Exploratory Data Analysis Project Report
Exploring MIMIC-III for Collaborative Disease Insights

CISC 839: Topics in Data Analytics, Winter 2024
Drishti Sharma (21ds128@queensu.ca),
Jasmine Mishra (18jvm@queensu.ca),
Jing Tao (21jt35@queensu.ca),
Veronika Grigoreva (veronika.grigoreva@queensu.ca)

1 Abstract

Diagnostic errors, occasionally made by physicians during the process of determining a patient's condition, can have severe consequences. These errors may worsen the patient's illness or, in more tragic instances, even result in fatalities. Beyond the harmful impact on individual health, such errors impose significant financial burdens on patients, their insurance providers, and governmental healthcare systems. Furthermore, they can damage the reputation and career prospects of the healthcare provider involved, particularly in cases where incorrect medications are prescribed or incorrect diagnoses are made.[14] In our project, we have used data mining methodologies to explore past patient records comprehensively, uncovering important insights to enhance the diagnostic procedure. Through examination of disease symptoms, our methodology seeks to precisely predict type of disease afflicting a patient. This effort will greatly improve how accurately diagnoses are made and make the process more efficient. It'll give strong backing to healthcare workers in their efforts to diagnose illnesses.

In this project, we're using the MIMIC-III medical dataset, which contains information about patients and the notes doctors wrote during their assessments. Our goal is to analyze these notes to identify the diseases or health issues patients might have. This analysis could provide valuable insights into the patients' health status and help in their medical treatment and care. Through in-depth analysis of the MIMIC-III database, we have carefully screened and integrated data related to the top 10 most frequently occurring diseases. The core objective of this study is to perform in-depth mining of the recorded texts of these diseases, focusing on extracting key symptom information from them. Through this process, we have collected a comprehensive symptom dataset and used it to train and test several innovative disease prediction models.

2 Introduction

In healthcare industry, the accuracy of medical diagnoses stands is a fundamental step for effective treatment and patient care. However, despite advancements in medical science and technology, diagnostic errors persist, posing significant challenges to both patient well-being and healthcare systems worldwide. Research indicates that annually, medical misdiagnosis impacts approximately 12 million individuals globally. This equates to an average misdiagnosis rate of one in 20 patients, with a significant portion—ranging from 10 to 20%—affected while in critical condition.[14]These errors, occasionally made by physicians during the intricate process of interpreting a patient's condition, can yield grave consequences. In some instances, they worsen the patient's illness, leading to prolonged suffering, while in more tragic scenarios, they culminate in fatalities. Every year, an estimated 40,000 to 80,000 individuals succumb to fatalities resulting from diagnostic errors made by medical professionals. Notably, this issue disproportionately impacts women and minority communities, manifesting in mortality rates that are 20 to 30% higher than those observed in other demographic groups.[14] The repercussions of diagnostic inaccuracies extend beyond individual health, putting a substantial financial toll on patients, their insurers, and governmental healthcare frameworks. Moreover, such errors cast shadows over the reputations and career trajectories of healthcare providers, particularly when they involve erroneous medication prescriptions or incorrect diagnoses.

Due to the above challenges, innovative approaches are important to reduce the presence of diagnostic errors and enhance the effectiveness of medical diagnoses. Various modern data mining methodologies present a promising method to explore the wealth of historical patient records, extracting actionable insights, and refining the diagnostic process. Thus, in our project we are using data mining techniques to comprehensively explore past patient data. Our primary objective is to unravel pivotal patterns and associations within these datasets, with a keen focus on explaining disease symptoms and categorizing illnesses with precision.

A notable aspect of our project is the classification of patient diseases from doctor's notes.

Initially, we were targeting four overarching categories: Respiratory, Cardiovascular, Mental Health, and Others to categorise our symptoms but for our final project, we decided to go a step further and classify not only in four categories but predict the actual disease based on symptoms.

Expanding disease categorization beyond just four categories to include a wider range of specific diseases enhances disease prediction by providing a finer level of granularity, precision, and flexibility in the analysis. This approach allows to capture unique disease characteristics and patterns which can lead to more accurate predictions specific to individual health conditions. By including the diversity of health conditions beyond the four disease categories that we initially decided to work upon, the model can better adapt to the complexities of clinical data and offer comprehensive coverage of diseases present in the dataset.

The subsequent sections of this paper are structured as follows: Section 3 provides some background to the problem statement and includes Literature Survey. Section 4 elucidates the Data Analysis Method. Section 5 outlines the Methodology and ensuing discussions. Section 6 talks about the Result and Analysis. Lastly, section 7 draws Conclusions.

3 Background

3.1 Problem Statement

Misdiagnoses can be large financial and health burdens for patients, physicians, insurers, and governments worldwide. Some of the repercussions of misdiagnoses include patients receiving incorrect prescriptions and therefore experiencing more negative health effects. Doctors can face potential career setbacks and legal consequences, or increased stress during their workday. Insurers face financial penalties, while governments have to deal with drug shortages and excessive spending on imports or production of medications. Having timely detection and classification of diseases is very important for saving lives and addressing these other issues within the health industry. Accurate disease prediction does not only reduce costs and save time in hospitals but also enhances the situation for all stakeholders, meaning that there can be an overall more effective healthcare management system in place.

3.2 Literature Survey

In recent times there have been many global studies that have investigated forecasting diseases, exploring treatments, and advancing drug discovery. Various data mining methods have been used in disease detection, each used for a different purpose with different results. Below, we outline these studies across methods including big data, frameworks for visualization, uses of our selected database, classification methods, and applications of our learnings.

Big Data has been used to predict a patient's diagnosis through the use of patient complaints. Silahtaroglu and Yilmazturk (2021) used natural language processing techniques along with probabilistic neural networks and random forest decision trees for text mining the patient complaints to predict their diagnosis based on what complaints were being made. [16] Another big data study used online medical inquiries to simulate disease trends within society for map-based disease risk prediction. [13] Another proposed approach within research with big data for investigation of misdiagnosis is to sue the SPADE (System-Disease Pair Analysis of Diagnostic Error) framework. First proposed by Liberman and Toker (2017), the SPADE method uses patient information over time to determine the relationship between symptoms and diagnosis for accurate predictions. [11]

Some frameworks have been proposed for better data visualization and analytics of healthcare data for misdiagnosis purposes. One such framework as proposed by Widanagamaachchi et al. (2022) suggests using a Sankey diagram to make a comprehensive visualization of a diagnosis path – which can help healthcare professionals understand how diagnoses might change over time. This type of visualization could be used in predictive models to see how different symptoms may present in several types of disease. [20] Another method for diagnostic prediction with data analytics is to use clinical event sequences in an unsupervised adversarial domain. This method, named ADADIAG, [21] uses a trained language model then predicts the domain origin. This method is more focused on addressing issues with domain shifts with data within the medical community but is still able to be used with diagnostic purposes.

