# Abuse Detection in Medical Claims Using NLP and Deep Learning Techniques

A Thesis

Presented in Partial Fulfillment of the Requirements for the

Degree of Master of Science

with a

Major in Computer Science

in the

College of Graduate Studies

University of Idaho

By

Bushra Alkomah

Approved by:

Major Professor: Frederick Sheldon, PhD.

Committee Members: Ananth Jillepalli, Ph.D.; Jia Song, Ph.D.

Department Administrator: Terry Soule, Ph.D.

May 2022

# Abstract

Our research describes effective data mining based on the Named Entity Recognition (NER) technique for medical claims fraud / abuse detection system. Fraud and abuse in medical claims have become a major concern among health insurance companies in Saudi Arabia in recent years, not only by faking prices and numbers, but by assigning inaccurate ICD-10 codes to diseases that do not match the diagnosis of the claim. Handling medical claims is backbreaking manual work performed by a few medical experts who are responsible for approving, modifying, or denying applications for grants within a limited period of time from receipt. The proposed screening system uses a NER detection tool for each of the claims involved and classifies whether it is a fraud or not. Our research study has carried out an analysis for medical claim data from a private insurance company collected from different databases. Results from our fraud detection system show clearly that the number of suspicions of abuse is excessively high compared to the number of no suspicion of abuse.

# Acknowledgment

At first, thanks to God, who facilitated the difficulties and granted me to complete postgraduate studies in the Department of Computer Science.

I would also like to convey my appreciation to my supervisor, Frederick Sheldon, who gave me a chance to do this incredible project and advised me throughout this project. His recommendations and guidance enlightened my mind. I am sincerely grateful and appreciative of his efforts and time spent on my development.

Side to my advisor, I would like to thank my committee members, Doctor Ananth Jillepalli and Doctor Jia Song, for accepting them to participate in the success of my research. Thanks to you, doctors.

Finally, I would like to extend a special thanks to my parents, Saeed and Hessa Al Qahtani, for their support and love.

## Dedication

*Dedicated to my family for their faith and their advice, bless them. Mainly devoted to my father and sister Al hanouf, you are my life, happiness, and everything. May Allah keep you all safe and happy.*

# Table of Content

# List of Tables

# List of Figures

# CHAPTER 1: INTRODUCTION

## 1.1 Overview

Natural Language Processing (NLP) is a subdomain of Machine Learning (ML) that deals with extracting patterns from human language terms and texts to learn tasks that humans can perform independently, such as text categorization [1], emotion classification [2], and sentiment analysis [3].

With the increase of data in all industries, Natural Language Processing is often used for tasks that make people's lives easier. Natural Language Processing is also used in the medical field as it provides the ability to search, analyse, and learn patterns in patient records. Natural language processing with machine learning offers incredible insights into quality of understanding, improved methods, and better outcomes for patients, including improving clinical documentation [4], supporting clinical decisions [5], and medical claims [6]. Since there are huge amounts of unstructured and unlabelled text in the medical field, trained models created using libraries such as spacy, nltk are very useful for extracting information and labelling the dataset without training procedures.

Medical claims data, also called administrative data, are electronic records of information about medical appointments on a larger scale. Claims data is collected at appointments where the records are provided by doctors. This is very important to ensure the reliability of the records. Since the records consist of diagnoses, diseases, prescriptions, and dosages of medications, researchers can examine the data to create various analyses, such as medical conditions and rare diseases. With the popularity of Natural Language Processing [6] tasks in the medical field, studies using medical data are also increasing. In medical data, doctors also enter ICD-10 codes for diagnosing diseases, which are the 10th ICD is the abbreviation for International Statistical Classification of Diseases and Related Health Problems. It enables the systematic collection, analysis, and interpretation of medical claims independent of countries and regions. It also ensures semantic interoperability and reusability of the data collected for various use cases beyond just health statistics, including decision support, resource allocation, reimbursement, guidelines, and more. There are studies that have been conducted using ICD-10 codes to identify and characterize diseases. In medical claims, doctors enter the diagnosis using ICD-10 codes, which may be incorrect due to human error. Therefore, automatic ICD-10 coding is also one of the current topics in the medical field using rule-based methods, KNN machine learning methods, and deep learning methods.

In this study, we propose a rule-based method combining the trainee model with defined rules to find data leakage in medical services. We use the spacy tool [7] to extract named entities in the data, which are defined as keywords to categorize ICD-10 codes. We compare the ICD-10 codes we find to the ICD-10 codes in the dataset. If they do not match, we classify the dataset as misuse.

## 1.2 Problem Statement

Different data mining strategies were used in the field to assist stakeholders in medical claims in decision making. Financial statement fraud is a troubling issue for both insurance organizations and community regulators. Data leak in medical claims is potentially considered as a financial fraud, because it leads at the end to refund inappropriate amounts of money by insurance. Mistakes made by doctors during the claim not only produce fake financial data with the intention of inflicting heavy financial losses on a large scale, but also lead to a downturn in the financial outcomes. During a general audit, regular audit exercises can focus primarily on fact-checking. In medical claims, most fraud detection research limits its investigation to statistics that are very numerical. From real data, fake wrong disease code that does not match the diagnosis description with correct prices in financial data can hardly detect the data leak case. It can be difficult to assess the general lack of scope for reporting medical claim fraud, in the way it occurs and the manner in which it is typically characterized by industry-conscious people who market their fraud.

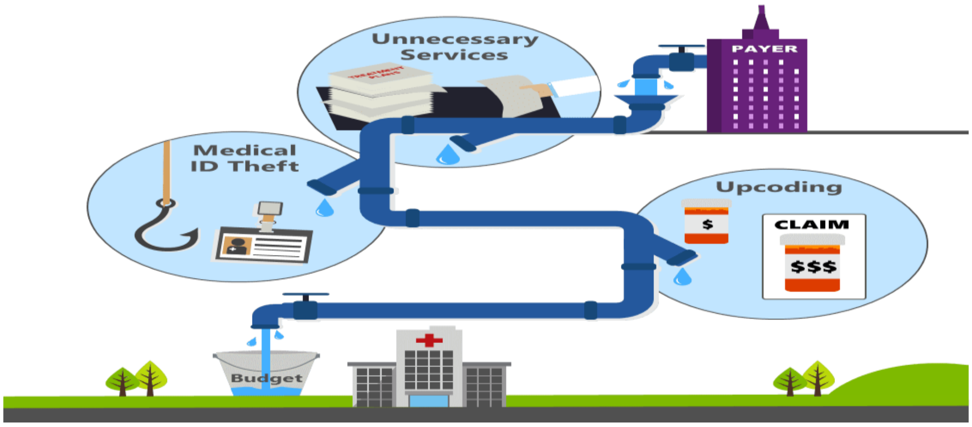


*Figure 1:Problem statement in medical claim (simplified) [13]*

## 1.3 Goals and Objectives

The goals and objectives of this study will help to build an effective solution to the problem statement.

AIM:

❖ **Identify the wrong assigned disease code “ICD10” based on NER (Named Entity Recognition) and Named Entity Linking.**

The research aims to detect the value of the wrong ICD10 in the claim based on the given diagnosis description. This smart detection can assign a characteristic to the claim [Fraudulent, Non- Fraudulent]

## 1.4 Objectives

The objective of this research would help in fulfilling the aim that is mentioned.

- ➔ Find reliable insurance companies to collect good data eligible for this research.
- ➔ Prepare Machine Learning Model (NER) to classify our Data.
- ➔ Evaluate the models.
- ➔ Classify the claims into suspicious/non-suspicious.

- ➔ Summarize the challenges and limitations.

## 1.5 Dissertation Organization

The proposed study focuses on the importance of detecting fraud in medical claims using Machine Learning and Named Entity Recognition. The dissertation involves the research, the data collection and development of the Python code based on the multiple analysis performed of the model. The structure of the thesis is as follows.

