A Model for the Automated Detection of Fraudulent Healthcare Claims using Data Mining Methods.

by

Obodoekwe Nnaemeka Chukwudi Fortune

DISSERTATION

Submitted in fulfilment of the requirements for the degree

MASTER OF SCIENCE

in

INFORMATION TECHNOLOGY

in the

FACULTY OF SCIENCE

at the

UNIVERSITY OF JOHANNESBURG

SUPERVISOR: PROF DUSTIN TERENCE VAN DER HAAR

OCTOBER 2018

Abstract

The menace of fraud today cannot be underestimated. The healthcare system put in place to facilitate rendering medical services as well as improving access to medical services has not been an exception to fraudulent activities. Traditional healthcare claims fraud detection methods no longer suffice due to the increased complexity in the medical billing process. Machine learning has become a very important technique in the computing world today. The abundance of computing power has aided the adoption of machine learning by different problem domains including healthcare claims fraud detection.

The study explores the application of different machine learning methods in the process of detecting possible fraudulent healthcare claims fraud. We propose a data mining model that incorporates several knowledge discovery processes in the pipeline. The model makes use of the data from the Medicare payment data from the Centre for Medicare and Medicaid Services as well as data from the List of Excluded Individual or Entities (LEIE) database. The data was then passed through the data pre-processing and transformation stages to get the data to a desirable state. Once the data is in the desired state, we apply several machine learning methods to derive knowledge as well as classify the data into fraudulent and non-fraudulent claims.

The results derived from the comprehensive benchmark used on the implemented version of the model, have shown that machine learning methods can be used to detect possible fraudulent healthcare claims. The models based on the Gradient Boosted Tree Classifier and Artificial Neural Network performed best while the Naïve Bayes model couldn't classify the data. By applying the correct pre-processing method as well as data transformation methods to the Medicare data, along with the appropriate machine learning methods, the healthcare fraud detection system yields nominal results for identification of possible fraudulent claims in the medical billing process.

Acknowledgements

I would like to thank my parents (Emmanuel and Gladys), my Uncle Timothy and his Wife Philile as well as my aunt Edith for their support and patience throughout my studies. I would like to express my gratitude to my supervisor Dr Van Der Haar for his guidance and motivation throughout my research. Lastly, I would also like to dedicate this work to my sister Jennifer for her continuous belief in me even at my most trying times.

Table of Contents

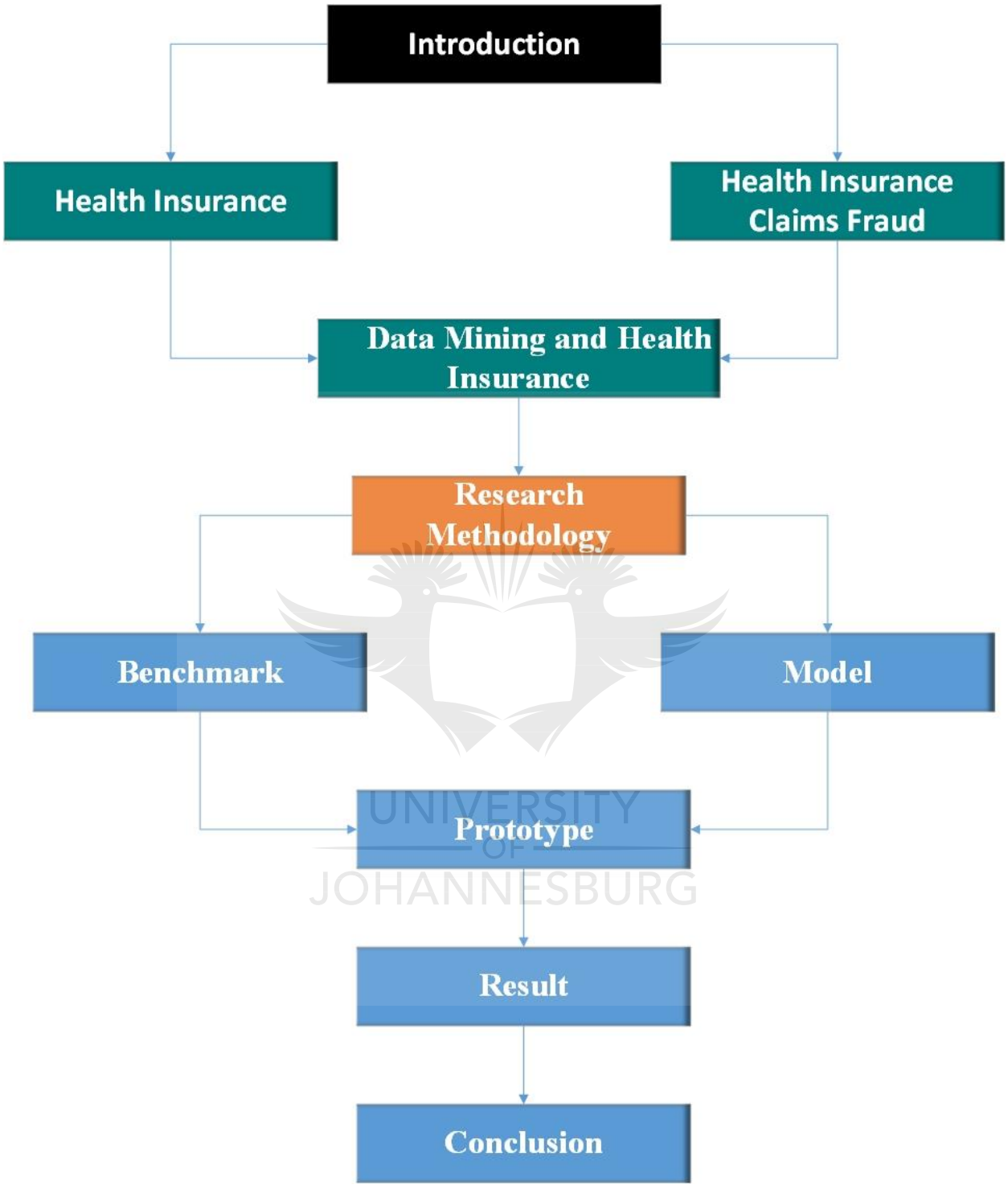
List of Figures

List of Tables

List of Abbreviation

LEIE	List of Excluded Individuals or Entities
POS	Point of Service
PPO	Preferred Provider Organisation
HMO	Health Maintenance Organizations
MSA	Medical Savings Account
ICD	International Classification of Diseases
CPT	Current Procedural Terminology
OCR	Optical Character Recognition
PCA	Principal Component Analysis
LSI	Latent semantic indexing
NHI	National Health Insurance
HIRA	Health Insurance Review and Assessment
LOF	Local Outlier Factor
CMS	Centre for Medicare and Medicaid Services
CPT	Current Procedural Terminology
GP	General Practitioner
NPI	National Provider Identifier
SVM	Support vector machine
PCA	Principal Component Analysis
NMF	Non-Negative Matrix Factorization
LDA	Latent Dirichlet Allocation

Dissertation roadmap



Chapter 1 Introduction

1.1 Introduction

A critical component of most people's lives is healthcare and as such families should be able to afford decent healthcare services. The need for healthcare is particularly more important in the elderly and infant population. The cost of healthcare forms a crucial part of individual and family expenditure. The rising cost of healthcare means that healthcare has become more of a luxury to some households. One of the factors that have led to the increase in the cost of healthcare is due to the money lost to fraud, waste and abuse in the healthcare system. The current global average loss rate of 6.99% - a running average taking account of 15 years of data - when taken as a proportion of global healthcare expenditure for 2011 ($6.97 trillion, £4.48 trillion or €5.38 trillion), equates to $487 billion, £313 billion or €376billion [1].

Fraud is a problem in the society and the health insurance industry has also been affected by fraud. Fraud causes loss of money in the insurance industry with an increase in insurance premiums being a resulting effect of the revenue lost from fraud. Fraud also leads to an increase in the time spent in assessing claims as the insurer must carefully analyse and process these claims to ensure there are no irregularities. This study attempts to address the problem of healthcare fraud and abuse by building an automated model to detect healthcare claims fraud.

The chapter starts with establishing a concrete definition of the research hypothesis that is addressed in the study in section 1.2, subsequently followed by the objectives set out for the study in section 1.3. The assumption and constraints set out in the study are then discussed in section 1.4. We show how we solve the problem by discussing details regarding the research design and the research methods used in section 1.5. We then give an overall outline of the study and then end with a conclusion.

1.2 Research Hypothesis

For many years, the problem of vetting insurance claims to pick out which claims are fraudulent has been a recurring issue, leading to the massive loss of revenue by the insurers. The loss of revenue through fraud is not because of an incompetent insurance system as there are trained actuaries and assessors that work to identify the fraud in the system. These actuaries and assessors specialize in analysing the several claims which the insurance provider receives from its customers to pick out the fraudulent ones. The problem lies in the fact that the number of claims coming in to be analyzed by the insurer has grown such that the manual process does not suffice anymore [2].

The manual process of analyzing each health insurance claim is expensive for insurance agencies and creates the possibility of a couple of high-value fraudulent claims going undetected. The use of manual fraud detection no longer suffices in the presence of a large volume of claims data to be processed as well as novel fraud patterns. Advanced logical solutions, using machine learning models and calculations can be utilized to analyze data to viably identify, foresee, oversee, and report irregular activities.

The use of machine learning to identify trends and patterns in data leads us to the hypothesis that machine learning methods can be used to detect health insurance claim fraud in an automated manner (**H1**). Machine learning methods have been used in other domains such as credit card, financial and tax claim to detect fraudulent transactions by detecting anomalies and patterns that have been identified as fraudulent activities.

1.3 Objectives

The aim of this study is to build an automated model that can help detect healthcare claims fraud. By detecting these fraudulent claims, the insurer can then avoid losing their revenue through payments for services never performed or paying for services not fairly priced as well as other fraudulent activities. In building the automated model, we would be contributing to the domain of health insurance claim fraud security research by providing a review on how data mining methods can be applied when designing an insurance claim fraud detection system.

To achieve automated healthcare claims fraud detection, we first need to identify what features make up a fraudulent claim (**O1**). Identifying these features would help in narrowing down the claims which are irregular. The features that would be used would be identified from the literature of previous works done and from the analysis that we will perform on the dataset. Experimentation, analysis and pilot studies would be used to ensure that only the most relevant features are considered. These features can be used to isolate the claims which are fraudulent by classifying and grouping them into clusters.

Once we have identified the features, we then build a model for applying the features to create a health insurance claim fraud detection system (**O2**). The model would serve to show how we create a solution for an automated health insurance claim fraud detection. The model will provide an alternative solution to the problem of detecting fraud in the healthcare claims process.

The next objective after developing the model is validating the model by implementing a prototype and benchmarking with a dataset (**O3**). By doing developing a prototype, several deductions and findings can be made. The benchmark will define metrics that can be used to evaluate the model. The benchmark done can potentially lead to results and recommendations.

1.4 Assumptions and Constraints

To carry out research in an area of study, there are certain assumptions that are made, and constraints established that are integrated with the arguments posed throughout the study. They also help define the scope of the study and ensure that due diligence is maintained throughout the study. The following subsections describe these assumptions and constraints.

1.4.1 Assumptions

Throughout the research carried out for the study there are specific assumptions that have been made and we have remained cognisant of these assumptions throughout the study:

1. We assume that the majority of healthcare claims fraud occurs through the medical practitioner, we make use of data that considers the medical practitioners as the sole source of healthcare claims fraud.

2. We also assume that the medical claims submitted by medical practitioners found in the List of Excluded Individuals or Entities (LEIE) database used for ground truth labels are likely fraudulent.

1.4.2 Constraints

The constraints are the operating limitations the study is carried out under and may restrict research efforts and these constraints need to be addressed in order to achieve the objectives set out. The initial analysis carried out on the study shows that the likely constraints that will affect the study are the following:

1. The capturing of claims data manually can introduce errors which may lead to noise in the data.
2. The LEIE data used does not fully correspond to the Medicare dataset as there are no unique identifiers between the two datasets.

1.5 Research Methodology

In every research, there is a need to establish a way in which the research is conducted. This section analyses the design strategy, research paradigm subscribed to and finally the methods or instruments used in achieving this goal.

1.5.1 Research Designs

The popular research designs include quantitative, qualitative and mixed designs. The qualitative approach is mainly concerned with gaining knowledge about the underlying reasoning and what motivated lines of actions. It is subjective and measures the different ways individuals view the world around them. When generating a hypothesis, setting up a problem, the qualitative approach can be used. In this approach, the aim is a very detailed and complete analysis of what is being observed. Its main purpose is to contextualize, interpret and understand perspectives [3].

The quantitative research design was mainly concerned with the measurement of quantity. It can be used in situations where the facts can be expressed as quantities [3]. This is more statistical in

terms of the analysis and is also objective as it aims for a precise measurement. The aim is always to explain what is observed or validate a proposed observation or hypothesis.

The mixed research design combines or integrates quantitative and qualitative research designs in a research study. The mixed methods mainly used in the social and health science research domains combines the strength of both the quantitative and qualitative research designs [4].

For this research work, the quantitative design is chosen as it allows for a statistical analysis of the data which has been collected in a structured manner and would be used for causal explanation of fraud detection in an insurance claim. We also use the quantitative approach as a result of the need to validate the hypothesis.

1.5.2 Research Paradigm

The research paradigm can be described as the entire process of thinking used in the research. It refers to known research traditions used in a philosophical framework or discipline. The research paradigm used determines the set of beliefs that would guide the actions in the entire research process [3]. There are two research paradigms: The positivist which believes in a stable reality that can be analyzed and dissected in an objective viewpoint. The interpretivist which has a viewpoint that only by intervening and by having a subjective approach can the reality be properly understood.

For this research, we have chosen positivism as the research paradigm because we would adhere to only observations and measurements to gain factual knowledge. We would also have roles limited to interpretation and collection of data using an objective approach as mentioned in subsection 1.5.1 and having quantifiable and observable research findings. We are also using the positivist approach as we would depend on quantifiable observations that can be analyzed. We would collect the historic insurance claim data and analyze it to make deductions without interfering with the insurance claim procedure.

1.5.3 Research Methods

The research methods used in this work are a literature review, experimentation, model creation and the development of a prototype. The research methods are the tools that we use in order to achieve the objectives of the research. The section unpacks each of the research methods used in the study to achieve the research objectives.

The literature review represents the systematic search of relevant work that has been published to gain an understanding of what is already known about the topic of intended research. It serves the purpose of establishing why the research is needed, broadening the knowledge of the researcher, and helping the researcher understand possible contributions the proposed study will make. The literature review was chosen as a research method because it will enable us to gain a better and deeper understanding of the research domain [3]. It would also enable the identification of gaps in the previous study.

Another research method used is the model which helps researchers relate concepts more accurately to reality. The model provides a framework to understand and describe the problem as well as a potential solution. The above-mentioned description of what a model motivates the use of the model for this project as the model will be used to test the facts and assumptions in this research. Being an abstraction of the reality, it can present a simpler case than the actual reality making it easier to work with.

The next research method used is prototyping which involves creating a preliminary implemented version of the proposed system. The prototype serves as part of a larger system of an application. It does not require much resource commitment. A prototype is used as a research method as it would be a preliminary way to test out some propositions and algorithms used to determine the effectiveness.

1.5.4 Population

The data that would be used for this research would consist of a combination of fraudulent and legitimate health insurance claims data. The dataset would be retrieved from Medicare online dataset repository. The Medicare dataset contain information about general payment made to

physicians in the United States of America. We also made use of data from the List of Excluded Individuals or Entities to establish ground truth labels for the Medicare dataset.

1.5.5 Data Sampling

For this study, the data sample comprises data about the Physician, Patient and Payment data. The data collected about the physician include location, specialty, service performed. The physician's data collected include National provider identification (NPI), last name, first name, zip code, provider type. Payment data include an average charge per service, charge for service, number of procedures performed per provider, number of distinct beneficiaries per day services, an average of the charges that the provider submitted, and the amount paid to the provider for services performed.

1.5.6 Research Plan

To gain a better understanding of how the research would be conducted, a research plan is provided below to give proper guidance. This study would first review the health insurance and health insurance fraud, traditional health insurance fraud detection the available machine learning algorithms applicable to insurance claim fraud detection. This would be done as part of the literature review. Once the literature review is done, a pilot study which entails data collection using the planned methods with the aim of testing the proposed approach. Using the pilot study, we are able to identify any other details that need addressing before the main collection of data. Next is model creation and then the building of a prototype to test the model. The results are collected, and an analysis is performed on the result. Once the analysis is done, the findings or deductions made are then reported.

1.5.7 Research Publications

During the course of the study, various research manuscripts were prepared for peer-reviewed conferences and the manuscripts were published at the respective subject matter conferences. The publications cover the review of previous works that have been carried out in the use of machine learning methods to detect healthcare fraud. The paper also presented a study of how some the machine learning methods were applied to the Medicare dataset to detect fraud in the claims

process. The review and feedback we got from the conference were applied in the final outcome of the dissertation. A copy of each research paper has been included in the appendix section of the dissertation. The research topic for the papers published include:

1. "*A Comparison of Machine Learning Methods Applicable to Healthcare Claims Fraud Detection*" presented at ICITS 2019, Springer International Publishing, Volume 918 page (548-558) in Quito, Ecuador.
2. "A Critical Analysis of the Application of Data Mining Methods to Detect Healthcare Claim Fraud in the Medical Billing Process." Ubiquitous Networking. UNet 2018. Lecture Notes in Computer Science, vol 11277, Chapter 29. Springer, Cham, Hammamet, Tunisia.