Some researchers have used the MIMIC-II and MIMIC-III dataset, both these datasets are similar in their uses in research. In one study by Dai et. al. (2020), they were able to analyze disease from adult patients. By finding the various characteristics of the diseases found in the dataset, the researchers were able to learn a lot about each disease. They performed an analysis using tableau and navicat. [2] One group of researchers created a web-based data visualization tool using the MIMIC-II database to allow for easier query of the large amount of data for researchers who may not be familiar with data analytics or coding. [9] The users can do basic visualizations with demographic information, administrative information, patient outcomes, vital signs, lab results, interventions, and some miscellaneous information. We can use this tool to get more familiar with the dataset during our project. The MIMIC-II database has also been used outside of diagnosis. For example, analyzing various treatments used with patients – in one study it was found that calcium supplementation can improve the outcome of intensive care unit patients. [22] Another study used the MIMIC-III database for mortality prediction of patients using unsupervised and semi-supervised clusters. [12] Databases such as the MIMIC-III can give data analysts a wide range of medical knowledge.

For our project, we are focused on classification of data for disease prediction. Multi-class classification models have been used in several contexts in research in data analytics. Multi-class classification is the type of classification task where each input has one output class. In the context of medical classification, there have been many uses of multi-class classification models. In one study, different categories of lung cancer sub-types were identified using an independent sub-task learning method for feature selection and construction, [10] combined with a Bayesian prior for improved classification. Another study used multi-class classification with a Boltzmann machine, k-nearest neighbors, and information-entropy datasets to feed a classifier system in cancer research classifications.[18] Another study used multi-class classification for the classification of blood cells from images, which are later helpful in testing the shape of blood cells for symptoms blood related diseases. [3]

Another form of classification/categorization in research for data analytics is multi-label categorization. Mutli-label categorization is different from multi-class classification in that each input can have many output classes. Multi-label categorization can be used for a wide variety of purposes including for categorization of DNA sequences [4], death certificates [5], images and videos [17], web pages [1], and like our project: categorization of diseases. [15] For our project, multi-label categorization can prove useful in the categorization of the many diseases diagnosed in the MIMIC-III database.

Applications of studying the diseases with the use of datasets such as MIMIC-III are far and wide. One such application is to use the various characteristics of diseases to be used in future diagnostic technologies. One technology that has been proposed includes the use of wearable sensors and artificial intelligence to avoid the misdiagnosis of Parkinson's disease. [19] These researchers used wearable technology to continually assess the patient's symptoms to make a correct diagnosis. Such advancements as this one show the power of using data-driven approaches in healthcare, ultimately benefiting patients and healthcare professionals.

By looking at the many ways that researchers have analyzed medical data in the past, we are able to make more informed decisions in our own analysis.

4 Exploratory Data Analysis

4.1 Mimic-iii Dataset

[6] MIMIC-III (Medical Information Mart for Intensive Care III) is a large, publicly available database developed by the MIT Laboratory for Computational Physiology at the Massachusetts Institute of Technology (MIT) in collaboration with Beth Israel Deaconess Medical Center. MIMIC-III contains hospital inpatient records for more than 40,000 patients between 2001 and 2012, covering patient vital signs, laboratory test results, medical procedures, medication information, nursing notes, diagnostic information and many more types of health data.

Table 1: Overview of Papers Surveyed

Surveyed Papers			
No.	Title	Journal	Relevancy
[16]	Data analysis in health and big data: A machine learning medical diagnosis model based on patients' complaints	Communications in Statistics - Theory and Methods	Current research with Misdiagnosis
[13]	Can the development of a patient's condition be predicted through intelligent inquiry under the e-health business mode? Sequential feature map-based disease risk prediction upon features selected from cognitive diagnosis big data	International Journal of Information Management	Current research with Misdiagnosis
[11]	Symptom-Disease Pair Analysis of Diagnostic Error (SPADE): a conceptual framework and methodological approach for unearthing misdiagnosis-related harms using big data	BMJ Quality & Safety	Current research with Misdiagnosis
[20]	A flexible framework for visualizing and exploring patient misdiagnosis over time	Journal of Biomedical Informatics	Frameworks for diagnosis
[21]	AdaDiag: Adversarial Domain Adaptation of Diagnostic Prediction with Clinical Event Sequences	Journal of Biomedical Informatics	Frameworks for diagnosis
[2]	Analysis of adult disease characteristics and mortality on MIMIC-III	PLOS ONE	Uses of the MIMIC dataset
[9]	A web-based data visualization tool for the MIMIC-II database	BMC Medical Informatics and Decision Making	Uses of the MIMIC dataset
[22]	Calcium supplementation improves clinical outcome in intensive care unit patients: a propensity score matched analysis of a large clinical database	SpringerPlus	Uses of the MIMIC dataset
[12]	Mortality prediction on unsupervised and semi-supervised clusters of medical intensive care unit patients based on MIMIC-II database	Informatics in Medicine Unlocked	Uses of the MIMIC dataset
[10]	Multi-Label Image Categorization With Sparse Factor Representation	IEEE transactions on image processing	Multi-Class Classification
[18]	Multi-Class Classification of Medical Data Based on Neural Network Pruning and Information-Entropy Measures	Multi-Class Classification	
[3]	Microcell-Net : A deep neural network for classification of microscopic blood cell images	Expert Systems	Multi-Class Classification
[4]	Enhancing Taxonomic Categorization of DNA Sequences with Deep Learning: A Multi-Label Approach	Bioengineering (Basel)	Multi-Label Categorization
[5]	Multi-label Categorization of French Death Certificates using NLP and Machine Learning	ACM	Multi-Label Categorization
[17]	Multi-Label Image Categorization With Sparse Factor Representation	IEEE transactions on image processing	Multi-Label Categorization
[1]	Multi-label incremental learning applied to web page categorization	Neural computing & applications	Multi-Label Categorization
[15]	Hybrid Multi-Label Classification Model for Medical Applications Based on Adaptive Synthetic Data and Ensemble Learning	Sensors (Basel, Switzerland)	Multi-Label Categorization
[19]	Avoiding Misdiagnosis of Parkinson's Disease With the Use of Wearable Sensors and Artificial Intelligence	IEEE Sensors Journal	Medical Data Analysis Application