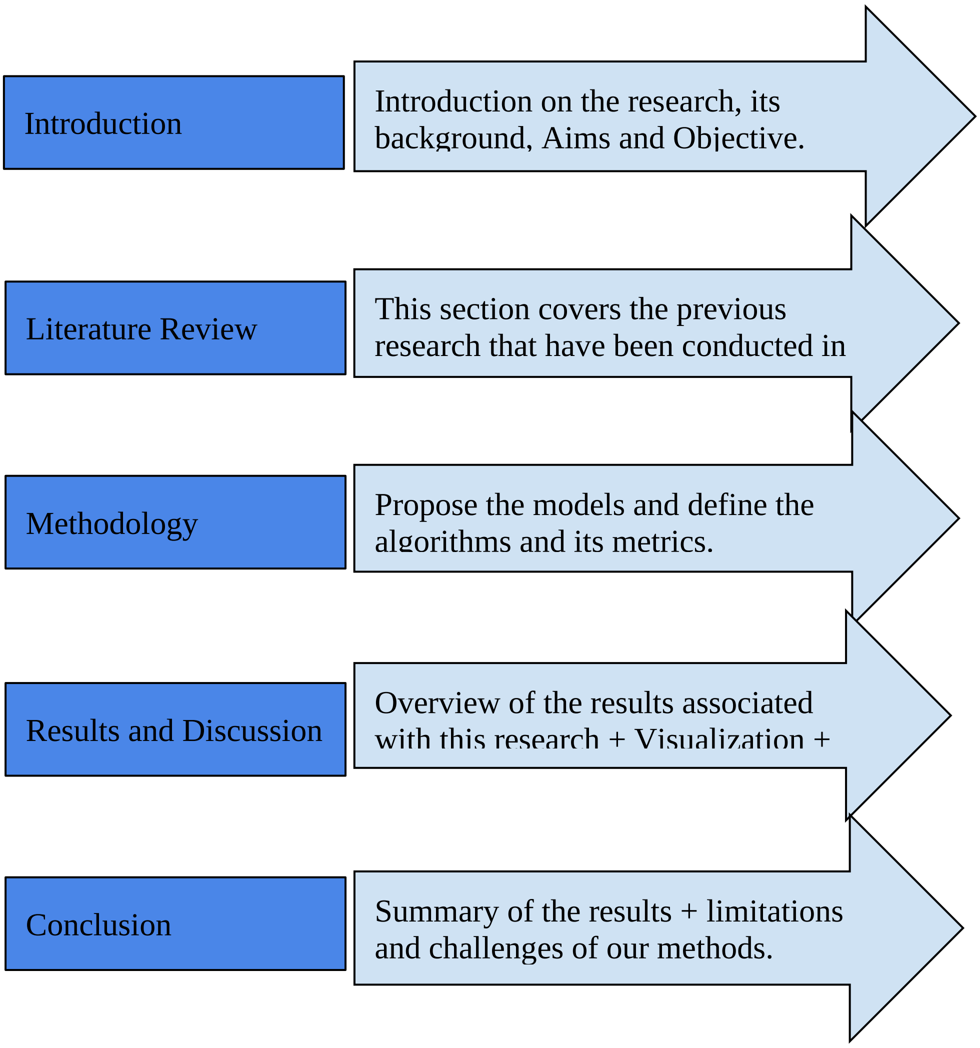

Introduction
Introduction on the research, its background, Aims and Objective.
Literature Review
This section covers the previous research that have been conducted in
Methodology
Propose the models and define the algorithms and its metrics.
Results and Discussion
Overview of the results associated with this research + Visualization +
Conclusion
Summary of the results + limitations and challenges of our methods.

*Figure 2:Dissertation Organization*

# CHAPTER 2: BACKGROUND

## 2.1 Classification of Diseases and Related Health Problems (ICD)

The ICD10 code which is overseen by the World Health Organization (WHO) [6], gives an all-inclusive way to deal with mortality coding that in excess of 100 nations use. Its complete name is the International Statistical Classification of Diseases and Related Health Problems, and as its name shows, it has been created to allow the interpretation of mortality and dismalness data into a normalized factual organization [6]. The ICD is presently in the fifth release of its tenth amendment, and this is the rendition as of now suggested for use by nations. Progressing updates to the ICD include remedies and handy solutions to distinguished issues, the inclusion of additional subtleties, the adding of new terms and the harmonization of content. Combined updates then, at that point, lead to the giving of another version. Interestingly, the total modification of the ICD is a significant cycle that includes primary changes and the presentation of new parts – for instance, the continuous amendment of ICD-10 to ICD-11 [6]. The reason for the ICD is to permit the orderly recording, investigation, understanding and correlation of mortality and grimness information. The ICD is utilized to interpret composed analyses of diseases and other medical conditions into alphanumeric codes in an interaction known as clinical coding. Once coded, reasons for death can be aggregated and measurements delivered for capacity, recovery and investigation.

## 2.2 Named Entity Recognition (NER)

Medical named entity recognition (NER) is a region where clinical named elements are perceived from clinical texts, like sicknesses, drugs, medical procedure reports, physical parts, and assessment archives [8]. Clinical named entity recognition (NER) is a significant procedure that has as of late got consideration in clinical networks in removing named elements from clinical texts, like illnesses, drugs, medical procedure reports, physical parts, and assessment records. electronic wellbeing records, clinical preliminaries are having diverse organization dependent on the nation’s guidelines.

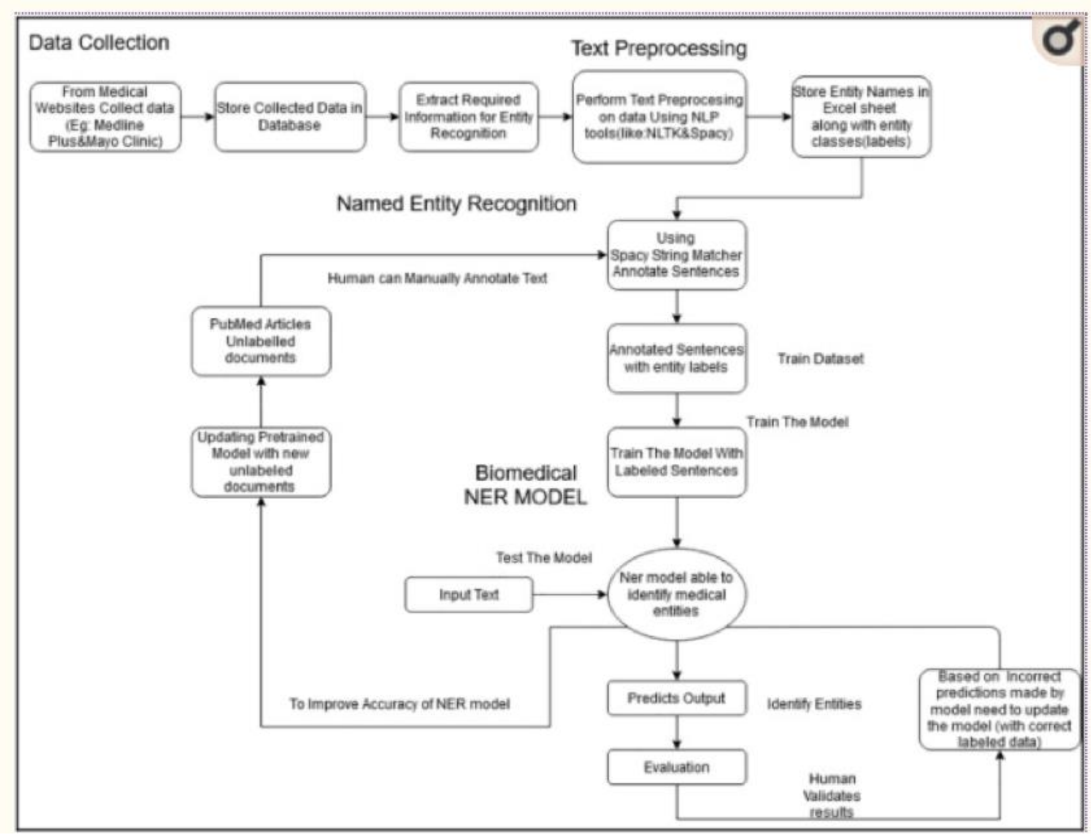

Data Collection
Text Preprocessing
From Medical Websites Collect data (Eg: Medline Plus&Mayo Clinic)
Store Collected Data in Database
Extract Required Information for Entity Recognition
Perform Text Preprocesing on data Using NLP tools(like:NLTK&Spacy)
Store Entity Names in Excel sheet along with entity classes(labels)
Named Entity Recognition
Using Spacy String Matcher Annotate Sentences
Human can Manually Annotate Text
PubMed Articles Unlabelled documents
Annotated Sentences with entity labels
Train Dataset
Train The Model
Updating Pretrained Model with new unlabeled documents
Biomedical NER MODEL
Train The Model With Labeled Sentences
Test The Model
Input Text
Ner model able to identify medical entities
To Improve Accuracy of NER model
Predicts Output
Identify Entities
Based on Incorrect predictions made by model need to update the model (with correct labeled data)
Evaluation
Human Validates results

*Figure 3:Biomedical NER Model [40]*

Such extraction can prompt critical reserve funds of difficult work and limit the time taken to get another medication to advertise. Late progressions in AI exploit the enormous text corpora accessible in logical writing just as clinical and drug sites and train frameworks for long-time errands going from text mining to responding to questions [8]. One of the critical difficulties in preparing NLP-based models is the accessibility of sensible estimated, excellent explained datasets. Further, in an average modern setting, the general trouble in earning critical time from space specialists and the absence of apparatuses and methods for successful explanation, alongside the capacity to audit such comments to limit human blunders, influences exploration and benchmarking of new learning procedures and calculations [8]

### 2.2.1 NER Algorithms

Named entity recognition (NER) is an important part of many natural language processing (NLP) technologies such as community detection, information extraction, information retrieval, and many other fields. This section is based on Yonghui Wu and others previous work [8] which describes the development of the AL-CRF model, which is an NER approach based on active learning (AL).

