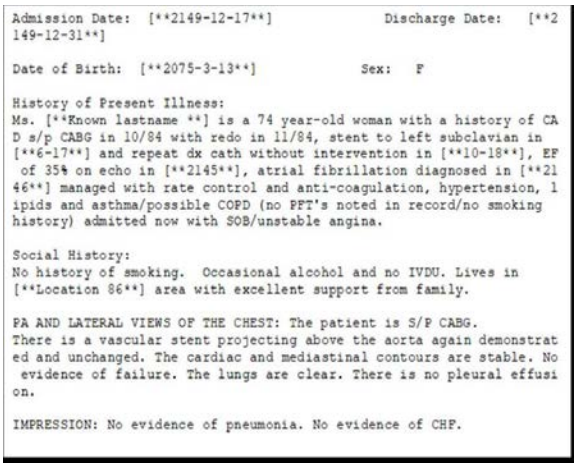
A. EHR Anatomy

EHR systems can be considered as operational systems which stores information of a patient, right from demographics to vitals to diagnosis and to medication. Not just this structured data but it also stores unstructured clinical data.

A major challenge when dealing with EHR data is its heterogeneity. Listed below are some of the data types which poses difficulty.

- Some quantities like body mass index which are numerical in nature.
- Date and time data type, as it occurs while storing information like date of admission, date of discharge, etc.
- Codes which are used to for coding of diagnosis, drugs, procedures etc.
- Natural language which is used while writing of clinical notes.
- Time series data which can be vital sign signals.

An example of Clinical Note from MIMIC- III is given in Fig. 1.



```
Admission Date:  [**2149-12-17**]              Discharge Date:   [**2
149-12-31**]

Date of Birth:  [**2075-3-13**]             Sex:   F

History of Present Illness:
Ms. [**Known lastname **] is a 74 year-old woman with a history of CA
D s/p CABG in 10/84 with redo in 11/84, stent to left subclavian in
[**6-17**] and repeat dx cath without intervention in [**10-18**], EF
 of 35% on echo in [**2145**], atrial fibrillation diagnosed in [**21
46**] managed with rate control and anti-coagulation, hypertension, l
ipids and asthma/possible COPD (no PFT's noted in record/no smoking
history) admitted now with SOB/unstable angina.

Social History:
No history of smoking.  Occasional alcohol and no IVDU. Lives in
[**Location 86**] area with excellent support from family.

PA AND LATERAL VIEWS OF THE CHEST: The patient is S/P CABG.
There is a vascular stent projecting above the aorta again demonstrat
ed and unchanged. The cardiac and mediastinal contours are stable. No
 evidence of failure. The lungs are clear. There is no pleural effusi
on.

IMPRESSION: No evidence of pneumonia. No evidence of CHF.
```

Fig 1 : Clinical Note Sample from MIMIC

B. Common Tasks

The common task that can be performed on EHR data are listed down in Table I.

TABLE I. COMMON TASKS ON EHR

Task	Description
Concept extraction	Drug names, treatment names, medicines name
Entity representation	Vector representation of patients, drug, diagnosis, medical concepts
Mortality prediction	Death of the patient after hospitalisation
Disease prediction	Prediction of one disease or multiple diseases
Length of stay prediction	Length of stay of patients in the ICU or hospital
Readmission	Possibility of re admission and after how many days
Phenotyping	New discovery of traits of diseases
Disease clustering	Clusters of diseases as per some characteristics
Patient clustering	Clusters of patients as per some characteristics
Patient trajectory modelling	Course of illness over time and also the patient's actions
Clinical adverse event detection	Drug adverse effect, hospitalisation, heart failure, death etc.

IV. LITERATURE REVIEW

The review of various tasks that can be performed on the EHR data has been mentioned below.