1.6 Conclusion

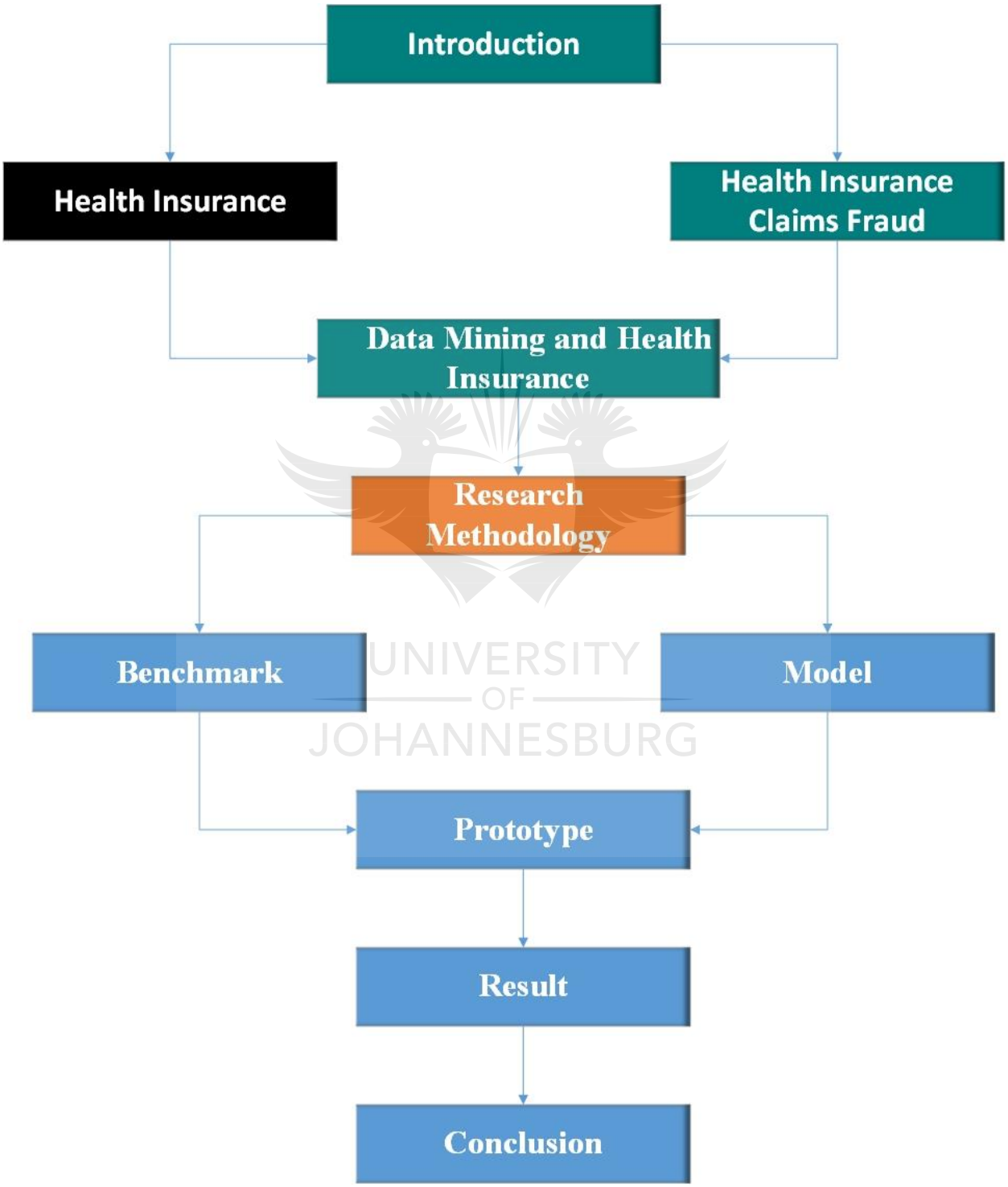
The chapter starts with an analysis of the research hypothesis in section 1.2. Rooted in the progression of healthcare fraud detection systems from a manual detection mechanism to a machine learning based fraud detection model, we discuss the steps taken to improve a manual healthcare fraud detection model to automated healthcare claims fraud detection model using machine learning methods.

To address the hypothesis, the study makes use of three objectives that must be addressed in determining the viability of healthcare claims fraud detection model using machine learning methods. The first objective we set is to identify what features make up a fraudulent claim (**O1**). The next objective set is to use the features identified to create a model for applying the features to create a health insurance claim fraud detection system (**O2**). The finally objective set out is to validate the model by implementing a prototype and benchmarking with a dataset (**O3**). The proposed model will encapsulate these objectives and can be designed, implemented and evaluated to know the feasibility in the healthcare environment.

Next, we defined the constraints and assumptions that were made during the course of the study. The assumptions and constraints help to align arguments and act as guidelines when carrying out an investigation. We then took a pragmatic approach in choosing a research methodology that yields a sustainable designed model. The methods chosen are a literature review, model creation and the development of a prototype.

With an understanding of the study approach, the study may then be further pursued based on the outline specified in the research plan. The next chapter introduces health insurance, the medical billing process as well as how fraud occurs in the healthcare claims process.

Dissertation roadmap



Chapter 2 Health Insurance

2.1 Introduction

According to Maslow's hierarchy of needs, the 2nd most basic need of every individual is the safety need. The safety needs of individuals consist of basic needs such as food, water and good health [5]. For the full functioning of any individual, the individual first needs to be in good health. To be in good health one needs to have access to the adequate medical treatment when needed. The problem lies in the continuous increase in the cost of the medical care, making it more of a luxury than a basic need in recent times [6]. Health insurance was introduced to help manage this cost and make healthcare more affordable for everyone.

Before we get into the depth of health insurance a definition of insurance is necessary. Insurance is a contract represented by a policy which a person or entity gets financial protection or reimbursement against potential losses from an insurance provider [7]. The insurance company pools client risks in order to make payments more affordable for its insured member. Health insurance is a contract between a group of individuals or person with an insurer stating that an individual pays an agreed premium for a specified health insurance cover [7]. Health insurance relates to an insurance type that covers medical and surgical expenses of a member. Health insurance can reimburse an insured member for expenses made in the treatment of injury or illness or it can directly pay the service provider for services provided to an insured member.

The health insurance system has been impacted by fraudulent activities with several parties involved in the process, trying to gain illegal advantage or benefits. The impact of fraud in the health insurance system results in massive loss of revenue, excessive time spent on reviewing these claims by the insurance service provider thereby leading to delayed feedback on reimbursements.

In this chapter, to gain a contextual knowledge of health insurance, we discuss the history of health insurance from a global perspective (in section 2.2). We also then look at the properties of health insurance (in section 2.3) by analyzing the different entities that make up the health insurance eco-system with the roles they play (in section 2.4). Next, we discuss the health insurance claim process (in section 2.5) and analyze how the entities interact with each other in processing a claim. Finally,

(in section 2.6) we look at the advantages and disadvantages of the existing medical billing process, identifying areas that can be improved.

2.2 History of Health Insurance

Section 2.2 gives an in-depth discussion of health insurance by discussing the history of health insurance. The history of health insurance is analyzed from a global perspective. In this section, we are going to see the different time periods applicable to the development of health insurance. We start by taking a look at the inception of health insurance and how what is known today as medical care came into existence and then the growth that has been experienced over the years.

2.2.1 1900-1920: Health Insurance and Sickness Insurance

The years between 1900 and 1920 marked a defining period in the history of health insurance as health insurance was enacted in this period. The period between 1900 and 1920 was mainlycharacterized by the initial low cost of medical care and the unwillingness of insurance companies to see health as an insurable commodity.

In the early 1910s, the cost of medical care was not a major concern in comparison to the loss of income due to the sickness of several workers. There was limited medical care in this era which mainly involved preventing diseases by personal hygiene, good diets, basic surgery, and prayers. The state of medical technology which was rudimentary at the time meant that the cost of medical treatment was initially low. There was a slow adoption of medical insurance as there was no perceived value of medical insurance, rather the only worry was the wages households lost by being sick and missing work [8]. So, cost of medical expenditure was from the wages lost at work when sick rather than the actual payment for medical treatment.

There was a low demand for health insurance as well as an unwillingness to offer medical insurance cover to individuals. The unwillingness was due to the fact that insurance companies did not see human health as an insurable commodity as human beings can change behaviors after purchasing insurance [9]. The insurance companies could also not define ways to accurately measure risk associated with individuals and then subsequently write insurance premiums (an amount paid for an insurance cover).

The insurance companies believed that the possibility of fraud in health insurance upset all statistical calculations since health and sickness being vague terms open to endless construction. Even though death is defined, it is very difficult to calculate what constitutes a loss of health or sickness to justify a level of compensation [9].

Healthcare remained cheap in the early 1900s because of the lack of advancement in medical science and technology. The mid-1900s saw advancements in medicine with several inventions being made and discovery of treatments for several previously incurable diseases

2.2.2 1920-1960: Birth and Growth of the Health Insurance Market

Section 2.2.2 shows how growth in the health insurance market occurred. It starts with one of the major health breakthroughs such as the discovery of a treatment for diabetes which subsequently led to increased demand for medical care and the higher cost of medical care. The growth in medical demand for medical care meant that the healthcare sector generated revenue which was not ignored by fraudsters as the number of fraudulent activities grew with the healthcare sector.

The 1920s saw advancement in medicine as Best, Macleod, and Banting at University of Toronto 1923, changed diabetes from a death sentence to a treatable disease through the discovery of active ingredients in insulin [3]. The transition in medical care from using simple medical procedures to more sophisticated and effective medical methods led to the increase in household expenditure on medical care. The quality of physicians also increased as several changes were implemented by the American Medical Association (AMA) to improve the standard of medical care.

The new revolution in medicine due to advancements in treatments led to an increased demand for medical services as well as the rise in the cost of medical service. Soon medical expenses took the center stage in household expenditure and the cost of healthcare gradually became unaffordable when needed in cases of emergency. The Blue Cross and the Blue Shield were the first organizations to offer hospital and treatment plans that allowed subscribers to contribute to a pool of funds to ensure health coverage was available whenever needed.

The success of Blue Cross and Blue Shield meant that the initial doubts most insurance companies had about insuring health were suddenly fading away. There was a growth in the commercial market supplying medical insurance in the 1940s. The growth in the medical insurance market was

aided by the fact that commercial companies could afford to offer groups of people with relatively low premiums.

Another major driver for growth in the healthcare sector was the role of government during World War II. There was competition for the already scarce labour market, so companies tried to use salary offerings to compete with each other to attract workers. The government stepped in to regulate the salary range but mandated that companies could now offer health insurance to employees as an incentive to potential employees [3]. With government influence, health insurance was introduced as a corporate package. The inclusion of health insurance as a corporate package also influenced the tax of the employee providing services to the employer as the employees had their yearly taxation reduced based on their medical scheme contributions. The influence from the government further increased the use of medical schemes as most organizations were made to put in medical schemes as part of the benefits package to cut down on taxes.

Advances in medicine have led to an increase in the cost of medical care making health insurance a necessity in modern times. The growth in health insurance has led to a transformation in the health insurance system. The healthcare system has transformed from a simple to a complex but well-structured system. Section 2.3 discusses the properties that characterize the health insurance system.

2.3 Properties of Health Insurance

Section 2.3 discusses health insurance and the different entities that make use of the health insurance ecosystem namely: the service provider, the insurance subscriber, and the insurance carriers [7]. As discussed earlier in section 2.1, health insurance is a type of insurance which covers the cost of medical care for an insured individual, with the medical care can be either basic medical care or more complex medical care such as surgical treatments [6].

In some countries such as Canada and Austria, medical care coverage is every citizen’s right and is provided by the government. In some other countries, health Insurance is partially provided by the state.

There are two broad types of health insurance schemes we will be discussing, the private and the public health Insurance. Private healthcare is funded from contributions by individuals and companies as well as subsidies on tax granted to organizations by the government. Public health insurance is almost completely funded through government tax.

2.3.1 Public Health Insurance

Public health insurance often uses the primary care services with clinics at the small community level and more serious cases being directed to the referral regional, district and central medical centers for treatments. The public health sector is designed in such a way that clinics at the local government level are responsible for the minimal medical care such as preventive care and environmental health services while the provincial hospitals are responsible for curative care. As earlier mentioned, most of the funding for the public health sector comes from government taxes but funding can come from donors [7].

2.3.2 Private Health Insurance

Private health insurance grew alongside economic growth and was created by organizations trying to protect their employees from the cost of healthcare and maintain a healthy workforce [10]. Funding for the private health sector comes from health insurance companies, direct out of pocket payments. Indirect funding can come from the government by subsidizing the tax paid by an entity (individual or organization) based on the medical scheme contribution of the entity.

The private medical insurance schemes offer a wide range of benefits to the members including dentistry, maternity care, optical care, mental care, and hospital cover. The cost of medical cover has been on the rise and has increasingly become unaffordable for households. The unaffordability of medical has put financial pressure on insurance subscribers which have subsequently led to more fraudulent activities in the health insurance industry.

The private and public health insurance service different classes of a country's population and they both complement each other to ensure a healthy population. We have briefly introduced the different entities that make up the health insurance ecosystem which are the health insurance subscribers, health insurance carriers, health insurance service providers and the health insurance

clearinghouses We discuss all the different entities and how they interact with each other in the health insurance ecosystem (in section 2.4).

2.4 The Health Insurance Ecosystem

In section 2.4, we discuss the different interacting entities that make up the insurance ecosystem. Figure 2.2 gives a summary of the different individuals or organizations involved in the functioning of the health insurance system. The subsections that follow describe each of the entities found in figure 2.2 in more detail.

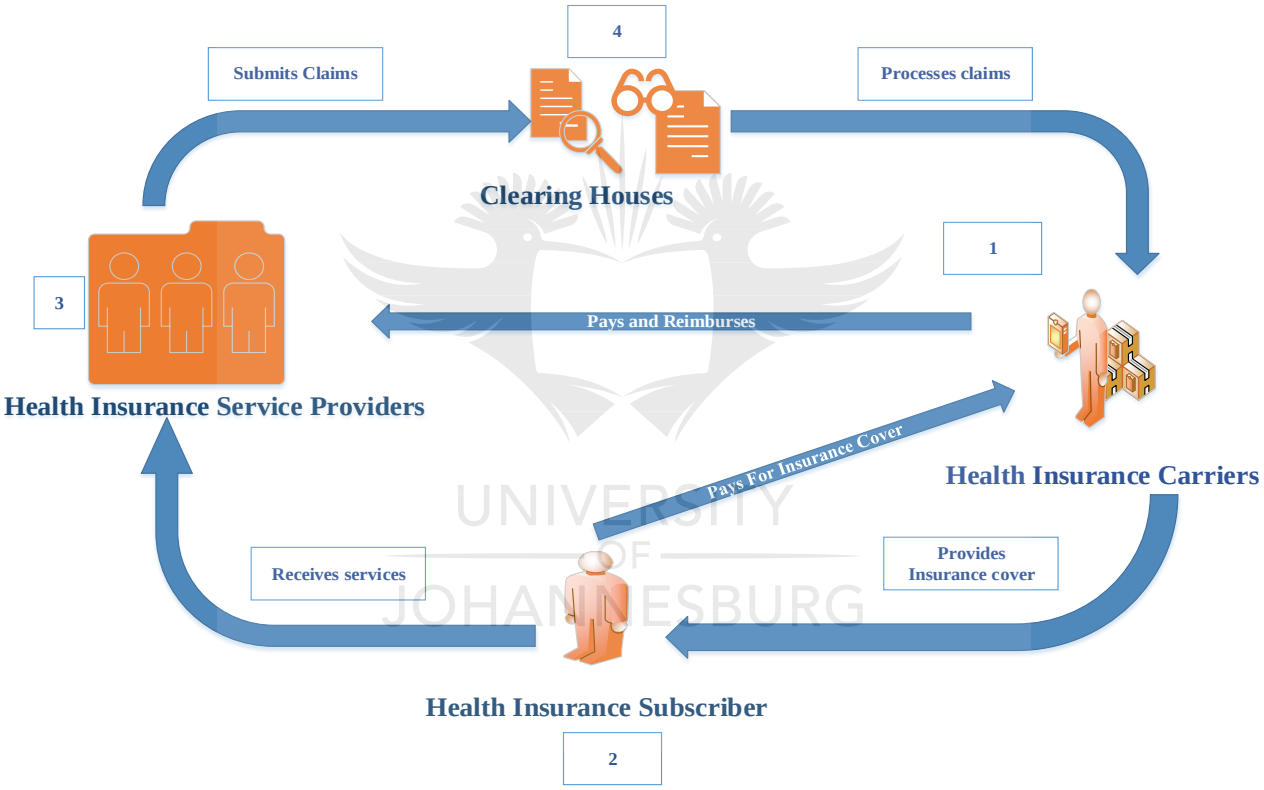


Figure 2.1 A summary of the health insurance ecosystem (as illustrated by the author)

2.4.1 The Insurance Carriers

The insurance carriers are the organizations that manage claims and pay benefits for the healthcare plan of a member. They receive the insurance premiums on a regular basis from the insurance subscribers defined (in subsection 2.4.2). They make payments on behalf of the insurance

subscriber to the service providers for healthcare costs. In some instances, the insurance carrier can be a government institution that pays for the healthcare expenses incurred by its citizens.

2.4.2 The Insurance Subscriber

These are the individuals that pay for the services that are offered by the insurance carrier and services from the service providers. The insurance subscriber includes the patient and the patient's employers. The insurance subscriber makes a monthly fee payment called insurance premium to the provider to obtain cover for medical costs. The health insurance subscriber receives healthcare benefits from health insurance providers. Depending on the contribution made by the subscriber, the health insurance subscriber can receive coverage for varying levels of care which can range from basic hospital care to comprehensive care. The different levels of care make up the managed care plan.

2.4.3 The Service Providers

The service providers are the entities that provide the required services to the insurance subscriber. They include doctors, ambulance companies, hospitals, pharmacists, and laboratories. The various service providers make up the provider network. The provider network is a framework that forms an integral part of managed care plans that serves the function of managing the cost of medical services received by the insurance subscriber. There are three main types of provider networks which include the health maintenance organization (HMOs), Preferred provider organizations (PPOs) and the Point of Service (POS). Each of these we briefly discuss below.

The first is the Health Maintenance Organizations (HMO) that utilize service providers who are contracted to render services to insurance carriers within specific geolocation. Members are required to only use services of the providers who are either employed by or contracted to the HMO.

The second type of provider network is the Preferred Provider Organizations (PPOs). They contract with both "in network" and preferred service providers. Member can receive care from

healthcare service providers outside the network although the health insurance carrier would cover more of the cost of treatment when care is received from "in network" service providers.

The final provider network is the Point of Service plans (POS). This takes a hybrid approach by combining different aspects of both PPO and HMO. They allow members to see any service provider with or without referrals.

2.4.4 Insurance Clearinghouses

In the health insurance billing process, these are the companies whose function is to serve as intermediaries between the service providers and the insurance carriers. They forward claim information from healthcare service providers to the insurance carriers. They pre-process the claims by performing claim scrubbing which involves checking a claim for error such as typographic errors and incorrect biographical data entry. Claim scrubbing is the process of validating the combination of data on a healthcare claim form [11]. Claim scrubbing edit ensures that the claims are processed faster and reduces risks of errors.

Each insurance carrier chooses the clearinghouse it wants to use and then pays the clearinghouse per claim processed. The clearinghouses then submit payment to either payer directly or through other intermediaries using the payer's software.

2.4.5 Contributions

The contribution system used by medical schemes is 'pay as you go' as the contributions are paid monthly but not accumulated. A member loses access to the benefits once he or she discontinues from making the monthly contributions. The size of your monthly contribution depends on which of the medical scheme options is chosen.

There are two main types of contributions, the one that is based on a monthly payment and the newer form of payment called the Medical Saving Account (MSA). The monthly payment which is more popular involves every principal member contributing to the scheme's pool of funds. The funds are then used to settle claims by members of the scheme for the benefits listed as part of the core benefit packages such as hospital care and chronic illness.

On the other hand, the MSA which is available in some medical schemes makes use of the idea of a savings account which member pay money into for their monthly health expenses such as the purchase of drugs from the pharmacist, dental care and eye care which are not covered by the core benefit packages. The money in the MSA is accumulated and can be carried over from year to year.

Now that we have discussed the properties of health insurance and the roles within the health insurance ecosystem, we would unpack the health insurance claim process which shows how the entities identified in (Section 2.4) interact with each other.

2.5 The Medical Billing Process

The section analyses the traditional way a health insurance claim is processed. It does this by looking at what happens at the individual phases of claim processing. The purpose of the health insurance claim process is for service providers to receive reimbursements for the services they provided. The reimbursements can be paid to patients or insurance providers. Figure 2.3 shows the flow of activities in the health insurance claim process. We elaborate on the steps identified in figure 2.3 to give an understanding of what happens in the entire process. For further clarity, we use a fictional representation of each of the entity: Mr. Kyle is the insurance subscriber, Dr. Brad is the Healthcare Provider, Next Insurance the insurance carrier.

The health insurance claim process begins at the point where Dr. Brad who is a medical practitioner, treats Mr. Kyle and sends the cost of the treatment to the designated carrier. The carrier, in this case, is Next Insurance. Next Insurance then analyses the claim and determines which services were provided and if the provider needs to be reimbursed.