### 2.2.2 CRF Model

A CRF model is a model that expresses the conditional probability of a random variable Y with a random variable X. This model has different shapes including linear chain shape, matrix shape etc. In the NER process, the CRF model is generally simplified, meaning that the random variables have the same graph structure as shown in Figure 1. The function of the CRF model is to predict the conditional probability of Y by training the parameters of the model, and the method of calculation is shown in Equation below [9]:

$$P(y \mid x) = \frac{1}{Z(x)} \exp\left( \sum_{i,\mathrm{k}} \lambda_{\mathrm{k}} t_{\mathrm{k}} (y_{i-1}, y_i, x, i) + \sum_{i,\mathrm{l}} \mu_{\mathrm{l}} s_{\mathrm{l}} (y_i, x, i) \right), \qquad (1)$$

where $Z_X$ is a normalization factor, whereas tk is an eigenfunction defined on the edge k, which is called transfer feature. It depends on the current position and on the previous position.

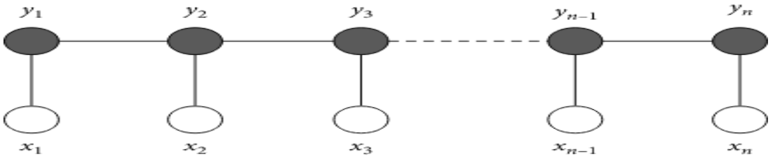


*Figure 4:The structure of CRF model [9].*

### 2.2.3 AL-CRF Model

The AL-CRF model takes active learning as the infrastructure and CRF as the basic classifier of AL. A hierarchical sampling method based on the grouping of k-means is used in the paper of Han Huang and others [9] to select the training set for initial training. An information entropy based SSS is adopted to select samples for an iterative process. The condition to stop recurrence is based on a defined rate of change of the F value. The structure of the algorithm is shown in Figure below.

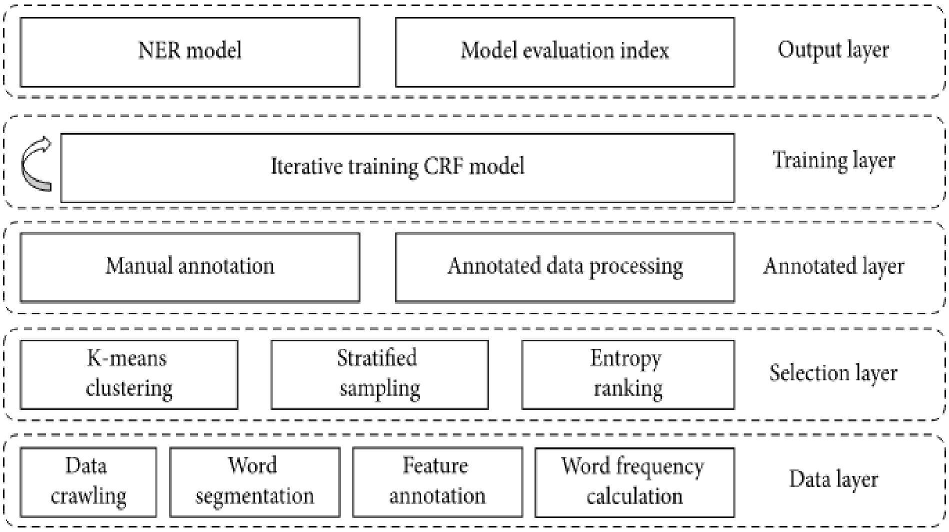


*Figure 5:The algorithm framework of AL-CRF [9].*

.

## 2.3 Machine Learning Algorithms

We initially consider the conditions wherein machine learning might be particularly valuable as for two components of a wellbeing result:

- The attributes of its analytic models.
- The configuration wherein its indicative information is typically put away inside EHR frameworks [10]. In the principal aspect.

This previous study recommends that for wellbeing results with demonstrative rules including numerous clinical variables, ambiguous definitions, or emotional understandings, machine learning might be valuable for displaying the complex indicative dynamic cycle from a vector of clinical contributions to recognize people with the wellbeing result. In the subsequent aspect, we suggest that for wellbeing results where demonstrative data is generally put away in unstructured configurations like free text or pictures, machine learning might be helpful for separating and organizing this data as a feature of a characteristic language handling framework or a picture acknowledgment task.

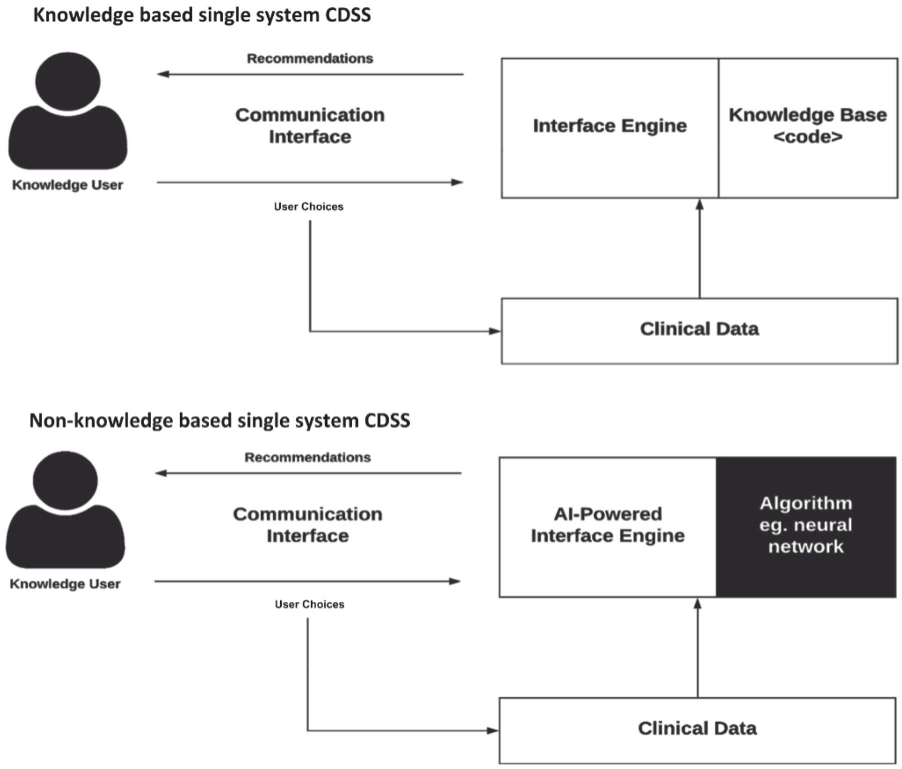


*Figure 6:Clinical decision support system [11].*

We then, at that point, consider these two aspects together to characterize four normal situations of wellbeing results. For every situation, we talk about the likely uses for machine learning – first expecting precise and complete EHR information and afterward loosening up these suppositions to oblige the restrictions of true EHR frameworks [10]. We delineate these four situations utilizing substantial models and portray how late investigations have utilized machine learning to distinguish these wellbeing results from EHR information. Machine learning can possibly work on the precision and productivity of wellbeing resulting in recognizable proof from EHR frameworks, particularly under specific conditions [10]. To advance the use of machine learning in EHR-based phenotyping errands, future work ought to focus on endeavours to build the mobility of machine learning calculations for use in multi-site settings. Wellbeing results examination could profit from the utilization of more adaptable, information driven machine learning techniques, as a considerable lot of the assignments being handled utilizing these methodologies in different areas bear striking similarities to the difficulties specialists face when utilizing EHR information to recognize wellbeing results. Hence, the reason for this survey is to introduce a bunch of normal situations that scientists might view as accommodating for pondering when and for what errands machine learning might be valuable for distinguishing wellbeing results from EHR information. We start by giving a concise outline of a few machine learning strategies that have been regularly used to gauge wellbeing results from electronic wellbeing information. Then, we think about two elements of a wellbeing result and recognize the conditions in each aspect where machine learning might be particularly valuable. We then, at that point, consider these two aspects mutually to make four normal situations of wellbeing results and examine the likely uses for machine learning in every situation – first accepting exact and complete EHR information and afterward loosening up these suppositions to oblige the restrictions of genuine EHR frameworks. We represent these four situations utilizing substantial models and depict how late examinations have utilized machine learning to distinguish these wellbeing results from EHR information. In the entirety of our conversations, we allude to machine learning as the utilization of calculations that exist far along the machine-driven finish of the 'machine learning range.