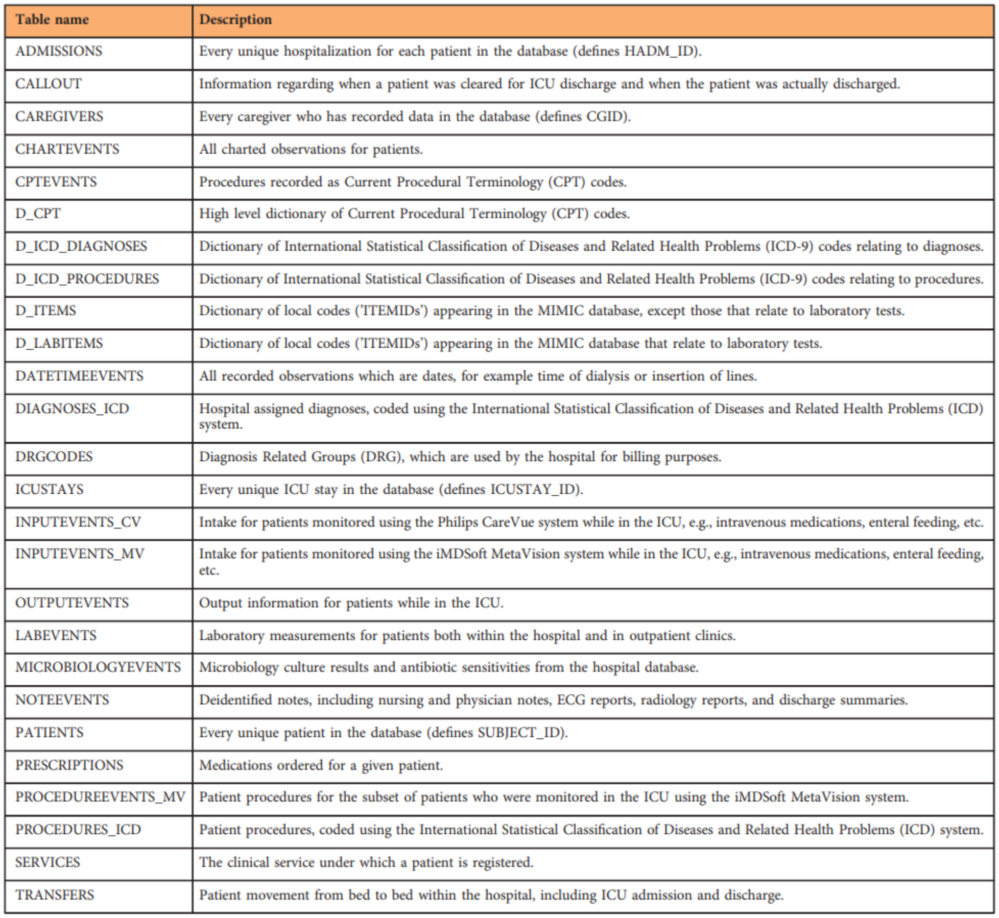


Table name	Description
ADMISSIONS	Every unique hospitalization for each patient in the database (defines HADM_ID).
CALLOUT	Information regarding when a patient was cleared for ICU discharge and when the patient was actually discharged.
CAREGIVERS	Every caregiver who has recorded data in the database (defines CGID).
CHARTEVENTS	All charted observations for patients.
CPTEVENTS	Procedures recorded as Current Procedural Terminology (CPT) codes.
D_CPT	High level dictionary of Current Procedural Terminology (CPT) codes.
D_ICD_DIAGNOSES	Dictionary of International Statistical Classification of Diseases and Related Health Problems (ICD-9) codes relating to diagnoses.
D_ICD_PROCEDURES	Dictionary of International Statistical Classification of Diseases and Related Health Problems (ICD-9) codes relating to procedures.
D_ITEMS	Dictionary of local codes ('ITEMIDs') appearing in the MIMIC database, except those that relate to laboratory tests.
D_LABITEMS	Dictionary of local codes ('ITEMIDs') appearing in the MIMIC database that relate to laboratory tests.
DATETIMEEVENTS	All recorded observations which are dates, for example time of dialysis or insertion of lines.
DIAGNOSES_ICD	Hospital assigned diagnoses, coded using the International Statistical Classification of Diseases and Related Health Problems (ICD) system.
DRGCODES	Diagnosis Related Groups (DRG), which are used by the hospital for billing purposes.
ICUSTAYS	Every unique ICU stay in the database (defines ICUSTAY_ID).
INPUTEVENTS_CV	Intake for patients monitored using the Philips CareVue system while in the ICU, e.g., intravenous medications, enteral feeding, etc.
INPUTEVENTS_MV	Intake for patients monitored using the iMDSoft MetaVision system while in the ICU, e.g., intravenous medications, enteral feeding, etc.
OUTPUTEVENTS	Output information for patients while in the ICU.
LABEVENTS	Laboratory measurements for patients both within the hospital and in outpatient clinics.
MICROBIOLOGYEVENTS	Microbiology culture results and antibiotic sensitivities from the hospital database.
NOTEEVENTS	Deidentified notes, including nursing and physician notes, ECG reports, radiology reports, and discharge summaries.
PATIENTS	Every unique patient in the database (defines SUBJECT_ID).
PRESCRIPTIONS	Medications ordered for a given patient.
PROCEDUREEVENTS_MV	Patient procedures for the subset of patients who were monitored in the ICU using the iMDSoft MetaVision system.
PROCEDURES_ICD	Patient procedures, coded using the International Statistical Classification of Diseases and Related Health Problems (ICD) system.
SERVICES	The clinical service under which a patient is registered.
TRANSFERS	Patient movement from bed to bed within the hospital, including ICU admission and discharge.

Figure 1: An overview of the data tables comprising the MIMIC-III (v1.3) critical care database.

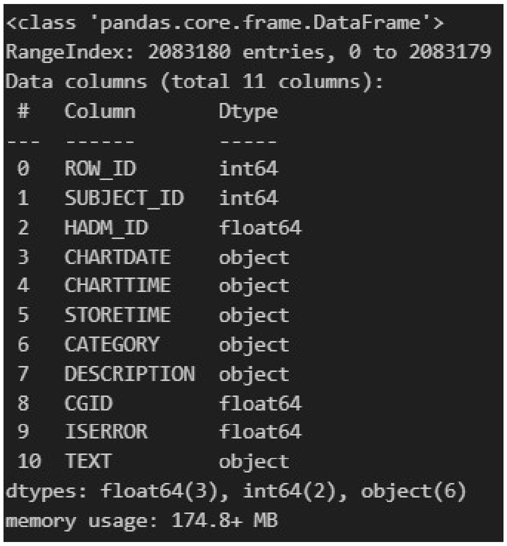
In this project, we performed an in-depth analysis of several main tables in the MIMIC-III medical record dataset: NOTEEVENTS, DIAGNOSES_ICD, and D_ICD_DIAGNOSES, which are linked by specific identifiers, which usually have the suffix 'ID'. For example, SUBJECT_ID represents a unique patient identity, while HADM_ID points to a particular hospitalization record. This design enables cross-table queries and data integration, providing a powerful way to track and analyze a patient's entire hospital stay.

Using the NOTEEVENTS table, we can access the detailed diagnostic texts doctors leave for patients in each hospitalization record. These texts record the doctor's observations, diagnostic thoughts, and assessment of the patient's condition and are an essential basis for understanding the patient's condition and developing a treatment plan. We aim to extract key symptom terms and analyze this rich text data in-depth. These symptom keywords will be used as input data for subsequent disease diagnosis model training, aiming to improve the accuracy and efficiency of diagnosis through machine learning methods.

Next, we relied on the patient's unique identifier, SUBJECT_ID, and the corresponding hospitalization record identifier, HADM_ID, to locate all diagnostic codes of the patient during this admission in the DIAGNOSES_ICD table. These diagnostic codes are coded according to the International Classification of Diseases (ICD) standards, which reflect the patient's primary and secondary diagnoses and provide us with a standardized way of identifying the disease. With these diagnostic codes, we can query the D_ICD_DIAGNOSES table for the detailed disease names that correspond to them, a step that is key to enabling cross-table correlation and in-depth data analysis.

