Leveraging Natural language Processing and Machine Learning Techniques to find Frailty Deficits from Clinical Dataset

by

Mehrnoosh Bazrafkan

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Statistical Machine Learning

Department of Mathematical and Statistical Sciences
University of Alberta

Abstract

Introduction Frailty is a syndrome that is often associated with aging. It can be identified through specific frailty scales or a comprehensive assessment by a healthcare provider. In Alberta, it appears that there are no specific billing or diagnostic codes for frailty. So, healthcare providers may use specific assessments or codes related to conditions such as muscle weakness or decreased physical activity to identify frailty. **Purpose** This project aims to leverage Natural Language Processing algorithms to extract frailty keywords from structured and Unstructured clinical datasets to identify frailty deficits and classify patients into frail and non-frail classes using Machine Learning algorithms. **Methods** The dataset included 450 patients over the age of 60, medical information related to diseases, and clinical frailty scales. We first clean medical notes using NLP techniques and removing negation terms, then extract keywords from clinical notes and structured datasets, and finally, we use resampling techniques to deal with imbalanced clinical datasets, and we feed these extracted keywords into machine learning classifiers to classify patients as frail or not frail. **Results** There are many different types of machine learning classifiers that have been used for this task, Random Forest and Decision Three with 0.95 performed better than LR, KNN, NB, SVM, and neural network models. **Conclusion** Natural Language Processing algorithms can effectively extract frailty keywords using Electronic Medical Record (EMR) notes. Moreover, comparing the results shows that using both structured and unstructured data gives better results than using only structured data.

Acknowledgements

I would like to express my gratitude to my primary supervisor, Professor Linglong Kong, for his guidance throughout this project. I would also like to express my sincere gratitude to Dr.Marjan Abbasi of the Department of Family Medicine at the University of Alberta, who assisted me with my research. I greatly appreciate the assistance I received from the Department of Medicine and its Data Administrators in obtaining access to the dataset. As well, I would like to acknowledge my family for their support and insight, which made this study possible. I would like to express my deepest gratitude to my mother and my husband, for their unwavering support and encouragement throughout my years of study and research. I would also like to thank my daughter and son for showing me the beauty and opposite side of life and relaxing my mind during the hard times of my thesis.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Frailty is a medical syndrome that typically affects adults over 65. It is defined as a higher likelihood of unfavorable outcomes following sickness or injury, even when age, other diseases, and medical treatment are considered [3]. Frailty is linked to higher healthcare expenses[15], a higher risk of adverse events during and after surgery, significantly lower quality of life, and a more significant load on family carers of frail patients[3]. Studies conducted in 2018 imply that 1.5 million Canadians are living with frailty, and this number is expected to increase to 2 million by 2025[3]. However, research has shown that some therapies, including nutrient supplements and increase unfavorable, can either delay or ameliorate frailty [45]. According to the research, early detection of frailty may provide better care for those who have been identified, such as assuring early intervention initiation [29] and guiding advanced frailty screening tests [9], which may result in fewer hospitalizations and lower downstream costs [29]. There are many ways to evaluate frailty, and the prevalence frailty estimates depend on the frailty instrument employed; hence, there is currently no accepted definition or method for measuring frailty[14]. Since there is no standard definition of frailty, in the Canadian Study of Health and Aging (CSHA), the Clinical Frailty Score (CFS) was developed to provide a summary of an older adult's overall degree of fitness or frailty after assessment by a skilled clinician [40]. To determine an individual's overall health status, CFS gathers clinical information on elderly patients. In [40], researchers ranked patients on a frailty scale from 1 to 7 based on their research. They considered a group of 40 descriptions for each disease, like "head and neck problems" or "sucking problems," and they determined some descriptions and medical keywords for each disease. They used this information to construct the 70-item CSHA frailty index. The highest CFS grade (level 7) included both extreme frailty and life-threatening illness. Later, it became clear that researchers were required to differentiate between recognizable categories that were initially included together on the original scale—namely, the terminally ill, the very severely sick, and the severely frail—as clinically distinct groups requiring distinct care strategies. In 2007, the CFS was increased from a 7-point scale to a 9-point scale, and it has been widely used in that format since then[37, 41]. Electronic medical records (EMRs) are a valuable source of clinical data for primary care research. Several studies have used machine learning algorithms to predict or index frailty for each patient using EMR.

In [2], the authors detected several concepts from doctors' notes in EMR. They utilized these keywords as inputs to a predictive frailty model. In [42], they considered 58 outcomes or variables like age, housing conditions, number of hospitalizations, number of medications, education level, and anemia that are not directly frailty factors. For this scenario, just 11 terms were taken from the data and used in the machine-learning model.

The specific objectives of this study are as follows: 1. Using NLP techniques to extract keywords from the unstructured and structured dataset 2. Using resampling techniques to handle the nature of an imbalanced clinical data set 3. Feeding extracted keywords (deficits) into Binary Machine Learning classifiers. 4. Evaluating the performance of the classifiers 5. Comparing the performance of structured and combination of structured and unstructured clinical datasets using different ML classifiers

1.0.1 Research Probelm

In order to speculate on a potential diagnosis, doctors and nurses make notes that include symptoms and signs of diseases. These notes provide vital details about a patient's health and information. The health record narrative commonly refers to the potential diseases that can rule out and symptoms the patient does not have. Clinical notes made by healthcare professionals are often unstructured and noisy, which makes it difficult for standard NLP tools to operate effectively. These notes frequently contain spelling errors, abbreviations, and unfinished sentences. Also, due to the special nature of the notes, using pre-trained text summarizers to preprocess clinical notes may encounter problems because they are trained on text with full-sentence documents or corpora. Additionally, automatic keyword tokenizers may have difficulty if they are not trained in medical texts, especially notes with disjointed sentences. In this study, we use NLP algorithms to clean data by replacing abbreviations with original words, and identifying negation terms and removing them, and extracting valuable clinical knowledge to help clinical experts make decisions.

Chapter 2

Clinical Dataset

2.1 Data Description

We aim to show how medical notes and generally text based datasets can help us to identify frailty deficits in this study. In the following sections, we discuss structured and unstructured databases.

An Electronic Medical Record or EMR is a digital version of a table containing unstructured notes. In EMR, Local-FI-EMR is the most important table consisting of text-form medical records and every visit is recorded using the SOAP (Subjective, Objective, Assessment and Plan) approach.

The Canadian Primary Care Sentinel Surveillance Network (CPCSSN) is a structured database that routinely collects and stores de-identified patient EMR data from eight provinces and one territory across Canada, with approximately 1.8 million unique patients in its database over a thousand primary care providers [20]. The CPCSSN database consistently collects diagnoses, billing codes, prescribed medications, measurements, and patient status information. In addition, each database entry includes a unique patient ID and a timestamp, which can be used to aggregate the data by patient or visit. In this study, we will use the Medication, billing, and patient tables as the structured tables. Also, we use the disease case indicator table to determine the clinical frailty score.

2.1.1 DiseaseCaseIndicator, Medication, Patient Tables

The DiseaseCaseIndicator table 2.1 aggregates data from all tables, including lab results, medications, diagnosis, etc., to identify deficits for each patient.

Patient_ID	Disease	IndicatorValue	TableName	DateCreated
2051	Dementia	493	HealthCondition	2016-05-18
2051	CKD	26	LAB	2016-08-28
2051	Dementia	290.0	Billing	2017-09-20
2051	Depression	211	Billing	2018-08-19
2051	Diabetes Mellitus	Gliclazide (A10BB09)	Medication	2019-09-11

Table 2.1: This table shows an example of a DiseaseCaseIndicator table for a single patient.

The medication table contains information about the medications prescribed for patients during every visit. There is information about the patient's date of birth and gender in the patient table. The following section provides some observations regarding patient gender and age distribution, as well as frailty expressions. According to Figure 2.1, 440 patients are active, 8 are inactive, and 2 are deceased. In this data set, 268 patients are female, and 182 are male 2.2. The patient's age range from 61 to 105 years old2.3. Each patient's frailty score is recorded and represents the patient's frailty during their last visit to the clinic. As shown in Figure2.4, the majority of patients have a frailty score below three, indicating that the majority of patients in this data set are relatively frail or "pre-frail." As a result of this distribution, we have significantly more information about patients who are not currently frail. We have a somewhat imbalanced data set, and our data set is slightly skewed towards pre-frail patients. We depict the distribution of frailty scores by gender in Figure 2.5. In this particular sample, at the population level, there are more females than males in all levels except for level 6. Frailty distributions by age can also be seen in Figure 2.6. Patients of all ages are likely to have frailty scores of 2 or 3. A frailty score of 7 does not occur in patients 61 to 80, and the majority of frailty scores 7 occur in those over 83. As a result, frailty scores increase with age. We notice that even without further testing to verify, age distribution, and statistically relevant, there are few differences

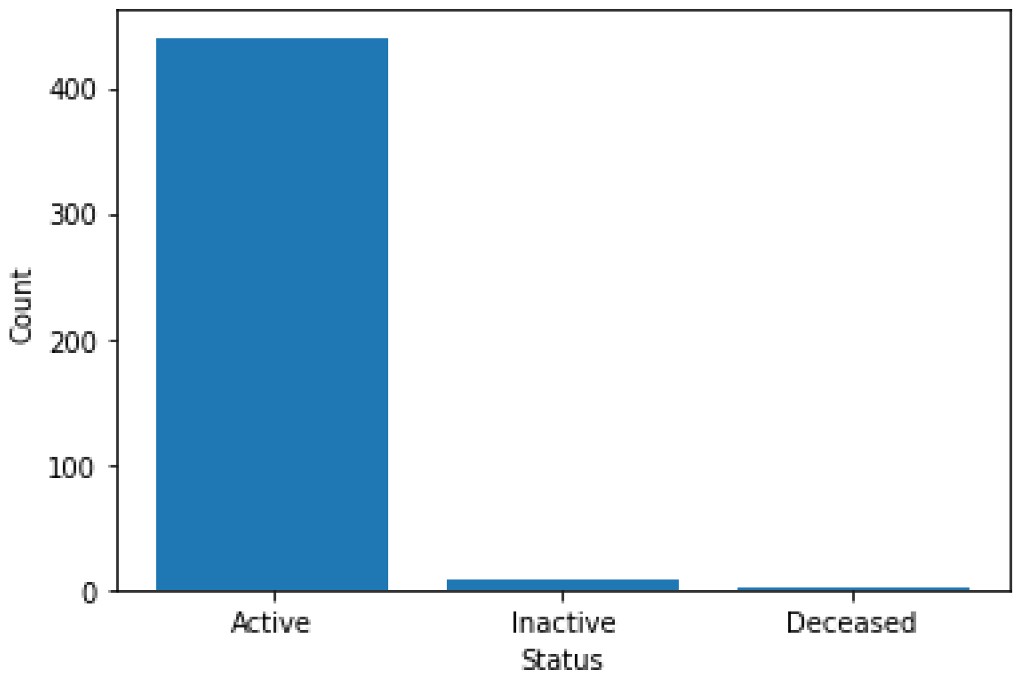

Figure 2.1: This figure displays the distribution of patients by status.

between them. However, we need additional data to generalize our results.

Our analysis of the "DiseaseCaseIndicator" table reveals that several patients were diagnosed with certain disease indicators. This table provides valuable information about diseases that may be considered indicators of frailty. These bar charts in Figure 2.10 illustrate the prevalence of each disease among these patients. According to the figure, some diseases, such as Dysplimedia, Hypertension, and Osteoarthritis, are more common among these patients, while others, such as Parkinson's, dementia, Herpes Zoster, COPD, and Epilepsy, are not. This table gives us valuable information about disease indicators, and in our research, we will use this information to compare our gained results with this table. However, some of their indicators aren't included in our frailty indicators, such as dyslipidemia, pediatric diabetes mellitus, and pediatric asthma.

In this study, 36 deficiencies are taken into account to predict frailty. Those who are frail will be assigned to class one, while those who are not frail will be given to class zero. Due to the lack of class information provided by the

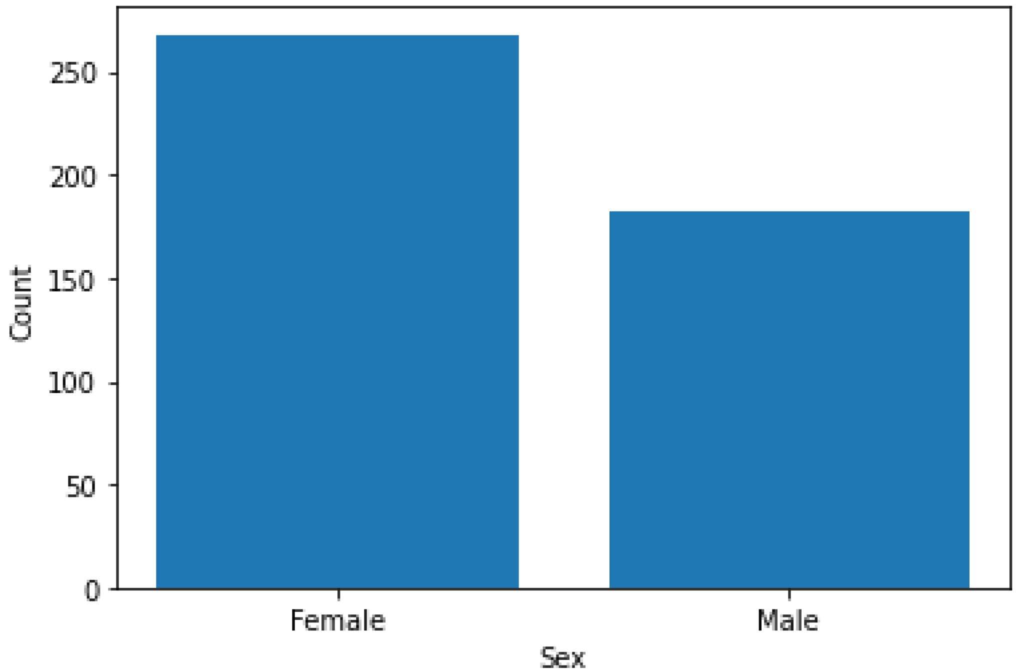

Figure 2.2: This figure illustrates the distribution of patients by sex.