## 2.4 Evaluation Metric

### 2.4.1 Overview

Model assessment is an essential part of building an effective machine learning model. There are several scoring metrics such as confusion matrix, cross-validation, AUC-ROC curve, etc. Different scoring metrics are used for different types of problems.

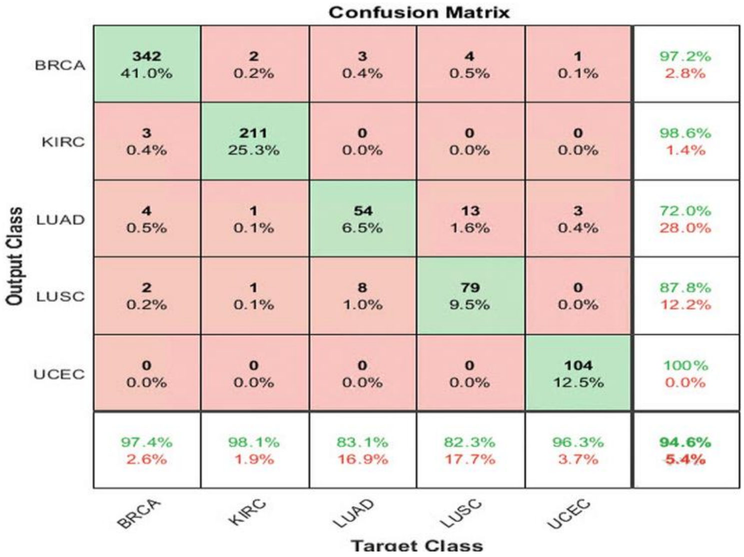
Model evaluation is a key part of building an efficient machine learning model. There are several diagnostic indicators such as confusion matrix, cross-validation, and AUC-ROC curve. The idea of creating a machine learning model works on the principle of constructive feedback. Build the model, get feedback from your metrics, make improvements, and keep going until we get the desired accuracy. Diagnostic metrics describe the performance of our model. An important aspect of the matrix is the ability to differentiate between model outcomes. Even many analysts and ambitious data scientists have not confirmed the strength of their models. Once they become models, they rush to map predictions on hidden data. This is the wrong way. Making a prediction model is not your only motivation. This is to create and select a model that provides high accuracy from the sample data. Therefore, it is important to check the accuracy of the model before calculating the prediction values. Looking at different types of matrices to test machine learning and matrix industry models. The choice of metric depends entirely on the type of model and the model implementation plan.

### 2.4.2 Confusion Matrix

The confusion matrix is the NxN matrix. Where N is the expected number of classes. For the problem at hand, N = 2, so we get a 2 x 2 matrix. Here are some definitions to keep in mind about the confusion matrix:

- Accuracy: Percentage of the total number of correct predictions.
- Positive prediction or accuracy: Percentage of correctly identified positive cases.
- Negative Prediction: Percentage of Correctly Identified Negative Cases.
- Sensitivity or Recollection: Percentage of actual positive cases identified correctly.
- Specificity: Percentage of actually negative cases identified correctly.

We usually rely on one of the metrics defined above. In a pharmaceutical company, for example, they will be more concerned with a minimal false-positive diagnosis. Therefore, they will be more concerned with high specificity. Attrition models, on the other hand, would be more concerned with sensitivity. The confusion matrix is generally used only with the class output model (Fig 5).



Confusion Matrix

| Output Class | BRCA | KIRC | LUAD | LUSC | UCEC | |
|---|---|---|---|---|---|---|
| BRCA | 342<br>41.0% | 2<br>0.2% | 3<br>0.4% | 4<br>0.5% | 1<br>0.1% | 97.2%<br>2.8% |
| KIRC | 3<br>0.4% | 211<br>25.3% | 0<br>0.0% | 0<br>0.0% | 0<br>0.0% | 98.6%<br>1.4% |
| LUAD | 4<br>0.5% | 1<br>0.1% | 54<br>6.5% | 13<br>1.6% | 3<br>0.4% | 72.0%<br>28.0% |
| LUSC | 2<br>0.2% | 1<br>0.1% | 8<br>1.0% | 79<br>9.5% | 0<br>0.0% | 87.8%<br>12.2% |
| UCEC | 0<br>0.0% | 0<br>0.0% | 0<br>0.0% | 0<br>0.0% | 104<br>12.5% | 100%<br>0.0% |
| | 97.4%<br>2.6% | 98.1%<br>1.9% | 83.1%<br>16.9% | 82.3%<br>17.7% | 96.3%<br>3.7% | 94.6%<br>5.4% |

Target Class

*Figure 7:Example of confusion matrix for 60% training set and 40% training set. [12].*

### 2.4.3 F1 Score

The importance of choosing a use case fit / recall basis. For use cases, what if you try to get the highest precision and recall at the same time? The F1-score is the harmonic mean of the precision and recall values of the classification problem.

The formula for F1-Score is:

$$F_1 = \left( \frac{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}}{2} \right)^{-1} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}.$$

The obvious question that comes to mind here is why take the harmonic mean instead of the arithmetic mean. Indeed, HM punishes more extreme values. Let's understand this with an example. There is a binary classification model with the following results:

precision: 0, recall: 1

Here, if we take the arithmetic mean, we get 0.5. Clearly the above result comes from a silly classifier who simply ignores the input and infers only one class as the output. Now, if we take HM, we get 0, which is precisely because this model is useless for all purposes. That sounds easy. However, there are situations for which the data scientist wants to give a higher value / weight percentage for accuracy or recall.

In this chapter we have covered most of the algorithms that we are going to use in our thesis to detect the fraud in medical claim data, starting from basic ML algorithms to the NER which is a technique and not an algorithm, but it is based on ML algorithms. We presented the evaluation metrics mathematical formulas and how it works, later we will use them to evaluate our result.

# CHAPTER 3: LITERATURE REVIEW

## 3.1 Introduction

With the constant advancement of technology, new and more different types of fraudulent behaviours are being uncovered, so to combat these behaviours more advanced fraud detection techniques are being developed. The Machine Learning (ML) technique is part of an Artificial Intelligence (AI) system that can engage in complex analyses which can not only replace human input but become even better than it. Machine learning applies its AI and gives the system the ability to learn and adapt from experience, with no additional programming.

## 3.2 Benefits of Machine Learning in Fraud Detection

The machine learning system regarding fraud detection works by analysing the customer’s data patterns and transaction records. It can analyse these patterns and behaviours much faster and efficiently than any human could do, which can show any potential discrepancies from the norms. The system adapts in real-time to new data which allows it to create better and more accurate predictions.

A major benefit of Machine learning is the higher degrees of accuracy since the problems relating to human error in processing and analysis of data is completely eliminated. Furthermore, it can allow the system to process massive amounts of data without fear of minor errors and delays.

Machine learning is a relatively cost-effective detection technique for companies. Since data can be processed and analysed in shorter amounts of time, which means that there are no manual checks and reviews every time new data is acquired.

## 3.3 Potential Indicators of Health Insurance Fraud

Firstly, we need to identify the potential indicators that could lead to insurance fraud in the healthcare industry. A major indicator is where healthcare providers bill customers for highly paid services than those that were actually performed, which is also known as upcoding [14].  Customers may receive bills for non-covered services which could be another indicator of fraud occurring [15]. There is also a possibility of medical insurance claims containing procedures that are medically unnecessary for the customers, which would be a major fraud indicator leading to potentially a fraud investigation [16]. Other potential fraudulent behaviour that may be carried out by the healthcare providers includes the modification of prescriptions, reimbursement claims for treatments that haven’t happened, and creating ‘ghost patients’, among others.