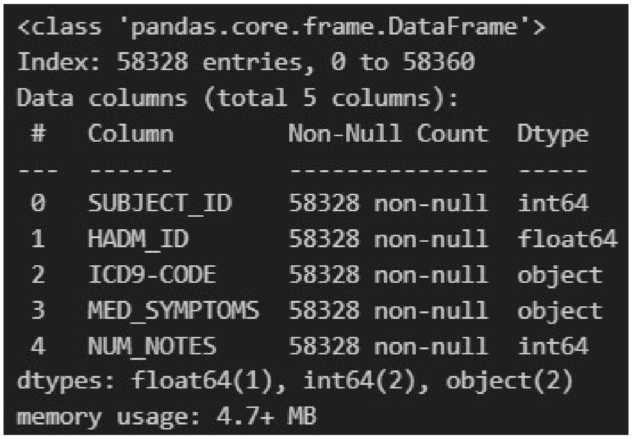
4.2 Data processing and analysis

The MIMIC-III dataset is a massive database of 46,146 patients and 2,083,180 diagnostic texts. To extract valuable medical information from this wealthy but complex textual data, we employed the MedCAT medical extraction tool, explained more in depth in 5.1.1. The efficient capabilities of MedCAT enable us to identify and extract critical medical terms precisely and refine and integrate the raw data into 58,361 structured records. It is worth noting that since some patients had multiple hospitalizations in the dataset (i.e., various HADM_IDs), they may correspond to multiple records.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2083180 entries, 0 to 2083179
Data columns (total 11 columns):
 #   Column       Dtype
---  ------       -----
 0   ROW_ID       int64
 1   SUBJECT_ID   int64
 2   HADM_ID      float64
 3   CHARTDATE    object
 4   CHARTTIME    object
 5   STORETIME    object
 6   CATEGORY     object
 7   DESCRIPTION  object
 8   CGID         float64
 9   ISERROR      float64
 10  TEXT         object
dtypes: float64(3), int64(2), object(6)
memory usage: 174.8+ MB
```

Figure 2: Information about the NoteEvents data stream