There are several steps involved in processing a health insurance claim. A brief review of the steps involved in the medical insurance claim process involves Mr. Kyle who is the patient receiving care from the licensed practitioner Dr Brad. In step 1 from figure 2.3, Mr Kyle is admitted for treatment and the admission process involves capturing of Mr Kyle’s biodata and retrieving his medical records by Dr Brad. Mr Kyle’s insurance details are also collected during the admission process to determine the type of medical plan Mr Kyle belongs to. The information is used to determine what treatments are covered by the provider and the treatments that are not covered.

In step 2 of figure 2.3, Dr. Brad attends to Mr. Kyle. He performs all the necessary diagnosis and treatments. Dr. Brad keeps a record of all the services provided as he would be required to provide the record of treatment when submitting a claim for all the services he provided Mr. Kyle. Dr. Brad then updates Mr. Kyle's medical records to reflect the treatments carried out.

Step 3 involves Dr. Brad recording all the services provided to Mr. Kyle and the relevant International Classification of Diseases (ICD) codes if the diagnosis were made or Current Procedural Terminology (CPT) codes if Mr. Kyle was treated [11] are attached. The ICD code is a system of coding used by physicians for classifying all diagnoses and symptoms. The CPT code, on the other hand, is a medical code set that physicians use in reporting medical or surgical procedures carried out, Mr. Kyle's biodata such as age, name, and sex, along with his insurance information are also captured and added to the bill for claim processing.

After these claims have been prepared by the healthcare service provider, the next phase is the pre-processing of the claims by the clearinghouses as seen in step 4. The clearinghouses serve as the third-party or intermediate between Dr. Brad and Next Insurance. The Clearinghouses act as the central hub where all the insurance claims are brought to be sorted and sent through to the various insurance carriers [7]. The clearinghouses are necessary because of a large number of claims that the healthcare services providers submit daily, and all these claims go to different healthcare insurance carriers.

Once the clearinghouse is done with pre-processing the claim. The claims are sent to the correct insurance provider which in this case would be Next Insurance. Processing these claims involves taking the details of the claim form and then lining it up to a policy and then ensuring that these claims correspond to a rule set out in the policy related to the healthcare plan. Processing a claim can be a tedious and difficult task when done manually as these claims come in different formats. Certain technologies such as Optical Character Recognition (OCR) have been used to improve the speed and accuracy of claim processing. Software systems have been used to capture health insurance claim, thereby reducing the possibility of unreadable information and reduce the risk of error in retrieving the information on the claim form. OCR equipment has also been used to process hard copy claims for efficiency and to achieve more accuracy [11].

Next Insurance has an internal team of actuaries who analyze these claims to make sure they are legitimate and calculate the payback amount. The actuaries develop heuristics around fraud indicators such as claims history, financial distress and over utilization of services. Using these indicators, decisions are now made whether a claim is fraudulent. The claims which are picked out as fraudulent are sent through for further investigation, so it can be validated if a claim is fraudulent or not.

These claims are scored according to the likelihood of being fraudulent. These scores are grouped together with the claim value and the score given to each claim to determine if the case would be investigated further. The criteria for scoring claims are updated periodically to reflect the change in the environment and to guarantee that the healthcare claims process adjusts to feedback from investigations.

Once the claims have been verified, Next Insurance initiates the process of reimbursement for the services rendered. At the point where Dr. Brad, receives the reimbursement for services, then the claim process is finalized. Fraudulent claims can then be sent for further investigations and if found fraudulent, the prosecution of the perpetrator can subsequently follow.

Now we have described the processes involved in the healthcare claim process by identifying each entity in the health insurance ecosystem and then the unique role each entity plays within the system. The next section discusses what benefits exist when the traditional claim process is followed, and software use is limited. We also look at the disadvantages of the traditional claim process in the section that follows.

2.6 Advantages of the Traditional Healthcare Claim Process

In section 2.4, we described the traditional healthcare claims process as the activities that take place from the point a patient is admitted by a healthcare provider, the healthcare provider then renders service to the patient and then gets reimbursed for the services rendered by the insurance carrier. There are several factors that can be highlighted that makes this process effective. In this section, we would be discussing the advantages of the traditional medical billing process.

The traditional healthcare claims process is a mature system. The maturity in the healthcare claims process means that there are available and trained resources to carry out the tasks involved effectively. It is a working system and has served the purpose of reimbursing the healthcare service provider for the services they provided to the health insurance subscribers. The traditional healthcare claims process also effectively reimburses service providers. The evidence of growth in the health insurance industry as illustrated in figure 2.1 shows that this system is a working system.

The traditional approach to the health insurance claim process is simple and it is already integrated into the workflow of the actuaries. The actuaries have been trained over time for this process, hence there is an available skill set to carry out the work required. It is not a novel process as most insurance firms already have departments and structures set up to carry out the tasks required for the medical billing process to complete.

The next advantage is in terms of cost. The traditional billing process has some elements of automation in it as described in step 5 of figure 2.3. This has led to the saving of cost as the electronic systems used for submission of claims has reduced the amount of paper used and the cost involved in posting these claim forms. Claims can now be stored and transferred online, eliminating the need for physical copies to be sent through the post as the claim form goes from one entity within the healthcare ecosystem to another entity.

In terms of access to the patient's medical files, the medical billing process has made medical records as patient medical records are now stored online. Medical specialists and other professionals can have a single view of the patient's medical record since the medical records are stored in a central repository. The single view of medical records is more reliable, unlike the paper-based system which can be fragmented between different doctors.

We have seen the advantages of the traditional health insurance claims process. However, there are several disadvantages associated with this process. These disadvantages are discussed in the section that follows.

2.7 Disadvantages of the traditional approach to health insurance claims process

The disadvantages of the traditional health insurance claims process can be analyzed from two perspectives. We can look at the challenges with implementing the system and then the effect of the human factors (such as the actuaries and assessors) on the fraud detection process.

2.7.1 Challenges with implementing the traditional health insurance claims process

The first challenge with the traditional approach to a health insurance claim is the cost involved in purchasing the system initially can be very high as the necessary software and hardware need to be purchased. Another area of high cost is the cost of resources. Actuaries are highly involved in processing healthcare claim and acquiring resources with the necessary experience in actuarial science can be very expensive.

The danger of information loss or claim form damage can also occur when software used to submit the patient's files are not compatible with the health insurance carrier's software. There are several software products available in the market used by the different individuals involved in the claim process. These software products need to be compatible to avoid a disruption in the process. For instance, the software used by the service provider in completing the claims needs to be compatible with the one used by insurance clearinghouses to avoid loss of claim information that can arise from file corruption. Using the wrong software can lead to the healthcare service providers not getting reimbursed or the healthcare service provider being wrongfully reimbursed.

The medical billing process has grown into a complex and specialized process over time as can be seen later in figure 3.1, creating demand for an individual with specialized domain knowledge to maintain the system and these resources are not readily available. The actuaries that carry out healthcare claims processing are trained to have very strong statistical and mathematical skills. The knowledge required to carry out the claims processing are highly specialized and the supply of the actuaries with these skills is limited.

Now we have seen the problems that impact the implementation of a fully functioning healthcare claims system. The next set of problems we discuss are the issues that arise when the actuaries try to process the claims and distinguish between legitimate and fraudulent claims.

2.7.2 The impact of the manual intervention on the health insurance claims process

The traditional health insurance claim process faces challenges as there is a heavy reliance on human intervention. Actuaries and other human elements play a major role in the medical billing process, hence any human possible error that can when the actuary or human resource carries out his/her duties consequently filters down to insurance claim process. The manual intervention leads to limitations such as rigidity in the ways the process works, difficulty with incorporating changes to reflect the environment and slow feedback on claims.

The actuaries are compelled to work with a constrained arrangement of known parameters using prior knowledge while staying aware that a portion of the different characteristics could likewise influence choices. The process has been predefined and the parameters that influence the process also mapped out by the actuaries. The traditional health insurance claim process assumes that the health insurance environment is static and does not adjust or accommodate the fact that some outlier behaviors can occur within the process, hence missing out on quick identification of market opportunities or threats [12].

Another limitation is the potential inability to comprehend context specific connections between parameters (geology, client portion, insurance sales process) that may not reflect the ordinary picture. Insights that can be generated from this process are limited since most computation and analysis are human-dependent [12].

As discussed (in section 2.5), the 5^{th} step of the medical billing process, claims are given scores based on how likely they are to be fraudulent. The scoring is updated periodically to guarantee that it adjusts to feedback from investigations. The process of incorporating these feedbacks into the existing system can be challenging since it has a manual dependency on human resources to do the calculations and recalibration.

The process is also slow, as the number of claims to be processed by individuals is very high since it is highly dependent on human beings. The time spent processing claims increases thereby delaying feedback to the client. With the growth in the insurance market as can be seen in section 2.2.2, the number of subscribers has also increased leading to more claims being submitted. This means that the number of claims to be processed by an actuary has also increased resulting in an increase in the time taken for the actuary to give feedbacks on the claim.

2.8 Conclusion

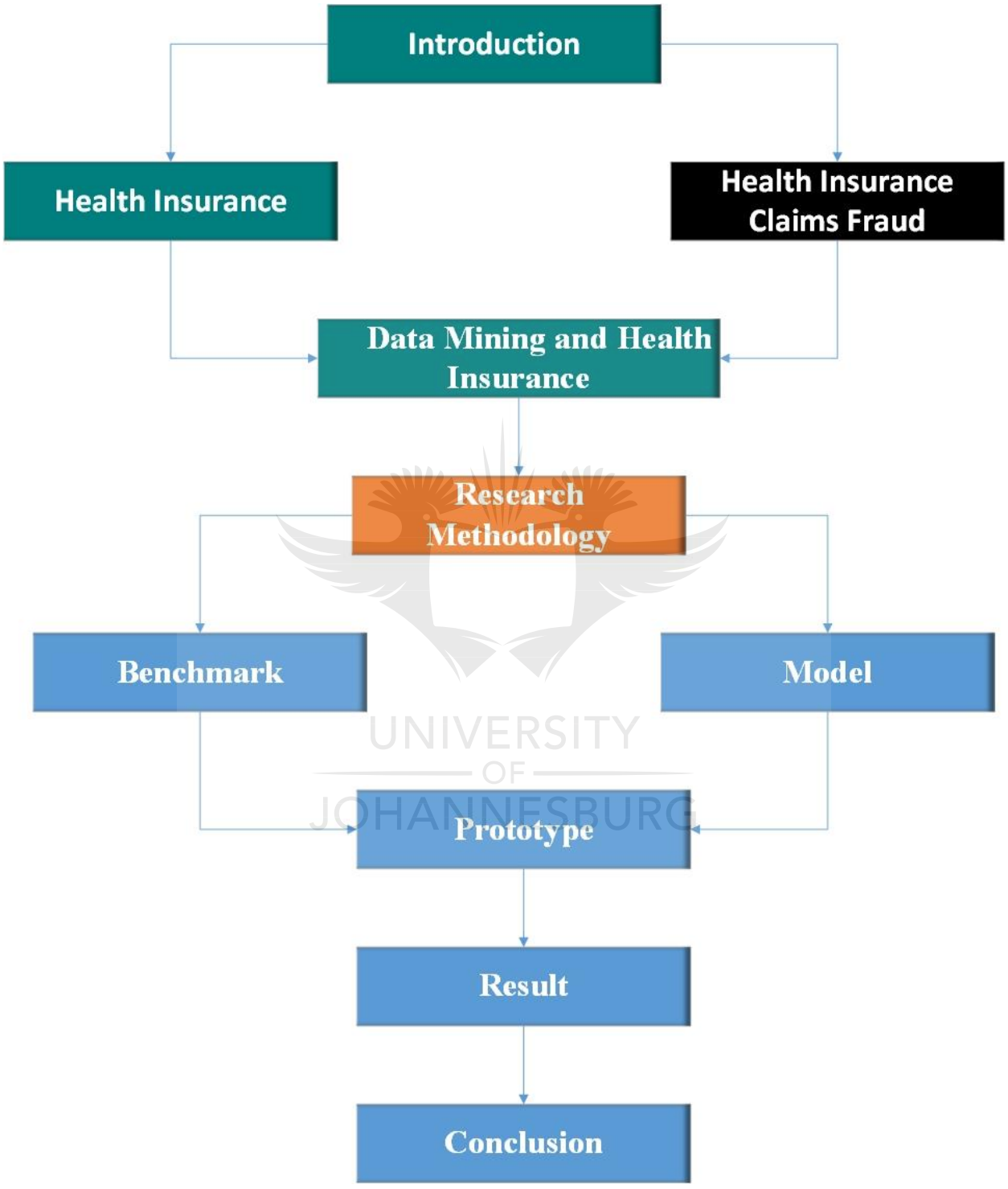
Health insurance is a type of insurance product that covers medical and surgical expenses of the member. The medical insurance scheme process hasn't been always streamlined but took several modifications to get to what we know today. We saw how the cost of medical care grew as advances in medicine were made and the demand for medical care also grew. At the initial phase, there was no need for medical insurance due to the low cost of hospital care. However, the advances in medicine led to the rise in the cost of medical care. The rise in the cost of medical care then led to the need for medical insurance. For any individual who doesn't have enough money saved, getting access to funds at the eleventh hour can be a very challenging task. The need for health insurance cover has continued to rise, leading to the continuous expansion of the insurance industry.

Next, we analyzed several entities that interact with the health insurance ecosystem. They include the service provider, the insurance subscriber, the health insurance carrier and the clearinghouses. The medical billing process shows the interaction between these entities. The medical billing process starts at the point where a patient receives treatment from the service provider. The service provider then prepares a claim to get reimbursement for the service he or she rendered to the insured patient. The claim information is then sent to the clearinghouse. The clearinghouse acts as the middleman between the service providers and the health insurance carriers. They sort and pre-process these claims before sending them to the relevant insurance carrier. The insurance carrier then processes the claim to determine if the claim is fraudulent or not, before making payments.

Finally, we discussed the advantages and disadvantages of the traditional health insurance claim process. The process is mature leading to a vast awareness of the steps required to effectively process healthcare claims. The rigidity of the process, when compared to the ever-changing environment, has led to difficulty in keeping the process updated to reflect the change.

The medical billing process has been plagued by many fraudulent activities as the several entities involved in the process can be victims or perpetrators of the fraud. Several measures have been put in place by the insurance company to pick out these fraudulent claims, but they have not been very effective as several fraudulent claims have gone undetected. The next chapter discusses health insurance claim fraud and we specifically touch on how complexity in the health insurance industry has facilitated fraud. We also look at the impact of fraud in the health insurance industry and the current measures in place to combat fraud in the healthcare industry.

Dissertation roadmap



Chapter 3 Health Insurance Claim Fraud

3.1 Introduction

In the past, the cost of healthcare was not seen as an important or significant household expenditure. Healthcare was not as advanced as it is today and due to the simplicity of healthcare processes, the cost of rendering healthcare services was inexpensive. The healthcare payment process used to be a basic process that simply involved a patient receiving treatments from a physician and then making payments for the services.

In recent times, healthcare cost has increased due to legal, socioeconomic and demographic changes. This increase in cost has influenced both the government and the private insurance systems [13]. The increase in the cost of healthcare can be attributed to the evolution of the healthcare system. The evolution of the healthcare system has led to the multiple processes and several other systems illustrated in figure 3.1. Due to the addition of the new systems and processes, there has been an increase in the complexity of the healthcare process. The increase in the complexity of the healthcare system has led to an increase in the cost of healthcare [14].

As discussed, the increase in cost and complexity of the healthcare coupled with other factors such as the economic state of the country has created an avenue for fraudulent activities to be introduced in the healthcare system [15]. Chapter 3 builds on the discussions from chapter 2 on the health insurance ecosystem to show how fraud can occur in the medical billing process. We discuss in section 3.2 an overview of fraud in the medical billing process followed by a definition of what fraud is, to gain a better understanding of the concept of fraud in section 3.3. We then analyze the impact fraud has on the health insurance industry in section 3.4. Finally, to gain an in-depth understanding of how fraud occurs in the health insurance industry we identify the different parties that can be involved in these fraudulent activities in section 3.5.

3.2 The definition of health insurance fraud

Fraud has affected many facets of life and from the discussion above, the healthcare sector is no stranger fraudulent activities. To gain a clearer understanding of what healthcare fraud is, a

distinctive comparison is given between the terms fraud, waste and abuse as these terms are sometimes misused.

Health insurance waste in healthcare is most times unrelated to fraud as it is mainly the provision of unnecessary health services. Waste can only be fraud and abuse when the act is intentional. Waste can occur when services are over-utilized and then results in unnecessary expenditure. When there is an unjustified consumption of services or unnecessary expenditure without adequate return then waste has occurred [16]. The key criteria that differentiate waste from fraud and abuse are the waste is non-intentional but once there is a motive behind healthcare waste, then it becomes fraud and abuse. For example, waste can occur when a physician provides medically unnecessary services without the intention of exploiting the system [16].

Health insurance abuse describes the billings of practices that either directly or indirectly, is not consistent with the goals of providing patients with services that are medically necessary, meet professionally recognized standards, and are fairly priced [16]. Abuse occurs when the practices of the service provider are not in line with sound business practices. It involves medical practices that may result in incurring an unnecessary cost to the insurance carrier, reimbursement for medically unnecessary services, or substandard services that do not meet the professionally recognized benchmark. Abuse is different from fraud in the sense that for abuse there is no requirement to prove that the abusive practices were carried out knowingly, willfully and intentionally [17]. Examples of abuse include but not limited to billing for a service that is non-covered by the insurer, misusing codes on the claim form in a way that does not comply with the general coding guideline.

Health insurance fraud is purposely billing for services that were never performed and/or supplies not provided, medically unnecessary services as well as altering claims to receive higher reimbursement than the service produced [16]. Fraud is when healthcare is paid for by the insurance subscriber, but healthcare services are not provided or a situation whereby reimbursements for services are paid to the service provider while no such services were provided. Fraud in healthcare can also be described as situations where healthcare service providers receive bribes or patients trying to receive prescriptions that are potentially harmful to them (seeking

prescriptions to satisfy addictions) and when service providers prescribe services that are unnecessary [18].

Now we have discussed the differences between waste, abuse, and fraud. The next section discusses what leads to fraud. We look at the drivers for fraud in the healthcare industry and how these factors have affected the health insurance industry.

3.3 Causes of fraud and abuse in health insurance

Over the years, the insurance industry has placed itself as one of the basic pillars of our modern society. A large volume of transactions is made daily in the health insurance industry. The health industry is today a global industry that has multiple interdependencies with other industries. The growth in popularity and importance of the health insurance industry has led to greater exposure to the health insurance industry to fraudulent activities [19]. In this section, we analyze the various factors that create avenues for fraud to occur in the health insurance process. We analyze how the pressure on individuals, the volume of transaction, growth in complexity of the healthcare process and technological advancements have contributed to the growth of fraudulent activities in the healthcare sector.

An individual can be involved in fraud due to financial pressure. The pressure faced by individuals can be internal or external. Pressure can arise internally because of family problems or externally due to the struggling economy. Financial problems can put pressure on individuals leading them to seek opportunities to commit fraudulent activities to gain unlawful profits.

The growth in the health industry has created an opportunity for fraudsters to commit fraud. The growth in the health sector has led to a large volume of transactions of high monetary value being carried out. Aside from the pressure on individuals and growth in financial transaction volume in the health industry, another factor that facilitates fraudulent activities is the complexity of the entire healthcare process. Looking at the medical billing process described previously in (section 2.5), there has been a certain measure of complexity introduced compared to the system in place in the early ages of health insurance. The complexity of the healthcare billing process comes from the evolution of the healthcare sector over time from a simple process with few interacting entities to a complex system with several interdependent entities.