A. Concept identification, extraction and representation

In the research paper[7], deep learning has been used to process clinical notes and extract entities. A combination of local and global context is used to predict the clinical entities. The methodology that is used is Combination of CNN, Bi-LSTM, and CRF. The dataset used are 2010 and 2012 versions of i2b2 which have given following results, on I2b2- 2010, F1 score- 93.57 and on I2b2- 2012, F1 score- 86.11. It performs entity identification only. In the research work [8], the authors have developed framework for recognition of entities from clinical notes. The methodology that are used is BiLSTM, BERT. The dataset used is CMeEE (Chinese), CMR which is not for public use. The results obtained are F1 score is 0.9320, Precision is 0.9509, Recall is 0.9138. The research specified in[9] have developed a framework for rare disease identification which do not have ICD codes in the data, by employing weak supervision with customized rules and ontology without the annotation from domain experts. The methodology that is used are Ontologies, Rules for weak labelling, and BERT. The dataset used is MIMIC-III (discharge summaries). The results obtained are Precision= 63.7, Recall= 78.1, F1 score= 70.2.

B. Disease Prediction

The authors in research paper[10] have developed a model for identification of patients with chronic cough using deep learning, NLP and traditional ML techniques. The model/ methodology that is used is BERT and attention mechanism for interpretability.The dataset used is : EHR data MIMIC III (medication and diagnosis) including Clinical notes. The results obtained are Sensitivity is 0.856, specificity is 0.866. After using clinical notes, sensitivity was 0.952, specificity was 0.930. The authors in research paper[11] have employed deep learning model based on time i.e. transformer architecture to predict future diagnoses of depression. Bidirectional representation learning has been performed on EHR data. The model that is used is BERT for EHR. The dataset used is their own EHR dataset,by Bioengineering department, University of California, Los Angeles. The results obtained are Precision recall was 0.76 and the future directions are that the future studies could be based on BERT on clinical notes, to improve the performance further, There is no multiclass support, and also no lab Results were included. The work specified in research paper[12] is about the model to identify the onset of the disease by making use of initial symptoms. The methodology that is used are Vector space representation technique Doc2Vec and topic modelling. Other techniques like MLP, ConvNet, LSTM, Bi-LSTM, Conv-LSTM, Seg-GRU are compared. The dataset used is MIMIC-III including clinical notes. The results obtained are that AUPRC is improved by 19.34 percent and AUPRC is improved by 5.41 percent but the study did not include real-time clinical data.

In the research paper[13], the authors have used patient's historical data (EHR) to forecast diagnoses, and predict zero or more labels and also probability for every disease. The model that is used is self-attention (multi-head), positional encoding, and masked language model (MLM). The dataset used is Clinical Practice Research Datalink (CPRD), UK, linked with Hospital Episode Statistics, age and diagnosis codes. The results obtained are APS=0.216, AUROC =0.904 but the attention-visualization tools were not present.

C. Length of Stay Prediction

The work done in research paper[14]is development of a framework to predict length of ICU stay. Datasets like eICU Collaborative Research Database and MIMIC-III were used for development of model and validation respectively. Four machine learning techniques such as random forest, SVM, deep learning, and gradient boosting decision tree (GBDT) were used, out of which GBDT gave the best performance. No clinical notes were used.

D. Readmission

The authors of research paper[15] have developed a model named as ClinicalBert. As it can be noted out from the name itself the work is on clinical notes. Hospital readmission is predicted, and it has given better performance than other models to predict readmission.