There is a chance that the patients may also be involved in suspicious fraudulent behaviours. This generally concerns providing inappropriate information to insurance providers. They may provide incorrect medical history, false geographic information or show fake financial information to get better insurance coverage. Some examples regarding fraudulent behaviours by patients include, submitting claims for ineligible claimants, filing claims for services that were never actually received, and using a different person’s financial information to enhance their own portfolio.

## 3.4 Methodology to Detect Insurance Claim Frauds

With regards to insurance companies, differentiating between the legal and fake claims successfully is one of the ways to ensure the economic well-being of the insurance providers and will allow them to efficiently protect their customers. Failure to do so would create significant costs and extra workload on the insurance companies leading to interruptions in processes and costs for customers with any hidden frauds causing added premiums to the honest customers.

The most efficient strategy would be to implement a computerized system for fraud detection that analyses consumer patterns and transactions faster than a human could. This ‘Machine learning’ system eliminates human error while it processes a massive amount of data. The machine learning process keeps learning and adapting to new consumer patterns to create even better predictions. Finally, machine learning is a relatively cost-effective fraud detection technique for companies. Massive amounts of data can be analyzed in seconds while the human analysts are unburdened and focused on more strategic tasks.

The following section shows the use of analytics in health insurance. The project has modelled the deceitful claims in the framework of a health claims group. According to the European Insurance and Occupational Pensions Authority (EIOPA), claims management is the most vital part of the insurance ecosystem, in which tools such as machine learning algorithms can be highly beneficial [17].

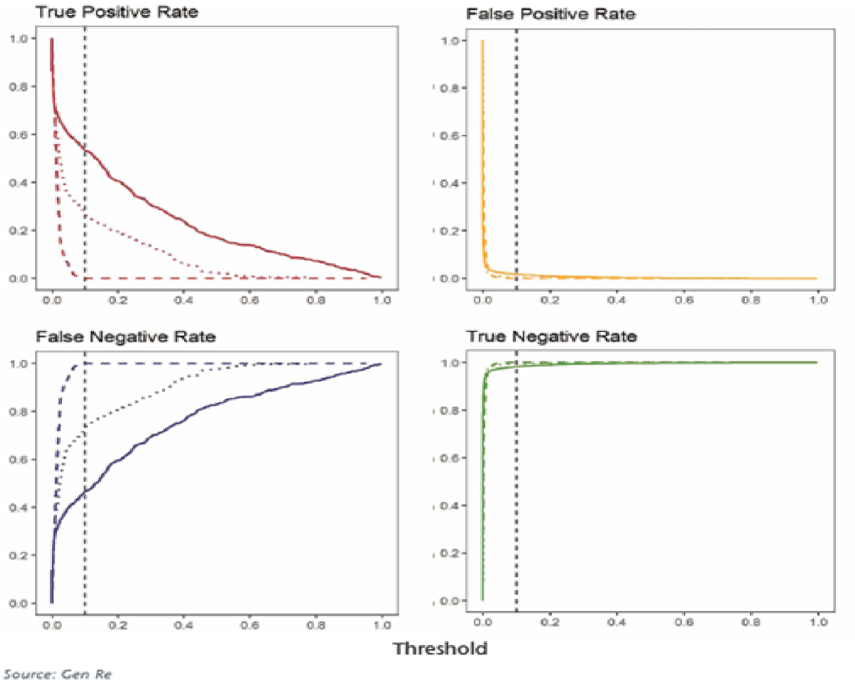
## 3.5 The Data being used for Project

The claim specialists have evaluated and have categorized the claims as either normal or flagged the claims as probably being fraudulent. The claims would be flagged because of either suspicious policy, individual claims or any institution related fraudulent activity. These highlighted claims are said to be approximately 0.3% of the total number of claims which is very rare. This makes it highly imbalanced as it is very challenging to apply a model that would distribute an accurate response.

## 3.6 Research Algorithms Used in the Fraud Detection Analysis

Previous studies aimed to make a machine learning model, a supposed binary classifier, that can help detect the two labels as suitably as possible. Throughout the analysis, they assigned 0 (negative) for normal claims and 1 (positive) for flagged claims. If a high threshold was chosen, then there is a possibility for many claims to be forecast as normal as their predicted chance is less than the high threshold. This could cause many of the flagged claims to remain undetected. On the other hand, choosing a threshold of zero would imply that most of the claims are flagged as its predicted probability is greater than zero which we are aware is false. Consequently, they are aiming to find an optimal low threshold for the basis of this analysis.

They have used three different algorithms: Generalized Linear Models (GLM), Gradient Boosting Machines (GBM) which is a group of decision trees and Neural Networks (NN) which is an algorithm consisting of many layers. By comparing the performance of models inside these algorithms, the most suitable model can be found.



Source: Gen Re

*Figure 8:Components of the confusion matrix for three models. The applied GLM is shown by dashed line, the GBM by dotted line and the optimized NN by solid line [18].*

In Figure 8, at the top left panel, it is shown that the GLM model (red line) hasn't been able to detect any flagged claims successfully, essentially giving a zero positive rate through the assortment of thresholds. The blue lines situated in the bottom left diagram shows a huge number of false negative rates, meaning that claims are being falsely spotted as normal even smaller thresholds. Therefore, the GLM model is too modest for this specific problem, because of its direct structure while not having the necessary complexity to process this data.

The GBM models are usually complex enough to perform while studying comprehensive patterns. In this case, the GBM improves the appropriately spotted flagged claims (true positive rate) which is highlighted by the red dotted line in Figure 1.

By applying a NN model with the default constraints, we amazingly don't see a lot of development as compared to the GBM results. Afterwards, it has been discovered that only by using thorough modification of the parameters and implementing tweaks in the cases of unnecessary data, we could improve these results to reach an optimal level. This is shown by the straight red line in the top left panel seen in Figure 1.

The study reveals that in order to detect fraud, the methods with the bigger datasets would be using SVMs, potentially combined with CNNs to have a more stable and reliable performance. For the smaller datasets, the combined approaches of SVM, Random Forest and KNNs can provide good developments. Convolutional Neural Networks (CNN) usually outclasses other deep learning methods such as Autoencoders, RBM and DBN.

### 3.7 Limitations of Using These Methods

The models used above can help in choosing an optimal threshold to categorize between normal and flagged claims but that will only be possible if there is effective communication between the business and the analytics teams. These models are complex to develop and even harder to bring into production and with the different parameters that they have to deal with, if they are not evaluated by the best analysts, it could not give the organization the data they need. Having the right set-up, both locally and in the cloud, environment is also a significant limiting factor that can halt the use of these models. Therefore, projects like these need to be tailor made for the needs of the consumers which is not something every organization can provide.

### 3.8 Research Methods Used for Healthcare Claim Fraud Detection

Healthcare fraud is being perceived as one of the more serious concerns as it has a relatively consequential effect for individuals, companies and governments. Generally, healthcare fraud detection relies on the knowledge of the specialists in this particular field, which could be expensive and time-consuming. This has led several researchers to develop more fraud detection systems. The aim of these systems is to spot and account for frauds just as they appear in the system [19].

Healthcare fraud consists of different fraudulent activities and behaviors that vary according to the situations. The kinds of frauds can be divided based on which individuals or groups are involved in the fraud [20]. These frauds could be perpetrated by the providers of the insurance service, the subscribers, the carriers or conspiracy fraud which involves more than one party.

The raw data obtained for healthcare fraud detection is usually claims which comes from various sources of insurance claims data, such as data obtained from doctors, prescriptions given by the doctors, date of the medicines prescribed and the bills and general transactions. Every country has its own unique requirements and procedures regarding their healthcare system data [21].