Complexity can be defined as a state of being intricate or complicated. The advances made in the healthcare process have led to a more complex system as seen in figure 3.1 which shows the growth of complexity in the medical billing process by looking at what the healthcare process looked like, before 1965 and what it looks like today. Prior to 1965 patients in need of treatment went to the doctor who assessed and treated the patient. The patient then makes direct payments for the services rendered by the doctor. Presently the system is no longer as simple as it used to be [14]. The introduction of health insurance has increased the number of entities involved in the medical billing process, therefore, increasing the complexity of the process. From figure 3.1, we see that post-1965, when a patient was sick, he or she visits the physician as seen in step 1, he or she might need a referral to be treated by a physician. After the patient receives the treatment from the physician, the physician then needs to document and attach the relevant code to the treatments carried out.

There are now several parties introduced in processing payment such as the government payers, managed care payers and commercial payers. The payment system now requires the physician to complete a claim form based on the healthcare carrier's guideline to receive reimbursements for services. The healthcare system has evolved from what was a simple system before 1965 to a complex system today. The complexity of the healthcare system today has created more opportunities for fraud to occur as there are now more parties involved and more processes that can be exploited by fraudsters.

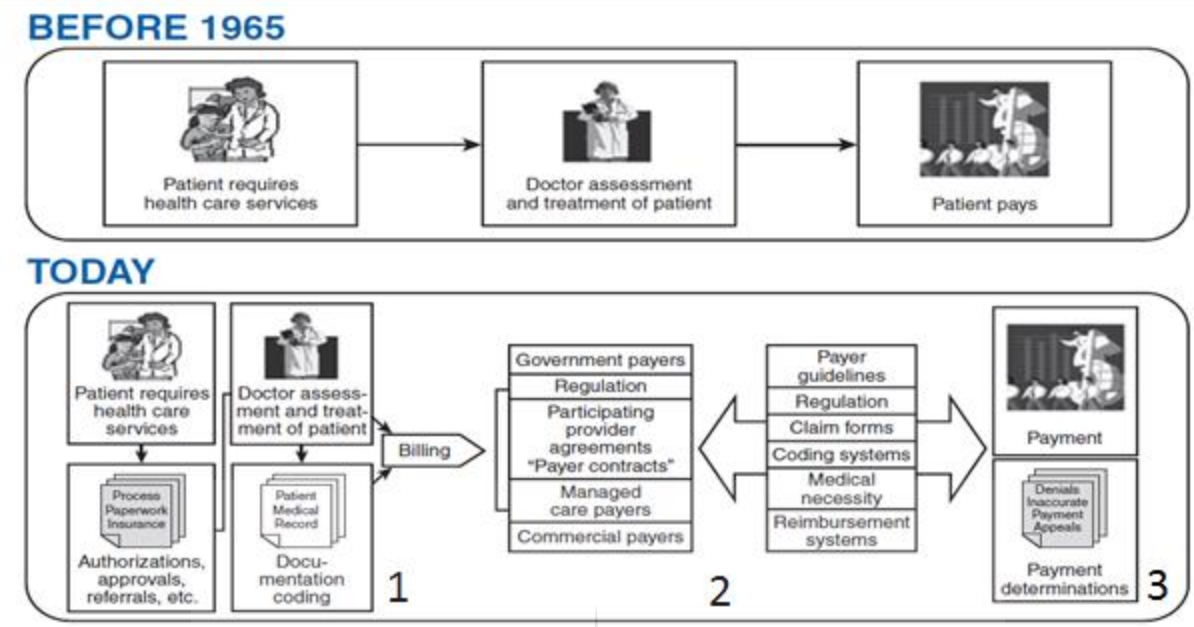


Figure 3.1 A comparison showing the growth in Complexity in Medical Billing Process (Taken from [14])

The technological advancements in the healthcare billing process have also played a role in the increase in fraud occurring in the healthcare billing process. Although technological advancements were made to improve the quality and efficiency of the healthcare billing process, the electronic data exchange and other technological advances have created further exposure to fraudulent activities. The reason for the exposure to fraud is the introduction of electronic claim transactions has subsequently led to an increased volume of claims and has created an avenue for these fraudsters to use technology to carry out fraud with a minimal chance of being detected [14].

Health insurance fraud occurs at different levels with different individuals involved. On each level of fraud in the healthcare process, there are different entities involved. The subsequent section describes the different levels of fraud and the various kinds of fraudulent activities that can occur at each level.

3.4 Levels of health insurance fraud

Health insurance claim fraud occurs on multiple levels with different individuals involved. These fraudulent activities in the healthcare industry can be grouped into irregular activities with the

following individuals possibly involved: staff, suppliers, managers, medical professionals, and affiliates. According to the party involved in the fraud, healthcare fraud can be categorized as follows: healthcare service provider fraud, healthcare insurance subscriber fraud, and healthcare carrier fraud [20]. These levels of fraud are subsequently discussed.

3.4.1 Healthcare service provider's fraud

Healthcare service provider's fraud occurs when the healthcare service providers (as described in section 2.4.3) use illegal and non-ethical means to exploit the health insurance system to gain some profit. They exploit the healthcare system by filling in dishonest healthcare claims. The healthcare service provider fraud takes different forms which include upcoding, unbundling, falsifying medical records and some other forms of fraud which we elaborate on subsequently.

Claiming for payment of services that were not performed: The healthcare service provider can do this by using legitimate patient information which can be retrieved legitimately or through identity theft, to create a false claim or by including additional non-rendered services or procedures that did not occur.

Unbundling is a type of fraud that occurs when a service provider claims for each treatment stage like it was a separate treatment. It can be seen as exploding medical reimbursements by the service providers. Service is broken down into individual separate services resulting in higher payment [21].

Upcoding occurs when more expensive services are billed rather the ones performed [21]. It involves a service provider billing the health insurance carrier using a CPT which was defined (in section 2.5) code for a more expensive treatment than the treatment that was carried out.

The physician can exploit the insurance system by performing medically irrelevant treatments with the aim of increasing claim value. There have been situations where the service provider decides to carry out treatments that are not necessary just to generate insurance repayments.

The physician can exploit the insurance system by misrepresenting uninsured treatments as insured treatments. They misrepresent treatments that are not covered as medically essential treatments to receive insurance payments. Misrepresenting treatments is prevalent in the cosmetic surgery

schemes where some treatments such as nose jobs which aren't covered by the scheme as deviated septum repairs

Intentionally misdiagnosing patient or falsifying a patient's medical history to be able to justify irrelevant treatments. Falsifying medical records endangers the patient as the subsequent treatments and diagnosis would be based on the wrong information. Now we have seen how the health insurance service provider can be involved in fraudulent activities, the next subsection discusses how fraud can be perpetrated by the health insurance subscriber.

3.4.2 Health insurance subscriber fraud

The insurance subscriber can be involved in committing fraud directly or indirectly. The health insurance subscriber can be directly involved when he or she directly commits a fraudulent activity without the influence or the help of a third-party. The health insurance subscriber indirectly commits fraud when a third party e.g. a medical practitioner carries out fraudulent activities with the knowledge and approval of the health insurance subscriber. We discuss the different fraudulent activities the health insurance subscriber can be involved in such as identity theft, claiming for services not performed and obtaining false prescriptions and submitting fake eligibility records.

The health insurance subscriber can fake employment or eligibility records to receive a lower premium rate. Insurance is a contractual relationship which participants are always supposed to act in good faith. This is not always the case as there are several instances where the subscriber might supply false information to the insurer to obtain lower premiums [22].

The health insurance subscriber can claim for services that were not actually received. The health insurance subscriber can also collaborate with the healthcare provider to submit claims for services that were not received. In doing so, he or she exploits the system to gain unlawful profits.

Claiming insurance benefits by illegally using another person's membership card. Using another person's membership details illegally is known as a form of identity theft [21]. The uninsured individual steals the insurance details of an insured individual and impersonates the insured individual in order to gain the benefits of the insured individual.

The health insurance subscriber can obtain false prescriptions with the purpose of reselling them or giving the drugs to someone else. They can obtain false prescriptions by inappropriately getting multiple prescriptions for drugs which can include narcotics as well as painkillers. Although the health insurance service providers and the subscribers are the main culprits when it comes to fraud in healthcare, the health insurance carrier can in some ways be involved in fraud. We look at how the insurance carrier can be involved in fraud in the next subsection.

3.4.3 Health insurance carrier fraud

It is very uncommon for the health insurance carrier to be involved in healthcare fraud nevertheless there are several actions which the healthcare carrier might perform that can be categorized as fraud which we discuss below.

The health insurance carriers can *fake reimbursements*. Although this is a rare occurrence, the insurance companies claim to have reimbursed service provider for services provided by creating fake proof of reimbursements.

Creating fake account statements for benefits/services to justify payments can occur as a result of the health insurance carrier trying to embezzle funds or evade tax. The health insurance carrier then provides fake proof of payment to justify the expenditure made.

The insurance carrier can intentionally fail to provide authorization for medical services required by members. They may deny the authorization because they want to avoid making payments for the treatment rendered and save the cost [22].

There are several situations where claim fraud is carried out by more than one party and that is known as conspiracy fraud. For example, when a physician conspires with the patient to claim for services not rendered. Their collaboration to intentionally defraud the healthcare system is known as conspiracy fraud.

Insurance claim fraud has been on the rise with the continuous decline in the economy [2]. Fraud is a growing problem as rising costs would lead to an increase in illicit activities for financial gain. From a global perspective, 6.99% of healthcare expenditure is lost to fraud on the average every year [23]. We would be unpacking the full impact of healthcare fraud in the next section.

3.5 The impact of fraud in health insurance

Initially, when health insurance started, fraud existed but the impact was minimal and could be easily ignored. But with the continuous growth in the size of the health insurance industry, the impact of fraud on the health insurance industry can no longer be ignored. As the money spent grew, the impact of fraud continued to grow. The number of criminal organizations targeting the healthcare industry has also increased [24]. The impact of healthcare fraud can be analyzed from two viewpoints: the impact on the lives of people and the financial impact.

Fraud in healthcare billing insurance can have safety issues which could come as a surprising factor, but the impact of fraud has effects that can subsequently affect the safety of the patient involved. Health insurance fraud has severe and real consequences on the quality of healthcare being rendered [24]. For example, when the insurance provider is involved by submitting a fraudulent claim to the health insurance carrier, the service provider who in most cases is the physician, falsifies the patient's health records to support the fraudulent claim. Over time, if these falsifications are not corrected, the patient's life may be endangered as future references to the patient's record would be inaccurate, thereby misguiding future diagnosis or treatments. Sadly, the fraudulent healthcare claim fraud can be difficult to identify and they can take a long time to rectify. Sometimes as in the case of falsifying medical records can lead to the death of patient [24].

The financial impact of healthcare fraud on a patient can occur when the patient is being charged for services not rendered. Financial loss can occur as a result of identity theft. For example, in 2016 a woman in Jacksonville arrived at the hospital for treatment for a critical injury and was told to her surprise that a thief already used up her insurance to have her leg amputated. The resulting effect was that she received a $200,000-dollar bill for the treatment she never received [25].

The next category of people affected is taxpayers. The public health sector is mainly funded by the taxpayer's money, meaning that most of the money lost to fraud in the health insurance claim process is the taxpayers' money. The taxpayers' money ends up going to criminals rather than being used in the effective development of the society.

Insurance companies also suffer because of fraudulent activities as they bear the majority of the financial brunt. It has a huge impact on the company's finances and ends up making the cost of

premiums higher which leads to restricted benefits to subscriber [25]. The loss of revenue to fraudulent claims then result in the insurance companies increasing costs across the board. Another impact is on the reputation of the insurance company. The subscribers tend to lose faith in insurance companies that have been subjected to data breaches that led to fraudulent transactions.

Although the doctors and other healthcare providers are mainly responsible for healthcare fraud, they can also be a victim in some instances. For example, in a situation where a doctor's medical license information or the doctor's National Provider Identification (NPI) is compromised, the doctor's identity can be used for unlawful activities. The NPI number is a unique identifier for healthcare service providers. The NPI is used by the clearinghouses and insurance carrier to identify the healthcare provider during administrative and financial transactions. The doctor can receive very high taxes for revenue never earned. The doctor might also be asked to return overpayments for services he or she did not provide. The process of rectifying a case of identity theft can take a while and can lead to a loss of license to practice and have an impact on the reputation of the practitioner [25].

The impact of fraud in the healthcare claim process can have an effect, from the practitioner losing his or her practice license to the insurance company losing revenue which in turn leads to higher premiums charged to providers. The most severe impact of healthcare fraud which is inadequate healthcare that can have a safety impact from falsification of a patient's medicals record by the practitioner [26].

From the discussion (in section 3.4), one can easily see that there are more cases of fraudulent activities in the service provider space. Hence several types of research are being carried out in this area. The focus of this research work would be mainly focused on the health insurance service provider fraud. We identify ways in which we can detect the following fraudulent health insurance service providers activities: unbundling, upcoding and misdiagnosing patient to justify payment for treatments. We tackle these problems specifically as they lend themselves towards automation. In this research, the healthcare carrier fraud and health insurance subscriber fraud do not form part of the scope.

3.6 Conclusion

Health insurance has transitioned from an industry of little significance to a very significant pillar of society. The need for health insurance grew as the demand and cost of healthcare increased. The health insurance industry like other sectors of society has been exposed to fraud and abuse. We provide a distinction between waste, abuse, and fraud to gain a clearer understanding of what health insurance fraud entails. We found that waste in the healthcare system was because of performing medically unnecessary treatments or over-utilization of services. Abuse involves medical practices that may result in incurring an unnecessary cost to the insurance carrier and reimbursement for medically unnecessary services. Fraud is when healthcare is paid for by the insurance subscriber but not provided or a situation whereby reimbursements are paid to the service provider while no such services were provided. We found that while waste was an unintentional act, fraud and abuse were performed intentionally.

The introduction and growth of fraud in healthcare have been promoted by factors such as financial pressure on the individual which can come from family or from external sources such as the economy of the society. The growth in the volume of transactions carried out in the health insurance industry also created an opportunity for fraudsters. The growth in the healthcare system has led to a rise in the complexity of what used to be a relatively simple process. The increased complexity then led to the introduction of several new processes to the healthcare billing process. With the increase in complexity, there was a greater avenue for fraud and abuse to be carried out in the health insurance process.

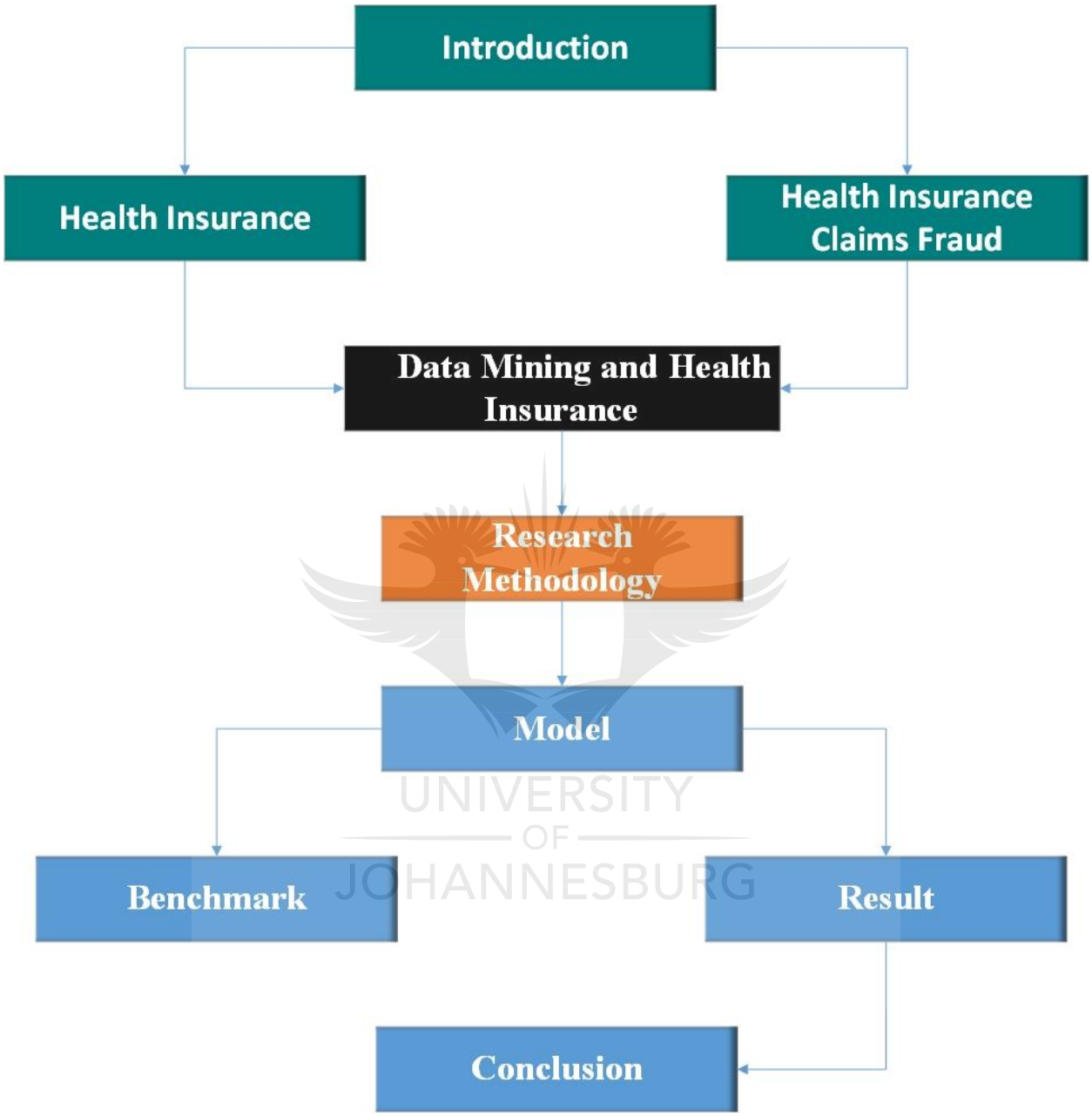
With an understanding of health insurance fraud and the causes of healthcare fraud, we looked at the different levels of healthcare fraud. For each level, we identified who can be involved in fraudulent activities and how fraud can be perpetuated. The healthcare service providers had more instances of fraudulent activities when compared to the other levels of fraud.

Finally, we analyzed the impact of health insurance fraud. Health insurance fraud can have both safety and financial impact. Health insurance fraud that involves falsification of a patient's medical record can lead to future misdiagnosis which is a safety concern for the patient. The financial

impact results from the money lost by insurance companies to fraudsters which then results in higher premiums.

As much as fraud has continued to grow in the health insurance industry, several measures have been set up to combat new and innovative ways fraudsters try to use in defrauding the health insurance claim process. In the next chapter, we would be looking at how data mining has been used to fight health insurance fraud.

Dissertation roadmap



Chapter 4 Data Mining and Health Insurance

4.1 Introduction

Technology has played a major role in the Health insurance industry. The impact of advances in technology can be seen in the various aspects of health insurance including the medical billing process. In the past, the medical billing and coding process was carried out with a paper and pen, then submitted using the mail system but today the professionals in the insurance field have found their workspace in the virtual and electronic world [27].