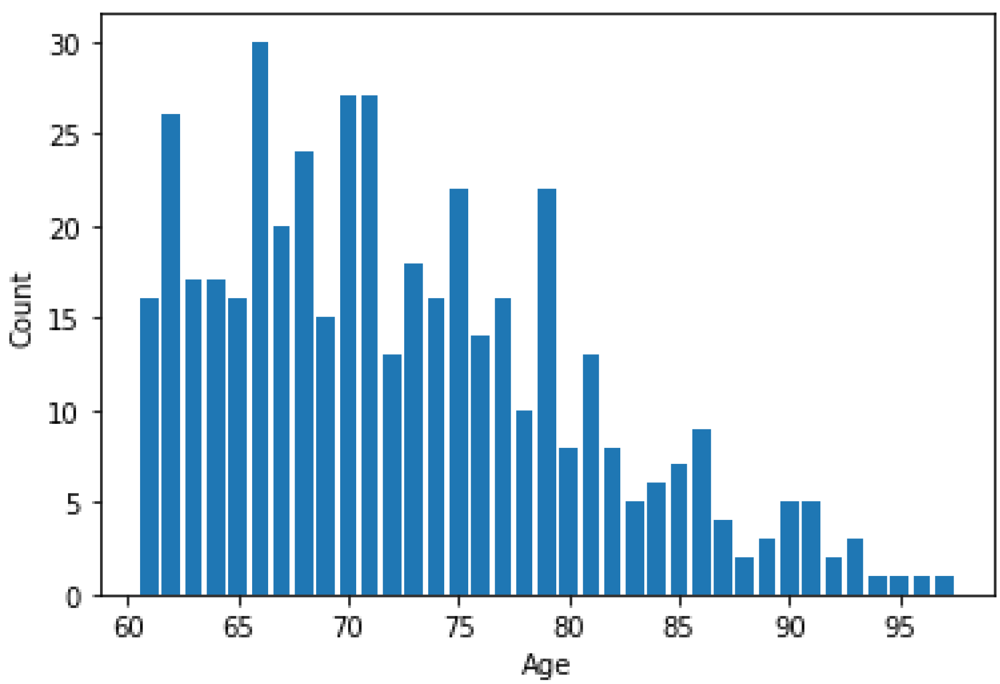

Figure 2.3: This figure shows the distribution of patients by age.

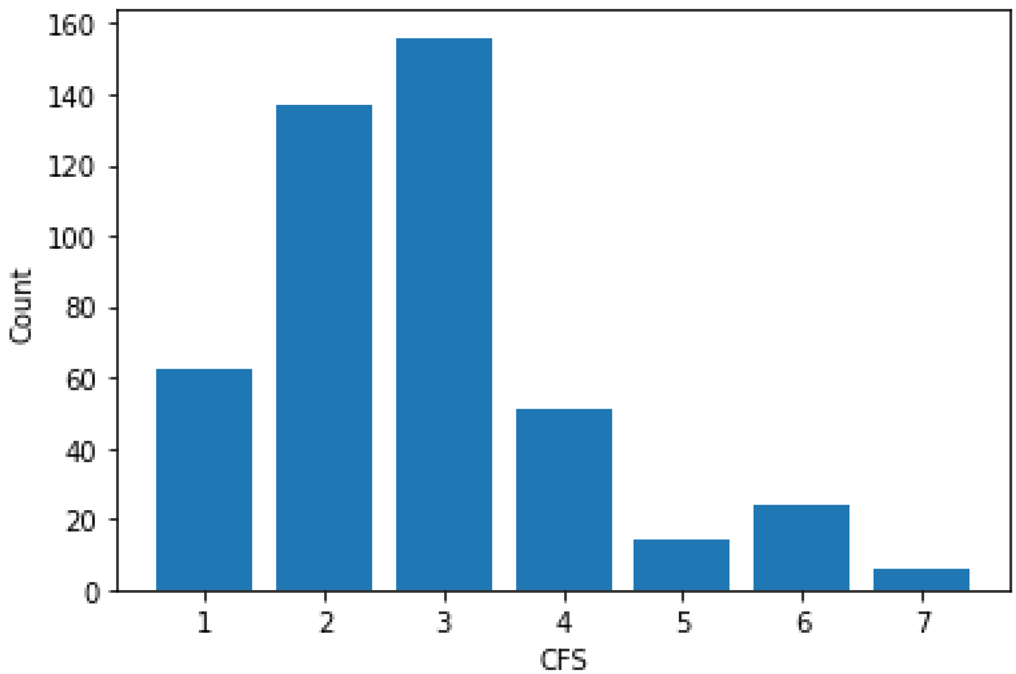

Figure 2.4: This figure shows the Clinical Frailty Scale with a range of 1 to 7.

medical texts that comprise our data, patients with a CFS of less than 4 were considered non-frail, and those with a CFS of 4 or higher were considered frail2.4.

According to figure 2.11, patients are split into frail and non-frail categories. The histogram shows 95 patients are frail and 355 patients are not. The frail group includes 33 men and 62 women. According to data from patients considered frail, only 0.072 percent are men and 0.14 percent are women. Patients between the ages of 60 and 80 are less likely than those above 80 to be classified in the frail class. According to the graph2.8, patients above 80 are more likely to be classified as frail.

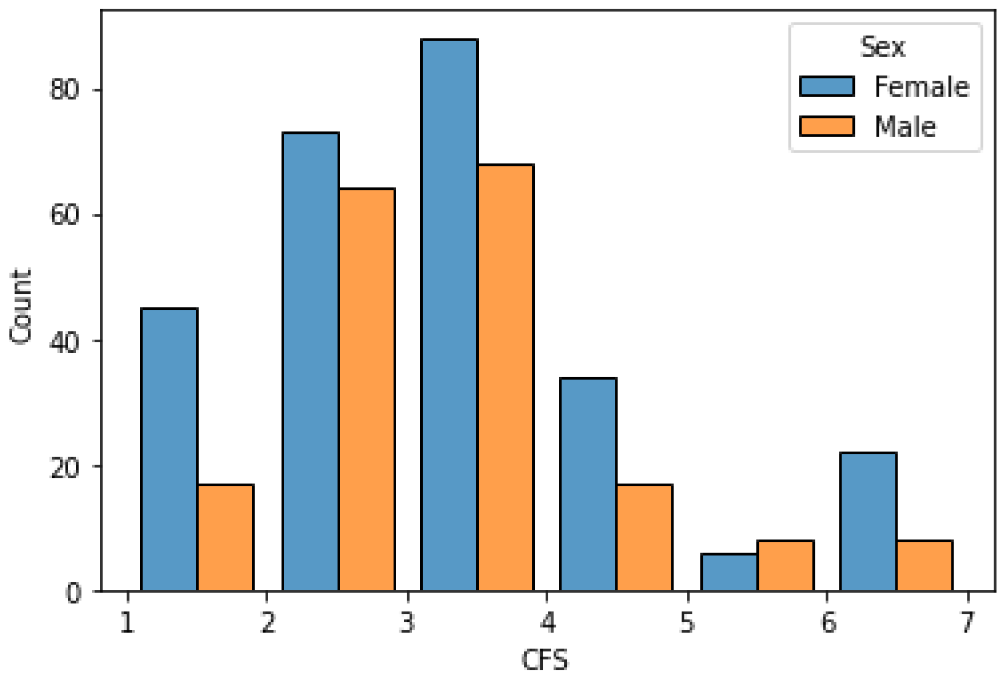

Figure 2.5: This figure shows the relation between Sex and CFS.

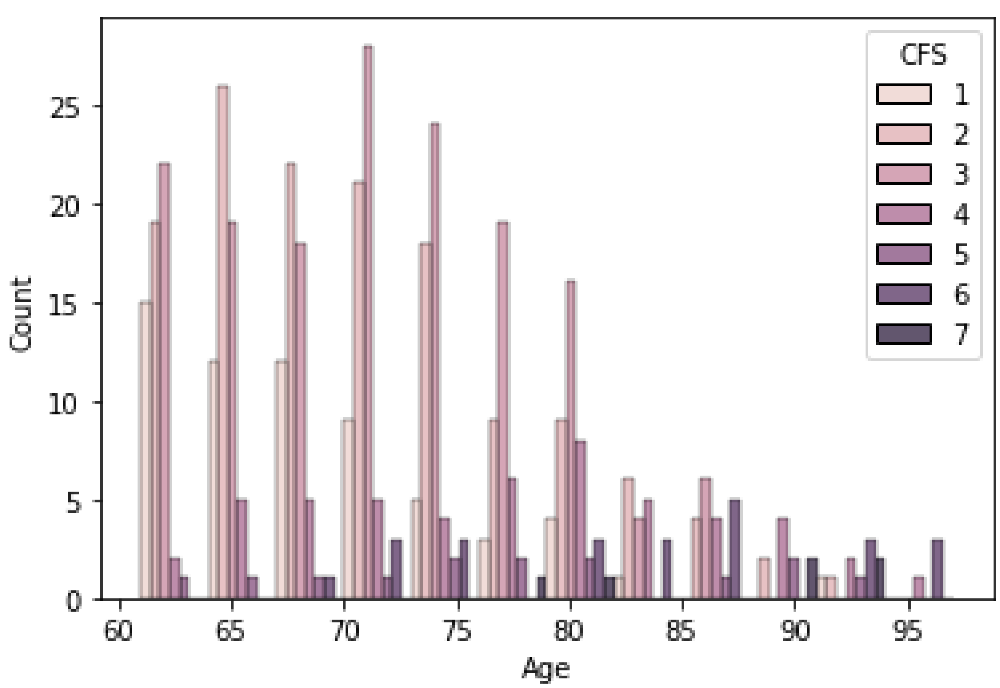

Figure 2.6: This figure shows The relation between Age and CFS.

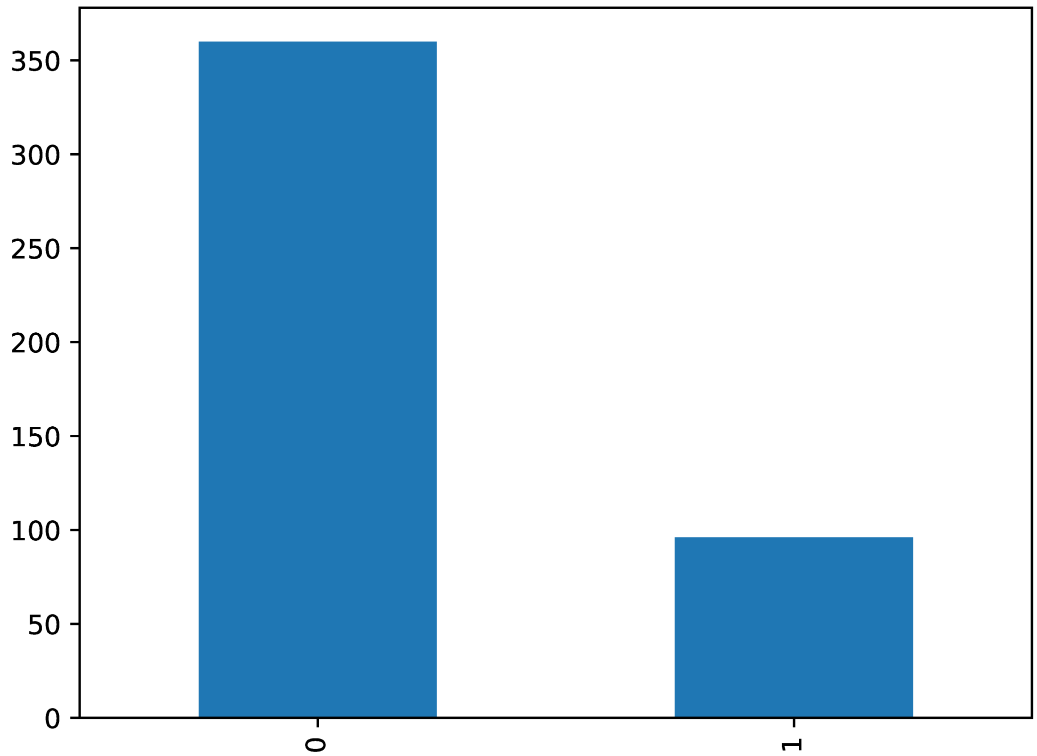

Figure 2.7: This graph divides patients into frail and non-frail classes according to their ages.