Traditionally, only a small number of auditors used to tackle large amounts of health care claims. But generally, only the more experienced auditors would handle this important matter regarding the detection of frauds. But because of the large amounts of data involved, this method becomes time-consuming and less efficient as time goes on. The developments made in regard to data mining and machine learning tools have become more cost-effective and less time consuming for fraud detection [22]

For detecting anomalies and detecting fraud communicative patterns that are based on techniques (machine learning) are used. So, for the purpose of this, behaviour patterns of every individual tangled in the healthcare database, is organized to check for any deviation from the norms [23]. Another method that has been used, considers an anomaly detection technique by the application of Rule-based Data Mining, which is more of a non-supervised technique, based on the insurance claims that were acquired from the Medicare data. After analysing the health insurance claims that has power over big data for the detecting of fraud. The medical insurance claim irregularities were detected by means of these applications that requires private health insurance providers to identify any unseen cost projections that the systems aren’t able to detect [24]. Another study, which uses the 2013 Centre for Medicaid and Medicare Services (CMS) dataset, has made a machine learning model that detects when the doctors are exhibiting suspicious behaviour in their medical activities. Its purpose is to determine whether or not, doctors are acting outside of their respective specialty, which could lead to fraud, or problems regarding billing procedures [25].

The research regarding Fraud Detection has been going on for over 20 years now and has used several methods from manual examination to customer end verification. Machine learning models have also had great successes in this field. Deep learning models have been recently implemented in many applications aided by the rise in higher computation power and cheaper computing cost.

After careful consideration of the various studies on the fraud detection of healthcare systems, we can therefore conclude that the frauds that are occurring in healthcare systems can have many rare patterns. To detect these anomalous patterns, more work needs to be done using advanced techniques using machine learning. There is a need to develop improved methods that consider the little details of healthcare data. In order to achieve this, different forms of healthcare data need to be taken into account.

# CHAPTER 4: METHODOLOGY

## 4.1 Overview

In healthcare, diagnosis codes are defined to easily categorize diseases, symptoms, disorders, poisonings, adverse effects of drugs, chemicals, injuries and other reasons for patient encounters. They facilitate the identification of diseases, symptoms, disorders, poisonings, adverse effects of drugs, chemicals, injuries and other reasons for patient encounters with the appropriate codes. Diagnosis codes are used in the clinical coding besides intervention codes in medical classification. Diagnosis codes used in medical classification models are labeled by a clinical coder or Health Information Manager [26].

In the literature, there are several systems for classifying diagnoses worldwide. The International Statistical Classification of Diseases and Related Health Problems (ICD) is one of the used coding systems in classification systems. It is also used for mortality and morbidity data [27]. Each number that follows the International Statistical Classification is the version of the International Statistical Classification. Various versions have been used and the latest version is ICD-10.

## 4.2 ICD-10 Code

The ICD-10 code entered in the report with the diagnosis for each patient is very important. Therefore, there should be a system to match the ICD-10 code entered with the patient information. If the ICD-10 code does not match with the correct ICD-10 codes that can be provided by https://www.icd10data.com/, we can assume that there is a misuse of the information entered. The WWW.ICD10Data.com Is the famous and free knowledge base for the Icd10 code. When the researcher writes Icd10 code in google, the first website that appears to the researcher is WWW.ICD10Data.com.

The aim of this study to identify misuse related to ICD-10 codes by comparing the ICD-10 code in the dataset with the predicted ICD-10 code (that is the correct ICD-10 code extracted from https://www.icd10data.com/ with the key words) using a rule-based method. There are several features that can be used in the rule-based method, and named entities are the most popular in the biomedical field [28,29,30].

## 4.3 Named Entity Recognition (NER)

Named Entity Recognition (NER) is a widely used Natural Language Processing (NLP) task that involves identifying and categorizing key entities from a given text [31,32]. The detected entities are classified into given categories. Named Entity Recognition is generally used to extract an entity with a real word from a text, e.g., a person, an organization, or an event. The task is also commonly used in biomedical to address the challenges related to the biomedical field [33,34,35] by extracting entities such as RNA, protein, cell-type, cell line, DNA, drugs and diseases. Samples from NER tagging in biomedical can be found in Table 1.

*Table 1:Samples from different datasets for Named Entity Recognition task*

| **Id** | **Text** |
|---|---|
| #1 | Total content of T$_{\text{B-cell}}$ lymphocytes$_{\text{I-cell}}$ was decreasd 1.5-fol in peripheric blood |
| #2 | We observed patients treated with gentamic sulfate$_{\text{Gentamicins (D005839)}}$ or tobramycin sulfate$_{\text{Tobramycin (D014031)}}$ for the development of aminoglycoside$_{\text{Aminoglycosides (D000617)}}$-related renal failure$_{\text{Renal Insufficiency (D0501437)}}$. Gentamicin sulfate$_{\text{Gentamicins (D005839)}}$ decreased renal function$_{\text{Renal Insufficiency (D0501437)}}$ more frequently than tobramycin sulfate$_{\text{Tobramycin (D014031)}}$. |

## 4.4 Machine Learning Algorithms (ML)

- Naive Bayes Algorithm: It is a simple technique for creating classifiers: models that assign class labels to problem examples, which are presented as vectors of feature values, where class labels are generated from a limited set. - There is no algorithm to train such classifiers, but a family of algorithms based on a general principle: all naive Bayes classifiers assume that the value of a particular attribute is the value of another attribute. Is independent of the value of, which is given a class. The variable is gone. For example, if some fruit is red, round and about 10 cm in diameter, it can be considered an apple. A simple Bayes classifier considers each of these characteristics to contribute independently to the possibility that the fruit is an apple, regardless of any possible correlation between color, roundness, and diameter characteristics. -

- Random Forest Algorithm: It has been a flexible and easy-to-use ML algorithm that produces excellent results in most cases without hyperparameter adjustments. This is one of the most commonly used algorithms because of its simplicity and versatility.
- K Neighbors Classifier: It is a simple and easy to implement supervised ML algorithm that can be used to solve both classification and regression problems. Break! Let's unpack this.

Various machine learning algorithms have been applied in the literature including Hidden Markov Models (HMM), Decision Trees (DT) [36], Support Vector Machines (SVM) [37] and Conditional Random Fields (CRFs) besides Deep Learning algorithms [38, 39] which are the most recent studies. One of the libraries for NLP tasks with biomedical data is scispaCy [7]. scispaCy is an open-source Python library for Natural Language Processing (NLP). The library provides statistical machine learning algorithms for various Natural Language Processing (NLP) tasks, including Named Entity Recognition (NER), non-destructive tokenization, sentence segmentation, etc. for processing biomedical, scientific or clinical texts. There are several pre-trained models for different entity types.

*Table 2:Model name trained for entity types*

| **Model** | **Entity Types** |
|---|---|
| en_ner_craft_md | GGP, SO, TAXON, CHEBI, GO, CL |
| en_ner_jnlpba_md | DNA, CELL_TYPE, CELL_LINE, RNA, PROTEIN |
| en_ner_bc5cdr_md | DISEASE, CHEMICAL |
| en_ner_bionlp13cg_md | CANCER, ORGAN, TISSUE, ORGANISM, CELL, AMINO_ACID, GENE_OR_GENE_PRODUCT, SIMPLE_CHEMICAL, ANATOMİCAL_SYSTEM, IMMATERIAL_ANATOMICAL_ENTITY, MULTI-TISSUE_STRUCTURE, DEVELOPING_ANATOMICAL_STRUCTURE, ORGANISM_SUBDIVISION, CELLULAR_COMPONENT |

In Table 2, the model’s name indicates the name of the dataset used to train the model. Since our dataset was created to identify DISEASE, we used the model "en_ner_bc5cdr_md" and used a rule-based method in our study.

The workflow of the study is shown in Figure 9. We explain each step below:

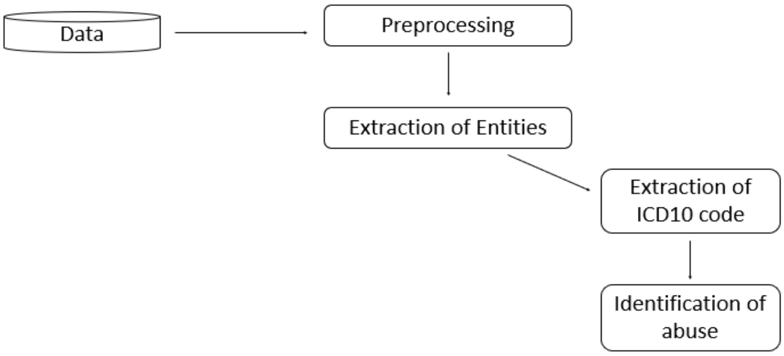


*Figure 9:The workflow of the proposed system*

## 4.5 Data Analysis

### 4.5.1 Pre-processing

In pre-processing, we remove only new line marks ("\n") and remove the lines where no diagnoses are given. Although punctuation marks are usually removed in Natural Language Processing (NLP), we did not remove them because the names may contain punctuation marks. In addition, we did not lowercase the diagnoses because they are specific disease and chemical names.