E. Other Predictive tasks

The work mentioned in[16], is development of framework for ventilation prediction, which also makes use of clinical notes. Deep latent representations are generated for clinical notes and then BiLSTM is used. The model has given better performance as compared to other work with AUROC=0.773. The future work includes combining multiple sources of clinical notes. The researchers in paper[17] have developed a framework which is used to predict various tasks like mortality prediction inside the hospital, Procedure prediction, Diagnosis prediction, Length of stay prediction. The dataset used is MIMIC-III dataset. The future scope includes working on negation of terms and numerical values. The research work in paper [18] is a classification task to recognize if a patient reveals some specific symptoms by studying his/her former clinical records. The methodology that is used is Feature extraction using bag-of-words, word embeddings, BiLSTM is compared with CML models (SVM, KNN, Random Forest, Na ıve Bayes), Ensemble learning approach is employed to further improve the results. The dataset used is n2c2 dataset, released by Harvard Medical School. It also includes Clinical notes. The results obtained are DL Approach gave Average F1 score of 87.53 and Ensemble Approach gave Average F1 score of 98.5. The authors in paper [19] developed a framework to find out correlation between symptoms of cancer and type 2 diabetes, race, and smoking status. They have considered colorectal cancer and breast cancer. Techniques like BioWord2Vec, Hierarchical Clustering and UMLS are used. The research work by the authors in paper[20] is to estimate mortality of diabetes patients. They have used ML models, NLP approaches, and UMLS which has yielding an AUC of 0.97.

F. Phenotyping

The work mentioned in research paper[21] aimed at developing EHR phenotypes to pin down the deceased patients with higher grade cancer or higher stage chronic kidney disease . The sensitivity for EHR phenotype for cancer was 99.5 percent. The future scope is to include EHR phenotypes that can be designed to prioritize specificity over sensitivity.

V. MATERIALS AND METHODS

A. Architecture

The architecture of the model has been displayed in Fig 2.

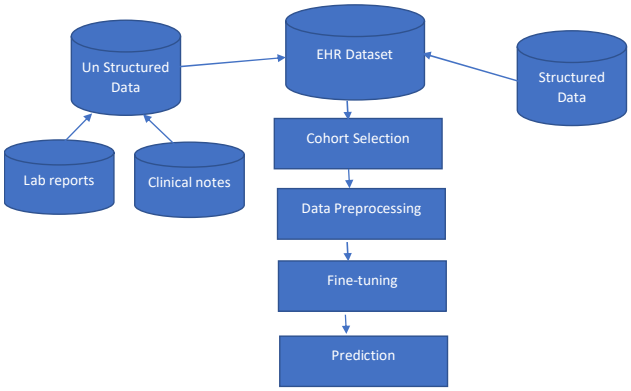


Fig 2 : Architecture of the fine-tuned model

The EHR data consisting of both structured as well as unstructured data is used. Cohort selection is performed to retrieve only the relevant data and filter out the remaining. The data is then pre-processed as per the requirements of the BERT model. The models i.e. BERT and DISTILBERT are fine tuned to perform the classification task.

B. Methodology

The dataset that has been used is MIMIC- III. Structured as well as un-structured data from the dataset were used to train the model. It consists of 26 tables but only relevant tables for the study were used. This study included various forms of clinical notes.

The primary step was to select the cohort. The work focused on predicting diseases like asthma, renal diseases, heart diseases and arthritis. So, records of selected ICD9 codes were retrieved. The tables included are NOTEEVENTS.csv, D_ICD_DIAGNOSES.csv, DIAGNOSES_ICD.csv. The NOTEEVENTS table contains the unstructured clinical notes. The other tables contain the structured data.

The relations are extracted as follows:

DIAGNOSES_ICD $\bowtie_{\text{DIAGNOSES\_ICD.SUBJECT\_ID} = \text{NOTEEVENTS.SUBJECT\_ID}}$ (NOTEEVENTS)

Records with ICD9 codes of '49302','71691','586','4243' were used. The total clinical notes that were retrieved were 2154. After cohort selection, pre-processing was performed on the clinical notes as per the requirements of the BERT model. The pre-processing steps like lowering the case of data, removing duplicates, dropping values with no clinical notes, and applying regular expression to filter out the special characters were performed.

The Deep Learning models are promising in the sense that there is no need of feature engineering on the data set for training purposes like its machine learning counterparts. They themselves do the job of feature extraction, which implies less preprocessing of data. The deep learning techniques that can be used are:

CNN- CNN is a common technique used while handling image data. Its prevalence has been seen in handling textual data, but the research community is now moving towards more sophisticated solutions.

