Classification of Cancer Progression by Standardizing Unstructured Clinical Data

Chennuri Prateek, Girish Chandar G, Gowtham Chitipolu, Rahul Challa
Indian Institute of Technology Gandhinagar
{chennuri.prateek,girish.chandar,chitipolu.gowtham,rahul.challa}@iitgn.ac.in

Abstract

According to PubMed Statistics, most of the papers on the prediction and classification of cancer progression are based on Machine Learning and Radiological techniques, with the drawback being that one cannot predict cancer before the development of the tumor since it is largely dependent on radiology reports such as X-rays. This problem can be addressed by making use of Electronics Health Records(EHRs) since it has information about the general medical history of patients. In most of the EHRs 80 percent of records are unstructured and thus, they are captured and leveraged using Natural Language Processing systems. We made use of Deep Learning Algorithms on the resultant structured data to predict the probability of cancer. After applying the machine learning techniques we achieved an accuracy of 68.90 percent and an F1 score of 0.669. We have open-sourced the codes pertaining to our paper[1].

1 Introduction

Cancer has been characterized as a heterogeneous disease consisting of different sub types. Once cancer crosses a particular stage, then there is no treatment in this world that can cure this deadly disease. Terminal cancer victims have no other option than to wait for their inevitable death. As per the statistics provided by the World Health Organization(WHO), cancer stands second as the leading cause of death globally responsible for 9.6 million deaths in 2018. Prognosis of cancer stages has become a necessity in cancer research, as it can facilitate the subsequent clinical management of patients. Hence, reducing the delay in cancer prognosis and late-stage presentation plays an important role in increasing the survival of cancer victims. The cause for the delay in prognosis can be broadly classified into two reasons: 1. Lack of public awareness about cancer and its related symptoms 2. Diagnostic delay by physician. We try to address the issue of delayed prognosis as resolving the issue can significantly increase the life expectancy of cancer patients. Once we find an efficient and feasible solution for the prognosis of cancer in patients, we can develop similar methods which can be implemented for prognosis of various other diseases using structured Electronic Health Records.

An Electronic Health Record is a digital version of patient's prescriptions. EHRs depict the real time, patient related records that make medical information available instantly at any point of time to the authorized users. A typical EHR contains the following information:

- Medical history of the patient.
- Diagnosis of the patient.
- Medication and treatment plans related to patient's diagnosis.
- Laboratory and test results of the patient.

Maintaining Electronic Health Records for each patient enables doctors to have complete information about the patients which help improve care, reduce safety risks and take decisions quickly. Moreover, it also increases the privacy and security of the patient data.

In past cancer research, prediction of cancer was done by applying machine learning techniques on radiological reports of the patient. But, the radiological reports of the patient show abnormality (or) signs of disease at later stages of cancer. In

[1]The link to the GitHub repository: https://github.com/girish1511/Cancer-Prognosis-using-NLP

this paper, our main aim is to make prognosis of cancer at an early stage itself. We plan to make this possible by structuring of Electronic Health Records.

Structuring of EHR data can be considered as an information extraction (IE) task, which automatically extracts and encodes clinical information from text. An Information Extraction application generally involves one or more of the following subtasks: concept or named entity recognition that identifies concept mentions or entity names from text (e.g., person names or locations), co-reference resolution that associates mentions or names referring to the same entity, and relation extraction that identifies relations between concepts, entities, and attributes (e.g., person-affiliation and organization-location).

The remaining of the paper is organised as follows. Related work is outlined in section 2. Section 3 describes the dataset used in this paper. A detailed Work Pipeline is described in section 4. Results and Discussion are presented in section 5. Finally, the concluding remarks and acknowledgement are provided in section 6 and section 7.