In an increasingly digitally connected world, technology has created the opportunity for the health industry to make changes in operations and go for a more streamlined system. The new form of the medical billing process has made it possible to easily track a patient's treatment records. In general, the introduction of technology in the medical billing process has reduced the amount of paperwork involved [28]. The medical billing process has also improved the number of successful treatments as it allows medical practitioners to look at the general population when treating an epidemic. Despite all the advantages technology has brought to the medical billing process, each of the stages of the medical billing process as described in figure 2.3, is reliant on human input and creates an avenue for fraudulent activities.

The adoption of technology in the medical billing process has led to increased sophistication in health insurance claim fraud. Most organizations that exchange money with the insurance carrier, service providers or customers are potentially exposed to the risks related to fraudulent activities. As can be seen (in section 3.4), health insurance companies are affected by fraudulent activities and they also lose a portion of their revenue through fraudulent activities.

Data mining techniques have in recent times been applied in the health insurance domain to detect these fraudulent claims. Data mining involves extracting, discovering or mining knowledge from a large amount of data. The availability of data and the advancement in technology allows for the

design of data mining systems that can extract previously unknown knowledge and insights from the available data [29].

In section 4.2, we would be looking at what data mining means and how it can be applied to health insurance claim fraud detection by reviewing the relevant literature on the different systems that have been built. An introduction to what data mining is given and then we look at the general applications of data mining in other domains to get more clarity on how data mining works and how it has been applied. We narrow these applications down to how data mining has been used in the insurance industry from customer profiling to fraud detection. Finally, we review work done on health insurance claim fraud detection.

4.2 Overview of data mining

A large amount of data available in the information technology world as well as the continued exponential growth of data has led to the need for this data to be converted into useful information. The growth and availability of data stored in several repositories such as files and databases have created the need to generate insights on the data as organizations and businesses seek to gain value from the data being stored. Knowledge can be generated from data through powerful means of analysis and pattern interpretation of the data which can subsequently aid in decision-making using knowledge gained from the data [30].

The terms data mining and knowledge discovery are used interchangeably. It is the nontrivial extraction of implicit, previously unknown and potentially useful information from data in a repository [31]. Data Mining is the extraction of information from a large volume of data. It refers to the extraction of knowledge from large datasets. Data mining enables us to filter through immense volumes of data to find unknown or hidden patterns that can give new perceptions [32]. Although the terms data mining and knowledge discovery in the databases are synonyms, data mining forms a part of the knowledge discovery system as can be seen from the figure below. In the context of the dissertation, we are going to use the terms knowledge discovery in database and data mining interchangeably.

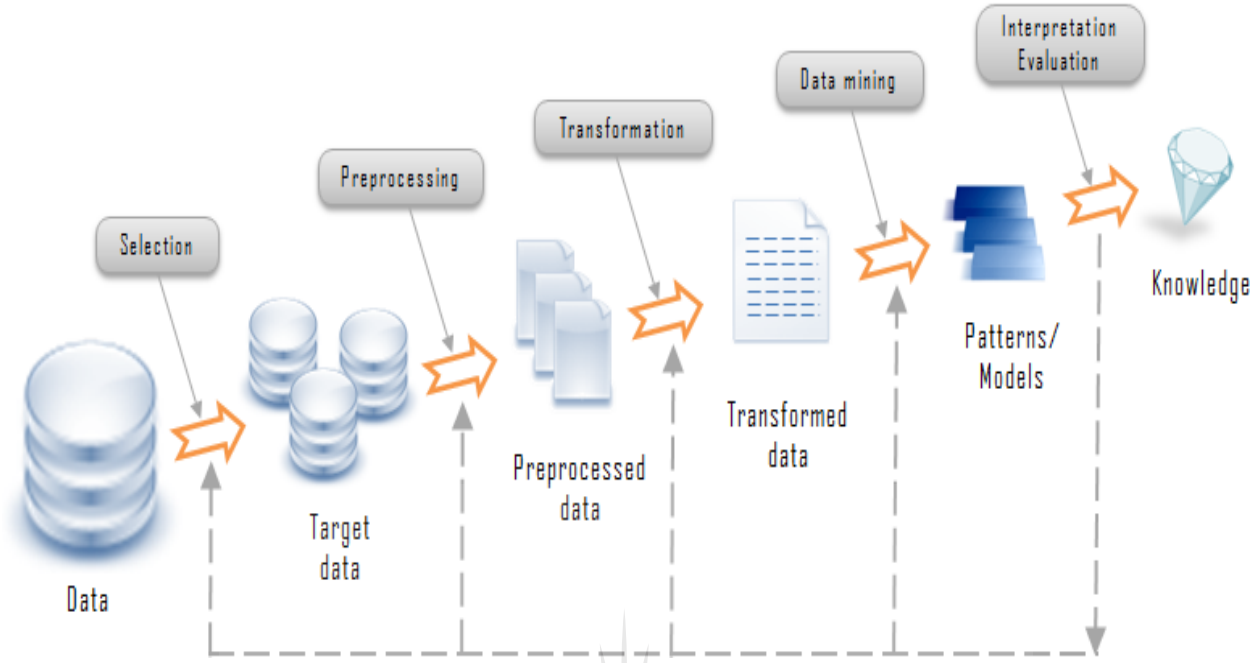


Figure 4.2-1 Diagram illustrating the processes in the data mining system adapted from [33]

Figure 4.1 shows what a data mining system looks like. In a data mining system, we need some methods to capture the data for processing. Once the data has been captured and store, the next step is cleaning the data to get it to a desirable state. Due to the volume of the data and potential noise, we also need to pre-process data. Pre-processing the data to get it to a transformed state involves feature extraction and selection. The next is the application of data mining methods and algorithms to the data. Once we have applied the data mining methods and algorithm to the data, we can then benchmark the prototype. In the following section, we elaborate more on the different steps in figure 4.1.

4.3 The data mining system

In the process of data mining, the end goal is to discover knowledge or useful insight but before data mining can occur, the data needs to be collected, pre-processed and transformed before data mining algorithms can be applied [30]. In this section, we discuss the different stages of the knowledge discovery system.

4.3.1 Data capturing

Data can have several formats. Data can be structured, unstructured and semi-structured [34]. The form of data determines the methods used to capture and process the data. Structured data consist of data that can easily be persisted in a database. Structured data follows a predefined schema. Structured data lends itself to easy processing and navigation of content. Structured data contains a relational key which can be used to map it into a pre-designed field in the database. Examples of structured data include an excel spreadsheet and relational database table data.

Unstructured data can be in the form of text or multimedia. Data is referred to as unstructured when there is no identifiable kind of structure within the data. Examples of unstructured data include video, photography, scientific data, and mobile data. Although unstructured data may have an internal structure, they are still categorized as unstructured as they cannot easily fit neatly into a relational database. To fit unstructured data into a relational database would involve a significant amount of processing power.

Semi-structured data is a class of data that consists of data that cannot be stored in a relational database, but the data has some organizational properties that allow for easier analysis. Semi-structured data can be processed to allow storage in a database through data transformation. Examples include CSV, JSON and NoSQL database.

The different types of data require different methods of collection and extraction. For structured data, it is easy to extract the data from the database using SQL queries. For unstructured and semi-structured data, some transformations need to be done to allow for storage in a relational database. The data collected and stored most times does not come in the desired format needed. We need to apply some modifications and enhancements to the data. The next subsection discusses how this can be achieved.

4.3.2 Pre-processing

Pre-processing data is a very important activity in data preparation. Real-world data is not presented in the perfect format that we need it to be. Data in the real world is “dirty”. Real-world data can be incomplete, noisy and inconsistent. Data is incomplete when certain attribute values are missing, or data is lacking some attributes of interest. Inconsistent data contain discrepancies in the features. All these characteristics of data lead to the need for data to be pre-processed before use. The tasks involved in pre-processing data include data cleaning, data integration, and data transformation [35].

Data cleaning is the process of identifying and removing corrupt or incorrect records from a dataset. Data cleaning allows the detection of irrelevant parts of data and then carrying out several actions such as replacing, modifying or removing the dirty data [35]. We can replace missing values by ignoring the tuple, using the attribute mean, or by predicting missing values. We also use data cleaning to identify and smooth out noisy data and we can correct data inconsistency by using domain knowledge.

Data transformation and integration can be done through normalization i.e. scaling attribute value to fit into a certain range. Data integration is done to combine data from several sources into a more coherent data source [35]. Data transformation involves consolidating data into forms that are appropriate for mining. Data transformation can be done by normalizing, aggregating and generalizing data. Once the data transformation is complete and we have the data in an acceptable form, the next challenge faced is that in many cases the data can have several features which are not all important. The next subsection discusses how we can retrieve the features that are relevant to what we are doing.

4.3.3 Feature Extraction

Insurance claims data has grown over time as the health insurance industry grew. The extraction of claims data is done from several large volume databases. The database size continues to grow with every new claim submitted. As data continues to grow exponentially, we have limited time, space and processing capabilities to work on the data. Feature extraction methods are used to

retrieve only the required information. Feature extraction is a form of dimensionality reduction and it forms an important step in the data mining process. The feature extraction technique is used to retrieve a subset of new features derived from the original set by means of some functional mapping retaining as much information in the data as possible [36]. The process of feature extraction creates new variables as a combination of the existing variables to reduce the dimensionality of the selected features.

Given that the health insurance claims information may contain some data that are not required, we use feature extraction to extract some pertinent features of data before we run any pattern discovery algorithm on it. Features that can be used include many indicators such as the kind of services rendered, the number of services rendered and the provider's specialization.

One of the most commonly used feature extraction techniques is the Principal Component Analysis (PCA). Other feature extraction methods include latent semantic indexing, clustering methods and a bag of words. These feature extraction methods will be fully addressed in section 6.5. Once we have the features, we can then further improve the quality of the data by reducing the redundant features present. Feature selection enables us to reduce redundant features and we discuss it in more details in the next subsection.

4.3.4 Feature Selection

Feature selection is a mechanism in data mining used for dealing with redundant features. Redundant features may be irrelevant and further increase the complexity of data [37]. Hence it is important to eliminate the redundant features. Features should be selected such that performance is not reduced and the output remains the same. Three categories of feature selection algorithms which include filters, wrappers, and embedded techniques [37]. Filters perform feature classification without using any learning mechanism. On the other hand, wrappers work by applying learning techniques and they perform better than filters. The embedded methods are a combination of several approaches [38].

At this stage of the data mining process, we have collected and persisted the data, pre-processed the data to remove errors, reduce the dimensionality of features and eliminate the redundant

features. The data we have now exists in the desired format that allows for the application of data mining techniques to gain knowledge from data.

4.4 Data mining model formation

The data mining model creation is a crucial phase that involves the application of techniques to extract potentially useful insights. We analyze the data mining model in this section by analyzing how learning occurs in the data mining model and the different types of data mining models. We finally discuss some of the common methods that are used in data mining operations.

4.4.1 Learning for data mining

In data mining, there needs to be some form of learning. Data mining learning methods can be classified in different ways. The classification method used depends largely on the kind of data being processed, the type of knowledge to be discovered and the algorithms utilized. Machine learning experts have divided these classification methods into two categories, the supervised and unsupervised methods. Supervised methods try to find the relationship between input variables and output variables, while unsupervised methods are applied when there no prior knowledge of the output (dependent) variables [39].

Supervised methods can be employed when the aim is to classify or predict outcome using statistical methods such as regression analysis, Bayesian networks, discriminant analysis, neural network and Support Vector Machine. The unsupervised methods are mainly used for descriptions including association mining rules like Apriori algorithm and segmentation methods like clustering and outlier detection [39].

In the context of detecting fraud in the medical billing process, supervised data mining learning methods utilize sample datasets of previously identified fraudulent and non-fraudulent claims. Using these two known classes of datasets, a model can then be constructed which permits the addition of new records to each category based on the observations. A measure of confidence is required when adding to the dataset classes. They are useful when trying to detect fraudulent acts that take the form of previously known patterns hence a continuous update of the model is required to reflect new patterns of fraudulent behaviors.

One limitation the supervised data mining learning methods have is that they can only detect known fraud patterns. The fraudsters keep mutating their techniques of fraud to beat the fraud detection system, thereby creating unknown fraud patterns. Unsupervised data mining learning methods detect fraud by comparing similarities or difference between records. It can detect sequence and association between records and pick up similar records within a cluster or the anomalies.

As earlier discussed, data mining is the discovery of unknown patterns from data. There are several tasks that can be performed in data mining such as classification, neural networks, regression, clustering. In the subsequent subsection, we will be discussing some of these data mining tasks.

4.4.2 Types of data mining tasks

Data mining models vary from one application domain to another. The data mining models deal with all kinds of patterns. Data mining uses different kinds of techniques and methods to identify a different possible pattern. Depending on the type of pattern we are looking for, we can make use of classification, cluster and regression tasks to solve the problem at hand.

Classification is a commonly used data mining technique. It uses a sample set of the pre-classified dataset to build a model that can classify the entire population of data [40]. Detection of credit card fraud and risk analysis would be well suited to this analysis. Before classification can occur, learning needs to take place. In the learning phase, we make use of training data which is analyzed by the classification algorithm. Test data are used to benchmark the accuracy of the classification algorithm. Once an acceptable accuracy is reached, the classification rules are then applied to the entire population. The types of classification models that exist include classification by decision tree induction, association rule-based classification, Bayesian classification, neural networks and support vector machines [41].

Clusters indicate similarity and in data mining clustering involves the identification of similar classes of objects. Using clustering we can discover how the entire data set is distributed and see the different sparse and dense regions in the dataset. Although classification algorithms can be used to detect similar classes, it becomes expensive over time leading to the use of clustering as a prepossessing algorithm for the selection of attribute subset and classification. The clustering

methods that can be used include density-based, model-based, grid-based, partitioning and hierarchical agglomerative [42].

Regression analysis can be used for prediction. Using regression analysis, we can build a model that describes the relationship between dependent and independent variables. Independent variables are known variables used to predict the response variables. Types of regression models include linear regression, multivariate linear regression, nonlinear regression and multivariate nonlinear regression [43].

When searching for frequent items among large datasets we make use of **association** rules and correlations. Association describes togetherness or link between several entities. It reveals the existing relationships between different objects [44]. Association rules can be used to discover interesting correlations between objects in a database. One example of an association rule is in the discovery of frequent items in a shopping basket. The identification of frequent items can be used for catalog design and analyzing customer buying behavior. The types of association rules include multilevel association rule and quantitative association rule [45]. Association rules can be used to predict the cause-effect relationship between antecedent and consequent. The association rule shows that the occurrence of the antecedent implies the occurrence of the consequent [46].

For each of the data mining tasks discussed in this subsection, there are several data mining methods and algorithms that can be used for classification and regression. The next section discusses some of the common data mining techniques available.

4.4.3 Data mining methods and algorithms

Data mining leads to the discovery of previously unknown knowledge. To discover the knowledge or patterns within datasets we make use of several data mining methods to carry out the data mining tasks that eventually leads to knowledge discovery. In the following subsections, we discuss some commonly used data mining algorithms and methods such as neural networks, decision trees, Bayesian network, and random forest.

Bayesian Network

The Naïve Bayes classifier is a simple classifier that learns by assuming that the features are independent given class. Although the simple idea of independence is generally a poor assumption, practically the Naïve Bayes competes well with the more sophisticated classifiers. The Bayesian network belongs to a class of probabilistic graphical model that is used for knowledge representation of an uncertain domain [47]. The nodes that make up the Bayesian network contains random variables while the edges between the nodes represent the probabilistic dependencies among the random variables. These probabilistic dependencies are determined by the application of statistical and computational methods [47].

The Bayesian networks which are sometimes called the belief networks are graphical models for representing probabilistic relationships a collection of variables. They can be used to represent knowledge about an unknown domain. Bayesian networks combine knowledge from graph theory, statistics, the theory of probability and computer science [41]. In a Bayesian network, each node of the graph represents a random value and the probabilistic dependencies between the nodes i.e. the corresponding random variable represent the edges.

The nodes and edges of a Bayesian network form a directed acyclic graph. The Bayes' rule applied to the graph makes it possible to efficiently compute belief propagation algorithms. Bayesian networks provide a representation of the potentially complicated world in a succinct graphical form. The compact representation reduces the number of parameters required when describing a complex world state.

Bayesian networks are not black box models as humans can be involved in the construction of the network. A subject matter expert is usually involved in defining the random variables as well as the topology of the network. The computational cost of doing full Bayesian learning is extremely expensive and they perform poorly with small datasets. They also do not perform well when modeling relationships between random variables.

Logistic Regression

In the task of binary classification, the logistic regression is a common method used. A binary response typically takes the form of 1/0, with 1 generally implying true or success and 0 implying false or failure. The value of 1 and 0 can vary widely depending on the goal of the study. For instance, in a study about the odds of a failure in an academic setting, the number 1 may represent a pass and the number 0 may represent a failure. The use of a normal linear regression to model a binary classification problem introduces a substantial bias into the estimated parameters. The linear regression makes the following assumptions: the error and response terms are either Gaussian or normally distributed, the variance is always constant across all observations and these observations are independent. Modeling a binary variable using the linear regression the assumption that error and response terms are either normally or Gaussian distributed and that the variance is always constant.

Similar to the linear model which is based on the Gaussian probability distribution function, the binary classification is based on Bernoulli distribution. The binary logistic regression model is derived from the canonical form of the Bernoulli distribution. Logistic regression is similar to linear regression in the way it works but logistic regression has a binomial response variable. The logistic regression model allows the use of continuous explanatory variables and can easily handle simultaneously two or more explanatory variables.

The logistic regression model will show the probability of an outcome depending on the characteristics of individual entities. Since chance is a ratio, we define the logarithmic of the chance by the formula below:

$$log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m \qquad (1)$$

where π indicates the probability of an event, and β_i are the regression coefficients associated with the reference group and x_i the explanatory variables [48].

Logistic regression has its own pitfalls. Dealing with continuous explanatory variables or variables that have more than 2 levels can be more difficult. We think of what happens when the explanatory variables are no longer binomial, which leads to the need to create n-1 binary variables, with n=

number of levels of variables. Another difficulty is in identifying independent variables. Logistic regression uses independent variables to predict outcomes. If the incorrect independent variable is chosen, the model will have very little predictive value. The logistic regression model are also prone to over-fitting. Finally, logistic regression can predict categorical outcomes as well as multinomial outcomes. However logistic regression cannot be used to predict continuous outcomes.

Ensemble Learning

Ensemble learning methods make use of a combination of weaker learners to create a stronger learner. They combine multiple hypotheses, with the aim of deriving a better hypothesis than the original best hypothesis alone [49]. A weaker machine learning, in this case, is one the can consistently arrive at better predictions than random predictions. Ensemble learning trains several of these weak learners and then combine their weighted results.

An example of an ensemble learning model is the random forest classifier which combines multiple decision trees. The forest comprises multiple trees that makes use of multiple random data features as input. The resulting prediction in the random forest classifier is decided by weighted voting.

The random forest classifier makes use of a small number of parameters and has considerable resistance to over-fitting [49]. With the random forest classifier, there is no need for feature selection as they can use many attributes. The variance of the random forest model decreases as the number of trees increases, while bias remains the same. Random forest models are difficult to interpret and lose performance due to correlated variables.

Random Forest

The random forest machine learning method is a supervised classifier based on ensemble learning algorithm that creates numerous individual learners. They are built from a combination of various predictions of multiple trees each of the trees trained in isolation. The difference between the random forest algorithm and boosting is that while the boosting method trains base models and combines them using a sophisticated weighting technique, the trees in the random forest are trained independent of each other and they are combined by averaging the predictions [50].

The random forests model is a very popular ensemble learning method that makes use of a combination of tree predictors. In the random forest model, each tree that makes up the model uses values of a random vector distribution that is independently sampled combined with the same distribution for all trees that make up the forest. It uses the idea of bagging to build up a randomized set of data for constructing a decision tree [51].