RNN and its variants - RNN can be used with sequential data. They manage time dependencies of its input. Vanilla RNNs are not used due to vanishing gradient and exploding gradient issues and its improved versions like LSTM(Long Short-Term Memory), GRU(Gated recurrent Unit) are often used when dealing with sequential time series data. They perform additive updates whereas RNN performs multiplicative updates and hence results in vanishing gradient problem. They use the concept of memory cells and gates and hence can retain information for a longer period of time. BiLSTM- the initial letters ”bi” stands for bidirectional and it is a type of LSTM. It is named so, as it can work on the input sequences from both directions which can be used to better understand the context of the sentences but it is not truly bi-directional.

Transformers - like RNNs, transformers too handles sequential data and are in use for NLP related tasks. Transformers are promising solution to advanced natural language processing applications.

BERT stands for Bidirectional Encoder Representations from Transformers and is developed by Google. BERT was trained on a large corpus and it can be fine-tuned for a specific task. BERT contains stacked encoders. BERT learns a language by training on two tasks namely Masked Language and Next Sentence Prediction. In masked language, some of the words are masked and the model predicts the masked words. In next sentence prediction, it determines if the second sentence follows the first. It is truly bi-directional and understands the context better than any other model.

DISTILBERT is a lighter version of BERT, technique called as distillation was used to reduce the number of layers, thus making it lighter.

Fig 3 depicts the state diagram of the clinical note. The clinical notes are filtered, then they are joined with the structured EHR data. These notes then undergo pre-processing as required by the BERT model and then they are used for training.

Fig 3: State diagram of a clinical note

The dataset was first split into training and test data. In order to handle the class imbalance problem which often occur, class weights were assigned. The class weights were added to the model during the training process. The models i.e. BERT and DISTILBERT were then fine-tuned to perform the classification task. One of the hyper parameters of the model, i.e. learning rate was set to 1e-4 and then the models were trained for 11 epochs each.

VI. RESULTS AND DISCUSSION

TABLE II shows the metrics when BERT model was used and TABLE III shows the metrics when DISTILBERT was used.

TABLE II. METRICS WHEN BERT WAS USED

	Precision	Recall	F1-score
Heart disease	0.83	0.76	0.79
Asthma	0.88	0.84	0.86
Renal disease	0.75	0.81	0.78
Arthritis	0.84	0.87	0.86
Accuracy			0.82
Macro avg	0.82	0.82	0.82
Weighted avg	0.82	0.82	0.82

TABLE III. METRICS WHEN DISTILBERT WAS USED

	Precision	Recall	F1-score
Heart disease	0.72	0.89	0.80
Asthma	0.93	0.9	0.91
Renal disease	0.88	0.75	0.81
Arthritis	0.89	0.92	0.9
Accuracy			0.85
Macro avg	0.86	0.86	0.86
Weighted avg	0.86	0.85	0.85

The formulas of precision, recall and f1-score are given in equations (1), (2) and (3) respectively:

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

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

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

The first four lines in Table II and Table III show per class Precision, Recall and F1 score. The macro-average value, which is the arithmetic mean, is better in case of DISTILBERT and also the weighted average.

Also, it can be seen, all the values of precision, recall and f1-score are better when DISTILBERT was used. Another advantage of using DISTILBERT is that the amount of time taken for training significantly less which is near about one third of the time when BERT was used. This implies that DISTILBERT could be a preferred solution over BERT.

Fig 4, Fig 5, Fig 6, Fig 7 displays some of the examples of the output produced. When the sentence "The patient current symptomatology began last friday [**12-12**] when she developed URI symptoms including nasal congestion and cough. Since then she has noted increased dyspnea, increased episodes of her anginal pain including at rest and cough productive of yellow/celery colored sputum. Dyspnea has increased to point where she has trouble with stairs now where recently she has not. Has dry cough at baseline attributed to Mavik, but her current cough is different" was fed to the system the output displayed was "Asthma" which is a correct answer. Likewise other examples are fed yielding correct results.

