BERT- BASED DISEASE PREDICTION FROM ELECTRONIC HEALTH RECORD

AKHIL RAJ RAJAN POALO MISSIER - Supervisor

School of Computing, Newcastle University

INTRODUCTION

- Recent years, Witnessed the remarkable surge in disease prediction capabilities, driven by the integration of Artificial Intelligence (AI) and machine learning algorithms.
- AI-driven disease prediction has revolutionized early diagnosis, leading to better patient outcomes and substantial cost reductions.
- The Role of Electronic Health Records (EHR): Accessibility and utilization of EHR data have paved the way for ground-breaking research in disease prediction.
- In this project, Bidirectional Encoder Representation from Transformers (BERT)[1] is being harnessed to unlock the predictive potential in disease diagnosis.
- This study delves into the exploration of disease embeddings, analyzing intricate relationships between diagnoses to make informed medical decisions.

RELATED WORK

- This project mainly inspired from a works on predicting medical condition based on the model BEHRT [2] (BERT Model) which predicted the likelihood of 301 conditions in future visits
- It is trained and evaluated on CPRD data from nearly 1.6 million individuals and outperforms existing state-of-the-art deep EHR models by 8.0–13.2% in average precision scores for different tasks.
- One of the novel aspects in their study is the integration of age as an embedding.
- The model offers scalability and superior accuracy and BEHRT enables personalized interpretation of predictions.
- Its flexible architecture can incorporate multiple heterogeneous concepts (e.g. diagnosis, medication, measurements etc.) to enhance prediction accuracy.
- The training results in disease, and patient representations can be used for future studies (i.e., transfer learning).

DATA UNDERSTANDING AND EXPLORATORY DATA ANALYSIS

The data used in this project, collected as part of the AI-MULTIPLY project, comprises approximately 150,000 patient records extracted from electronic health records (UK Biobank). These records represent longitudinal medical histories, including diagnoses and drug prescriptions.

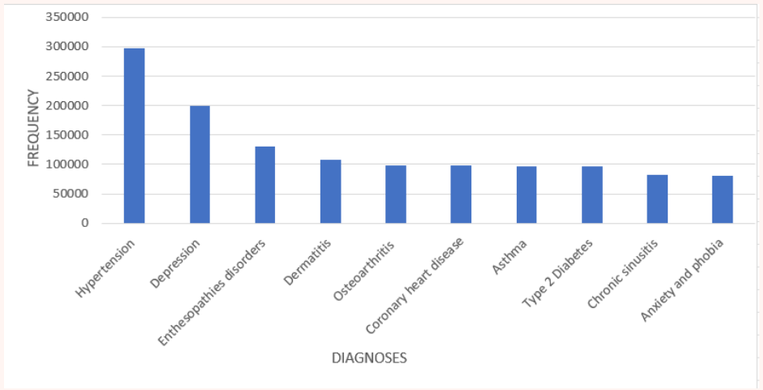


Figure 1. TOP 10 DIAGNOSES

- The exploratory data analysis (EDA) figure 1 indicates that Hypertension stands as the most prevalent diagnosis.
- Notably, individuals in the age range of 60-70 have been significantly impacted by various diseases.

METHOD

MODEL DEVELOPMENT

- Among the four models, one outstanding scratch-built model was selected.
- The novelty of this study lies in incorporating gender as a parameter in predicting diagnoses alongside the age embedding [2].
- The training parameters include a batch size of 32, a custom vocabulary of size 210, an embedding size of 200, and 6 transformer layers. ADAM optimizer is used with a learning rate of 0.0001.
- The training took place on AIM/BSU server (also knownas Melville) a server supplied by Newcastle University, equipped with a NVIDIA Quadro RTX 5000 GPU boasting 376 GB of memory. The entire training process spanned 2 days.
- A custom model and configuration were developed due to the available pretrained BERT model being trained on a different domain. Additionally, there was sufficient data accessible for training this specific model.