2 Related Work

Prediction of presence of various diseases in individuals has been a major area of interest for most of the researchers around the globe. Many works have been done in this area and many papers have been published about these works. These works use both rule-based as well as Machine Learning algorithms. Cancer has been the target disease in most of these papers because it is not only the most dangerous disease, but also has large number of categories as it is spread across various parts of the body. Machine Learning is being widely used in the field of cancer detection and diagnosis. According to the PubMed statistics, there have been around 1500 papers that are published on machine learning in cancer, but all of them are related to cancer diagnosis and detection. But there is relatively less work in the field of cancer prognosis and prediction. Although there are works on cancer prediction using Machine Learning techniques(Kourou et al., 2015), only few use Natural Language Processing approach. The NLP approach for the cancer prediction can be broadly divided into two categories : 1. Using EHR (Electronic Health Records), 2. Using Radiology Reports

Listgarten et al. (2004) developed a method to predict the occurrence of spontaneous breast cancer by using single nucleotide polymorphism(SNP) of steroid metabolizimg enzymes. They have collected data of 63 patients with breast cancer and 74 patients without breast cancer. They have reduced the sample-per-feature ratio to 45:1 from around 3:2 by reducing the number of SNPs considered from 98 to 3. On the reduced sample size, they have used various classifiers like Naive Bayes, decision tree and Support Vector Machine, with SVM doing the best. Over this they have also done extensive level of cross-validation.

There have been many works on standardizing unstructured data using Natural Language Processing methods. Ziemssen et al. (2016) work shows the significance of structured data by examining the importance of collecting structured clinical data on multiple sclerosis. The work of Yim et al. (2016) briefly describes the importance of NLP in extracting information from unstructured clinical data.

The work of Fu and Thirman (2016) focuses on automatic ICD coding by using EHRs available in the MIMIC dataset. They have tried to standardize the unstructured medical data by using NLP approaches such as bag-of-words. But, this approach has failed to understand the pragmatics behind the compound words like “urinary tract infection”, which are of least use when understood separately.

The work done by Yan is aimed at detecting cancer progression using radiology reports. This paper experimented with various machine learning models like SVM, Logistic Regression and Naive-Bayes classifier and extracted two types of features using Natural Language Processing, namely; single-word features and relation features. F1 score was used to evaluate the performance of each classifier by using both single-word features and relation features. Basic Naive-Bayes classifier gave better results on relation features outperforming other models. The drawbacks of this paper are lack of hyperparameter tuning and small dataset size with only 128 labelled data.

For our paper we have taken into consideration the NLP techniques used in the work of Fu and Thirman (2016) and the insights on the Machine Learning models used for cancer prognosis in the work of Yan.

Given the growing trend on the application of Machine Learning methods in cancer research, Kourou et al. (2015) presents a review of recent Machine Learning approaches employed in the modeling of cancer progression. The predictive models which are discussed, are based on various supervised Machine Learning techniques as well as on different input features and data samples.

The main focus of Sun et al. (2018) was on preprocessing of semi-structured and unstructured Electronic Medical Records (EMRs) and to emphatically analyze the key techniques. The paper discusses the methods of information extraction of EMR based on text mining and research status of named-entity recognition and relation extraction. Moreover, the paper also emphasizes on the applications of text mining on EMRs like and also the research issues for future work.

3 Dataset

MIMIC-III (Medical Information Mart for Intensive Care III) Johnson et al. (2016) Johnson (2016) is a freely-available database comprising de-identified health-related data associated with over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. The database includes information such as demographics, vital sign measurements made at the bedside (1 data point per hour), laboratory test results, procedures, medications, caregiver notes, imaging reports, and mortality (both in and out of hospital).

MIMIC-III is a relational database consisting of 26 tables. They are linked by `SUBJECT_ID` which refers to a unique patient, `HADM_ID` which refers to a unique admission to the hospital, and `ICUSTAY_ID` which refers to a unique admission to an ICU.

The `ADMISSIONS` table consists details of the admissions of the patients. It consists of both the admission and discharge time of the patients(in case of unfortunate deaths, death time is also mentioned). The details of date and time have been moved to the future to uphold the confidentiality of the patients but the duration between two events is preserved. It also has a column describing patients' diagnosis. Rather than searching for patients related to a particular disease through this diagnosis column of the `ADMISSIONS` table, the MIMIC-III dataset provides an easy way to extract the data pertaining to a particular disease.