Unlike the standard tree where the node is split using the best split among all variables, the random forest splits each node based on the best among a subset of predictors which are randomly chosen at the node [52]. When building up a random tree, there are three important choices to be made: the methods to be used in splitting the leaves, the choice of the predictor to be used in each half and the method to be used to inject randomness into the tree.

In order to specify the method to be used for splitting the leaves, there needs to be a selection of the shape of the candidate split as well as the method for determining the quality of each split. Splitting a leaf involves generating a collection of candidate splits and applying a criterion to evaluate which one to choose. The choice of the leaf after the split can be made uniformly at random or based on a whichever candidate maximizes the purity function [51].

The choice of the predictor to be used in each leaf can be derived from the average response over the training points that fall in the leaf. In order to inject randomness into the tree construction model, we can randomize the choice of the dimensions to split the candidates. We can also randomize the choice of coefficients for random combinations of features.

Gradient boosted machines

Before we go ahead to discuss gradient boosting, it is important to first understand what boosting is all about. Boosting is a technique which is used to convert weak or low performing learners into strong learners with high performance. Boosting makes use of trees with each tree fitted with a modified version of the original dataset. An introduction of the AdaBoost algorithm will aid in describing the gradient boosted machine. In AdaBoost, we start by training the first decision tree that has weights assigned to all observations [53]. After evaluating the first tree, some modifications are made. The observations that were difficult to classify are assigned higher weights while the easier ones are given lower weights. The modification is done to improve upon the predictions of the first tree. Next, the weighted data is then used to grow the second tree. The

model will now consist of a combination of two trees. The classification error of the new model is calculated, and a third tree is grown to predict the updated residuals. The process is repeated across many iterations. The implication of this approach is that newer trees help classify observations that were difficult to compute by previous trees. The predictions of the final ensemble model are the weighted sum of the previous predictions made by preceding trees.

The gradient boosting machine is similar to the AdaBoost model in that it trains multiple models in an additive and sequential manner. The two algorithms differ in the way they identify weak learners such as decision trees. We saw how AdaBoost identifies the shortcoming of weaker learners by using high weight data points. The gradient boosting machine achieves the same objective by making use of gradients in the loss function. The loss function is a good indication of how good a model is at fitting the training data [54]. Logically, we can explain the loss function based on the optimization problem that being tried to solve. For example, if we are trying to predict ticket prices for a football game using a regression, the loss function will be determined by the error between true and the predicted ticket prices for a football game. The gradient boosted machine allows for the optimization of a user-defined cost function instead of a loss function that gives less control and does not really relate to corresponding real-world problems. The ability to apply user-defined cost functions is the major motivation for the use of gradient boosted machines.

Artificial Neural Networks

Artificial neural networks are basic mathematical models that operate like the neural structure of the brain. It mimics the brain's ability to learn from experience. It is a computational model that functions similarly to the biological neural networks found in the brain [55]. The artificial neural network consists of interlinked artificial neurons with the capability of performing certain computations on the input data.

The neurons form the basic building block of the artificial neural network. The neurons consist of a simple mathematical function (model). The mathematical model consists of 3 simple actions: multiplication, summative rule, and the activation function. The entry point of the artificial neural net, the input data is weighted which involves multiplication of the data with individual weight. The middle layer uses a summation function to add up all weighted inputs and bias. At the exit of

the neural network, the sum all weighted inputs and bias is passed through the activation function. The activation function is also known as the transfer function.

Figure 4.1 shows a simple artificial neural net. The neurons in the first layer of the artificial neural network are activated by the input data received and the output passed to the second layer. Similarly, each layer in the network passes its output to the next layer of the network and the last layer outputs the result [56]. The layers in-between the input and output layers in the network are called the hidden layer.

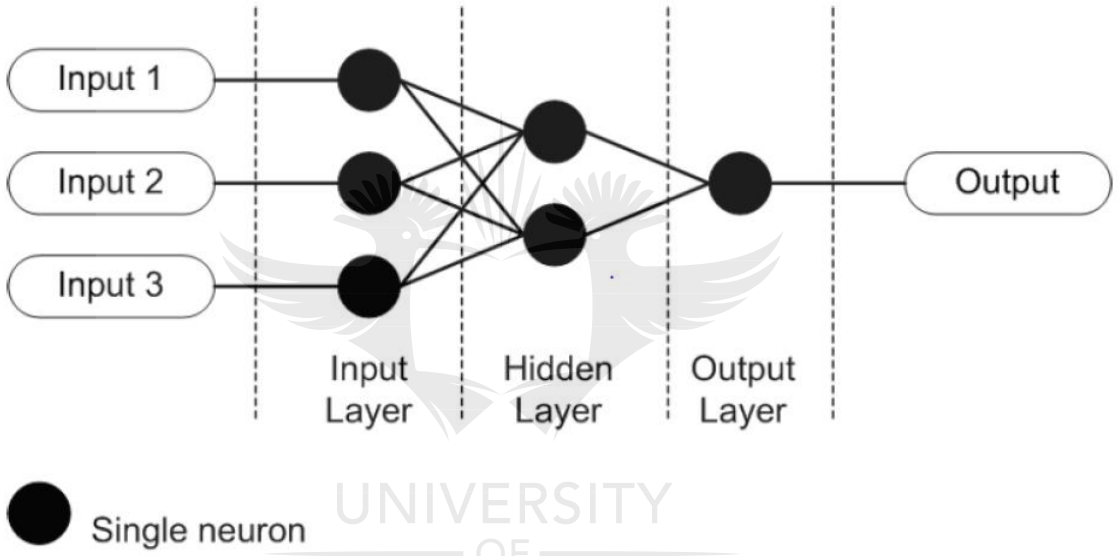


Figure 4.1 Example of a simple feed-forward multilayer perceptron artificial neural net as visualized by author

A **neural network** comprises a set of inputs or output connected with each other with each connection having a weight. Neural networks require a learning phase where the networks learn by adjusting weights to predict the appropriate class of labels of the input tuples. Neural networks are very efficient when trying to extract meaning from complicated data [57]. Neural networks can be used to extract patterns and trends from continuous valued inputs and outputs.

The simplistic nature, underlying principles and fundamental rules of the individual artificial neurons look like nothing special, we are able to exploit the full potential and calculation power of the model when the neurons are interconnected to form the artificial neural net. The artificial

neural nets are based on the principle that complexity can grow out of merely a few basic and fundamental rules.

Artificial neural networks perform well in pattern recognition but due to the resource requirements, artificial neural networks have been slowly adopted [58]. The development in the field of feedforward, recurrent and convolutional neural networks are gaining popularity and due to the processing power required, artificial neural networks are now commonly implemented on graphics processing units (GPU).

Once we are done applying the data mining algorithms to perform the data mining tasks, the next part of the data mining system is the interpretation of the patterns derived. The interpretation of the patterns discovered can be achieved using visualization tools which we discuss in the next section.

4.5 Visualization

Visualization is a very important step in the knowledge discovery process. Visualization helps display the patterns discovered in the data mining stage. Using a visual representation of the patterns a story can be told about the data. Knowledge can be discovered, and decisions can be made. Several visualization tools exist such as Tableau, PowerBI, Microsoft Excel, matplotlib.

Now we have understood the different activities and stages of the knowledge discovery process, we can now to proceed to see how data mining has been applied in the different areas. We look at how data mining has been used to solve problems in the retail industry, insurance, and financial industry.

4.6 Applications of data mining

Data mining has grown in popularity in recent times. There has been a positive movement towards the adoption of data mining techniques in the operations of organizations. Data mining has been used in the retail industry, financial sector, health sector, and the manufacturing industry. Below we discuss how data mining has been applied in the following areas.

4.6.1 Market basket analysis

In the retail industry, there is a need to understand the buyer's needs and customize a buyer's shopping experience based on your knowledge about the customer. Market basket analysis enables the retailer to understand the purchase behavior of the customer. It is based on the theory that if a customer purchases a certain basket of items, he or she is more likely to purchase another group [45]. This can help improve the customer's shopping experience and give more insights about products to the retailer.

It is very important to know the potential customer, the way to attract these customers and then how to retain the customers. Data mining goes beyond traditional customer segmentation to give a deeper understanding of the customers and improve market efficiency [45]. Data mining helps retailers understand how to align customers into distinct segments and then tailor the needs of customers according to insights generated.

4.6.2 Banking applications

Data mining has been used to detect patterns, causalities, the relationship between business information and changes in market price [45]. The volume of data generated by transactions makes it difficult for a manager to immediately discover these patterns. Information derived from data mining in financial banking can be used for better market segmentation along with identifying, acquiring, retaining and maintaining good relationships with the customers.

4.6.3 Fraud detection

Fraud in society has led to financial loss. The use of traditional fraud detection approaches does not suffice due to the complexity and sophistication of fraud. Data mining is now being used to detect fraud in different industries such as healthcare and credit card [45]. A system is built using data mining techniques to detect if a record or transaction is fraudulent or not.

4.6.4 Data mining for fraud detection in health insurance

The insurance industry generates a large amount of data, which can be used for risk analysis, marketing, customer profiling, and fraud detection. Data mining methods can be applied to insurance data to derive more value from the system. Data mining methods have been applied to the medical billing process to detect fraudulent transactions [59]

Several strategies are available for detecting healthcare fraud: audits, market signals, and electronic fraud detection. The audit is carried out in a targeted manner by fraud investigators or in a random manner. The market signals come from whistle-blowers who expose irregular activities. The whistle-blowers can be patients who visited the practitioner or other employees. The electronic form of fraud detection uses computation to detect anomalies which indicate fraud based on pre-existing rules or some statistical functions [59].

Electronic fraud detection is a relatively new field and has grown in popularity with the advent of larger databases. The use of data mining for electronic fraud detection has become a reality now with the development of cloud services, data warehouses, and big data. Advances in the field of information technology, digitalization of the medical billing process and the vast amount of research on healthcare fraud have created an avenue for the use of data mining and machine learning to fight fraudulent activities [59]. Data mining has gained popularity in researchers as a potential tool to combat healthcare fraud.

The healthcare industry has seen slow adoption of data mining technologies as a fraud detection method when compared to the rate at which comparative industries such as the credit loaning and telecommunications industries. The slow adoption can be attributed to the complexity of the medical billing process, the volume of the distributed storage of healthcare claim data and inadequate funding of fraud detection.

Now we have seen how data mining can be applied to solve different problems, the next section discusses similar systems that make use of data mining. We analyze these systems to understand how the different data mining methods and algorithms were applied to achieve the desired results.

4.7 Review of related works

There are several works that have been done in using data mining for fraud detection and this section explores some of those systems. The literature has categorized data mining and machine learning techniques into supervised, unsupervised or hybrid methods [60]. The supervised methods require labeling of data to build a training set, the unsupervised methods, on the other hand, deals with data statistical detection of outlier behavior. In the following subsections, we discuss some systems that make use of data mining to detect health insurance fraud. The section helps us to address the objectives of the study. A review of similar work will help us get a better understanding of the features that can be used to identify fraudulent healthcare claims. We also learn how to build the models z that can identify these fraudulent claims and the best way to benchmark and measure the performance of the model.

4.7.1 Systems based on supervised learning

The section discusses the systems that make use of supervised data mining learning algorithms to solve the problem of health insurance fraud. The systems are reviewed in chronological order according to the date of publication to see the progression of methods used.

The National Health Institute Taiwan, **Wan-Shiou, and San-Yih** in 2006 developed a process-mining framework which detected potential healthcare fraud [61]. Data from a regional service provider for NHI was used in the study. They created the datasets by filtering out noisy data, identifying medical activities using domain knowledge, identifying fraudulent instances and non-fraudulent instances with the help of domain experts. They made use of the filter model for feature selection as the filter model algorithm does not need to search through the entire space for feature subsets. It is very efficient for domains with many features such as the health insurance domain.

The model uses of clinical pathways which are defined as multidisciplinary care plan which diagnosis and therapeutic intervention are performed. In simple terms, it means the order which a physician is supposed to take when carrying out a treatment. They mined frequent patterns from clinical instances using the Apriori algorithm. They systematically identified practices that deviated from the pathway and flagged them as fraudulent. The approach was evaluated with a

real-world dataset retrieved from NHI Taiwan. The results showed that the proposed model could capture and identify several fraudulent and abusive cases that cannot be detected manually [61].

Liou et al. in 2008 Taiwan made use of supervised data mining methods to analyze claims for diabetic outpatients that were submitted to National Health Insurance Taiwan [62]. The data used in validating the model was from a random sample of diabetes patients' health insurance claims. The fraudulent claims in the dataset were identified by the termination of claim contract.

The fraud detection model was built by selecting nine expense related variables and a comparison of these variables was done in two groups of fraudulent and non-fraudulent claims. The expense related input variables used include average drug cost, average drug cost per day, average consultation and treatment fee, average diagnosis fee, average days of drug dispense, average claim amount, average medical expenditure per day and average dispensing service fee. They used three machine learning techniques including logistic regression, neural networks and decision tree for the fraud detection model [62]. They compared the three data mining techniques and discovered that all three were accurate, but the classification tree method had the best performance as it recorded a 99% accuracy in the overall identification rate. The model had a limitation in the sample data used. The sample data used only consisted of three fraudulent service providers whose contracts were canceled by NHI. The dataset did not include fraudulent service providers that padded claims but didn't face any penalty such as contract termination. This caused some data limitations and the result cannot be generalized [62].

Shin et al. in 2012 created a scoring model to detect abusive patterns in health insurance using the 3705 Korean internal medical clinics [63]. They made use of data from the Health Insurance Review and Assessment Services (HIRA). They examined the relationship between the intervention listed by HIRA and the indicating factors to get the data to a manageable size. They extracted 38 indicators which were further validated by HIRA domain experts. The 38 features extracted were used for identifying fraudulent claims using a simple definition of anomaly score.

The model comprised two aspects: scoring to rate the degree of abusiveness and a second part which is identifying the problematic providers by performing segmentation and finding similar utilization patterns. They made use of a decision tree to classify the providers.

The model performed well when presented with different payment arrangements in detecting abusive patterns. The system made use of the scoring model to alert payers of a potential fraudulent billing pattern [63]. The scoring model provides information on the attributes most dissimilar from the norm. Fraud keeps growing in sophistication and the patterns identified for fraudulent and non-fraudulent behavior quickly become outdated. The model proposed by Shin et al. which can be used to automate abuse detection is flexible, easy to use and update. The use of decision trees in the model improved the level of complexity in creating the model as preparing decision trees especially the ones with numerous branches can be difficult and time-consuming.

Branting et.al in 2016 carried out work on healthcare fraud risk estimation using graph analytics. The approach applies several network algorithms to graphs derived from the Medicare payment data along with a list of excluded individual and entities LEIE data and location as well as drug prescription data [64]. They made use of two groups of algorithms. The first group uses measurable health activities such as medical procedures and drug prescription to determine the relationships between known fraudulent and non-fraudulent healthcare providers. The other group of algorithms identified how risk can be propagated from fraudulent healthcare service providers through geospatial collocation.

They empirically evaluated the model and achieved a result of 0.919 f-score, a ROC curve of 0.960 with 10-fold cross-validation. They also identified through ablation analysis that collocation-based risk propagation formed the basis for the most predictive features.

Bayerstadler et al. created a Bayesian multinomial latent variable modeling for fraud and abuse detection in health insurance in 2016 [65]. The created a model to predict varieties of fraud and abuse probabilities for the new invoices. The took the approach of monitoring the changes or deviations in the behavior of the healthcare service providers and insured members from the usual patterns based on the medical claims data. The model is based on the Markov Chain Monte Carlo algorithm with the Bayesian shrinkage techniques. The made use of a fitting algorithm comprising a lasso parameter shrinkage variable that served the purpose of preventing overfitting of the training set. The result generated indicates that the Bayesian model used improves the accuracy of the prediction when compared to the other benchmarking methods.

In supervised machine learning, the data to be used requires an established ground truth label. Unfortunately, openly available medical care data with established ground truth label can be difficult to access. The next section discusses systems that make use of unsupervised learning methods to derive knowledge from unlabelled data.

4.7.2 Systems based on unsupervised learning

Supervised methods work well when used with labeled data. Sometimes the healthcare claim dataset does not have a clear distinction between the records that are fraudulent and the records that are not fraudulent. The section discusses the several approaches that used unsupervised learning to detect fraud and abuse in healthcare.

Yaminishi et al. in 2004 presented the SmartSifter, an outlier detection system that addresses the problem of outlier detection in data mining. They provided a theoretical basis or the Smartsifter and demonstrated that it was effective empirically. The smartsifter made uses a probabilistic model to represent the underlying techniques of data generation. Whenever a datum is fed through, the SmartSifter uses an online learning algorithm to update the model and then a score is given to each datum based on learning model, measuring how much the model has changed after learning. A high then indicates an increased chance that datum is an outlier [66]. They used the Sequentially Discounting Laplace Estimation for learning the histogram density for the categorical features and the Sequentially Discounting Expectation and Maximising algorithm for learning the finite features of the continuous domain. The SmartSifter model demonstrated a high level of accuracy and efficiency in computation time

In Canada, **Fletcher Lu et al.** in 2006 built an adaptive fraud detection system using Benford's law. Adaptive Benford's law specifies the probabilistic distribution of digits for many repeating phenomena, notwithstanding incomplete records [67]. The data used was retrieved from Ernst and Young which contained already audited claims data.

Their approach created a new fraud discovery approach by combining digital analysis with reinforcement learning technique. They made use of the Benford's distribution to benchmark the unsupervised machine learning methods used to discover new cases of fraudulent activities [67]. When this technique was applied to several records of naturally occurring events, the fraud

detection system finds the deviations from expected Benford's law distributions showing an anomaly in the behavior indicating a strong possibility for fraud. The system then searches for the root cause of the anomalous behavior by identifying the underlying attributes causing the anomaly.

Shan et al. in 2008 proposed a mining medical specialist billing pattern method paper that used an association rule in determining compliance for health service management. The data used for this model was retrieved from Australia's Medicare data warehouse. For the feature extraction and selection, the data were grouped in transactions [68]. One transaction consists of all items claimed or billed for one patient daily by one service provider. The transactions that contained only one item were removed.

The model made use of association rules mining. Association rule can be described as statements in the form of antecedents and consequences [68]. For example, if a patient is diagnosed with A then the physician would prescribe drug B and C with a likelihood of 95%. 215 of the association rules were identified. Using these predetermined association rules, the model could pick out the physicians who broke these rules and were flagged with a high likelihood of fraud. The study introduced the mining of negative and positive association rules and not just positive rules only, as found in the previous model by Liou et al. in 2008.

The results derived from the model were validated in different ways. The positive and negative rules discovered were validated by a subject matter expert and found that most of the rules discovered were clinically relevant. The model also outperformed the baseline classifier significantly, when compared together [68]. The model created may also be used to determine the severity of potentially fraudulent activity. There were some drawbacks in this model, such as the possibility that some of the rules discovered may be false alarms, hence further investigation needs to be carried out by subject matter experts.

Lin et al. in 2008 did work on detecting fraud in the Nation health insurance Taiwan general physicians' practice data, they made use of unsupervised clustering methods [69]. They applied PCA feature compression for dimensionality reduction of the feature space. They made use of 10 features in the clustering of physicians' data. The indicators used include the number of cases, average treatment fee per case, average fee per case, amount of fee, average consultation fee per

case, percentage of antibiotic prescriptions, number of visits per case, percentage of injection prescriptions, average drug fee per case, amount of prescription days and percentage of injection prescriptions.