### 4.5.2 Extraction of Entities

We extracted diseases and chemicals for each diagnosis included in the dataset. The samples from extracted entities and chemicals can be found in Table 3.

*Table 3:Samples of diagnosis with extracted entities and chemicals with the model*

| Id | NER extracted Diagnosis |
|---|---|
| #1 | Cough DISEASE 1monthcough 1month cough DISEASE 1monthAcute pharyngitis DISEASE, unspecified |
| #2 | HTN, c/o headache DISEASE HTN, c/o headache DISEASE HTN, c/o headache Essential (primary) hypertension DISEASE |

| | |
|---|---|
| #3 | Headache DISEASE, pain DISEASE back of neck, blocked nose, sneezing headache, pain DISEASE back of neck, blocked nose, sneezing headache DISEASE, pain DISEASE back of neck, blocked nose, sneezing Acute maxillary sinusitis |
| #4 | BACK PAINBACK PAIN BACK PAINLow back pain DISEASE |
| #5 | left sided chest pain DISEASE, left sided chest pain DISEASE left sided chest pain DISEASE, left sided chest pain DISEASE left sided chest pain DISEASE, left sided chest painChronic ischaemic heart disease DISEASE, unspecified |
| #6 | Pain DISEASE abdomen, distension, constipation DISEASE, pain DISEASE swelling around the anuspain abdomen, distension, constipation DISEASE, pain DISEASE swelling around the anus pain DISEASE abdomen , distension , constipation DISEASE , pain DISEASE swelling around the anusExternal thrombosed haemorrhoids |
| #7 | throat discomfort cough DISEASE breathless 1monththroat discomfort cough breathless 1month throat discomfort cough DISEASE breathless 1monthAllergic rhinitis DISEASE, unspecified |
| #8 | on and off dry cough DISEASE, 1 monthon and off dry cough DISEASE, 1 month on and off dry cough DISEASE, 1 monthCough |
| #9 | throat pain fever cough 1weekthroat pain fever cough DISEASE 1week throat pain fever cough DISEASE 1weekAcute upper respiratory infection DISEASE, unspecified |

### 4.5.3 Extraction ICD-10 code

In this part, we identified ICD-10 codes for the words in named entities in the dataset. We used the website “https://www.icd10data.com/” to search for words contained in named entities for each diagnosis in the dataset. We assigned the extracted codes as predicted ICD-10 codes and assigned them as predicted ICD-10 codes in the dataset.

### 4.5.4 Identification of abuse

In the final part of the study, we used the system given in Figure 10 to identify whether or not there was abuse. We compared each predicted ICD-10 code with the corresponding ICD-10 code given in the dataset. If the predicted ICD-10 code(s) match with the ICD-10 code given in the dataset, then there is no suspicion of abuse. Otherwise, there is suspicion of abuse.

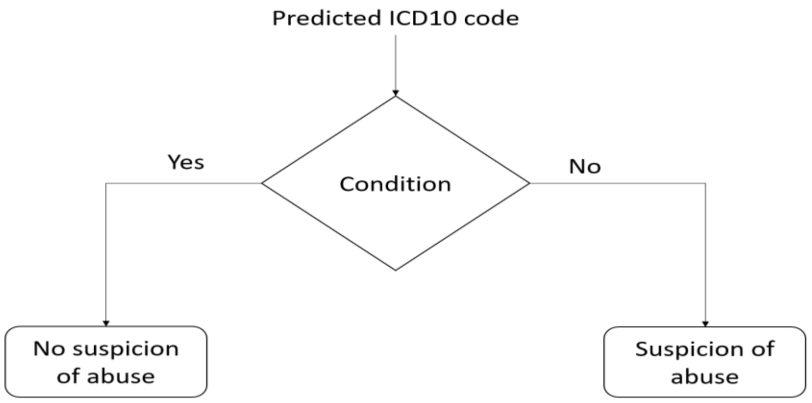


*Figure 10:System for identification of abuse*

# CHAPTER 5: EXPERIMENTS AND RESULTS

In this chapter we are going to present the details of the dataset with the results of identification of abuse.

## 5.1 Data Description

Here we explain two data sets. The first dataset is the dataset we used in the study and the second one is the dataset used to train the model “en_ner_bc5cdr_md”.

a. The dataset contains 334 samples with features Service_code, Services_Description, Inc_date_frm, QTY_Invoice, Specialty, Diagnosis, Temperature, Respiratory Rate, Blood Pressure, Height, Weight, Pulserate, ICD10_Code. Although there are 334 samples, there are only 58 unique disease definitions and 35 ICD10_Codes. The details of the dataset can be found in Table 4. The numbers extracted from the dataset are very important for understanding the dataset along with the results.

b. The NER model “en_ner_bc5cdr_md” that we used with the scispaCy library was trained with the BC5CDR dataset that is composed of 1500 PubMed articles with 4409 annotated CHEMICAL substances, 5818 DISEASES, and 3116 CHEMICAL-disease interactions.

*Table 4:Statistical details of the dataset*

| **Parameter** | **Value** |
|---|---|
| number of unique diseases | 58 |
| number of words | 1510 |
| Avg length of words | 26,03 |
| number of named entities | 222 |
| number of words in named entities | 347 |
| Avg length of words in named entities | 5,98 |
| number of words not in named entities | 1163 |
| Avg length of named entities | 3,82 |

### 5.2 Evaluation Metric

To evaluate the rule-based ICD 10 extraction model, we used accuracy as the evaluation metric.

$$Accuracy = \frac{numberofsampleswherepredictedICD10codeequalsICD10code}{totalnumberofsamples \in thedataset}$$

#### 5.2.1 Results

We have extracted the number of suspicions of abuse and the number of no suspicion of abuse and presented them in Table 5. It can be clearly seen that the number of suspicions of abuse is excessively high compared to the number of no suspicion of abuse. This indicates the qualification of the reports filed for patience's.

*Table 5:Results of identification of abuse*

| **Number of suspicions of abuse** | **Number of no suspicion of abuse** | **Total** |
|---|---|---|
| 266 | 68 | 334 |

We analyzed the result of the Number of suspicions of abuse to know why it was too high and presented them in Table 6. We found three reasons: first, the doctor miswrote the Icd10 code, which we did not find in the knowledgebase. Second, the doctor wrote the disease ICD10 code in general because their levels for card health insurance are A, B, and C for most health insurance companies. For example, Level A covers all cancer diseases, so the doctor wrote the disease ICD10 code in general. Third, the doctor did not write the diagnosis neatly, so the scispaCy model could not read the diagnosis to give us the icd10 code for the disease.

*Table 6 : analysis result of the Number of suspicions of abuse*

| **The doctor did not write the diagnosis neatly.** | **The doctor wrote the disease ICD10 code in general.** | **The doctor miswrote the Icd10 code** | **Total** |
|---|---|---|---|
| 122 | 96 | 48 | 266 |

We also calculated the matched ICD-10 codes as accuracy in the study to show the importance of the ICD-10 code in the report. Although the predicted ICD-10 codes are a list for each sample that can be either one item or multiple items, the result is shown in Table 7. It can be clearly seen that the accuracy result is very low for this data set.

*Table 7 : Accuracy of results of identification of abuse*

| Model | Accuracy |
|---|---|
| Rule based model | 0.20 |

In this experiment, we have presented a rule-based method to identify whether abuse is present or not. We used named entities extracted from the dataset using the scispaCy library and used the words in the name entities as key words used in https://www.icd10data.com/ website. We extracted the ICD-10 codes that match with the given key words. We assigned the predicted ICD-10 code as the correct ICD-code since they are extracted from the website. Finally, we followed the rule: if the ICD-10 code matched with the predicted ICD-10 code, we printed “There is no suspicion of abuse”. Otherwise, we printed “There is suspicion of abuse”.

Here, we extracted named entity features with rule-based matching. We used only the named entity recognition model from scispaCy. We did not focus too much on the pre-processing steps, although they are fundamental to achieve better results for the defined task.

We verified our results by putting every Icd10 code found in a column (icd10code) extracted from the Knowledge base inside https://www.icd10data.com/. We will get the exact name of the disease found in the column (extractedDisease), which the scispaCy library extracted.