Disease	**ICD-9 code**	**Patient Count**
Urinary Tract Infection	5990	6555
Thrombocytopenia	2875	3065
Pneumonia	2875	3065
Asthma	49392	2195
Rheumatoid Arthritis	7140	649
Cardiac Arrest	4275	1361
Anemia	2859	5406
Pulmonary Hypertension	4160	401
Acute Resp. Failure	51881	7497
Cancer	**140x-239x**	7361

Table 1: Distribution of patients across few major diseases in MIMIC-III dataset

ICD-9(International Classification of Diseases) has been used to code the patients' diseases. ICD is maintained by the World Health Organization that provides a standardized method to classify and record diseases. ICD-9 codes are employed in MIMIC-III dataset to ensure efficient retrieval of information corresponding to a particular disease. The `DIAGNOSIS_ICD` table of the MIMIC-III dataset connects the admission of a patient and the disease diagnosed during the admission by using three columns, namely: `HADM_ID`, `SUBJECT_ID` and `ICD9_CODE`. Muir and Percy (1991) give a summary of ICD-9 codes and the diseases/symptoms corresponding to the codes. To extract the patients diagnosed with a particular disease, first the corresponding ICD-9 is found and then using the ICD code we can extract the `SUBJECT_ID`s and `HADM_ID`s from the `DIAGNOSIS_ICD` table. One catch is that, for an admission multiple ICD codes can be assigned since symptoms are also coded using ICD. The `SEQ_NUM` in the `DIAGNOSIS_ICD` assigns prior-

ity to the ICD codes and during data extraction we ensure that the ICD code of the disease we are interested is the first priority. SUBJECT_IDs and HADM_IDs can further be used to retrieve specific information such as discharge summaries. Table 1 shows the number of patients in MIMIC-III dataset diagnosed with few major diseases. This information was extracted from the DIAGNOSIS_ICD table as mentioned above. ICD-9 codes 140x-239x correspond to any and all types of cancer and it can be seen from Table 1 that in the MIMIC-III dataset, 7361 patients are diagnosed with at-least one type cancer.

NOTEEVENTS and CHARTEVENTS tables consists most of the important data of a patient. CHARTEVENTS consists of charted events such as fluids intake of the patient during his admission in the hospital. We are interested in the data present in the NOTEEVENTS. It consists of HADM_ID and SUBJECT_ID to access the data of a specific patient for a particular admission. For every admission the NOTEEVENTS consists of textual data and the CATEGORY column specifies which category does the textual data pertain to such as discharge summary, nurse's notes or radiological reports. The discharge summaries are the Electronic Health Records that we are. We are only interested in the discharge summaries also known Electronic Health Records, to address the problem of cancer prognosis, as it contains all the information about a patient's stay in the hospital. The SUBJECT_IDs extracted from the DIAGNOSIS_ICD table using ICD codes, is used to extract discharge summaries corresponding to the cancer patients. Further processing on the dataset is mentioned in detail in Section 4.1.

4 Work Pipeline

In order to predict the probabilities of cancer we have divided the work pipeline into 3 stages: 1. Data Preprocessing, 2. Sentence Embedding, and 3. Machine Learning. Figure

4.1 Preprocessing

Kurniati et al. (2018) have done exploratory research on MIMIC-III dataset from the viewpoint of information extraction for oncology. Every time a patient gets admitted in the hospital, he/she is given a unique admission ID, HADM_ID(if the patient gets admitted multiple times he/she gets assigned new HADM_ID every time he/she gets admitted). As mentioned in Section 3, based on the diseases diagnosed during the admission, ICD codes corresponding to the diseases diagnosed is mentioned against the respective HADM_ID in the DIAGNOSIS_ICD table. The summary of classification of ICD-9 codes by Muir and Percy (1991) shows that, ICD-9 codes 140x to 239x are used to identify any type of cancer. These ICD-9 codes, along with DIAGNOSIS_ICD table, are used to extract EHRs pertaining to cancer. In the dataset used in this paper, **7361** out of 46520 patients were diagnosed with at least one type of cancer.