Expert opinions were employed to determine the impact for some of these features on the health expenditure. The opinions of the experts were then used to identify and rank critical clusters. The model successfully integrated the data mining process with the segmentation of the General Practitioner (GP)'s practice patterns [69]. The GP's practice pattern detected by applying clustering methods using the features of expenditures of the GP. The final step was then to illustrate managerial guidance based on these expert opinions. The model was benchmarked against real-world data and the results show that the model can effectively and accurately identify fraudulent abuse and behaviors in healthcare [69].

Shan et al. in 2009 carried out work on discovering inappropriate billings with local density-based outlier detection method [70]. Records of data on optometrists' billings were retrieved from the Australian Medicare database. The data initially had 26 features, but these features were then combined to get 12 unique feature combinations emanating from the original 26.

They applied a local density-based outlier detection method to analyze the Medical claims data for optometrists. They then calculated a single measure that indicates the degree of an outlier for each of the record called the Local Outlier Factor (LOF). The record with the maximum LOF values is the most important outliers [70]. The optometrist compliance history and expert feedbacks were utilized for validation of the model.

They validated the model in two ways: first, they matched the LOF model against compliance history and then the second validation was against domain expert knowledge. The insights from the validation method were that LOF can be used to identify high-risk individuals [70]. A high LOF value indicates a high-risk individual and vice versa. Since LOF method considers multiple factors, it can be more effective in identifying practitioners with multiple compliance issues.

Khoshgoftaar et al. 2016 in their work on creating a probabilistic programming approach for outlier detection in healthcare claims proposed a generalized outlier detection model [71]. The model was based on Bayesian inference with the Stan probabilistic programming language. The

difference between and the model presented by Bauder et al. and several other probabilistic models is that their model provided probabilistic distributions and not only point values.

They presented two use cases to demonstrate the model's effectiveness. The used temperature data that contained outliers in the first use case to compare several methods for outlier detection. The model detected all possible outliers and for further validation, the model provided a probability distribution for each value.

The second case study used by Bauder et al. made use of real-world Medicare claims data to detect the outliers. The data used was for specifically the dermatologist and optometrist. The model was able to detect possible fraudulent payment claims as well as provide valuable probability information. The results showed a successful outlier detection that indicates the possible presence of fraudulent activities which can provide meaningful information that can be used for further investigation.

Bauder et al. in 2017, created a novel method for fraudulent Medicare claims detection from expected payment deviations [72]. The approach taken by the study tried to identify a baseline for the values that reflect the payment for a particular service provider's specialty. The created baseline values for the expected payments for the service provider's specialty in comparison to the actual payment made by Medicare to the healthcare service provider for distinct medical specialties and healthcare services.

The model used support vector machines to analyze distinct groups of practices to generate results indicating possible fraud due to an abnormality in the normal behavioral pattern. They used several techniques such as automatic model selection, binning, grouping and adaptive threshold to optimize the flag rate of frequent behaviors.

We have seen the systems that used the supervised and unsupervised learning methods to solve the problem of health insurance fraud. The final class of systems we review are the systems that combine both the supervised and unsupervised data mining approach to detecting healthcare fraud.

4.7.3 Systems based on hybrid learning

Major and Reidinger in 2002 created an expert system that performs a task which experts cannot simply perform, and statistical techniques are not applicable [73]. They used data from the Electronic Fraud Detection (EFD) database which contains 22,000 providers in six metropolitan areas. They made use of 27 behavioral heuristics to review twenty thousand healthcare providers and compare them to similar providers. The used knowledge discovery on two levels: the performance and the development level. At the performance level, the system combined expert knowledge with statistical information assessments to identify the provider who was outliers to the defined pattern. For the development level, a score was assigned to a provider by calculations using these heuristics and then followed by a frontier identification method that selects providers as candidates for investigation. In the research, 900 providers were identified as suspicious but only 23 were further investigated. They found that with the applied set of heuristics, true positive rates are nearly 50%.

Rawte et al. in 2015 created a healthcare claims fraud detection model based on a combination of both the supervised and unsupervised machine learning methods [74]. The combined the advantages derived from using the supervised machine learning methods with the advantages of the unsupervised machine learning methods. They used the Evolving Clustering Method (ECM) for clustering processing because of the dynamic nature of the data used. They also used the Support Vector Machine (SVM) to classify the data. The ECM was used to cluster the data and it particularly performed well in handling the dynamic data as it modifies the size and position of the cluster as new data came in. Claims are submitted to the semi-supervised Framework wherein clustering Evolving Clustering Method is followed by classification Support Vector Machine to detect the fraudulent claims.

Taghi et al. in 2017 proposed a multivariate anomaly detection in Medicare using model residuals and probabilistic programming [71]. The took a hybrid approach by implementing a multivariate outlier detection method that consists of two processes. The first phase uses a multivariate adaptive regression splines model which fits the multiple predictor variables and outputs a single output variable. The corresponding residues generated by the multivariate adaptive regression splines model are sent as input to the second phase which uses a generalized fully Bayesian model which

applies the Stan programming language to detect anomalies. The multivariate adaptive regression splines model was able to identify nonlinearities between variables and how they interact since it is a non-parametric regression model. The result generated from the model showed that it is robust and does not depend heavily on the distribution of the underlying data when compared to the Mahalanobis distance.

Table 4-1 A summary of the similar systems discussed

Study Topic	Class of learning	Type of detected fraud	Data mining techniques
EFD: A Hybrid Knowledge/Statistical-Based System for the Detection of Fraud (US)	Hybrid supervised and unsupervised	Provider fraud	Outlier detection and rule extraction
Smartsifter	Unsupervised	Provider fraud	Outlier detection
A process- mining framework which detected healthcare fraud	Supervised	Provider fraud (Gynaecology services)	Classification based on associations algorithm, feature selection by Markov blanket filter
The adaptive fraud detection system	Unsupervised	Provider	Reinforcement learning
Detecting hospital fraud and claim abuse through diabetic outpatient services	Supervised	Provider	Logistic regression, neural network, and classification trees
Detecting fraud in the Nation health insurance (Taiwan)	Unsupervised	Provider	Outlier detection
A scoring model to detect abusive billing patterns in health insurance claims (Korea)	Supervised	Provider	Six statistical techniques — correlation analysis, logistic regression, and classification tree
Discovering inappropriate billings with the local density-based outlier detection method	Unsupervised	Provider	Local density-based outlier detection
Mining medical specialist billing patterns for health service management	Unsupervised	Provider	Association rules
Fraud detection in health insurance using data mining techniques	Hybrid supervised and unsupervised	Provider	Evolving Clustering Method and Support vector machine
Multivariate outlier detection in Medicare claims payments applying	Hybrid supervised and unsupervised	Provider	Multivariate adaptive regression, Bayesian approach

probabilistic programming methods			
A Novel Method for Fraudulent Medicare Claims Detection from Expected Payment Deviations	Unsupervised	Provider	Support vector machines (SVM)
Multivariate outlier detection in Medicare claims payments applying probabilistic programming methods	Unsupervised	Provider	Bayesian inference with the Stan probabilistic programming language
Bayesian multinomial latent variable modeling for fraud and abuse detection in health insurance	Supervised	Provider	Markov Chain Monte Carlo algorithm with the Bayesian shrinkage techniques
Graph Analytics for Healthcare Fraud Risk Estimation	Supervised	Provider	Graph analytics

4.8 Summary

Data mining is a process borne out of necessity to gain more insights from the large volume of data generated. Data mining is the nontrivial extraction of formerly unknown and potentially useful insight from data. Data mining forms part of the knowledge discovery process. Although the terms of knowledge discovery and data mining are sometimes used interchangeably, we found that data mining is a core component of knowledge discovery. The knowledge discovery system comprises the following sub-processes: Data capturing, data pre-processing, feature extraction, feature selection, data mining, and interpretation.

The first stage of knowledge discovery is the capturing of the data. Data capturing involves getting data from the source and storing it in a repository. We identified the different data structures which are unstructured, semi-structured and structured. For the different classes of data, we make use of different data collection techniques. Once the data has been captured the next step is pre-processing the data. Pre-processing the data is necessary as we want to eliminate all the noisy features i.e. reduce error in the dataset. The next knowledge discovery processes we looked at was the feature extraction process. Feature extraction is a dimensionality reduction technique. We make use of feature extraction to extract a subset of new features derived from the original set by means of

some functional mapping retaining as much information in the data as possible. Feature selection may follow so that we can eliminate redundant features in the data. Once we have finished transforming the data, we can then apply data mining techniques to the transformed data to perform the different data mining tasks required. We saw the different data mining methods that can be used to generate insights from a given dataset.

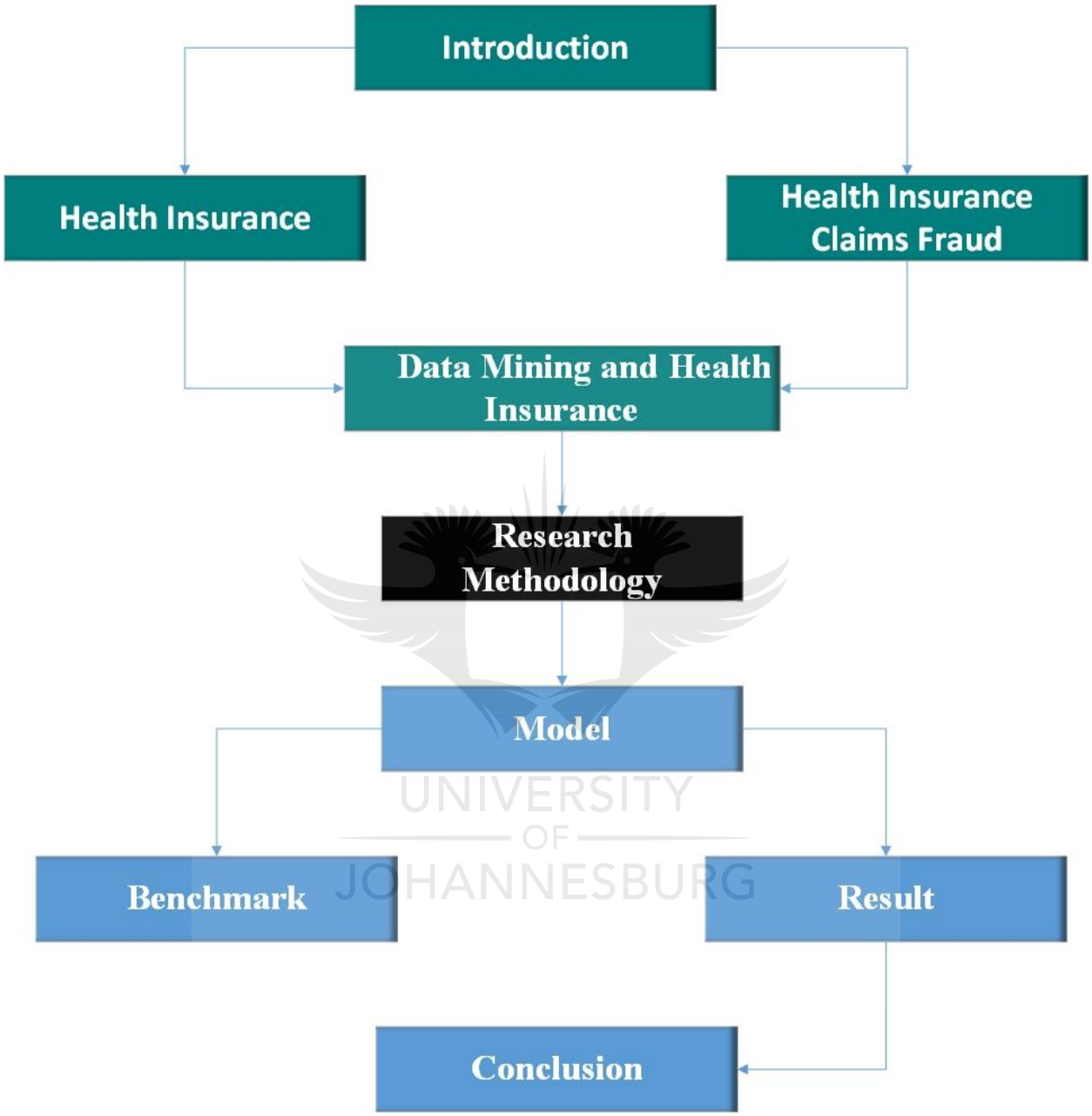
In section 4.5 we discussed the different applications of data mining. Data mining has been used in different areas such as the retail industry, financial industry, and the health industry. In the retail industry, data mining has been used for market basket analysis to understand customer's purchase behavior. It is also made use of customer segmentation to identify and retain customers. We also discussed how data mining has been used for credit card fraud detection in the financial industry. Another application of data mining for fraud detection we analyzed was in insurance claim fraud detection.

Finally, we analyzed similar fraud detection systems that use data mining. We saw how data mining has been used in the health insurance industry. For each of the systems, we examine what data mining algorithms were used. We unpacked the advantages and disadvantages of the approach each system took to address the problem.

The discussion on health insurance in chapter 2 created a better what the health insurance environment looks like, the discussion on health insurance claims fraud in chapter 3, as well as the application of data mining in the healthcare, claims fraud detection process. Through the study conducted in these chapters, we achieved the objective (**O1**) which was to identify the features that can be used to identify fraudulent healthcare claims.

To establish a way to conduct the research, we define a research methodology that forms the guide to solving the problem at hand. The next chapter unpacks the research methodology applied in the study. We discuss the steps and the different approach that was taken to conduct the research as well as achieve the objectives set.

Dissertation roadmap



Chapter 5 Research Methodology

5.1 Introduction

In every research, there is a need to establish a way in which the research is conducted. In order to fully address the objectives of the dissertation in an effective manner, we make use of the research methodology which defines the research design, research paradigm followed, and methods used. The research methodology forms a systematic approach to problem-solving. It shows how the problems posed in the research is going to be solved. It is a process that maps out the plans the researcher will use to solve the research problem.

Defining a research methodology to be followed in research is an important task. In the research methodology, the researcher analyses the question in a systematic approach to finding the answers that lead to a conclusion. In this chapter, we analyze the different types of research designs to determine which design is the most suitable for the research work in section 5.2. Next, we look at the different classes of research paradigms in section 5.3. With an understanding of the research paradigms, we then choose to analyze and choose a research method to follow for the dissertation in section 5.4. We also look at the data sampling and the population that will be used in this research in section 5.5. Finally, we define the proposed plan of work for this dissertation.

5.2 Research design

Research design describes the comprehensive strategy that is used for integrating all the different components in a study in a logical and coherent manner. The research design is a blueprint that describes the way data is collected, measured and analyzed to ensure that we effectively address the research problem in a logical and unambiguous manner [75]. The popular research designs include quantitative, qualitative and mixed designs.

5.2.1 Qualitative research

The qualitative approach is mainly concerned with gaining knowledge about the underlying reasoning and what motivated lines of action. It is subjective and measures the different ways individuals view the world around them. In the qualitative approach, the aim is a very detailed and complete analysis of what is being observed. Its main purpose is to contextualize, interpret and understand perspectives [3]. In qualitative research, we take an inductive and process-oriented approach to comprehend, interpret and develop a theory for a problem statement.

The qualitative research design approached was originally developed for researchers in the social science domain that need an understanding of the social and cultural issues [3]. Researchers that take the qualitative research design approach have a holistic view of the problem. The researcher's approach is person-centered as the researcher becomes immersed in the problem to be more familiar with it [76]. It takes a naturalistic approach to the subject matter, as the researcher observes how groups of people carry out activities in their natural setting. Qualitative research takes an interpretative approach. Qualitative research data sources include questionnaires, research impressions, fieldwork, interviews and documents, and text.

5.2.2 Quantitative research

The quantitative research methodology is mainly concerned with the measurement of quantity. It can be used in situations where the facts can be expressed as quantities [3]. The quantitative research is more statistical in terms of the analysis and is objective as it aims for a precise measurement. The quantitative research approach was initially developed for use in natural science study.

The modes of data collection in quantitative research include a questionnaire, through experiments and carrying out surveys. The data collected is revised and tabulated in a format that can allow for statistical analysis. Using statistical analysis, we can measure the variables of the sample data and find the relationships between these variables [76].

The major difference between the quantitative and qualitative research design the impersonal role the researcher lays in the quantitative research while in the qualitative research the researcher is immersed in the problem space and has an influence on the result. Another notable difference is

that while qualitative research is inductive as a hypothesis is not needed to begin the research while quantitative research is deductive.

For this research work, the quantitative research methodology is chosen as it allows for a statistical analysis of the data which has been collected in a structured manner and would be used for causal explanation of fraud detection in an insurance claim. We define the hypothesis and using statistical analysis of data, we are able to make deductions.

5.3 Research paradigm

The research paradigm can be described as the entire process of thinking used in the research. It refers to known research traditions used in a philosophical framework or discipline. The research paradigm used determines the set of beliefs that would guide the actions in the entire research process [3]. There are two research paradigms: positivist and interpretive.

5.3.1 Positivist research paradigm

The positivist believes in a stable reality that can be analyzed and dissected from an objective viewpoint. The positivist research paradigm is based on the theory that the most effective way to understand human behavior is through observation and reason [76]. The positivist sees true knowledge as only obtainable through experiments and observation. The positivist makes use of scientific processes to create a systematic knowledge generation process.

Using scientific means, the positivist finds the truth and presents the results with empirical means. The positivist research paradigm assumes that the researcher is detached and does not interact or influence the subject. When the researcher takes such a position, he or she can be more objective when performing the analysis. The researcher can make an analytical, non-personal interpretation of the data.

5.3.2 Interpretive research paradigm

The interpretive research paradigm aims to get a clearer understanding of people. The interpretivist which has a viewpoint that only by intervening and by having a subjective approach can the reality be properly understood. The interpretivist research serves to understand and interpret social

activities, structure and daily happenings. In interpretivism, social reality is a product of the actions and perceptions of those in society. Hence, they believe that the reality of the social environment is subjective and nuanced.

For this research, we have chosen positivism as the research paradigm because we would adhere to only observations and measurements to gain factual knowledge. We would also have roles limited to interpretation and collection of data using the objective approach as mentioned in section 5.4 and having quantifiable and observable research findings. We are also using the positivist approach as we would depend on quantifiable observations that can be analyzed statistically. We would collect the historic insurance claim data and statistically analyze it to make deductions without interfering with the insurance claim procedure.

5.4 Research methods

Research methods are essential tools used by the researcher to collect data, approach the problem and arrive at the desired solution. The research methods used in this work are a literature review, experimentation, model creation and the development of a prototype.

5.4.1 Literature review

The literature review represents the systematic search of relevant work that has been published to gain an understanding of what is already known about the topic of intended research. It serves the purpose of establishing why the research is needed, broadening the knowledge of the researcher, and enabling the researcher to gain an understanding of what has been done.

The literature review was chosen as a research method because it will enable us to gain a better and deeper theoretical framework of the research [3]. It would also enable the identification of gaps in the previous study. It will be useful in identifying the best variables to be considered in this research. The literature review shows what the current research landscape looks like by analyzing the problem domain of health insurance fraud, the current works being done in the health insurance fraud detection space and the different technologies that have been applied to solve the problem. The information gathered in the literature review forms the basis for how the model is designed and the model can also be compared to the other solutions found in the review.

5.4.2 Experimentation

Experimentation is a research method that tries to isolate and control all the relevant variables that determine the outcome of the events being investigated. Experimentation allows for these variables to manipulated and observations are made based on the outcome of the changes to the conditions [77]. The simplest form of experimentation is making changes to the independent variables to observe impact the changes have on dependent variables.