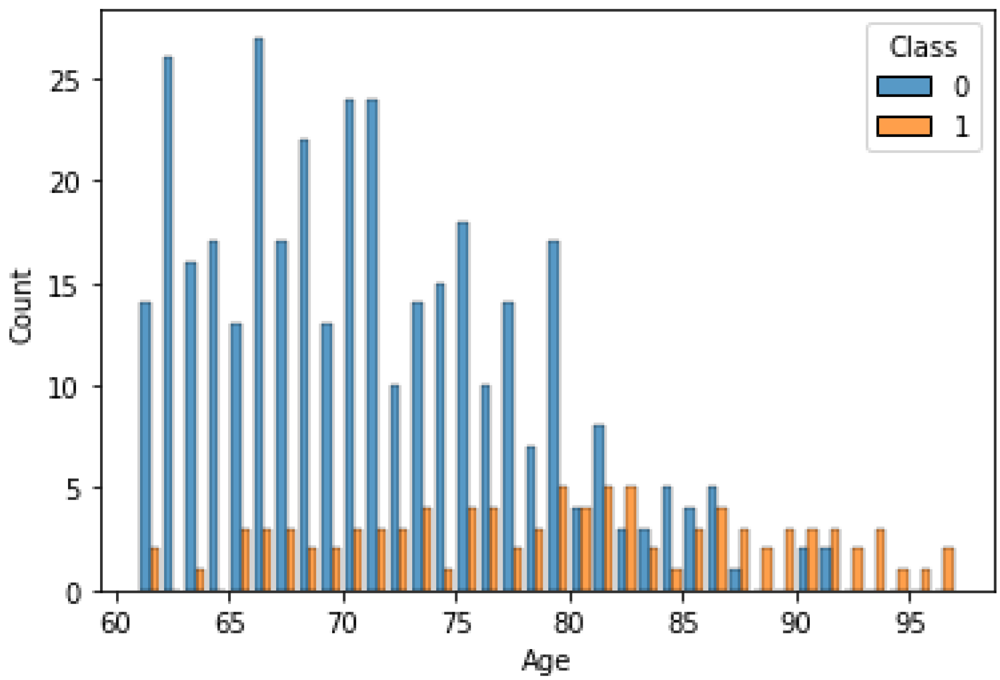

Figure 2.8: This figure shows the relation between Age and Class.

One of the deficiencies in determining whether a person is frail is the number of medications each patient takes. The second table we used as our dataset contains data on patients' medicines. As depicted in figure 2.12, each patient's range of medication use is between 0 and 74. Here, we divide patients into two groups—those without polypharmacy who take less than 15 drugs and those with polypharmacy who take more than 15 medications. Therefore, there are 141 patients with polypharmacy and 315 patients without polypharmacy 2.14. Our next step is extracting keywords from the local-fi-emr and billing dataset that we will consider later as other 35 frailty deficits.

2.1.2 Local-FI-EMR and Billing Tables

The primary dataset on which we will concentrate consists of text-form medical records of patient visits. These notes were created using the SOAP approach, which means there are three entries for each patient. These entries are divided into four categories: Subjective, Objective, Assessment, and Plan. The data

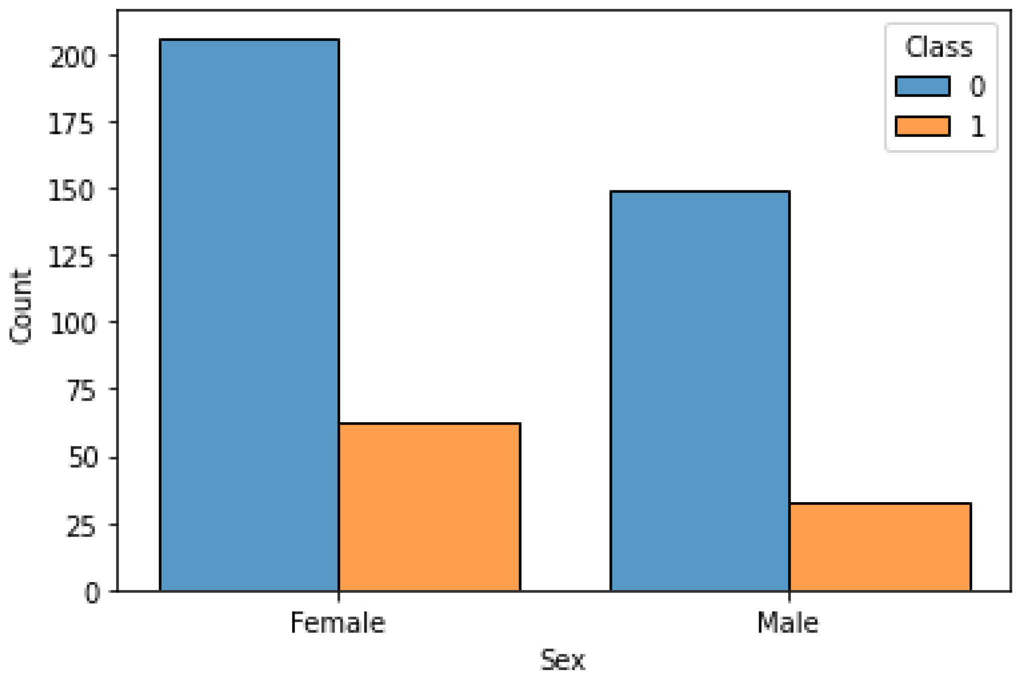

Figure 2.9: This figure shows the relation between Sex and Class.

local_FI_EMR_ID	Network_ID	Site_ID	Patient_ID	NoteType	NoteContent_orig	DateCreated
5084	2	6	2000XXX7	Subjective	CDM Nurse visit re management of weight	2014-06-26 15:30:00
2133	2	6	2000XXX7	Objective	Patient states she has had difficulty with mot...	2014-06-26 15:30:00
1	2	6	2000XXX7	Assessment	Discussion on the why weight program, review o.	2014-06-26 15:30:00

Table 2.2: An example of a single SOAP note for a single patient visit.

supplied for this project did not contain any "Plan" remarks. 2.2 illustrates how the Subjective and objective, and Assessment portions of a note are divided into three entries in the local-fi-emr database using notes from a single patient's visit as an example.

First, the medical notes must be edited since the information is both medically and personally sensitive. The Named Entity Recognition (NER) technique can be used to identify any personal information. It is a continuation of the work done by Alexander Tennant and his team, and they removed all personal information; we can skip this step. Therefore, we apply our techniques directly to local-FI-EMR datasets without personal information about patients. As well as the primary text-based table, several other tables from the

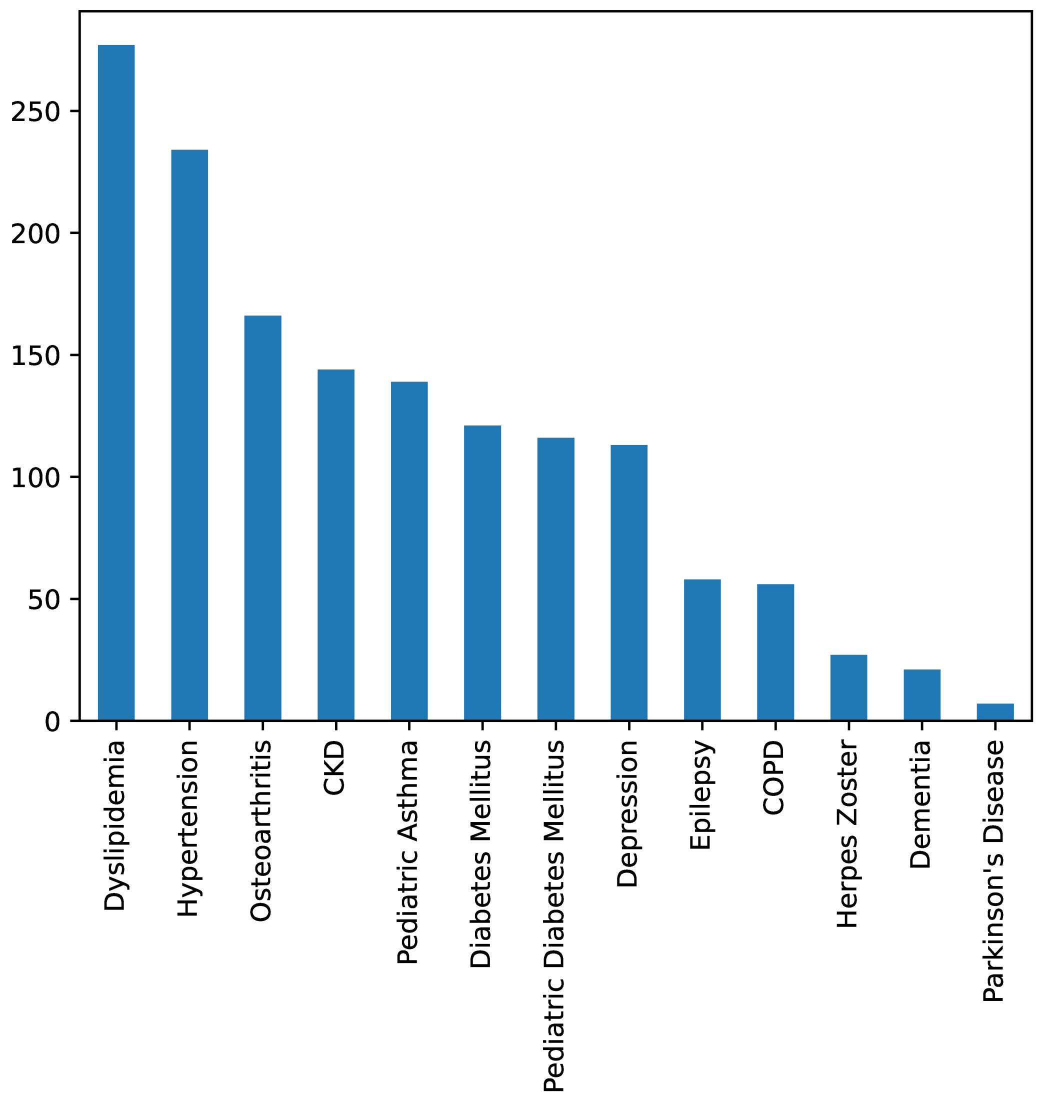

Figure 2.10: This figure illustrates the frequency of each disease in 450 patients

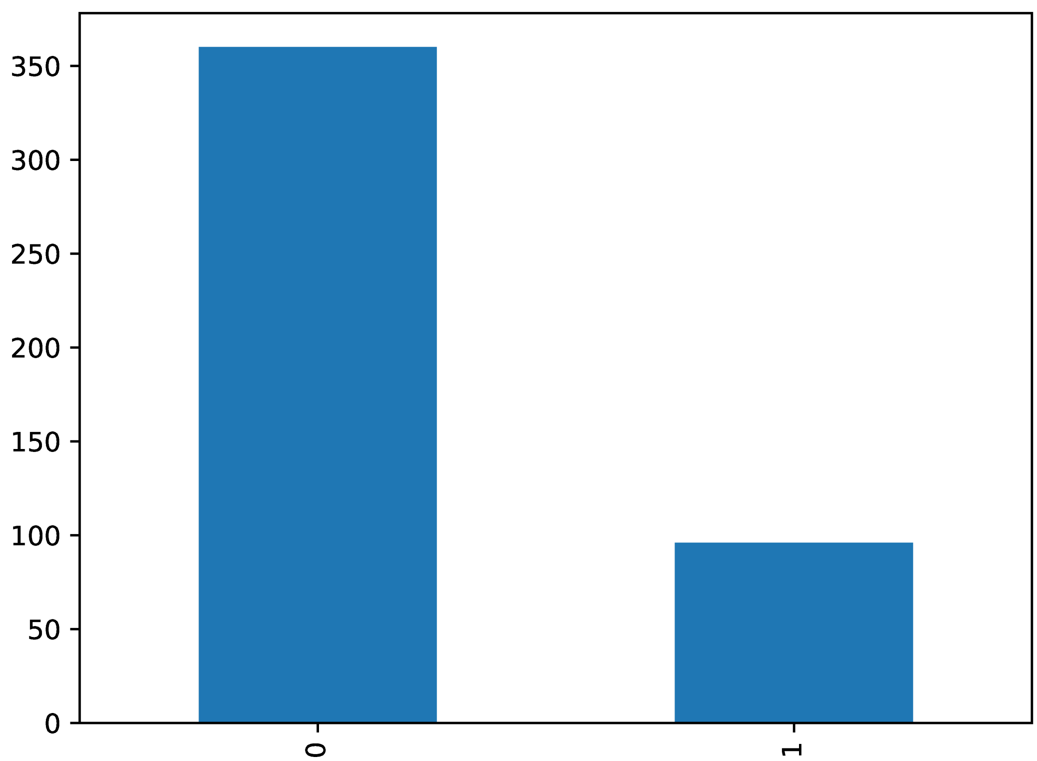

Figure 2.11: This figure shows number of patients that are frail and non-frail between 456 patients