Discharge summaries	Count
Corresponding to any disease	59456
Corresponding to cancer	11495
Corresponding to no-cancer	47961
Corresponding to before cancer	1949
Corresponding to after cancer	9546

Table 2: Distribution of discharge summaries

4.1.1 Division of Cancer related documents

The HADM_ID extracted from DIAGNOSIS_ICD table corresponding to ICD codes of cancer denotes the admissions during which the patients were diagnosed with cancer. More than half of the patients have been admitted to the hospital prior to being diagnosed with cancer. The discharge summaries corresponding to these admissions are not assigned the ICD code of cancer and therefore the HADM_ID extracted before do not contain the admission ID's corresponding to these admissions. As mentioned above the SUBJECT_ID extracted from DIAGNOSIS_ICD table corresponding to ICD codes of cancer is used to separate the whole dataset into cancer and no-cancer patients. Furthermore, the HADM_ID extracted from DIAGNOSIS_ICD table corresponding to ICD codes of cancer refer to the admissions after the patient has been diagnosed with cancer. These HADM_ID's are used to further divide the data of patients with cancer into before and after they have been diagnosed with cancer. To summarize, the dataset is divided into cancer and no-cancer data using the SUBJECT_ID and the cancer data is further divided using the HADM_ID into before cancer and after cancer data. As mentioned in the Table 2 the entire discharge summaries of the

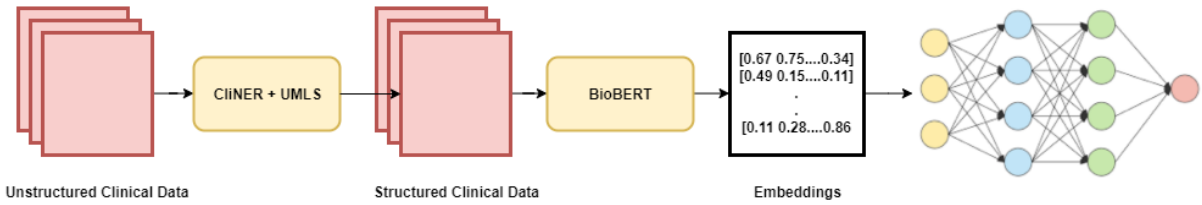


Figure 1: Pipeline

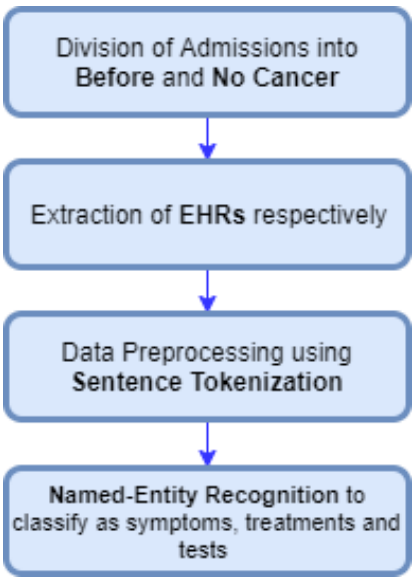


Figure 2: Flow of Structuring

MIMIC-III dataset is divided into **1949 before cancer**, **5412 after cancer** and **47961 no cancer** discharge summaries. Since we are interested in prediction of cancer, rather than detection, we exclude after cancer and only use no-cancer and before cancer data, thus the problem reduces to a binary classification problem. The discharge summaries(EHRs) of admissions corresponding to no-cancer and before cancer data are stored separately for facilitate further processing.

An EHR is a highly unstructured lengthy document containing many words and special characters, in which, most of the words do not add any value to the diagnosis of the patient, but rather would alter its meaning which is interpreted by the model. Therefore, we need to structure the EHRs so that we can use only the words and sentences which add value to the diagnosis of the patient.

4.1.2 Structuring of Electronic Health Records

Among various types of open-source document structuring softwares available, we used **CliNER** by Boag et al. (2015), a named entity recognition system specially designed for clinical health data. Moreover, a good advantage of using CliNER is that we can configure it along with Unified Medical Language System(UMLS) by Bodenreider (2004) for better results. **UMLS**(Unified Medical Language System) is a concise collection of many vocabularies related to the area of biomedical sciences.

As seen from the figure, it can be seen that CliNER structures each record broadly into 3 categories: 1. problem 2.test and 3.treatment. It identifies the phrases which belong to one of the three categories and assigns the corresponding tag to the phrase. Along with the tag, it also outputs the location of the word or phrase in the record. As we plan to use sentence embedding which will be explained in the next section, before structuring the data,we first format the EHR such that each line has only one sentence. (done using nltk.sent_tokenizer) so that we directly pick the sentence based on the location of the word(or phrase) given in the output.