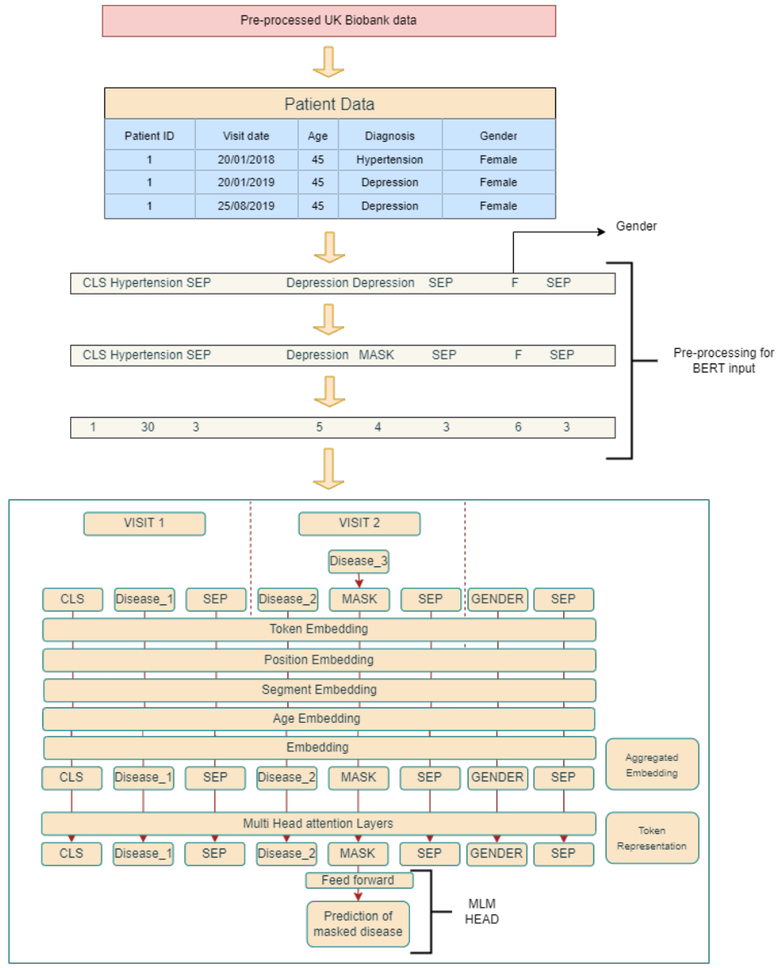


Figure 2. Overall Training Process

EVALUATION

The BERT-based model's performance in predicting the next diagnosis was evaluated using various metrics on an 80:20 split dataset.

Results

The model achieved the following performance on the test data set with an accuracy of 0.33 and AUC score of 0.52.

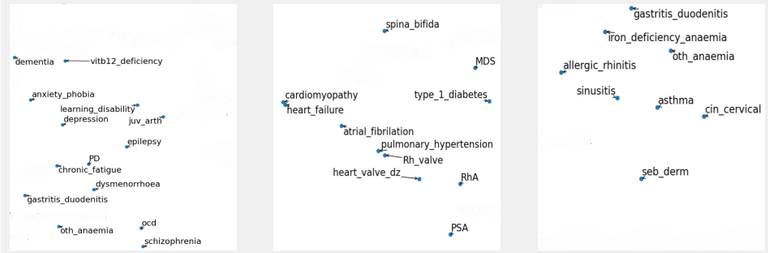
	Precision Score	Recall	F1 Score
Macro Average	0.18	0.17	0.16
Weighted Average	0.39	0.33	0.35

Table 1. Summary of Evaluation

- Analysis of the confusion matrix and loss curve from the training process suggests that the model has the potential to improve its performance through further training.
- The model's prediction capabilities are also affected by the imbalance in the classes of the data.
- The performance evaluation using the mentioned metrics may not accurately reflect the true performance due to its reliance on factors such as the source data, evaluation methods employed in the MLM model, and the characteristics of the test data.

Embedding Diagram

The embedding diagram, when reduced to a two-dimensional plane, offers valuable insights into the interrelationships between different diagnoses as shown below.



(a) Mental disorders (b) Cardiovascular disorder (c) Respiratory conditions

Figure 3. Diagrams showing similar diagnoses are clustered together

DEPLOYMENT

The deployment is accomplished through a Flask API with a JavaScript-based front end. With the model trained on 203 diagnoses, users can conveniently choose the disease from a drop down menu.

FUTURE SCOPE

The model's prediction capabilities can be enhanced by incorporating prescription (medicines) data, which reveals any conditions resulting from the side effects of medications.

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova.
Bert: Pre-training of deep bidirectional transformers for language understanding.
arXiv preprint arXiv:1810.04805, 2019.

[2] Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi.
BEHRT: Transformer for electronic health records.
Scientific Reports, 10(1), April 2020.