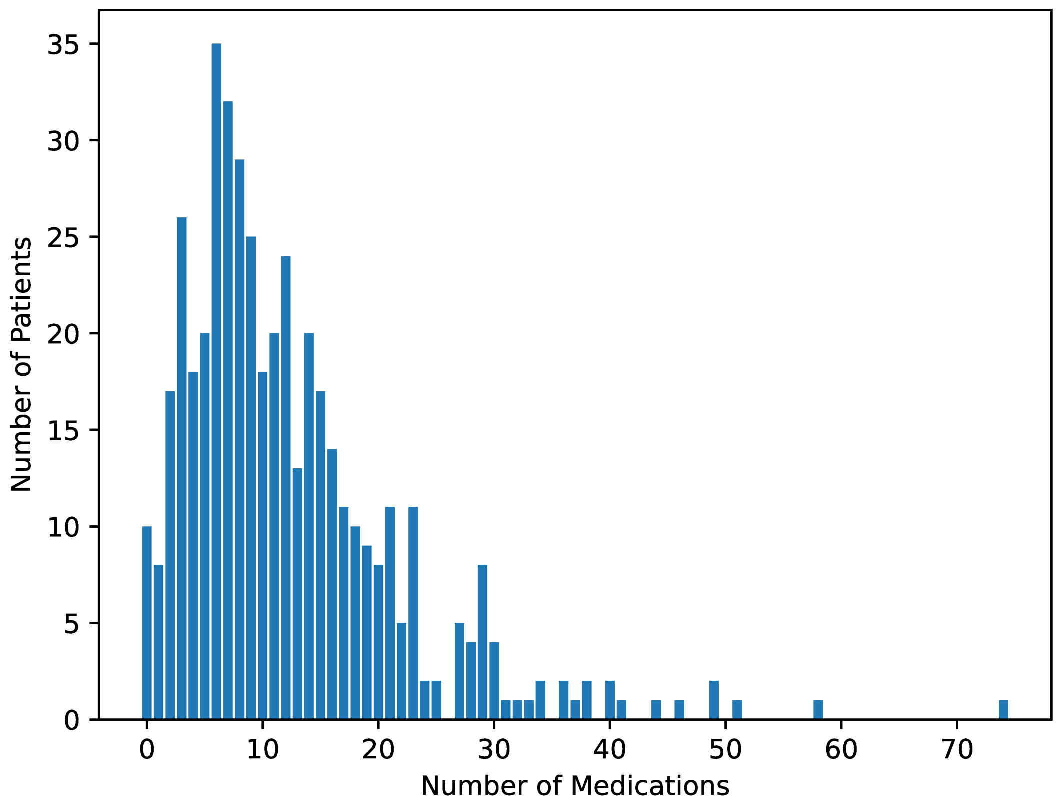

Figure 2.12: This figure shows the Number of Medications for each patient in 5 last years of their treatment

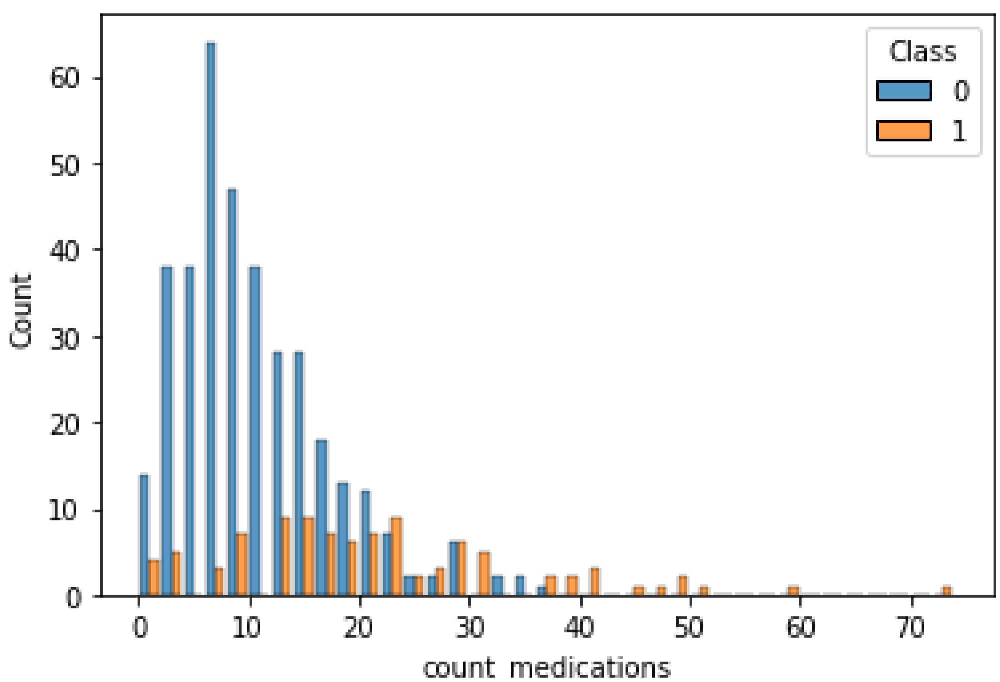

Figure 2.13: This figure shows the relationship between the number of medications and being Frail or Non-Frail.

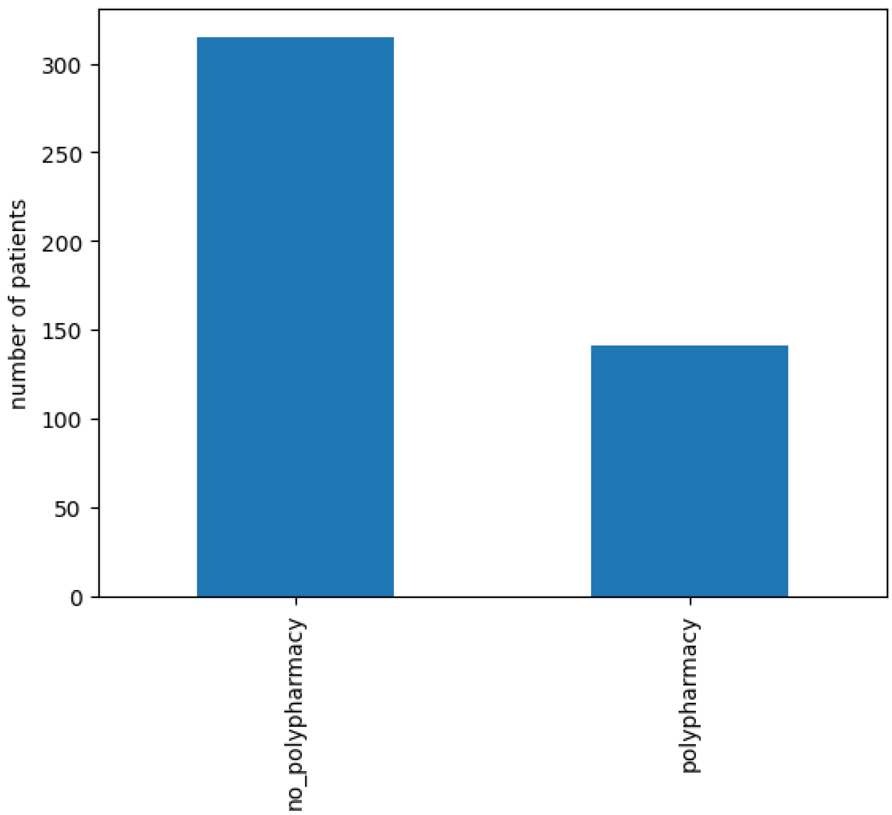

Figure 2.14: frequency of patients with or without polypharmacy

EMR database were also provided. There are several types of information in these tables, such as information from patient visits, lab tests, or other historical information. These tables contain different patient information such as risk factors (smoking and alcohol), vaccinations, diseases, BMI, age, medications, etc. Some data in these tables are fully free from text, some are semi-free from text, some are numerical information like age and weight, and others are categorical such as sex (Female or Male). Data in these tables can be used to predict frailty, especially information on age and gender. In this study, we first extract keywords from text-based tables such as billing and local-fi-emr. After that, we will discuss other factors to consider when calculating frailty indexes and classifying patients into frail and non-frail categories.

2.1.3 Clinical Notes preprocessing

Clinical notes provide vital details about a patient's health and information. In order to speculate on a potential diagnosis, doctors and nurses make notes that include symptoms and signs of diseases. Clinical notes are valuable information for the healthcare system. Consider the following sentence "She has pain radiating up the back of her neck on the right side." This is easy for human readers to understand the patient has radiating pain on the right side of her neck. However, it is quite challenging to develop a generalized technique that will enable us to have a computer program that is able to automatically extract this crucial information from unstructured text input. There are various NLP techniques that are available to make clinical text data available for analytics and modeling. The first and most crucial phase in model construction is data cleaning. Data cleansing depends on the type of data; it is especially important if the data is textual. There are many different text processing methods that may be used on text data, but we must be cautious while applying and selecting the processing steps. Here, the use cases determine the methods for processing textual data. The most common methods for cleaning text data step by step are:

- **Lowercasing the data**: The idea is to change input text to lowercase such that it changes "atrial Fibrillation" to "atrial fibrillation".

- **Removing Punctuations**: The removal of punctuation from text is the second most used text processing method. The clinical notes we are working on contain some punctuation marks like . $< br > < br < b\ and$ $+$) $xxxx$ ($-$? that we remove all of them.

- **Removing Numbers**: Depending on the note, numbers occasionally don't contain any essential information in the text. Therefore, getting rid of them is preferable to keep them. In this project, we first kept numbers as some keywords like B12 supplement contain keywords, then we found it doesn't work well for other sentences including numbers; therefore, we decided to remove numbers to get better results. For example, in the Deficit Description file (Alex), weight loss of more than 5 is considered a description for the patient with a weight loss deficit, but in keywords extraction, every number 5 in notes is considered as weight loss. So, it can cause wrong decisions in other project steps. It especially affects classification.

- **Replacing medical abbreviations**: Medical texts often use abbreviations; therefore, we must first replace all of these words with their equivalents to make them understandable and consistent. To do this, we created a list of abbreviations related to the project's terms, such as table 2.3. This file was created by reading medical notes related to more than 400 patients to effectively extract keywords in the following phases.

- **Removing Stopwords**: Stopwords are often used in text mining and natural language processing (NLP) to remove terms that contain no meaningful information. Stop words in English include "a," "the," "is," "are," and others. In this study, we removed stop words as a post-processing step.

- **Stemming and Lemmatization**: In Natural Language Processing (NLP), lemmatization refers to switching words to their base roots. For example, the term 'Caring' would return to 'Care'. As part of this study, NLTK was used to perform lemmatization. Moreover, Stemming removes the beginning or end of the word and considers a list of frequent

Acronyms	
&	and
++++	positive
R	right
L	left
abdo	abdominal
BPRV	benign paroxysmal positional vertigo
Diab mell	Diabete mellitus

Table 2.3: This table lists a selection of abbreviations frequently found in medical notations.

prefixes and suffixes that can be discovered in derived words. We assert that this strategy has some drawbacks because this trimming can be effective occasionally but only sometimes. For example, stemming the word 'Caring' would return to 'Car.' Various stemming algorithms can apply, but Porter stemmer is the most popular in English.

In this study, we used lemmatization and Stemming have been used as post-processing steps.

- **ALEX table**: Alex's file 2.4 includes 36 regular deficiencies in frailty and a few descriptions of those deficiencies. For instance, while describing Arthritis deficiencies, we took into account terms like "Arthritis, osteoarthritis, OA, rheumatoid arthritis, RA, DJG, degenerative changes, degenerative disc disease, knee pain-OA, knee OA, and hip replacement." Polypharmacy is one of these deficiencies, which we explained before in how we categorize patients.

Deficit	Description
Arthritis	Osteoarthritis Rheumatoid arthritis, Seronegative rheumatoid arthritis, Flare of rheumatoid arthritis, arthritis
Anemia and hematinic deficiency	Low hemoglobin abnormal blood count, hemoglobinopathy,Anemia, hematinic, blood disorder, anemia - iron deficient,

Table 2.4: This table shows a portion of Alex's file which contains 36 deficits and their description.

2.1.4 Negation Detection

The health record narrative commonly refers to potential diseases that can be ruled out as well as symptoms the patient does not have. Therefore, negation detection is crucial to identify the relevant conditions for valuable clinical decision support and knowledge extraction. The presence of negative entities in the clinical note could represent a wrong diagnosis. For example, in the following sentences, **No evidence of hypertension** it means the patient doesn't have "hypertension." According to my findings, if we only use pre-processing techniques to clean the data and then use Alex's file to find deficits, it will identify hypertension as a patient's deficit. However, if the dependency tree and negation algorithm are applied to this sentence, this deficit will be eliminated, which is the right thing to do. Negation algorithms perform their task well to find negative terms and eliminate them from notes, but there are occasionally some ambiguities in the understanding of sentences by machines. Another issue we encountered when using the negation algorithm to analyze medical notes was that, for example, for this specific sentence **"does not normally awaken her from sleep as lying in bed makes her feel better."** the physicist expects to extract the sleep disturbance from the notes since it contained negative words like **doesn't** the algorithm recognized it as a negation sentence and removed it from the note. Even if it didn't have this negative term, still negation algorithm was not able to find sleep deficiency for this patient because in Alex's file we considered as our reference to find deficit keywords for sleep disturbance: **Sleep disturbance, impaired sleep, sleep disorder, sleep apnea, sleep clinic referral** doesn't indicate directly to any terms in this sentence. This sentence has ambiguous for the computer to analyze it.