```
<class 'pandas.core.frame.DataFrame'>
Index: 58328 entries, 0 to 58360
Data columns (total 5 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   SUBJECT_ID    58328 non-null  int64
 1   HADM_ID       58328 non-null  float64
 2   ICD9-CODE     58328 non-null  object
 3   MED_SYMPTOMS  58328 non-null  object
 4   NUM_NOTES     58328 non-null  int64
dtypes: float64(1), int64(2), object(2)
memory usage: 4.7+ MB
```

Figure 3: Information about the Extracted Symptoms data stream

After completing the initial data extraction, we performed a thorough cleaning of the dataset, focusing on processing and removing missing values to ensure the accuracy and reliability of the analysis. Next, we performed an exhaustive statistical analysis of the diagnosis codes and their corresponding disease keyword occurrences, a step that is critical to understanding the patterns of disease distribution in the dataset. For our feature selection, we selected relevant features (symptoms) for classification.

Through this series of refinements, we successfully identified the 20 diagnostic diseases with the highest frequency of occurrence. These 20 diseases reflect the most common medical problems in the dataset and provide us an idea for a direction for our subsequent analysis.

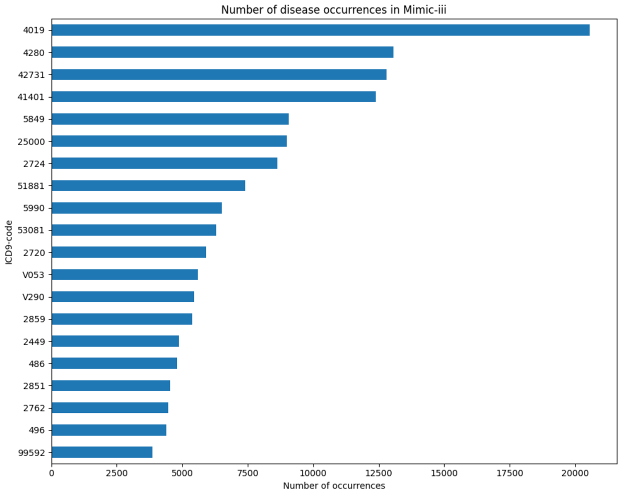
4.3 Challenges

Several complex challenges were encountered for this project, mainly in predicting diseases from a large amount of unstructured medical record text. Each record in the MIMIC-III database may be associated with multiple diseases, which poses the challenge of multi-label classification, which is considerably more complex than the traditional single-label task.

The large volume of EMR data also complicates model training. Due to the large amount of data processing, schemes that utilize large language models or techniques such as BERT require significant computational resources and long training times.

In addition, the variability and specificity of EMR texts, filled with medical terms and patient information, require highly adaptive models that can understand the nuanced data to classify disease labels accurately. Thus, we face both technical and performance challenges.

Overcoming these challenges for successful project implementation requires navigating through complex technical, data processing, and model training choices to achieve high accuracy in disease prediction.



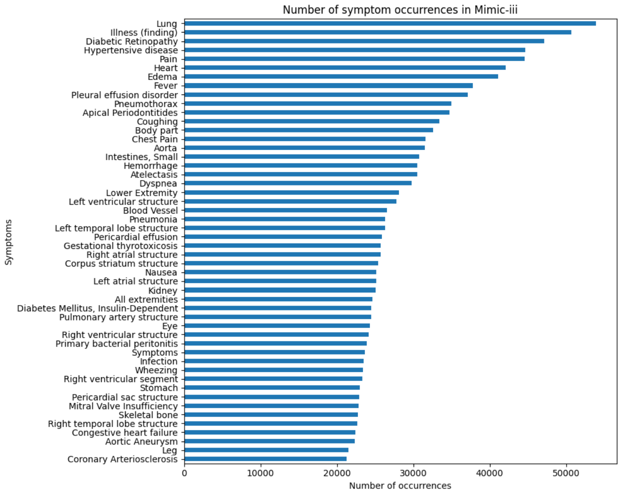


Figure 4: Number of disease occurrences in Mimic-iii Figure 5: Number of Symptom occurrences in Mimic-iii

This requires innovative strategies and technological advances to navigate the complexities of multi-label classification effectively.

To address these challenges, we have decided to perform both machine learning and deep learning techniques to evaluate the dataset for disease prediction. We also incorporate techniques from natural language processing to evaluate the large amount of language data extracted from the patient symptoms.

5 Methodology

5.1 Pre-processing

The first step of pre-processing was mapping the set of EMR notes to the set of diseases the patient has using MedCAT. First, we identified which notes would be relevant to diagnosing diseases. We extracted various keywords such as symptoms, observations, organs, and other medical related terms for diagnosis.

Then, we narrowed down the records to the top 10 diseases in the data. We used the top 10 diseases with highest frequency, which allows us to further prioritize during data extraction.

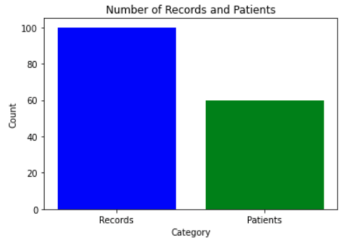


Figure 6: Disease and Patient Records

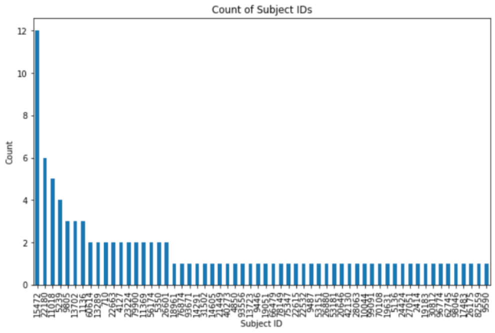


Figure 7: Subjects frequency

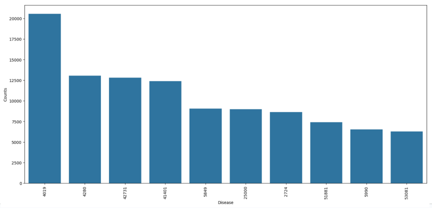


Figure 8: Top 10 Diseases plot

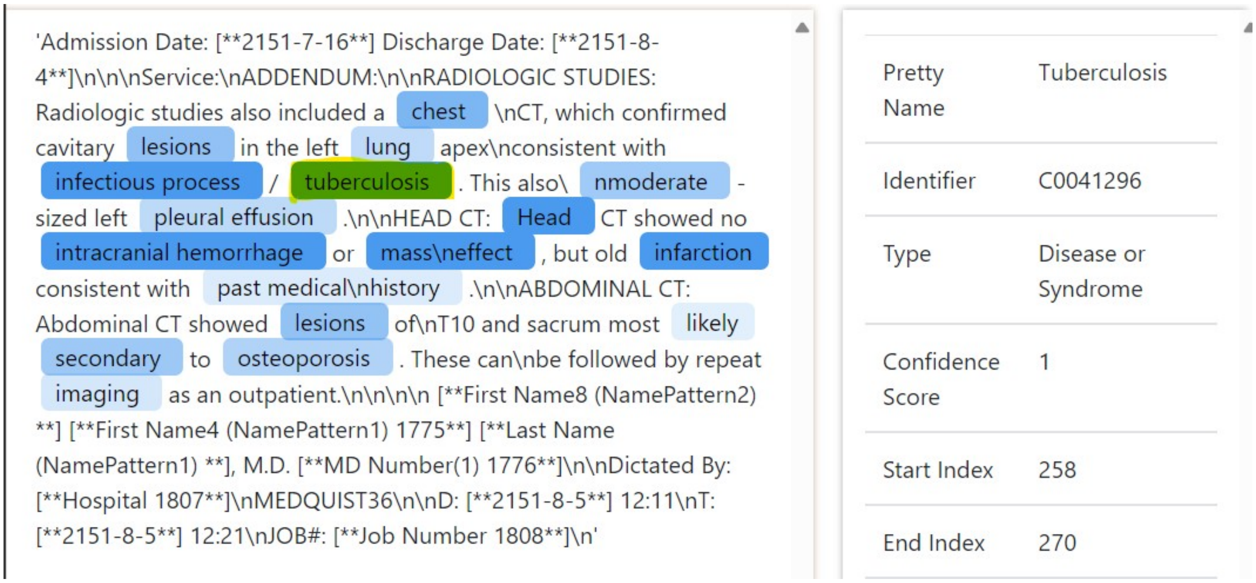
5.1.1 Medical Extraction Tool: Medcat

[7] MedCAT (Medical Concept Annotation Tool) is an advanced natural language processing tool designed to recognize and extract medical concepts and terminology from medical texts. It was developed by scientists and researchers in the UK's National Health Service (NHS) to support clinical text analysis and medical research. MedCAT combines both deep learning and natural language processing techniques to understand and parse many forms of textual material, such as medical records, research reports, clinical trial notes, and more.

The power of the tool lies in its deep understanding of medical domain knowledge, it is able to recognize a wide range of medical entities such as disease names, symptoms, medications, therapeutic procedures, etc., and match these entities with pre-existing medical terminology repositories or ontologies (e.g., UMLS - Unified Medical Language System). In this way, MedCAT recognizes specialized terms in the text and understands their relationships and contextual meanings, providing extremely high accuracy and rich insights for medical data analysis.

We used the tool to perform keyword symptom extraction on medical texts. Specifically, we included attributes such as 'Sign or Symptom,' 'Disease or Syndrome,' 'Pathologic Function,' 'Body Part, Organ, or Organ Component,' and' CRANIOCEREBRAL INJ' keywords were extracted as input datasets for subsequent models.

The following shows the Medcat visualization of the data extraction process.[8]