As a future goal, we want to extend the study by improving the pre-processing steps and expanding the feature set to improve the results of the model. Also, we would like to propose a new machine learning based model for this task.

# CHAPTER 6: CONCLUSION

The Research Data Leak in medical claims was based on the doctor input of ICD10 code in every claim, and the classification of the reliability of the claims was performed based on the data scraped. This research showed that Named Entity Recognition is an excellent technique in delivering the prediction value for the problem statement—the research intended to predict the historical data collected from different databases belonging to the same insurance company. Data is a key part of every data analysis research, which has a positive effect on the outcome. This chapter summarizes the methodological, administrative, and theoretical implications, the limits of research, and the future that applies to it. The aim of the research is to capture the importance of the Scrapy and NER techniques. The purpose of our research is to find out whether the claim is Fraudulent or not. The research focuses primarily on modules such as data collection, data pre-processing, web scraping and prediction. The predictive analysis of the research was started by collecting real data from different datasets, which was done inside an insurance company under an NDA for the confidentiality of the information. Scrapped data is analysed with NLP techniques, exactly with NER to see how different predictions for the target work and how the results come out with better accuracy and minimum error rate. This research helps us to understand how to deal with real data more effectively than pre-existing data recorded some time ago.

## 6.1 Advantages

- ➢ Such extractions can save a lot of manual labour and reduce the time to market for new claims. Recent advances in machine learning have leveraged the vast text corpus available in scientific literature, medical and pharmaceutical websites, and a train system of many NLP tasks, from text mining to answering questions.
- ➢ The specific problem we focused on in our experiments was to use NER to extract diagnosis from ICD10. As a solution, the extracted entities are processed further downstream, linking the entities and using dictionary-based techniques to flag inappropriate diagnosis against the ICD10 codes.
- ➢ One of the main challenges in training NLP-based models is the availability of high-quality, reasonably sized, annotated datasets. Additionally, in a typical industrial environment, the relative difficulty of securing a lot of time from subject matter

experts and the lack of tools and techniques for effective annotation, as well as the ability to view such annotations to minimize human error, impact research and benchmarking of new teaching methods. and algorithms.

## 6.2 Limitation

The result of the study is a NER model for detecting the value of diagnosis description versus their ICD10 codes. Although studies differ to obtain better results, several limitations are encountered in investigating medical claims.

1. Data availability: The data required for research is used by big insurance companies which is hard to get. Therefore, data availability is low.
2. Data reliability: It was a challenging task to find good data, we had to collect it from 3 different datasets.
3. This is just a part of Data Leak in Medical claims and does not replace the financial frauds that happen with prices and numbers.

4. Future work will use the new methodology and framework containing a new model to give us only two categories with no suspicion of abuse and suspicion of abuse.

# REFERENCES

1. Abdallah, A. M. (2016). Fraud detection system. *Journal of Network*.

2. Bauder, R. A. (2016). "Predicting medical provider specialties to detect anomalous insurance claims. *In Tools with Artificial Intelligence (ICTAI)*.

3. Bekkerman, Ron, and James Allan. *Using bigrams in text categorization*. Technical Report IR-408, Center of Intelligent Information Retrieval, UMass Amherst, 2004.

4. Danisman, Taner, and Adil A. (2008). "Feeler: Emotion classification of text using vector space model." *AISB 2008 convention communication, interaction and social intelligence*. Vol. 1.

5. EIOPA. (2019). *https://www.eiopa.europa.eu/sites/default/files/publications/eiopa_bigdataanalytics_thematicreview_april2019.pdf*.

6. Fukuda, Ken-ichiro, et al. "Toward information extraction: identifying protein names from biological papers." *Pac symp biocomput*. Vol. 707. No. 18. 1998.

7. Gridach, M. (2017). "Character-level neural network for biomedical named entity recognition." *Journal of biomedical informatics,* 70: 85-91.

8. Habibi., M, et al. (2017). "Deep learning with word embeddings improves biomedical named entity recognition." *Bioinformatics* 33(14): i37-i48.

9. Hazelwood, Anita C., and Carol A. (2003). Venable. *ICD-9-CM Diagnostic Coding and Reimbursement for Physician Services, 2004.* American Health Information Management Association.

10. Huang, H., Wang, H. and Jin, D. (2018). A Low-Cost Named Entity Recognition Research Based on Active Learning. Scientific Programming, 2018, pp.1–10.

11. Joudaki, H. A. (2015). "Using data mining to detect healthcare fraud and abuse. *Global journal of health science*.

12. Jyothsna, V. V. (2011). A review of anomaly-based intrusion detection systems. *International Journal of Computer Applications*.

13. Ju, Z., Jian W., and Fei, Z. (2011). "Named entity recognition from biomedical text using SVM." *2011 5th international conference on bioinformatics and biomedical engineering*. IEEE.

14. Kim, Yu J., et al. (2016). "International Classification of Diseases 10th edition-based disability adjusted life years for measuring the burden of specific injury." *Clinical and experimental emergency medicine* 3(4): 219.

15. Leser, U., and Jörg H. (2005). "What makes a gene name? Named entity recognition in the biomedical literature." *Briefings in bioinformatics* 6(4): 357-369.

16. Lewis, David D. (1991). "Evaluating text categorization i." *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991*.

17. Lin, Yi-Feng, et al. (2004). "A maximum entropy approach to biomedical named entity recognition." *Proceedings of the 4th International Conference on Data Mining in Bioinformatics*.

18. Liu, Q. a. (2013). Healthcare fraud detection: A survey and a clustering model incorporating Geo-location information. *World Continuous Auditing and Reporting Symposium*.

19. Mohit, B. (2014). "Named entity recognition." *Natural language processing of semitic languages*. Springer, Berlin, Heidelberg, 221-245.

20. Mukherjee, S. (2013). Medicare ‘upcoding’. *Think Process Organization*.Neumann, Mark, et al. "Scispacy: Fast and robust models for biomedical natural language processing." *arXiv preprint arXiv:1902.07669* (2019).

21. Nadeau, D., and Satoshi S. (2007). "A survey of named entity recognition and classification." *Lingvisticae Investigationes* 30.(1): 3-26.

22. Proux, Denys, et al. "Detecting Gene Symbols and Names in Biological Texts a First Step toward Pertinent Information Extraction." *Genome Informatics* 9 (1998): 72-80.

23. Robbins DB, A. A. (2011). Too much care? Stepped up medical necessity fraud litigation against hospitals. *Washington Healthcare News*.

24. Settles, B. (2004). "Biomedical named entity recognition using conditional random fields and rich feature sets." *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications (NLPBA/BioNLP)*.

25. Sohn, Sunghwan, et al. (2018) "Clinical documentation variations and NLP system portability: a case study in asthma birth cohorts across institutions." *Journal of the American Medical Informatics Association* 25.3: 353-359.

26. Soomro, Pir Dino, et al. (2017). "Bio-NER: biomedical named entity recognition using rule based and statistical learners." *International Journal of Advanced Computer Science. Appl* 8: 163-170.

27. Srinivasan, U. a. (2013). Leveraging big data analytics to reduce healthcare costs. *IT professional*.

28. Popowich, F. (2005). "Using text mining and natural language processing for health care claims processing." *ACM SIGKDD Explorations Newsletter* 7.1: 59-66.

29. Wong, J. (2019). Using machine learning to identify health outcomes from electronic health record data.

30. Wu, Y., Jiang, M., Xu, J., Zhi, D. and Xu, H. (2018). Clinical Named Entity Recognition Using Deep Learning Models. AMIA Annual Symposium Proceedings, [online] 2017, pp.1812–1819. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5977567/.

31. Wynia M, C. D. (2000). Physician manipulation of reimbursement rules for patients. *Journal of American Medical Association*.

32. Yang, W.-S. (2003). A Process Pattern Mining. *National Sun Yat-Sen University*.
33. Zhu, Q., et al. (2018). "GRAM-CNN: a deep learning approach with local context for named entity recognition in biomedical text." *Bioinformatics* 34(9): 1547-1554.

34. https://cloudblogs.microsoft.com/industry-blog/health/2017/03/14/the-most-common-types-of-healthcare-fraud/

35. https://europepmc.org/article/med/33723489

36. https://www.genre.com/knowledge/publications/ri20-1-en.html

37. https://www.nature.com/articles/s41746-020-0221-y

38. https://www.researchgate.net/figure/Confusion-matrix-for-60-training-and-40-testing-strategy_fig4_338909223.