2.1.5 Handling Negation Terms using Dependency Parsers

Many researchers use dependency parsers to analyze medical notes because they can process and analyze a sentence's grammatical structure to identify related words and the nature of their relationships. Dependency parsers represent these relationships and dependencies as a hierarchical tree. This algorithm divides a sentence into numerous components. This technique assumes that each linguistic component of a phrase has a direct link with the others. These relationships are called dependencies. Each connection consists of a head node and a dependent node that modifies the head. All connections have been labelled based on the dependency that exists between a head and a dependent. These labels exist at universal dependency relations 2.15 [1].

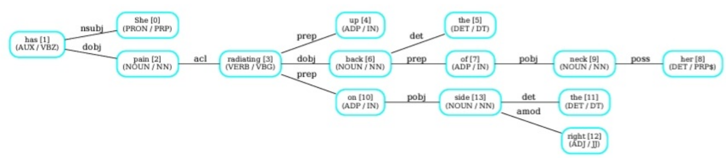

Figure 2.15: An example of dependency tree for the sentence "She has pain radiating up the back of her neck on the right side"

According to the "FI-EMR Final Report," to solve the negation problem in medical notes, they used a dependency parser to find the relationship between keywords and negation terms in a sentence. In their research, they considered various types of negations using dependency parsers that include:

- **Negation by level**: when the keyword and negation term are at the same level in dependency tree 2.16. **"Did not document any pain"**

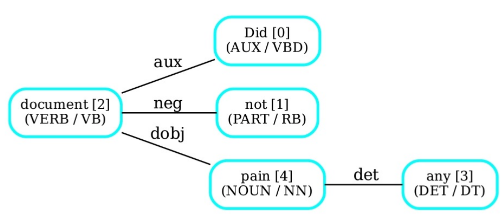

Figure 2.16: This figure shows the dependency parser for the sentence "Did not document any pain."

[1]The contents of this section have been rewritten from the FI-EMR final report written by Alexandar Tennant and his team at Cybera Inc. This project is the continuation of their project

- **Negation by child**: In this type of negation, the negation term will reveal as a child of the keyword 2.17. **She has arthritis in her wrists but no neck pain.**

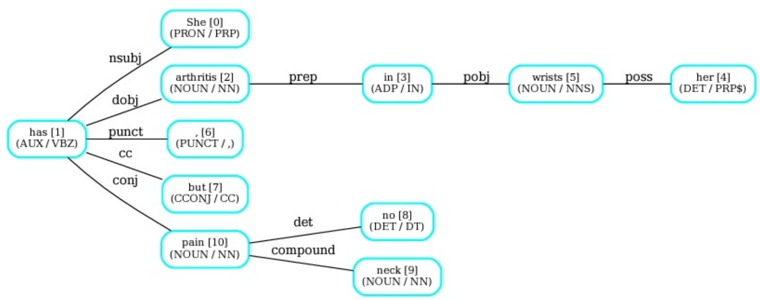

Figure 2.17: This figure shows the dependency parser for the sentence ”She has arthritis in her wrists, but no neck pain.”

- **Negation by parent branch**: when the keyword of interest and the negation sentence are separated by a descriptive noun. Consider the below sentence as an illustration: **There is no evidence of testicular cancer** According to the dependency tree, our keyword ”cancer” is a deep child of the noun ”evidence,” which is negated 2.18.

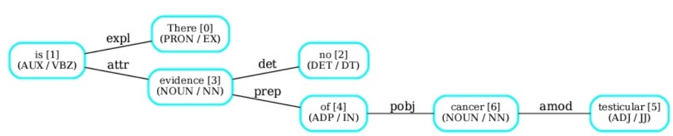

Figure 2.18: This figure shows the dependency parser for the sentence ”There is no evidence of testicular cancer.”

- **Negation by chain**: In sentences like **He denies having any pain in his left leg**, where the detected keyword is being negated by a verb like denies, this verb will typically be a parent node2.19.

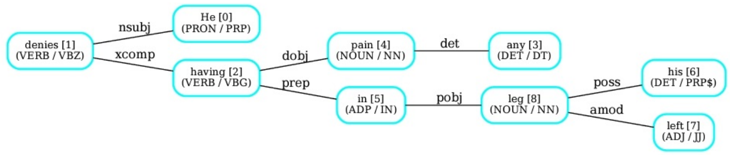

Figure 2.19: This figure shows dependency parser for the sentence "He denies having any pain in his left leg."

- **Negation by hanging root**: This is a unique instance of negation that frequently occurs when a keyword is connected to the negation of a keyword that is further specified by additional verbs and nouns. Take the following phrase, for instance2.20: **"Did not document any new indications of pain,"**

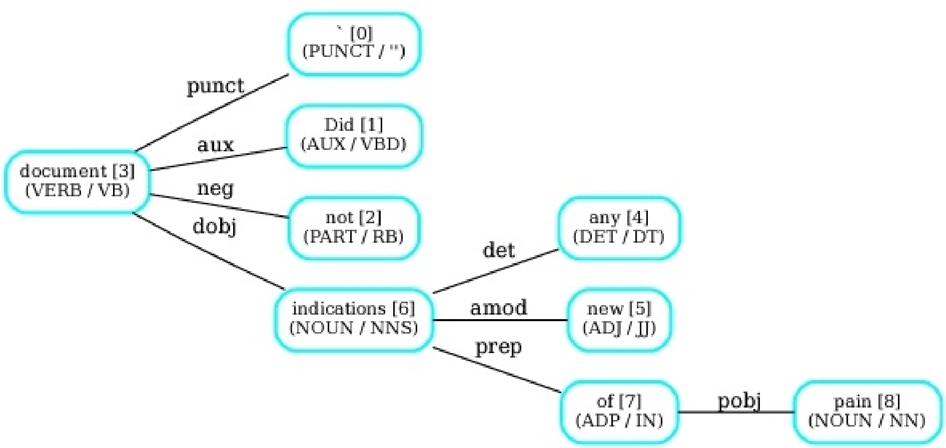

Figure 2.20: This figure shows dependency parser for the sentence "Did not document any new indications of pain," to describe negation by hanging root

- **Negation by resolution**: Negation by resolution is a special instance of negation by the chain. Consider the sentence **"His pain has resolved,"** This sentence describes a situation that formerly existed but has currently been resolved.

2.1.6 Outstanding Negation Algorithms problems

In this section, we discuss a few instances where the Dependency parser will always fail since the methods are not implemented.

- **Adjectival Negation**: First and foremost, there is the more ambiguous concept of adjectival negation or negation that is not explicit. For example, the sentence **"The patient is only a little tired,"**. This statement implies that the patient is not tired, although they have not stated this explicitly. They used adjectival negation to describe his situation. Fortunately, the official medical records won't use a lot of adjectival negation as mentioned in [33].

- **Double Negatives**: For example, consider the sentence "The patient cannot walk without pain", the term "pain" will be wrongly negated because pain is present, but the word "walk" correctly will be negated. In this specific instance, double negation implies a positive relationship. We haven't yet addressed the scenario of two negatives when implying a positive relationship.

This thesis does not use dependency parsers directly to find dependencies within a sentence. In the next section, we will describe various algorithms for handling negation terms in clinical notes. The unstructured nature of clinical notes taken by health providers makes it challenging for typical NLP tools to operate effectively. These notes frequently contain spelling errors, abbreviations, and unfinished sentences.

2.2 Handling Negation terms in Clinical notes

There are different approaches for negation detection in biomedical and clinical notes. In general, they can be divided into the following three categories.

- **Rule-based approaches:** This technique was developed for the detection of negations in clinical contexts. For example, in [11] developed a NegEx system to identify the negation of specific findings and diseases in narrative medical reports, the present version of NegEx uses 272 rules and utilizes regular expressions to match. According to the results, this system has a 95.93%, a 93.27% precision rate, and a 97.73% accuracy rate. In a similar manner, Negfinder [33] can be used to find negated concepts within medical notes. This algorithm using regular expression initially detects negation terms in the sentence. Then, they use a parser that uses a single-token look-ahead strategy and sends these words to identify negation concepts. Based on reported results, this system has a 95.27% recall rate and 97.67% is its precision rate. Another system which is designed by [18], employs the same methodology to find negation concepts in electronic medical records. This system was developed to find linguistic clues for negation in 41 clinical records, the recall and precision of the system were reported to be 97.2% and 91.2%, respectively.

- **Supervised-machine learning approaches:** These approaches can also detect negations. The authors [5] created an algorithm to recognize negative context patterns in medical narratives. A negative context pattern is learned through information gained by the algorithm. In [17], researchers consider four negation (Conditional Random Fields(CRF), Random Forest(RF), Support Vector Machine with the linear kernel(SVM-linear), XGBoost(XGB)) feature detection approaches based on supervised learning and compare their performance.

- **Hybrid approaches:** In [23], the authors present a hybrid approach for detecting negations in clinical radiology reports by combining regular expression matching with grammatical parsing. They proposed a hybrid approach to categorize negations in radiology reports according to syntactic groups. The classifier was created by manually examining 30 radiology reports. The classifier was then validated using 470 radiological reports. Results from the evaluation of 120 radiology reports showed that recall and precision were both 92.6 % and 98.6%, respectively.

2.2.1 Handling Negation terms Using NegEx

In this project, we only utilize the Negex system to find negation concepts, and here we will explain in detail this system. The NegEx is a popular algorithm that uses regular expressions (also known as triggers) to find negations. This algorithm relies on four types of lexical cues:

- **pseudo_negations:** Phrases that are false triggers, ambiguous negations, or double negatives. For example, the sentence **<u>no abnormal</u> heart rythms** is a double negative.

- **preceding_negations:** negation phrases that precede an entity such as **patient <u>denies</u> pain** .

- **following_negations:** negation phrases that follow an entity like the senetence **zoloft <u>refused</u> by patient**.

- **termination**: conjunctions that are used to terminate the scope of a trigger break a sentence up into sections in order to identify negations(.e.g., "but") like the sentence **<u>no</u> fever <u>but</u> low blood sever headache**.

2.2.2 Keyword Extraction

As shown in figure 2.21, we use preprocessing algorithms to clean medical notes by removing punctuation and replacing abbreviations with their equivalent words. All medical notes have been saved in the local-fi-emr dataset. After that, negation terms were removed using the NegEx algorithm. For post-processing, we used NLP algorithms, such as removing stop words and lemmatizing. Then, we examine our extracted keywords in Alex's description columns to find their relevant deficits.

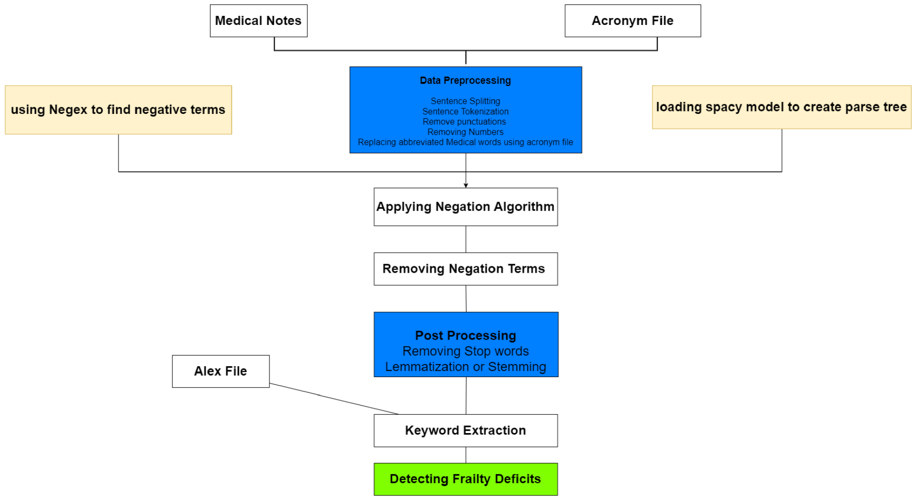

Figure 2.21: This figure illustrates the methodology used in this project for keyword extraction and eliminating negation concepts

We applied the steps in methodology to the local-fi-emr dataset. This dataset is entirely based on medical notes of the patients who have visited within the past 5 years. According to figure 2.22, respiratory disease, arthritis, Hypertension, ischaemic heart disease, and mobility and transfer problems are the most common frailty deficits. After that, all steps of the methodology are applied to the billing table. The billing dataset is a table in CPCSSN that contains diagnosis and disease information. According to figure 2.23, our model extracts frailty deficits from this table in this order: arthritis, hypertension, and social vulnerability as the most prevalent deficits. However, diseases like Activity limitations, Dyspnea, Falls, Housebound, Requirement for care, Mobility and transfer problems, peptic ulcer, and sleep disturbances are not seen among these samples.