'Admission Date: [**2151-7-16**] Discharge Date: [**2151-8-4**]\n\n\nService:\nADDENDUM:\n\nRADIOLOGIC STUDIES: Radiologic studies also included a chest \nCT, which confirmed cavitary lesions in the left lung apex\nconsistent with infectious process / tuberculosis . This also\ nmoderate - sized left pleural effusion .\n\nHEAD CT: Head CT showed no intracranial hemorrhage or mass\neffect , but old infarction consistent with past medical\nhistory .\n\nABDOMINAL CT: Abdominal CT showed lesions of\nT10 and sacrum most likely secondary to osteoporosis . These can\nbe followed by repeat imaging as an outpatient.\n\n\n\n [**First Name8 (NamePattern2) **] [**First Name4 (NamePattern1) 1775**] [**Last Name (NamePattern1) **], M.D. [**MD Number(1) 1776**]\n\nDictated By: [**Hospital 1807**]\nMEDQUIST36\n\nD: [**2151-8-5**] 12:11\nT: [**2151-8-5**] 12:21\nJOB#: [**Job Number 1808**]\n'

Pretty Name	Tuberculosis
Identifier	C0041296
Type	Disease or Syndrome
Confidence Score	1
Start Index	258
End Index	270

Figure 9: Example of Medcat Tool

5.2 Techniques Used

Using input embedding, we were able to use the various words from our data preprocessing in our analysis. Input embedding is a NLP technique that allowed us to process the data. An example of an input is e.g. ['Cardiac chamber structure', 'Body of vert...']. We also used one-hot encoding for the labelling of features which allowed us to transform our categorical labels into a binary representation to use in the machine learning models.

We evaluated the data using several models: machine learning models and deep learning models. The goal of these models was to predict the diseases based on the symptom data. The machine learning models that we used include Dummy baseline with stratified labels, K Nearest Neighbours, Random Forest, Catboost gradient boosting, Catboost + Optuna hyperparameter search, and Catboost as text. The deep learning models that we evaluated include Attention, Attention + Convolutional Neural Networks, and Attention + Bidirectional LSTM. Attention mechanisms allow for the model to focus on specific parts of the input sequence. Attention+CNN allows the model to capture features using Convolutional Neural Networks alongside the attention mechanisms. Attention+Bi-LSTM allows for the model to focus on features by processing the text in both directions. We used a variety of models on the data to compare their effectiveness.

5.3 Code Structure

In the provided repository, we have the following files:

- `medcat_mimiciii.csv`: This is the Mimic-iii dataset that we have used for our analysis.
- `sample-mimic-iii.rar`: This is the zip file containing the complete mimic-iii dataset.
- `final_839_project.ipynb`: This is the jupyter notebook that contains the project code.

Our implementation is also uploaded to a GitHub repository: `https://github.com/schorm/Predict_diseases`.

5.3.1 How to run?

In the "final_839_project.ipynb" notebook, a number of code cells are used to execute data analysis and model training and scoring. The cells can be executed to replicate the results (for example, by pressing the "Run all" button on the toolbar).

6 Results and Evaluation

6.1 Machine Learning Methods

We have run the experiments with the few machine learning models:

- The dummy model with stratified labels, as a baseline.
- KNN algorithm as a second, more nuanced baseline.
- Random forest as implemented in *sklearn* with default hyperparameters.
- *CatBoost* gradient boosting with default hyperparameters and one-hot encoding of symptoms.
- *CatBoost* gradient boosting with default hyperparameters and symptoms passed as text features.
- *CatBoost* gradient boosting with one-hot encoding of symptoms and hyperparameter search using the *optuna* framework.

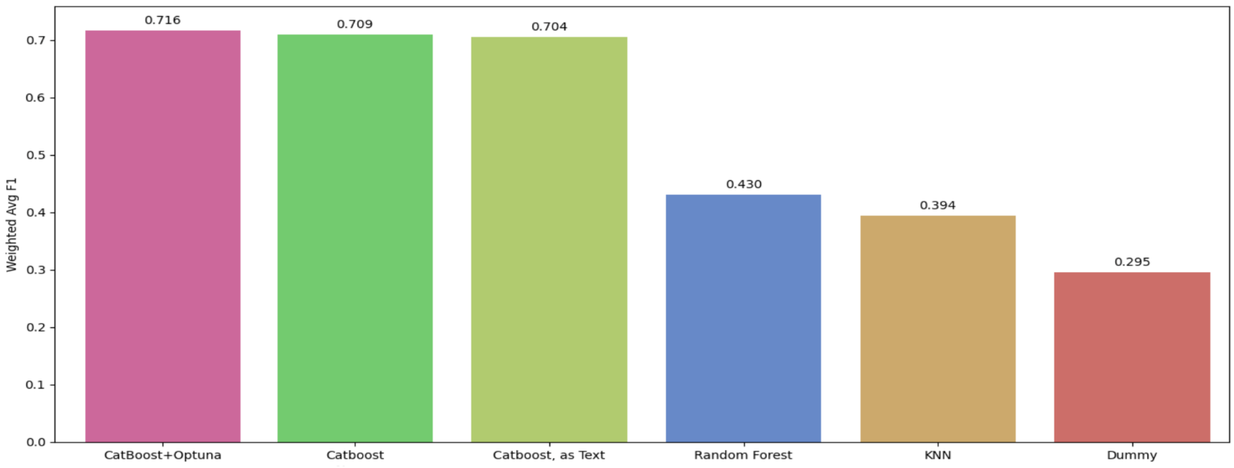


Figure 10: Results of various Machine Learning Models

The evaluation of these models showed that the gradient boosting algorithm performs significantly better. The top performer in this category was CatBoost with optimized hyperparameters with a weighted average F1 score of 0.716. Even in its default configuration, CatBoost demonstrates strong performance, achieving a weighted average F1 score of 0.709. In comparison, the performance of CatBoost with text features is slightly lower but still respectable, with a weighted average F1 score of 0.704. On the other hand, Random Forest, a widely-used ensemble learning method, exhibits relatively considerably performance with a weighted average F1 score of 0.430. Similarly, KNN falls short in this evaluation, obtaining a weighted average F1 score of 0.394. Finally, the baseline dummy classifier achieves the lowest performance among all models, with a weighted average F1 score of 0.295, demonstrating that the best-performing model far surpasses naive approaches. Overall, these results highlight the efficacy of CatBoost, particularly when coupled with hyperparameter optimization, in addressing the disease prediction problem.

6.1.1 Deep Learning results

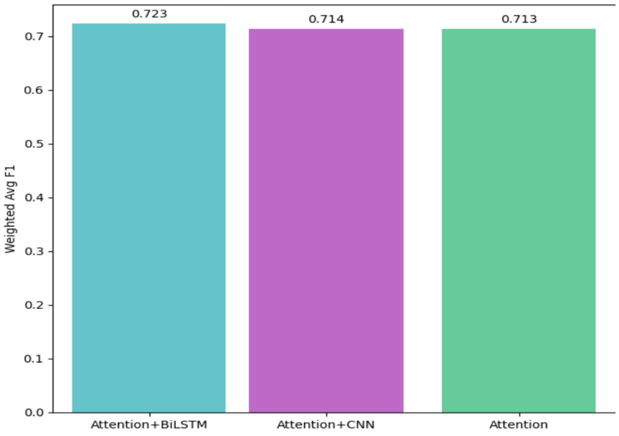


Figure 11: Results of various Deep Learning Models

Including deep learning models into the evaluation adds another layer of analysis to the comparison. The Attention + BiLSTM model emerges as the top performer among the deep learning architectures, boasting a weighted average F1 score of 0.723. Following closely behind is the Attention + CNN model, which achieves a decent weighted average F1 score of 0.714. Meanwhile, the standalone Attention model also shows strong performance, with a weighted average F1 score of 0.713. These results represent the effectiveness of attention mechanisms in enhancing the performance of neural network architectures. Also, they focus on the versatility of deep learning models in handling complex data and extracting meaningful patterns for classification tasks.

6.1.2 Comparison and discussion of all models

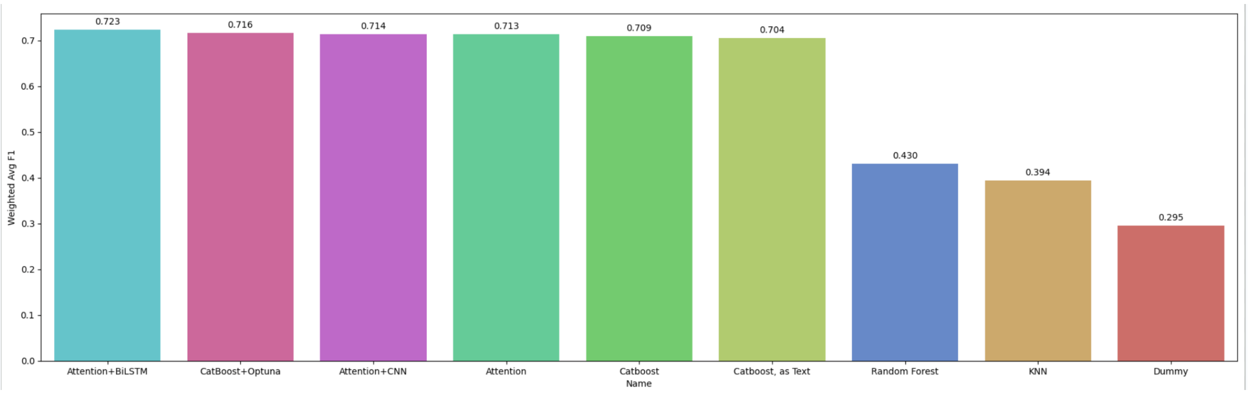


Figure 12: Comparison of Average F1 from all models