After each record is structured using CliNER which outputs the specific phrases describing either **problem**, **treatment** or **test**, we have collected the sentences corresponding to each category into separate text files. Therefore, now each EHR has 3 files, corresponding to problems, treatments and symptoms respectively.

4.2 Sentence Embedding

To apply Deep Learning model on texts, some sort of mathematical representation of the texts are required. Word/Sentence Embedding is the vectorization of the text data that is learnt on a defined corpus. The vector representation of similar words are clustered together in the vector space. BERT(Bidirectional Encoder Representations from Transformers) by Devlin et al. (2018) has

emerged to be a state-of-the-art embedding model for Natural Language Processing tasks. The power BERT is that it encodes words based on the context from both the directions of the word.

The effectiveness of a word embedding model for a given task can be increased by training the embedding model on a corpus specific to the task. For example, embeddings for medical data can be made more robust by training the model on medical corpus. BioBERT by Lee et al. (2019), is a modification of BERT pre-trained on PubMed corpus (Canese and Weis, 2013) for biomedical text mining tasks. The BioBERT encodes the text data in a 768 dimensional space. The maximum length of the input sentence to the BioBERT is 512, beyond that the model trims the input sentence. Lee et al. (2019) have published the pre-trained weights in their GitHub repository. In this paper we have used the BioBERT v1.1 for embedding the EHRs.

The embedding of a given EHR is taken to be the average of all the sentence embeddings of the EHR. Therefore each EHR file is represented as a 768 dimensional vector. As mentioned in 4.1.2, each EHR is structured using CliNER and separated into 3 files corresponding to the structuring elements, namely: problems, treatments or tests. Each of these files are embedded using BioBERT giving us three 768 dimensional vector for every EHR. When more than one structuring element is considered for training the machine learning model, the embeddings of the corresponding structuring elements are concatenated to represent a given EHR file. More details are mentioned in Section 4.3.2.

4.3 Machine Learning

As mentioned in the previous section, each category file of an EHR is embedded into a 768-dimensional vector. Our aim is to build binary classifier using the feature vectors obtained from the EHRs to predict the probability of cancer. As mentioned in Table 2 the dataset after preprocessing consists of 1949 before cancer patients' records and 47961 no cancer patients' records. Therefore, the data is highly biased. In order to decrease the bias of the dataset, we randomly picked 1949*5 = 9745 records of no cancer patients and used the corresponding vectors to predict the cancer probability.

We performed ensembling technique on six models. The six models include **five neural networks** and **one Gaussian Naive Bayes** model. All the five neural networks have the same architecture but the datasets used for no cancer patients are disjoint. As we have 9745 no cancer records and only 1949 before cancer records, to completely eliminate biasing problem in neural networks, we divided the no cancer records into 5 equal divisions, each containing 1949 records. However, the before cancer records remain the same for all the five datasets. In case of Gaussian Naive Bayes model, we randomly selected 1949 no cancer related records and trained it along with the before cancer related records.

4.3.1 Neural Network Architecture description

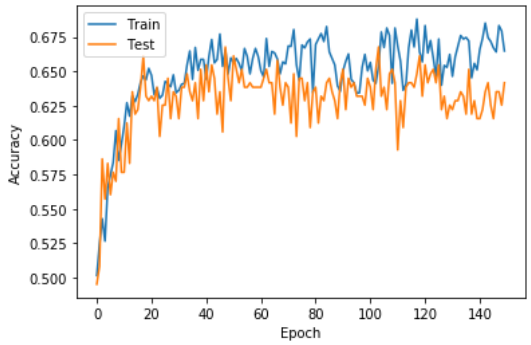
The architecture of the neural network used is described in the Table 3. We used ReLU activation function for all the hidden layers and input layer and softmax for the output layer in all five neural networks.
We used Adam optimizer as optimization function and sparse categorical cross-entropy as the loss function.In order to avoid over-fitting, we used dropout regularization technique in between the hidden layers.