The purpose of this project is to determine how much information can be extracted from medical notes using Natural Language Processing. According to frequency figures derived from medical notes and billing datasets, Natural Language Processing algorithms can extract much useful information from medical notes. Physicians can use this information to make better decisions about the patient's condition. Figure 2.24 shows the frequency of each deficit after combining the diagnosis column in the billing table with the NoteContent column in the local-fi-emr table.

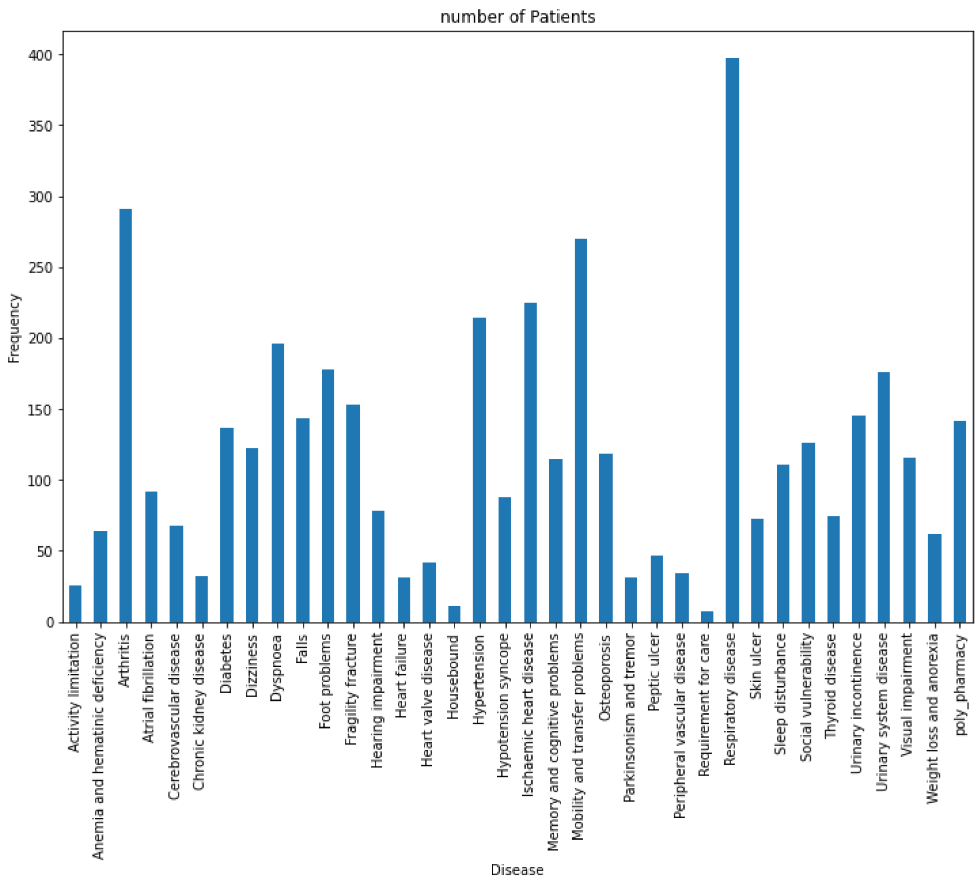

Figure 2.22: This figure shows the frequency of 36 deficits after applying Negex algorithm on the "local-fi-emr" dataset

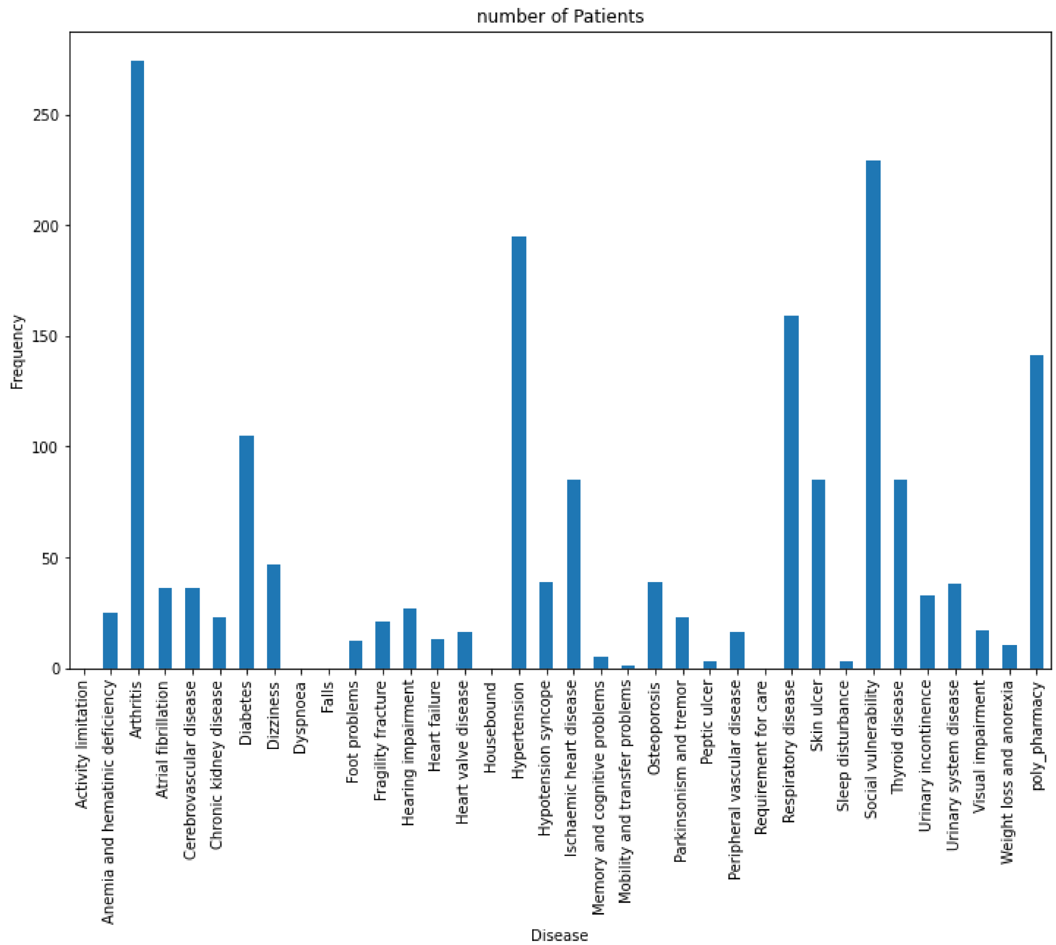

Figure 2.23: This figure shows the frequency of 36 deficits among 450 patients

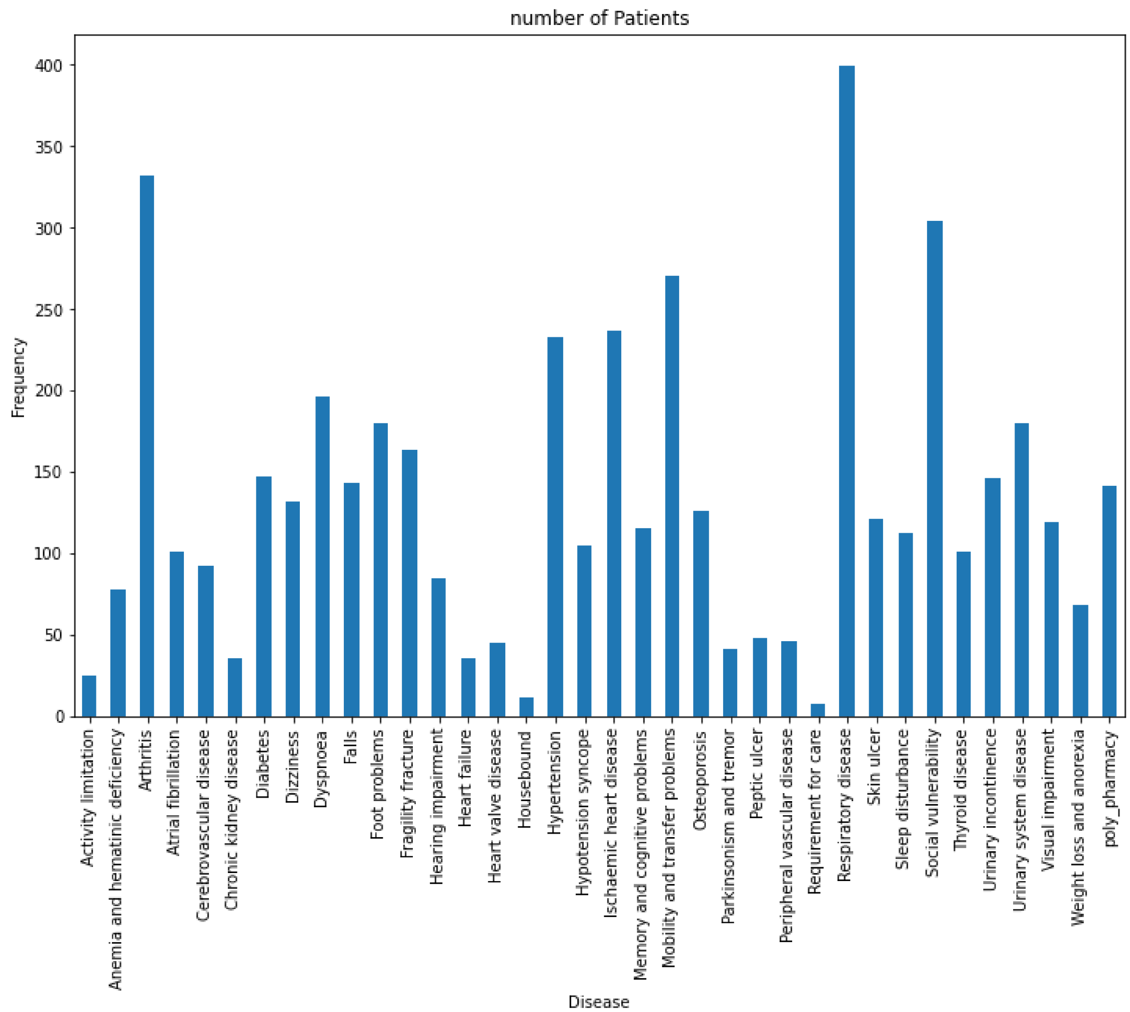

Figure 2.24: This figure shows the frequency of extracted disease after combining billing and local-fi-emr datasets

Chapter 3

Conclusion

3.1 Handling Imbalanced data set

We encounter the Imbalanced classification problem when our training data class distribution has a considerable skew. Even while the skew may not always be severe, we still consider imbalanced classification an issue because it can affect how well our machine-learning algorithms perform. This is important because we frequently show the most interest in the minority class. According to figure 2.11, the ratio of the majority to the minority class is 3:1, and the imbalanced proportions between the positive (frail) and negative (non-frail) classes show the data set is unbalanced.

There are various strategies to address imbalanced data, such as re-sampling [30] and cost-sensitive learning methods [27].

In this study, we used re-sampling techniques based on under- and over-sampling [31, 13, 12]. These techniques have the advantage of being classifier independent and being able to preprocess data, which can then be fed into any classifier as input. In order to balance the data, oversampling requires reproducing samples from the minority class. The drawback of oversampling is that it can result in an over-fitting issue because it duplicates the same instance and takes longer to execute than the under-sampling strategy. Therefore, it is advised when the dataset is relatively small in size. Another difficulty with oversampling is that it alters the class that we wish to identify, which may not be acceptable in some crucial real-time problems[1] since our goal was to find minority classes. By decreasing the number of samples from the majority class, under-sampling helps balance the unbalanced data. One drawback of the under-sampling strategy is that it can result in the loss of essential information or add bias to the data. According to studies, under-sampling tends to perform better than oversampling in particular contexts [48], while other studies show that oversampling outperforms under-sampling [34]. Under-sampling performs poorly in extremely small datasets, while oversampling performs worse in high-dimensional data[7]. Since we didn't have enough labeled data to work with, we used undersampling and oversampling to rebalance the sample distribution and compare results to show how well they performed. In this study, we only examined three different versions of near miss algorithm for undersampling. In addition, different types of SMOTE were used for oversampling, including borderline SMOTE and borderline SMOTE with SVM. Moreover, we compared the accuracy of predictive models based on random oversampling and undersampling and their combination.

3.1.1 UnderSampling Approaches

- **NearMiss undersampling**: This algorithm considers the distance between the majority and the minority class and then chooses data points based on this distance. There are three different versions of this algorithm: [52]

  - **version 1**: saves instances with a minimum average distance to the closest instances of the minority class.
  - **version 2**: selects rows with a minimum average distance to the furthest records of the minority class.
  - **version 3**: keeps examples from the majority class for each closest record in the minority class.