The analysis demonstrates that CatBoost, especially when optimized with Optuna, outperforms traditional machine learning algorithms, achieving a weighted average F1 score of 0.716. Deep learning models, particularly those incorporating attention mechanisms like Attention + BiLSTM, also show strong performance, with scores ranging from 0.713 to 0.723. These results highlight the effectiveness of advanced gradient boosting and attention-based deep learning architectures for the disease prediction classification task.

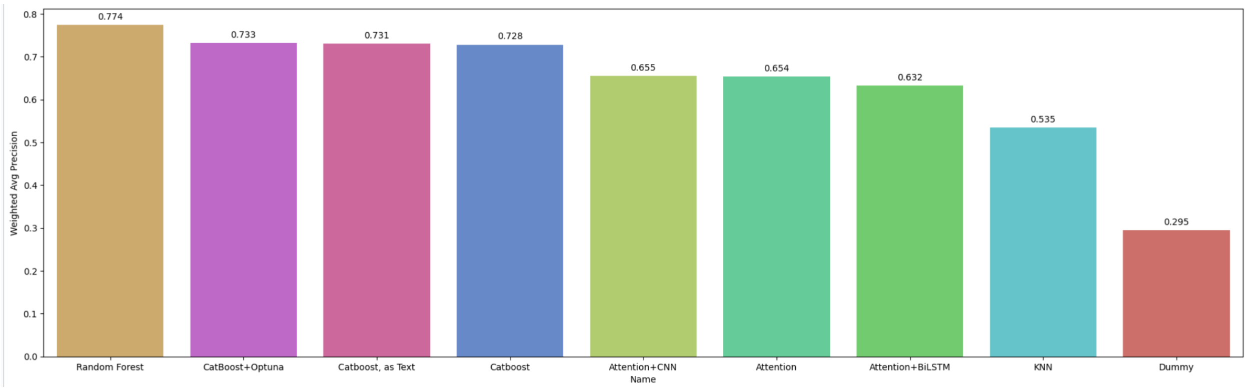


Figure 13: Comparison of Average Precision from all models

The results indicate various levels of precision across different machine learning models, with Random Forest exhibiting the highest average precision of 0.774, followed by CatBoost with Optuna tuning at 0.733, and CatBoost with text features at 0.731. This suggests that Random Forest is most effective at minimizing false positives and maximizing true positives in the classification task in comparison to other machine Learning models.

On the other hand, the deep learning models, including Attention + CNN, Attention, and Attention + BiLSTM, show slightly lower precision scores compared to the boosting models and random forest. This could be attributed to the complexity of these models and their sensitivity to parameter settings, as well as the nature of the dataset and the classification task. KNN exhibits even lower precision, suggesting its limitations in accurately classifying instances in this scenario.

The dummy classifier, which predicts classes based on simple rules, demonstrates the lowest precision among all models, as expected.

Overall, we can see from this comparison that the machine learning approaches show better average precision.

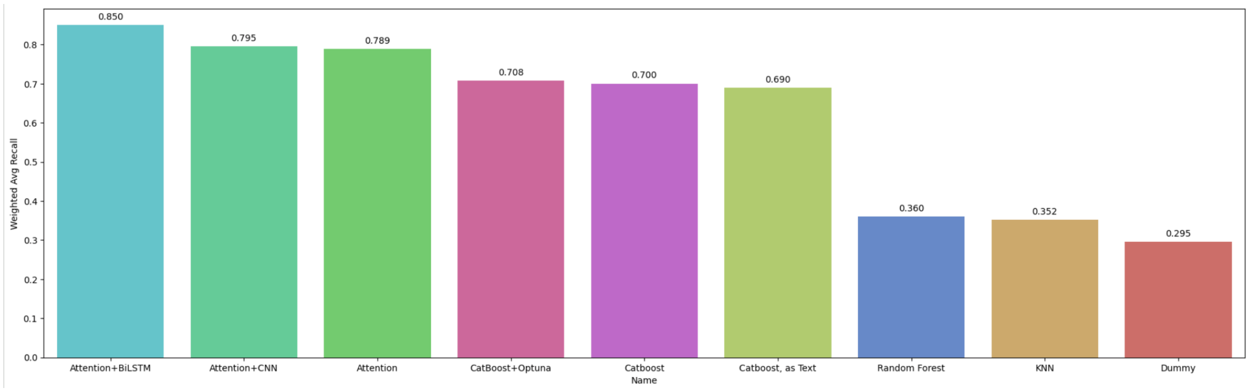


Figure 14: Comparison of Average Recall from all models

We can see from this comparison that the Deep Learning approaches show better average recall.

Deep learning models, especially Attention + BiLSTM and Attention + CNN, show high recall, effectively capturing positive instances. CatBoost variations also perform well in this regard, maintaining a balance between precision and recall. Random Forest and KNN exhibit lower recall, suggesting limitations in capturing positive instances. Dummy classifier shows the lowest recall. Overall, deep learning models excel in identifying positive instances, while CatBoost models maintained a balance, and traditional models show limitations in this aspect. As recall is crucial in medical tasks, we can conclude that selecting one of the top models in this comparison is adviced.

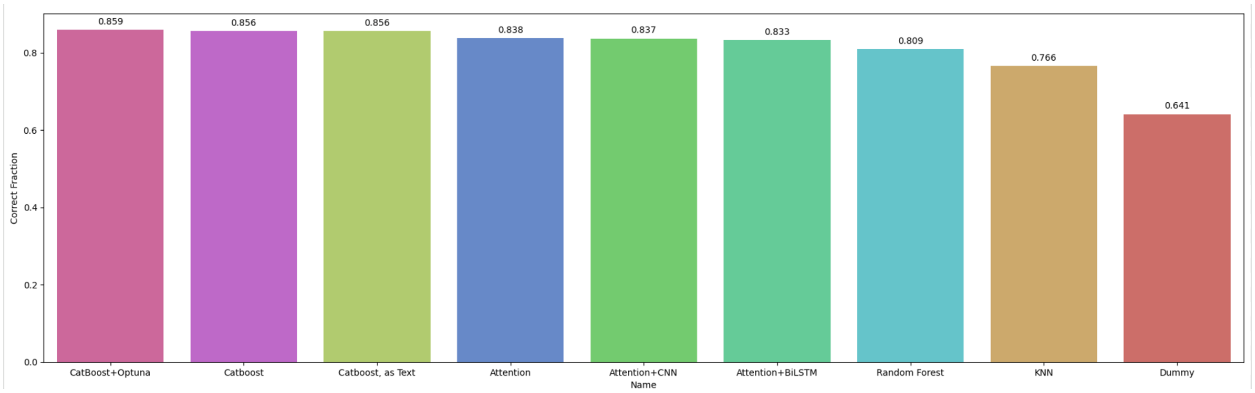


Figure 15: Comparison of the Fraction of Correct Labels from all models

The fraction of correct labels (that is similar to single-label accuracy) - *correctness*, for simplicity, varies across different machine learning models. CatBoost models achieve high correctness scores of 0.859 and 0.856, respectively, indicating their effectiveness in correctly classifying instances. Similarly, deep learning models, such as Attention + CNN and Attention, also demonstrate strong accuracy performance, with scores of 0.837 and 0.838, respectively. Attention + BiLSTM follows closely behind with an accuracy score of 0.833. Random Forest exhibits a slightly lower accuracy of 0.809, while KNN shows a further decrease at 0.766, suggesting limitations in accurately classifying instances compared to CatBoost and deep learning models. The Dummy classifier achieves the lowest accuracy score of 0.641, indicating its poor performance in accurately predicting labels. Overall, CatBoost models and deep learning architectures show superior accuracy in classification tasks, while traditional models like Random Forest and KNN lag behind.

Most approaches perform significantly better than 'naive' models: dummy model that predicts the most common label and KNN. For the natural language approach, BiLSTM with attention shows the best results overall, as the features are given in natural language. Depending on the importance of precision or recall - and in medical tasks recall is usually more important - different methods are best suited. In terms of performance, optimized gradient boosting shows scores close to the NLP method, as such, there can be a case for using it under resource constraints.

7 Discussion and Conclusion

7.1 Future Work

There is much future work that could be done with this dataset and with the results we obtained. Analysing specific diseases one at a time with machine learning techniques can lead to determining more intricate details behind the diagnosis. Classification of diseases in more broad categories would allow for pairing of patients with prospective professionals. The MIMIC-III database can also be used for work outside of disease diagnosis such as management of hospitals or other patient care topics. From our findings and exploration, we found that there is many opportunities for more research with the MIMIC-III database. Additionally, a different development vector would be to apply the same models we tested to a similar disease-prediction dataset and see if the results hold.

7.2 Conclusion

Throughout this project, we addressed the critical issue of misdiagnosis in the medical field, recognizing the large impact misdiagnosis has on both patients and healthcare systems, and tested the capabilities of ML and DL model in solving this task. Our literature survey allowed us to explore the various methods that researchers have used for automated diagnosis, and understand the different possible approaches to categorization and understanding the diseases with machine learning techniques. Building upon this approaches, we used the MIMIC-III database to conduct an exploratory data analysis to better understand disease patterns and diagnostic processes. We then were able to analyse the structure of the database, process it and use the MedCAT annotation tool to formulate a ML multi-label classification task based on the data. Lastly, we were able to perform a number of experiments comparing varios machine learning and deep learning techniques on the the dataset to predict the disease labels based on the symptoms. Our experiment have demonstrated that the BiLSTM model with attention mechanism achieves the best results due to the textual nature of the dataset.