Layer(type)	Output Shape	# Parameters
Dense Layer 1	(None,1024)	1573888
Dropout Layer 1	(None,1024)	0
Dense Layer 2	(None,512)	524800
Dropout Layer 2	(None,512)	0
Dense Layer 3	(None,256)	131328
Dense Layer 4	(None,512)	131584
Dropout Layer 3	(None,512)	0
Dense Layer 5	(None,1024)	525312
Dropout Layer 4	(None,1024)	0
Dense Layer 6	(None,2)	2050
Total Parameters		2888962

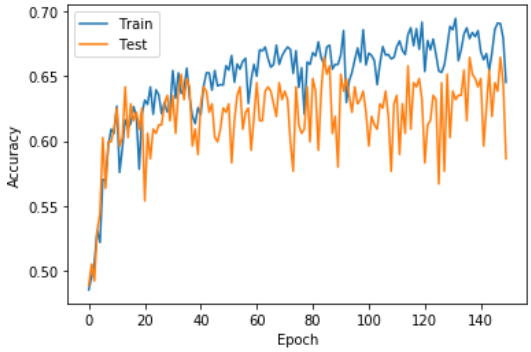
Table 3: Neural Network Architecture used for training the 5 models.

4.3.2 Training

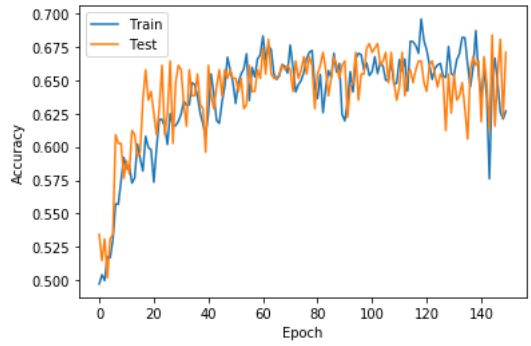
In this paper,we divided the dataset into 80 percent train, 10 percent validation and 10 percent test. We have done three types of training.In the first type, we took only problem related sentences only



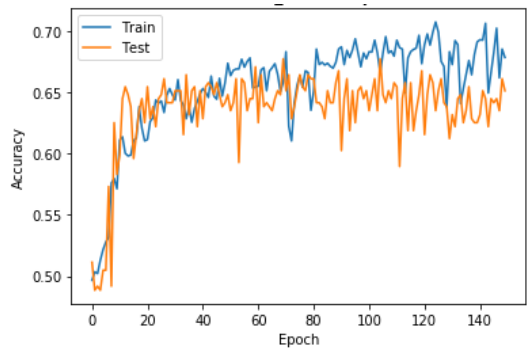
(a) Neural-1



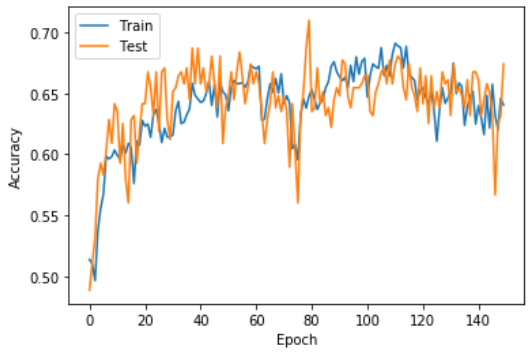
(b) Neural-2



(c) Neural-3



(d) Neural-4



(e) Neural-5

Figure 3: Training Accuracy of 5 Neural Models

in the dataset. In the second type, we considered problems and treatments related sentences only in the dataset. In the third type, we considered all the sentences obtained through structuring the dataset during training the models.

In the first type of training, only the problem related vectors of dimension 768 were used during training. However, in the second and the third type, we performed concatenation operation to combine the features of categorized sentences and considered 1536 and 2304 dimensional vectors respectively. We have considered 50 epochs for each model during training.

5 Results and Discussion

5.1 Results

While the training was being performed, we also captured the variation in the validation accuracy in each epoch. Figure **??** shows the variation in validation accuracy for all the five models trained. Table 6 shows the overall accuracy obtained by the ensemble of all the models of three types of training. Table 6 show the results of each model and Table 4 the confusion matrix obtained by training using the structuring elements problem+treatments.