- **Random Undersampling**: In order to decrease the number of examples in the majority class in the data set, this method aims to randomly pick and eliminate samples from the majority class. Huge amounts of data are (possibly) discarded in random under-sampling. The loss of such data can make it more difficult to understand the decision border between the minority and majority instances, which can lead to a decline in classification performance.[32]

- **Condensed Nearest Neighbour**: This algorithm algorithm follows the nearest neighbor rule to undersample the majority of examples [21]. The first sample is recorded in the STORE set at the beginning, and in subsequent iterations, new majority samples are classified according to the nearest neighbor rule by referring to the STORE set. In order to correctly classify the instance, it will be stored either in the GRABBAG set or in the STORE set. Iteratively, this process moves until termination, when the GRABBAG set is disposed of.
- **AllKNN**: The AllKNN [44] data reduction algorithm works by formulating a KNN classification model. The nearest neighbors of any seed minority sample are identified for a particular value of k. If the majority of the neighbors classify the seed sample incorrectly then it is considered noise, and the sample is dropped from the training set.
- **ClusterCentroids** The cluster centroids-based undersampling strategy [28]performs based on a centroid that is calculated by using a K-Means clustering algorithm. The algorithm replaces a majority class cluster (containing majority samples) with the centroid of that cluster that is returned by the K-Means algorithm in the previous step.

3.1.2 Oversampling Approaches

The goal of oversampling is to increase the samples of minority classes to boost the sample sizes for the minority class, it can also repeat the instances. Although it balances the data, it doesn't give the classification model any new information. It is therefore required to synthesize new samples using the proper methodology [19].

- **SMOTE**: refers to Synthetic Minority Oversampling Technique is referred. First, SMOTE randomly selects a sample from the minority class. Then it identifies its k nearest neighbours of this sample from the minority class. Then, one of these k neighbours would be selected randomly and drawn a line between these two samples. Finally, a convex combination of these two instances is used to create new synthetic examples.

- **Borderline SMOTE**: Since examples that are on the border or far away are more likely to be misclassified. this SMOTE variant identifies the minority class instance that was incorrectly classified With the k-nearest neighbour (KNN) classifier.

- **Borderline SMOTE SVM**: This Smote Variant, instead of using KNN, chooses the Support Vector Machine (SVM) to identify instances that were incorrectly classified.

- Adaptive Synthetic Sampling (ADASYN): This method generates new samples based on the density of instances in the minority class. When the density of instances in the minority class is low, it generates more samples, while when it is high, it produces fewer samples. In general, there is an inverse relationship between the density of minority class and the generation of new samples, as shown in the table **??**.

- Random Over-Sampling: A single instance may be chosen more than once because this technique selects random instances from the minority class with replacement and augmenting the training data with multiple copies of this instance. Consequently, it is possible to choose the same instance more than once.

- Combination of Random under and Over Sampling: The idea is to apply oversampling and undersampling algorithms on a small number of samples in minority and majority classes, respectively. In some cases, combining the two random sampling techniques can lead to overall better performance than when the techniques are used alone.

3.2 Predictive Models

The machine learning approaches selected for this study are Logistic Regression(LR), Support Vector Machine(SVM), K-nearest neighbors, Decision Tree, Random Forest(RF), Multi-Layer Perceptron(MLP), and Gaussian Naive Bayesian. An overview of various learning algorithms is provided below.

- **SVM**: This algorithm is a powerful classifier that can distinguish between two classes and choose an effective decision boundary from a large amount of training data. SVM is a powerful classifier that can distinguish between two classes and choose an effective decision boundary from a large amount of training data. SVM has been used in several research to predict diseases [26, 36, 22, 35]

  Support vector machines (SVMs) with radial basis function (RBF) kernels are frequently utilized for automatic disease diagnosis because of their high accuracy. In this study, we applied SVM with the radial basis function kernel with various gamma values and regularization parameters for each classification problem.

- **MLP-ANN**: Artificial neural networks, or ANNs, are modelled based on the operation of biological neural networks. ANNs are dense networks of connected artificial neurons that activate in response to inputs. ANNs are analytical methods that have successfully addressed classification issues across various fields [50, 38, 16, 8, 49]. This study uses the multilayer perceptron neural network (MLP), one of the most popular ANN models. One input layer, one or more hidden layers, and one output layer make up the MLPNN. The input nodes in MLPNN pass values to the first hidden layer, and the first hidden layer nodes transmit values to the second layer, and so forth, until outputs are produced. The key MLPNN parameters—activation function, solver, hidden layer size, and learning rate—are defined for each classification task.

- **Logistic Regression**: The most popular and well-known binary classifier is LR. This model is a specific kind of multivariate regression [24]. This model often uses binary variables, which represent the presence or absence of an outcome or predictor variable. It forecasts the likelihood of an event occurring by feeding a dataset into a logit function.

- **Decision Tree**: This algorithm constructs classification models using a tree-like structure [4]. ID3, C4.5, and the classification and regression tree [43, 46], which construct DTs utilizing the idea of information entropy, are the key methods utilized in DTs. In our study, the DT is built using the classification tree approach, with hyperparameters adjusted for each problem. In this study, LR has been employed to differentiate between frail and non-frail patients.

- **Random Forest**: This method operates as an ensemble learning and consists of a large number of DTs. Random forests are used for classification tasks, and their output is the class selected by most trees. For the regression tasks, Random forest returns the average prediction of each tree. The RF algorithm is known for its ability to predict medical domains[51, 6, 10]. For each classification task, all of the RF hyperparameters have been adjusted, including the number of trees in the forest, the maximum number of features that can be used to split a node, and the maximum number of levels in each DT.

- **Naive Bayes**: Naive Bayes classifiers are probabilistic machine learning models. The Bayes theorem is used to construct this classifier. Naive Bayes classifiers assume that features are independent of each other [47, 25].

- **K-Nearest-Neighbor**: This is a non-parametric machine learning algorithm for supervised classification. The KNN has been used in a variety of medical datasets[39]

3.3 Performance Metrics

A confusion matrix is a performance metric for classification problems in machine learning when the output can be two or more classes. It is a table with four separate sets of actual and predicted values. Recall, precision, specificity, accuracy, and—most importantly—AUC-ROC curves may all be measured with great success with this method. To interpret results based on this table, we need to know four terms: TP, FP, TN, and FN. TP stands for true positive, which means you predicted positive, and it is true. A True Negative means you predicted a negative outcome, and it came true. A false positive, also called a Type I error, occurs when you predict something positive, but it is negative. False negative is also known as type II error when you predicted negative, but it is true. These are four essential measures to scale the classifier's performance. The higher rate for all measures is ideal.

- Recall: This measure indicates how many of the samples in positive classes are accurately anticipated.

$$Recall = \frac{TP}{TP + FN}$$

- Precision: explains how many of the samples which are predicted as positive actually are positive.

$$Precision = \frac{TP}{TP + FP}$$

- Accuracy: This measure demonstrates How many samples from all positive and negative classes are correctly predicted.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

- F-Score: There is a challenge in Comparing two models with high recall and low precision or vice versa. So, F-score helps facilitate the evaluation of both recall and precision simultaneously.

$$F_Score = \frac{2 * Recall * PRecision}{Recall + Precision}$$

The accuracy of the proposed models is based on 5-fold cross-validation after oversampling and undersampling.

3.4 Results

We consider 38 features as input variables; 36 deficits extracted from the datasets and sex as the categorical variable, and age as a numerical variable. This binary classification model classifies patients into frail and non frail classes. According to table 3.1, F1-Score for class zero is 0.87 and for class one is 0. The accuracy of the model is generally 0.74. The number of frail patients in group 1 is three times less than the number of patients in group 0. Consequently, The model's accuracy, precision, and recall work well for class zero and worse for class one. Due to these problems, we use sampling methods to balance the data. we divide our data into train and test sets and cross-validation set. Here, we define 5-fold cross validation technique to avoid overfitting. Our next step is to tackle imbalanced data sets by applying different resampling algorithms.

SVM Without Sampling				
Class	**precision**	**recall**	**f1-score**	support
0	0.74	1.00	0.85	65
1	0.00	0.00	0.00	23
accuracy	0.74	88		
macro avg	0.37	0.50	0.42	88
weighted avg	0.55	0.74	0.63	88

Table 3.1: This table shows the results of SVM without Sampling

Further, the results of Logistic Regression and SVM, KNN, Decision Tree, Random Forest, Naive Bayesian, and MLP are shown after applying different over- and undersampling algorithms. The purpose of this study is to find out how much-unstructured data can be helpful in classifying frailty. For this purpose, we compared the performance of the proposed model on structured and combined structured and unstructured datasets separately. According to table 3.1, the highest accuracy is achieved by the Decision tree and Random

Structured Dataset													
	Undersampling						Oversampling				Random sampling		
Model	NearMiss V1	NearMiss V2	NearMiss V3	ALLKNN	ClusterCentroid	Condensed Nearest Neighbor	SMOTE	Borderline SMOTE	Borderline SMOTE SVM	ADASYN	Random Under sampling	Random Over Sampling	Combination of Random Over and Under
Logistic Regression	**0.84**	0.64	0.63	0.84	0.79	0.62	0.78	0.78	0.82	0.75	0.82	0.80	0.77
SVM	0.76	0.66	0.66	0.81	0.63	0.60	0.66	0.68	0.75	0.66	0.76	0.72	0.74
KNN	0.79	0.79	0.71	0.82	0.66	0.62	0.80	0.76	0.83	0.80	0.79	0.79	0.79
Decision Tree	0.74	0.71	0.55	0.84	0.61	0.55	0.83	0.76	0.82	0.79	0.74	**0.88**	0.76
Random Forest	0.79	0.82	0.66	**0.88**	0.74	0.55	0.83	0.83	0.83	0.83	0.79	**0.89**	0.78
Naive Bayes	0.74	0.66	0.68	0.81	0.84	0.57	0.50	0.51	0.64	0.50	0.63	0.63	0.79
MLP	0.50	0.56	0.79	0.80	0.71	0.57	0.78	0.80	**0.84**	0.79	0.78	0.78	0.73

Figure 3.1: This table shows the accuracy of models after applying over- sampling and undersampling on structured datasets

forest after Random oversampling on structured datasets. For the final step, we combine and evaluate all deficits from the notes(local-fi-emr) table and billing table as the combination of structured and unstructured datasets. Based on table 3.2, it is shown that using the combination of oversampling and Random Forest results in an accuracy rate of 0.95%. In general, Random Forest with all oversampling algorithms has the highest accuracy. In addition, Random Forest with ALLKNN undersampling with an accuracy of 0.91 has the highest accuracy among all undersampling algorithms. The results show that the combination of them has better performance than only using structured data, and based on the results, Random Forest and Decision Tree outperformed.

Structured and Unstructured Datasets													
	Undersampling						Oversampling				Random sampling		
Model	NearMiss V1	NearMiss V2	NearMiss V3	ALLKNN	Cluster Centroid	Condensed Nearest Neighbor	SMOTE	Borderline SMOTE	Borderline SMOTE SVM	ADASYN	Random Under sampling	Random Over Sampling	Combinatio of Random Ov Under
Logistic Regression	0.76	0.61	0.63	0.81	0.68	0.69	0.76	0.75	0.85	0.79	0.74	0.76	0.80
SVM	0.84	0.53	0.74	0.77	0.74	0.67	0.69	0.67	0.74	0.57	0.82	0.70	0.77
KNN	0.84	0.71	0.74	0.86	0.79	0.62	0.75	0.84	0.84	0.78	0.76	0.82	0.82
Decision Tree	0.66	0.63	0.63	0.78	0.58	0.57	0.79	0.82	0.81	0.86	0.71	**0.94**	**0.91**
Random Forest	0.84	0.66	0.76	**0.91**	0.68	0.69	**0.90**	**0.89**	**0.92**	**0.90**	0.74	**0.94**	**0.95**
Naive Bayes	0.79	0.66	0.61	0.88	0.79	.64	0.55	0.56	0.68	0.53	0.71	0.75	0.85
MLP	0.80	0.66	0.66	0.89	0.74	0.74	0.78	0.76	0.83	0.77	0.74	0.78	0.81

Figure 3.2: This table shows the accuracy of models after applying over-sampling and undersampling on structured and unstructured datasets

3.5 Conclusion and Summary

This study analyzes structured and unstructured data from the electronic health record to identify frailty. Despite advancements in both natural language processing and machine learning, extracting medically relevant text remains a challenging task. In the preceding sections, we outlined the methodologies considered for this task and followed by the approach we identified as the most promising. The NLP techniques for extracting main keywords or medical concepts and groupings of relevant description tokens have been investigated in great detail. Our proposed algorithms were used to extract a large volume of unstructured text tokens related to frailty. In addition to this effort, we performed automatic negation detection on medical text. By doing this, it was possible to distinguish between keywords and descriptions that were affirmative observations and those that were ruled out. Due to the fact that frailty in the elderly is still a relatively new field of research and development, many of the standard methods like using billing codes fail to properly support this diagnosis in Alberta. Over the past several decades, researchers have actively worked to establish a standard for frailty diagnosis and what factors contribute to frailty. These results are consistent with the initial analysis of the data extracted from the notes of the patients in our database. We observed and recovered some expected results and indications of frailty based on preliminary results identified from some basic statistics. For example, a higher frailty population is associated with an increased risk of falls, polypharmacy, Dyspnoea, Foot Problems, Mobility and transfer problems, and Ischaemic Heart Disease. We constructed our models using CPCSSN automated billing codes and local-fi-emr datasets to extract frailty-relevant concepts. Our proposed algorithms were used to extract a large volume of deficits related to frailty. We performed a negation detection algorithm on medical text to remove negated terms then, we created several classifiers which can predict the risk of frailty using extracted text data and the additional numeric and categorical information in the database. In addition, We resolved the imbalanced nature of the data through resampling techniques. We combined structured and unstructured data to find frailty deficits and our results show that the combination of them have better performance than only using structured data and the results show that Random Forest and Decision Tree outperformed.