References

[1] Patrick Marques Ciarelli, Elias Oliveira, and Evandro O. T. Salles. Multi-label incremental learning applied to web page categorization. *Neural computing & applications*, 24(6):1403–1419, 2014. Place: London Publisher: Springer London.

[2] Zheng Dai, Siru Liu, Jinfa Wu, Mengdie Li, Jialin Liu, and Ke Li. Analysis of adult disease characteristics and mortality on MIMIC-III. *PLOS ONE*, 15(4):e0232176, April 2020.

[3] Karnika Dwivedi and Malay Kishore Dutta. Microcell-net : A deep neural network for classification of microscopic blood cell images. *Expert Systems*, 40(7):e13295, August 2023.

[4] Prommy Sultana Hossain, Kyungsup Kim, Jia Uddin, Md Abdus Samad, and Kwonhue Choi. Enhancing Taxonomic Categorization of DNA Sequences with Deep Learning: A Multi-Label Approach. *Bioengineering (Basel)*, 10(11):1293–, 2023. Place: Basel Publisher: MDPI AG.

[5] Allaouzi Imane and Ben Mohamed. Multi-label Categorization of French Death Certificates using NLP and Machine Learning. In *Proceedings of the 2nd international Conference on big data, cloud and applications*, pages 1–4. ACM, 2017.

[6] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[7] Zeljko Kraljevic, Daniel Bean, Aurelie Mascio, Lukasz Roguski, Amos Folarin, Angus Roberts, Rebecca Bendayan, and Richard Dobson. Medcat–medical concept annotation tool. *arXiv preprint arXiv:1912.10166*, 2019.

[8] Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A Folarin, Angus Roberts, Rebecca Bendayan, Mark P Richardson, Robert Stewart, Anoop D Shah, Wai Keong Wong, Zina Ibrahim, James T Teo, and Richard J B Dobson. Multi-domain clinical natural language processing with MedCAT: The medical concept annotation toolkit. *Artif. Intell. Med.*, 117:102083, July 2021.

[9] Joon Lee, Evan Ribey, and James R. Wallace. A web-based data visualization tool for the MIMIC-II database. *BMC Medical Informatics and Decision Making*, 16(1):15, December 2015.

[10] Jinkai Li, Fan Song, Peng Zhang, Chenbin Ma, Tianyi Zhang, Yangyang Sun, Youdan Feng, Xiao Song, Shangqing Lyu, and Guanglei Zhang. A multi-classification model for non-small cell lung cancer subtypes based on independent subtask learning. *Medical Physics*, 49(11):6960–6974, November 2022.

[11] Ava L Liberman and David E Newman-Toker. Symptom-Disease Pair Analysis of Diagnostic Error (SPADE): a conceptual framework and methodological approach for unearthing misdiagnosis-related harms using big data. *BMJ Quality & Safety*, 27(7):557–566, July 2018.

[12] M.K. Lintu, David Raj Micheal, and Asha Kamath. Mortality prediction on unsupervised and semi-supervised clusters of medical intensive care unit patients based on MIMIC-II database. *Informatics in Medicine Unlocked*, 39:101264, 2023.

[13] Xin Liu, Yanju Zhou, and Wang Zongrun. Can the development of a patient's condition be predicted through intelligent inquiry under the e-health business mode? Sequential feature map-based disease risk prediction upon features selected from cognitive diagnosis big data. *International Journal of Information Management*, 50:463–486, February 2020.

[14] Sana Nazari Nezhad, Mohammad H Zahedi, and Elham Farahani. Detecting diseases in medical prescriptions using data mining methods. *BioData Mining*, 15(1):1–19, 2022.

[15] M Priyadharshini, A Faritha Banu, Bhisham Sharma, Subrata Chowdhury, Khaled Rabie, and Thokozani Shongwe. Hybrid Multi-Label Classification Model for Medical Applications Based on Adaptive Synthetic Data and Ensemble Learning. *Sensors (Basel, Switzerland)*, 23(15):6836–, 2023. Place: Switzerland Publisher: MDPI AG.

[16] Gökhan Silahtaroğlu and Nevin Yılmaztürk. Data analysis in health and big data: A machine learning medical diagnosis model based on patients' complaints. *Communications in Statistics - Theory and Methods*, 50(7):1547–1556, April 2021.

[17] Fuming Sun, Jinhui Tang, Haojie Li, Guo-Jun Qi, and Thomas S. Huang. Multi-Label Image Categorization With Sparse Factor Representation. *IEEE transactions on image processing*, 23(3):1028–1037, 2014. Place: New York, NY Publisher: IEEE.

[18] Máximo Eduardo Sánchez-Gutiérrez and Pedro Pablo González-Pérez. Multi-Class Classification of Medical Data Based on Neural Network Pruning and Information-Entropy Measures. *Entropy*, 24(2):196, January 2022.

[19] Aleksandr Talitckii, Ekaterina Kovalenko, Anna Anikina, Olga Zimniakova, Maksim Semenov, Ekaterina Bril, Aleksei Shcherbak, Dmitry V. Dylov, and Andrey Somov. Avoiding Misdiagnosis of Parkinson's Disease With the Use of Wearable Sensors and Artificial Intelligence. *IEEE Sensors Journal*, 21(3):3738–3747, February 2021.

[20] Wathsala Widanagamaachchi, Kelly Peterson, Alec Chapman, David Classen, and Makoto Jones. A flexible framework for visualizing and exploring patient misdiagnosis over time. *Journal of Biomedical Informatics*, 134:104178, October 2022.

[21] Tianran Zhang, Muhao Chen, and Alex A.T. Bui. AdaDiag: Adversarial Domain Adaptation of Diagnostic Prediction with Clinical Event Sequences. *Journal of Biomedical Informatics*, 134:104168, October 2022.

[22] Zhongheng Zhang, Kun Chen, and Hongying Ni. Calcium supplementation improves clinical outcome in intensive care unit patients: a propensity score matched analysis of a large clinical database MIMIC-II. *SpringerPlus*, 4(1):594, December 2015.