Ground Truth / Predicted	No Cancer	Cancer
No Cancer	306	156
Cancer	95	210

Table 4: Confusion matrix of ensemble of all models

5.2 Discussion

From Figure 3a- 3e it is clear that the training and validation accuracy are almost equal in every epoch. This is due to the fact that we have used unbiased dataset for every model and moreover applied dropout regularization to avoid overfitting.

It can be observed from the accuracies Table 5 for different types of training, using the structuring elements problems+treatments, shows maximum accuracy. This is because multiple patients having different diseases may have the same tests prescribed by the doctor(or)medical practitioner. Therefore, there are high chances that the model may get confused.

Structuring elements \ Metric	F1 Score	Accuracy	Precision	Recall
Problems	0.6327	0.5737	0.688	0.6259
Problems+Treatments	0.688	0.6475	0.663	0.655
Problems+Treatments+Tests	0.638	0.297	0.844	0.44

Table 5: Performance across different elements of structured data

Metric \ Model	Neural-1	Neural-2	Neural-3	Neural-4	Neural-5	Gaussian NB	Ensemble
F1 Score	0.59	0.663	0.647	0.592	0.614	**0.669**	0.655
Accuracy	0.655	0.636	0.667	0.675	0.657	0.583	**0.689**
Precision	0.519	0.63	0.642	0.495	0.578	0.885	0.6475
Recall	0.683	0.701	0.653	0.738	0.657	0.538	0.663

Table 6: Model wise results

6 Conclusion

We have shown through our work Electronic Health Records have ample amount of information that can be used for cancer prediction. To our knowledge our work is the first to use Electronic Health Records for the cancer prediction. We plan to work on improving the F1 scores and also try to apply the model on different datasets to make sure the model is not biased on the dataset. As an additional confirmation, we would be getting an opinion from doctors and get to the effectiveness of our project. MIMIC-III dataset hasn't used extensively for research and therefore we also plan to explore the dataset and try to make use of data other than the discharge summaries to improve the results.

7 Acknowledgement

We would like to thank Dr. Mayank Singh and Pratik Kayal for their constant support and providing resources for the successful completion of our project.

References

William Boag, Kevin Wacome, Tristan Naumann, and Anna Rumshisky. 2015. Cliner: a lightweight tool for clinical named entity recognition. *AMIA Joint Summits on Clinical Research Informatics (poster)*.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Kathi Canese and Sarah Weis. 2013. Pubmed: the bibliographic database. In *The NCBI Handbook [Internet]. 2nd edition*. National Center for Biotechnology Information (US).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Justin Fu and Daniel Thirman. 2016. Cs224n final project-medical record understanding.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035.

Pollard T. Mark R. Johnson, A. 2016. Mimic-iii clinical database.

Konstantina Kourou, Themis P Exarchos, Konstantinos P Exarchos, Michalis V Karamouzis, and Dimitrios I Fotiadis. 2015. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13:8–17.

Angelina Prima Kurniati, Geoff Hall, David Hogg, and Owen Johnson. 2018. Process mining in oncology using the mimic-iii dataset. In *Journal of Physics: Conference Series*, volume 971, page 012008. IOP Publishing.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Jennifer Listgarten, Sambasivarao Damaraju, Brett Poulin, Lillian Cook, Jennifer Dufour, Adrian Driga, John Mackey, David Wishart, Russ Greiner, and Brent Zanke. 2004. Predictive models for breast cancer susceptibility from multiple single nucleotide polymorphisms. *Clinical cancer research*, 10(8):2725–2737.

CS Muir and C Percy. 1991. Classification and coding for neoplasms. *Cancer registration: principles and methods. Lyon: IARC*, 81.

Wencheng Sun, Zhiping Cai, Yangyang Li, Fang Liu, Shengqun Fang, and Guoyan Wang. 2018. Data processing and text mining technologies on electronic medical records: a review. *Journal of healthcare engineering*, 2018.

Yu Yan. Detecting cancer progression in radiology reports.

Wen-wai Yim, Meliha Yetisgen, William P Harris, and Sharon W Kwan. 2016. Natural language processing in oncology: a review. *JAMA oncology*, 2(6):797–804.

Tjalf Ziemssen, Jan Hillert, and Helmut Butzkueven. 2016. The importance of collecting structured clinical information on multiple sclerosis. *BMC medicine*, 14(1):81.