Our algorithms are all available on a private GitHub repository. The repository will only be accessible to members of Cybera. Additionally, this repository contains instructions on running all algorithms discussed in this report and documentation in the code.

Bibliography

[1] Adnan Amin, Sajid Anwar, Awais Adnan, Muhammad Nawaz, Newton Howard, Junaid Qadir, Ahmad Hawalah, and Amir Hussain. Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study. *IEEE Access*, 4:7940–7957, 2016.

[2] Laura J Anzaldi, Ashwini Davison, Cynthia M Boyd, Bruce Leff, and Hadi Kharrazi. Comparing clinician descriptions of frailty and geriatric syndromes using electronic health records: a retrospective cohort study. *BMC geriatrics*, 17(1):1–7, 2017.

[3] Sylvia Aponte-Hao, Sabrina T Wong, Manpreet Thandi, Paul Ronksley, Kerry McBrien, Joon Lee, Mathew Grandy, Alan Katz, Dee Mangin, Alexander Singer, et al. Machine learning for identification of frailty in canadian primary care practices. *International journal of population data science*, 6(1), 2021.

[4] Stephane Aris-Brosou, James Kim, Li Li, Hui Liu, et al. Predicting the reasons of customer complaints: a first step toward anticipating quality issues of in vitro diagnostics assays with machine learning. *JMIR medical informatics*, 6(2):e9960, 2018.

[5] Mordechai Averbuch, Tom H Karson, Benjamin Ben-Ami, Oded Maimon, and Lior Rokach. Context-sensitive medical information retrieval. In *MEDINFO 2004*, pages 282–286. IOS Press, 2004.

[6] Juan José Beunza Nuin, Enrique Puertas Sanz, Ester García Ovejero, Gema Villalba, Emilia Condés Moreno, Gergana Koleva, Cristian Hurtado, Manuel Landecho, et al. Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease). 2019.

[7] Rok Blagus and Lara Lusa. Smote for high-dimensional class-imbalanced data. *BMC bioinformatics*, 14(1):1–16, 2013.

[8] Richard Bradley, Ilias Tagkopoulos, Minseung Kim, Yiannis Kokkinos, Theodoros Panagiotakos, James Kennedy, Geert De Meyer, Phillip Watson, and Jonathan Elliott. Predicting early risk of chronic kidney disease in cats using routine clinical laboratory tests and machine learning. *Journal of veterinary internal medicine*, 33(6):2644–2656, 2019.

[9] R Scott Braithwaite, David Fiellin, and Amy C Justice. The payoff time: a flexible framework to help clinicians decide when patients with comorbid disease are not likely to benefit from practice guidelines. *Medical care*, pages 610–617, 2009.

[10] Theodora S Brisimi, Tingting Xu, Taiyao Wang, Wuyang Dai, William G Adams, and Ioannis Ch Paschalidis. Predicting chronic disease hospitalizations from electronic health records: an interpretable classification approach. *Proceedings of the IEEE*, 106(4):690–707, 2018.

[11] Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310, 2001.

[12] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[13] Yu-Meei Chyi. Classification analysis techniques for skewed class distribution problems. *Department of Information Management, National Sun Yat-Sen University*, 2003.

[14] Rose M Collard, Han Boter, Robert A Schoevers, and Richard C Oude Voshaar. Prevalence of frailty in community-dwelling older persons: a systematic review. *Journal of the american geriatrics society*, 60(8):1487–1492, 2012.

[15] Tracy A Comans, Nancye M Peel, Ruth E Hubbard, Andrew D Mulligan, Leonard C Gray, and Paul A Scuffham. The increase in healthcare costs associated with frailty in older people discharged to a post-acute transition care program. *Age and ageing*, 45(2):317–320, 2016.

[16] Looket Dihge, Mattias Ohlsson, Patrik Edén, Pär-Ola Bendahl, and Lisa Rydén. Artificial neural network models to predict nodal status in clinically node-negative breast cancer. *BMC cancer*, 19(1):1–12, 2019.

[17] Lluıs Domınguez-Mas, Francesco Ronzano, and Laura I Furlong. Supervised learning approaches to detect negation cues in spanish reviews. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019), CEUR Workshop Proceedings, Bilbao, Spain*, 2019.

[18] PL Elkin, SH Brown, and B Bauer. a, husser, cs, carruth, w., bergstrom, lr, & wahner-roedler, dl (2005). a controlled trial of automated classification of negation from clinical notes. *BMC medical informatics and decision making*, 5(1):13.

[19] Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C Prati, Bartosz Krawczyk, and Francisco Herrera. *Learning from imbalanced data sets*, volume 10. Springer, 2018.

[20] Stephanie Garies, Richard Birtwhistle, Neil Drummond, John Queenan, and Tyler Williamson. Data resource profile: national electronic medical record data from the canadian primary care sentinel surveillance network (cpcssn). *International Journal of Epidemiology*, 46(4):1091–1092f, 2017.

[21] Peter Hart. The condensed nearest neighbor rule (corresp.). *IEEE transactions on information theory*, 14(3):515–516, 1968.

[22] Min-Wei Huang, Chih-Wen Chen, Wei-Chao Lin, Shih-Wen Ke, and Chih-Fong Tsai. Svm and svm ensembles in breast cancer prediction. *PloS one*, 12(1):e0161501, 2017.

[23] Yang Huang and Henry J Lowe. A novel hybrid approach to automated negation detection in clinical radiology reports. *Journal of the American medical informatics association*, 14(3):304–311, 2007.

[24] Sookyung Hyun, Susan Moffatt-Bruce, Cheryl Cooper, Brenda Hixon, Pacharmon Kaewprag, et al. Prediction model for hospital-acquired pressure ulcer development: retrospective cohort study. *JMIR medical informatics*, 7(3):e13785, 2019.

[25] Sukhminder Kaur, Azween Abdullah, Noran Naqiah Hairi, and Siva Kumar Sivanesan. Logistic regression modeling to predict sarcopenia frailty among aging adults. *Int. J. Adv. Comput. Sci. Appl.*, 12(8), 2021.

[26] Ioannis Kavakiotis. Tsave o salifoglou a maglaveras n vlahavas i chouvarda i. *Comput. Struct. Biotechnol. J*, 15:104–116, 2017.

[27] Salman H Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous A Sohel, and Roberto Togneri. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE transactions on neural networks and learning systems*, 29(8):3573–3587, 2017.

[28] Hyun-Jung Kim, Nam-Ok Jo, and Kyung-Shik Shin. Optimization of cluster-based evolutionary undersampling for the artificial neural networks in corporate bankruptcy prediction. *Expert systems with applications*, 59:226–234, 2016.

[29] Alethea Lacas and Kenneth Rockwood. Frailty in primary care: a review of its conceptualization and implications for practice. *BMC medicine*, 10(1):1–9, 2012.

[30] Paul H Lee. Resampling methods improve the predictive power of modeling in class-imbalanced datasets. *International journal of environmental research and public health*, 11(9):9776–9789, 2014.

[31] Wu Liu. Zhou, 2009 liu x.-y., wu j., zhou z.-h. *Exploratory undersampling for class-imbalance learning, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550, 2009.

[32] Yunqian Ma and Haibo He. Imbalanced learning: foundations, algorithms, and applications. 2013.

[33] Pradeep G Mutalik, Aniruddha Deshpande, and Prakash M Nadkarni. Use of general-purpose negation detection to augment concept indexing of medical documents: a quantitative study using the umls. *Journal of the American Medical Informatics Association*, 8(6):598–609, 2001.

[34] Mehdi Naseriparsa and Mohammad Mansour Riahi Kashani. Combination of pca with smote resampling to boost the prediction rate in lung cancer dataset. *arXiv preprint arXiv:1403.1949*, 2014.

[35] Laurence O'Dwyer, Franck Lamberton, Arun LW Bokde, Michael Ewers, Yetunde O Faluyi, Colby Tanner, Bernard Mazoyer, Desmond O'Neill, Máiréad Bartley, D Rónán Collins, et al. Using support vector machines with multiple indices of diffusion for automated classification of mild cognitive impairment. *PloS one*, 7(2):e32441, 2012.

[36] Liyan Pan, Guangjian Liu, Xiaojian Mao, Huixian Li, Jiexin Zhang, Huiying Liang, Xiuzhen Li, et al. Development of prediction models using machine learning algorithms for girls with suspected central precocious

puberty: retrospective study. *JMIR medical informatics*, 7(1):e11728, 2019.

[37] Mohammad Habibullah Pulok, Olga Theou, Alexandra M van der Valk, and Kenneth Rockwood. The role of illness acuity on the association between frailty and mortality in emergency department patients referred to internal medicine. *Age and ageing*, 49(6):1071–1079, 2020.

[38] Firdani Rianda Putra, Aldilas Achmad Nursetyo, Saurabh Singh Thakur, Ram Babu Roy, Shabbir Syed-Abdul, Shwetambara Malwade, and Yu-Chuan Jack Li. Prediction of clinical events in hemodialysis patients using an artificial neural network. In *MEDINFO 2019: Health and Wellbeing e-Networks for All*, pages 1570–1571. IOS Press, 2019.

[39] Ben Rahman, Harco Leslie Hendric Spits Warnars, Boy Subirosa Sabarguna, and Widodo Budiharto. Heart disease classification model using k-nearest neighbor algorithm. In *2021 Sixth International Conference on Informatics and Computing (ICIC)*, pages 1–4. IEEE, 2021.

[40] Kenneth Rockwood, Xiaowei Song, Chris MacKnight, Howard Bergman, David B Hogan, Ian McDowell, and Arnold Mitnitski. A global clinical measure of fitness and frailty in elderly people. *Cmaj*, 173(5):489–495, 2005.

[41] Kenneth Rockwood and Olga Theou. Using the clinical frailty scale in allocating scarce health care resources. *Canadian Geriatrics Journal*, 23(3):210, 2020.

[42] Adane Tarekegn, Fulvio Ricceri, Giuseppe Costa, Elisa Ferracin, Mario Giacobini, et al. Predictive modeling for frailty conditions in elderly people: machine learning approaches. *JMIR medical informatics*, 8(6):e16678, 2020.

[43] Olga Theou, Mario Ulises Pérez-Zepeda, Alexandra M van der Valk, Samuel D Searle, Susan E Howlett, and Kenneth Rockwood. A classification tree to assist with routine scoring of the clinical frailty scale. *Age and ageing*, 50(4):1406–1411, 2021.

[44] Ivan Tomek. An experiment with the edited nearest-nieghbor rule. 1976.

[45] John Travers, Roman Romero-Ortuno, Jade Bailey, and Marie-Therese Cooney. Delaying and reversing frailty: a systematic review of primary care interventions. *British Journal of General Practice*, 69(678):e61–e69, 2019.

[46] Jin Tsuchihashi, Shunji Koya, Keisuke Hirota, Noboru Koga, Hayato Narao, Manabu Tomita, Takumi Kawaguchi, Ryuki Hashida, Dan Nakano, Tsubasa Tsutsumi, et al. Effects of in-hospital exercise on frailty in patients with hepatocellular carcinoma. *Cancers*, 13(2):194, 2021.

[47] Tjeerd van der Ploeg, Robbert JJ Gobbens, and Benissa E Salem. Bayesian techniques in predicting frailty among community-dwelling older adults in the netherlands. *Archives of gerontology and geriatrics*, 105:104836, 2023.

[48] Byron C Wallace, Kevin Small, Carla E Brodley, and Thomas A Trikalinos. Class imbalance, redux. In *2011 IEEE 11th international conference on data mining*, pages 754–763. Ieee, 2011.

[49] Ben Wellner, Joan Grand, Elizabeth Canzone, Matt Coarr, Patrick W Brady, Jeffrey Simmons, Eric Kirkendall, Nathan Dean, Monica Kleinman, Peter Sylvester, et al. Predicting unplanned transfers to the intensive care unit: a machine learning approach leveraging diverse clinical elements. *JMIR medical informatics*, 5(4):e8680, 2017.

[50] Ashenafi Zebene Woldaregay, Eirik Årsand, Ståle Walderhaug, David Albers, Lena Mamykina, Taxiarchis Botsis, and Gunnar Hartvigsen. Data-driven modeling and prediction of blood glucose dynamics: Machine learning applications in type 1 diabetes. *Artificial intelligence in medicine*, 98:109–134, 2019.

[51] Jiangpeng Wu, Xiangyi Zan, Liping Gao, Jianhong Zhao, Jing Fan, Hengxue Shi, Yixin Wan, E Yu, Shuyan Li, Xiaodong Xie, et al. A machine learning method for identifying lung cancer based on routine blood indices: qualitative feasibility study. *JMIR medical informatics*, 7(3):e13476, 2019.

[52] Show-Jane Yen and Yue-Shi Lee. Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset. In *Intelligent Control and Automation*, pages 731–740. Springer, 2006.