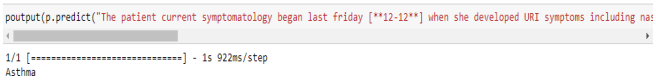


Fig 4 : Output displaying Asthma as the Disease

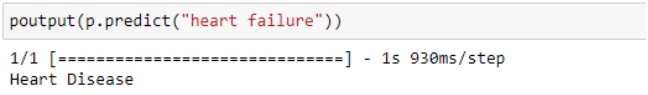


Fig 5 : Output displaying Heart Disease as the Disease

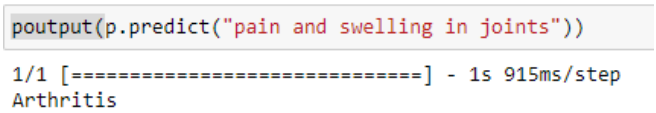


Fig 6 : Output displaying Arthritis as the Disease

Fig 7 : Output displaying Kidney Disease as the Disease

The training accuracy obtained after 11 epochs on BERT model is 0.9661, while the training accuracy obtained after 11 epochs on DISTILBERT model is 0.9831 whereas the validation accuracy is 0.82 in case of BERT and 0.85 in case of DISTILBERT.

VII. CONCLUSION AND FUTURE WORK

With many countries adopting Electronic Health Records, recent work has focused on exploring the power of EHR data. This has motivated researchers to extract entities, symptoms, classify diseases, identify characteristics, relations and predict diseases. Machine Learning and Deep Learning has been extensively used for these tasks. NLP plays a great role as data gets stored in the form of clinical texts too. BERT models have shown superior performance than traditional NLP methods. DISTILBERT which is lighter version of BERT gives better results as compared to BERT, also the time taken to fine-tune the model is significantly less, which implies that it could be used for processing of clinical notes.

The work can be extended by grouping out the ICD 9 codes of similar diseases. As this was a preliminary study, a smaller dataset was used, so a larger subset of the dataset can be considered for future work.

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] 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

[5] 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

[6] 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.

[7] S. A. Moqurrab, U. Ayub, A. Anjum, S. Asghar and G. Srivastava, "An Accurate Deep Learning Model for Clinical Entity Recognition From Clinical Notes," in IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 10, pp. 3804-3811, Oct. 2021, doi: 10.1109/JBHI.2021.3099755.

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

[9] H. Dong, V. Suárez-Paniagua, H. Zhang, M. Wang, E. Whitfield and H. Wu, "Rare Disease Identification from Clinical Notes with Ontologies and Weak Supervision," 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2021, pp. 2294-2298, doi: 10.1109/EMBC46164.2021.9630043.

[10] 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.

[11] 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.

[12] 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.

[13] 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

[14] 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

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

[16] Kexin Huang, Abhishek Singh, Sitong Chen, Edward Moseley, Chih-Ying Deng, Naomi George, and Charolotta Lindvall,"Clinical XLNet: Modeling Sequential Clinical Notes and Predicting Prolonged Mechanical Ventilation", In Proceedings of the 3rd Clinical Natural Language Processing Workshop, pages 94–100, 2019.

[17] Van Aken, Betty Papaioannou, Jens-Michalis Mayrdorfer, Manuel Budde, Klemens Gers, Felix Loeser, Alexander, "Clinical Outcome Prediction from Admission Notes using Self-Supervised Knowledge Integration", In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 881–893,2021.

[18] V. Kumar, D. R. Recupero, D. Riboni and R. Helaoui, "Ensembling Classical Machine Learning and Deep Learning Approaches for Morbidity Identification From Clinical Notes," in IEEE Access, vol. 9, pp. 7107-7126, 2021, doi: 10.1109/ACCESS. 2020.3043221.

[19] Luo X, Storey S, Gandhi P, Zhang Z, Metzger M, Huang K, "Analyzing the symptoms in colorectal and breast cancer patients with or without type 2 diabetes using EHR data", Health Informatics J. 2021 Jan-Mar;27(1):14604582211000785. doi: 10.1177/14604582211000785. PMID: 33726552

[20] 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.

[21] 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