We use experimentation as a research method because we take an exploratory approach to building the model as can be seen in section 8.3.3, where we experimented with different machine learning methods to determine which methods best solve the problem of healthcare claims fraud. For each of the method chosen, we try different values for each of the parameters and observe the output of the model with regards to the predefined metrics.

5.4.3 Model

Another research method used is the model which is an effective way to relate the research more accurately to reality. The model provides a framework to understand, describe and predict a complex problem. The outcome of the study is the model which encapsulates the findings from the literature study and then proposes a new structure to address these findings. The new structure created is analyzed and implemented to understand its true potential.

The above-mentioned description of a model motivates the use of the model for this project as it would be used to test the facts and assumptions in this research. Being an abstraction of reality, it can present a simpler case than the actual reality making easier to work with. We can test our hypothesis using the model and benchmark our results with the healthcare claims dataset.

5.4.4 Prototyping

Prototyping involves creating a preliminary version of the proposed system [3]. Prototype serves as part of a larger system of an application. It doesn't require much resource commitment. Prototyping allows for the creation of a representation of the proposed solution to the problem statement. Prototypes can be used to evaluate and test a design. The prototype we will use in this dissertation will give insight into how the model designed functions.

A prototype would be used as a research method as it would be a preliminary way to test out some propositions used, and algorithms used to determine the effectiveness. The prototype will allow the analysis of the different approaches to solving our problem statement. We can also use the prototype to determine how practical the proposed solution is.

For the study, the prototype implemented is a data mining system that applies machine learning methods to address the problem of health insurance claim fraud. The prototype developed for the study has different stages, with each stage accomplishing a particular data mining task. The prototype consists of processes that collect data, learn from the data to derive patterns and insights. The model is then validated with test data to determine how well the model understands the data base on predefined metrics. A full discussion of the prototype implemented for the study is discussed in Section 8.3 and the metrics that are used to benchmark the prototype is also discussed in chapter 7.

5.5 Population

We make use of a secondary dataset set to evaluate the model. The population consists of the physicians that render the services to the patients and submit the claim for reimbursement of the cost of treatment. The data that would be used for this research would consist of a combination of fraudulent and legitimate health insurance claims data from the USA Medicare dataset for the 2016 calendar year.

The dataset includes the data for providers that had a valid NPI and also submitted medical insurance claims for patient's treatment. The dataset would be retrieved from the Medicare online dataset repository. We also use the data from the List of Excluded Individual and Entities (LEIE) database to label the Medicare dataset. The LEIE dataset contains labels for identifying fraudulent providers. The LEIE exclusions are grouped according to rule numbers, which shows the severity as well as the length of time of each exclusion.

5.6 Data Sampling

For this study, the data sample comprises data about the Physician, Patient and Payment data. The data collected about the physician include location, specialty, service performed. Patients data

collected include Patient ID, Sex, Total number of physicians seen. Payment data include an Average charge per service, charge for service.

5.7 Research plan

To gain a better understanding of how the research would be conducted, a research plan is provided below to give proper guidance. The study would first review the available work done on using data mining methods for insurance claim fraud detection. This would be done as part of the literature review. Once the literature review is done, a pilot study which entails data collection, using the planned methods with the aim of testing the proposed approach. Using this we are also able to identify any detail that needs addressing before the main collection of data. After the pilot study comes to the main data collection. Next is the model creation and then the building of a prototype to test the model. The results are collected, and an analysis is then performed on the result extracted. Once the analysis is done, the findings or deductions made are then reported.

5.8 Conclusion

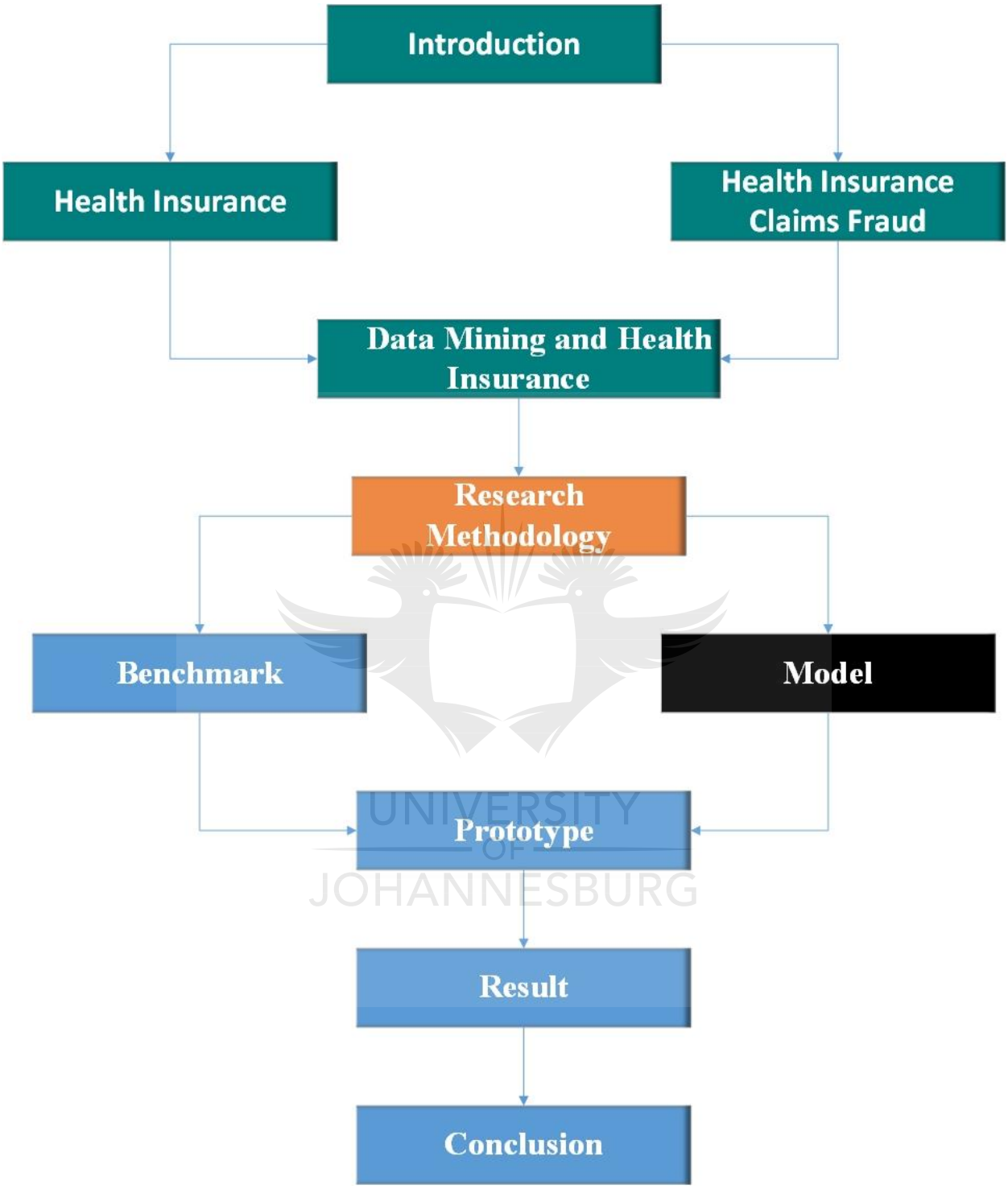
To achieve the objectives of the research in an effective manner, we defined a research methodology. We chose the quantitative research approach as it lends itself to the problem domain. Quantitative design allows for testing a hypothesis using scientific methods.

Next, we looked at the popular research design paradigms. We identified the positivist and the interpretivist research paradigms. While the positivist is objective in the research and does not influence the outcome, the interpretivist is more involved with the environment and can influence the outcome. We chose the positivist research paradigm as we will maintain an objective stance in the research and not have any influence on the outcome of results.

We also identified the different research methods that will be used in the dissertation, a Literature review is selected to gain an understanding of the problem domain and review the current solutions to the problems we are attempting to solve. We also make use of a model to create a conceptual abstraction of the proposed solution and then we implement a prototype to test the feasibility of the proposed model and then benchmark the results.

To be able to validate the model, we make use of population and data sample set. We make use of a dataset that allows for a repeatable result. Finally, we discussed an overview of the research plan and the resources required to execute the plan. The key areas highlighted by the plan are the literature study, model creation, prototyping, collecting results and benchmarking the results.

Dissertation roadmap



Chapter 6 Model

6.1 Introduction

In today's fast-growing, competitive computing industry, it is very important to develop software systems that are easily extensible to support updates and changes easily, portable, scalable and efficient. The data mining system requires as a framework is created to ensure that the system functions optimally.

Designing a data mining model involves a proper understanding of the individual processes that enable the derivation of knowledge from data. In data mining, the objective is to identify potentially useful, valid, understandable and valid patterns and correlations within a dataset [78]. The process of data mining consists of more than just data collection and data management but also comprises data analysis and predictions.

In chapter 6, we discuss the different methods that can be applied in the data mining process. Some of the methods discussed are discussed were not applied in the prototype for the research but are in general methods that are widely used for the different data mining tasks. For each of the data mining processed, we discuss methods that are widely used in the domain to achieve the task at hand.

Figure 6.1 shows a graphical representation of the proposed data mining model. The following subsections unpack each of the steps in the model. We start by discussing the preliminary process which is data collection and data preparation in sections 6.2 and 6.3 respectively. With an understanding of the data domain and the pre-processing of the data, we then perform feature selection in section 6.4 and feature extraction in section 6.5. Once the features have been extracted and selected, we discuss how the machine learning methods can be applied to the processed data in section 6.6 and finally in section 6.7, we discuss the different evaluation metric that can be applied to the data mining models.

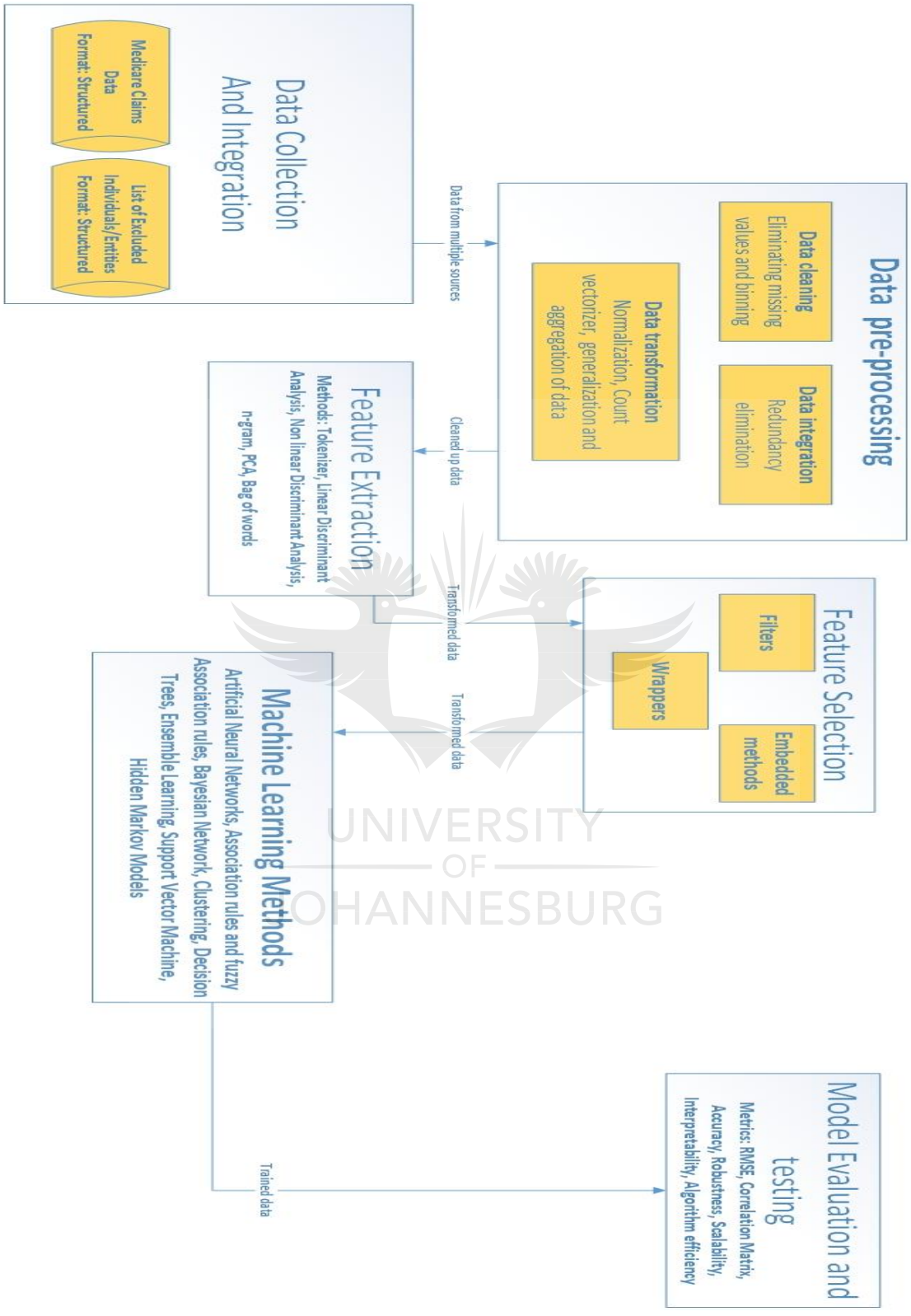


Figure 6.1-1 Diagram showing the different processes in the proposed model as illustrated by the author

6.2 Data Collection

A preparatory step in the data mining model is the gathering of the data that will be mined to discover knowledge. The process of selecting and creating a data set to perform data mining first requires a good understanding of the problem domain. Understanding the problem domain prepares the scene for understanding the data requirements and the various algorithms, as well as data transformations that will subsequently occur. Having understood the problem domain, we now proceed with defining the data set that will be used to discover the required knowledge. The data used in the study is mostly text-based. The process of selecting and creating the required dataset includes determining what data is available, retrieving additional necessary data and the integration of all the data to be used for the data mining process into one repository. The importance of this stage in the data mining process cannot be underestimated as the data mining methods used are only as effective as the quality of data used. The dataset forms the base for constructing and evaluating the model.

Data collection, integration and organization can be a very complex and expensive task. The task of data collection and organization takes into consideration the structure of data available. As discussed in section 4.3.1, there are three classes of data structures namely: structured, semi-structured and unstructured data. The structured data types are easier to process and navigate as they have a relational key which can be used to map to predefined database fields. Unstructured data has no identifiable structure or form and cannot easily fit into a database. Semi-structured data is a class of data that cannot be easily stored in a database, but the data has some organizational properties that allow for easier analysis.

The different types of data require different methods of collection and extraction. For structured data, it is easy to extract the data from the database using SQL queries. For unstructured and semi-structured data, some transformations need to be done to allow for storage in a relational database. The data collected and stored most times does not come in the desired format needed [79]. The Medicare dataset used in the dissertation comes in a structured format and can be easily imported into a relational database for further analysis. In the next section, we discuss how data can be further processed to get to the desired form.

6.3 Data pre-processing and cleaning

The quality of data used in the model determines the quality of results and knowledge we can derive from the model. Unfortunately, data in the real world does not come in the perfect format to be used for effective analysis. Data preparation forms an essential part of the data mining process and it is often very tedious and time-consuming in practice [80]. Data in the real world is usually incomplete, noisy and inconsistent. Data is incomplete when it lacks attribute characteristics of interest or containing mainly records that have been aggregated [79]. Data is noisy when it contains incorrect entries, errors or outliers such as "age= -10". Finally, data is inconsistent when it contains discrepancies in codes or names. Next, we will discuss the various data processing tasks that can be used to get data to the desired state.

6.3.1 Data pre-processing tasks

The process of transforming data from its raw form involves several tasks. The major data pre-processing tasks include data cleaning, data integration and data transformation [81]. The first step in data pre-processing is the data cleaning also known as data cleansing.

Data cleaning is the process of eliminating irrelevant and noisy data. The data cleaning process can handle missing values by eliminating or ignoring the missing values or estimating the missing values by using a global constant, mean, or predict the value based on features [82]. Noisy data which results in outliers in data can be handled by binning which involves first sorting the data and assigning data to bins based on equal frequency. Smoothening can be applied to the bins using mean, median, and bin boundaries. Regression, clustering as well as combined computer and human inspection are other important techniques that can be applied to handling noisy data [81].

Data integration combines data from several sources and persists the combined data in a coherent store. It integrates metadata from multiple sources. Redundancy can occur when integrating multiple sources. To detect redundancy in data, we can perform correlation analysis and apply more caution when integrating multiple sources [83].

Data transformation is another important phase of the data pre-processing task. It involves normalization, generalization and aggregation of data. Aggregation of data is the summarization

of data based on given criteria. Normalization is the scaling of data to fall within a range, relative to the entire dataset [81]. The common types of normalization methods used are min-max normalization and the Z-score normalization. The next subsection discusses how we can retrieve the features that are relevant to what we are doing.

6.4 Feature Selection

The process of selecting the subset from the pre-existing feature set based on the importance of the features of the data. Feature selection is a way of reducing data dimensionality. It can be used to reduce complexity which may arise from having irrelevant or redundant data in the model [84]. We use feature extraction to eliminate these irrelevant features and gain more information about the data. There are several considerations when selecting features for data mining. We need to ensure that the accuracy and performance of the system are not affected. The feature selection mechanism is successful, if it improves data visualization, creates a better understanding of data and reduces storage requirements [85].

The feature selection process consists of mainly 2 steps: the feature generation and the feature evaluation steps [86]. The feature generation creates subsets of features from the larger pre-existing feature set. We then evaluate the selected features to determine if the feature subset fits the requirements. The evaluation of the selected features can be done based on dependent and independent measures. When the evaluation of features is done by monitoring how the algorithms applied to feature set performs, we then classify that as a dependent measurement. Independent measures are used to evaluate the feature selection model without the application of any learning method. There are three major categories of feature selection algorithms. They include filters, embedded techniques, and wrappers [86].

Filters work without using classifiers. Filters technique makes use of several scoring methods to choose the top-N features with the highest scores. Although the filter techniques are generally faster than the wrappers, they have a limitation as they do not consider feature dependencies [86].

Wrappers perform better in feature selection and they use a learning model for evaluating the features. They take feature dependencies into consideration and uses a predictive model to evaluate and score subsets of features. They are computationally intensive but generate the best feature set

for the specified model. Embedded methods are similar to wrappers but are computationally slower than the wrappers [85]. With an understanding of the feature selection process as well as the feature selection algorithms, we proceed with a discussion of feature selection methods.

The first feature selection we discuss is the document frequency. **Document frequency** is a feature selection method that defines the frequency of documents in which terms are present [87]. For every term considered the document, the frequency is compared to a minimum threshold and if less than the threshold the term will be removed. The document frequency thresholding can scale with large data easily in a linear runtime.

Chi-square statistic is used to examine the independence of two variables. It is used to determine the difference between two categorical variables [87]. Two terms are independent of each other if

$$p(XY) = p(X)p(Y) \tag{6.4.1}$$

Best Term is also another feature selection method that works using a target class t and a BT algorithm [86]. Using these two inputs, it finds the documents that belong to the target class t and the determines the features that make up the target class t with high precision. Likewise, it will also find the documents that do not belong to, however, contain a minimum of one feature found in the preceding step. Combining the two feature sets yields a new feature set in the final step.

Ambiguity measure feature selection approach assigns higher scores to features that mainly appear under one category consistently [86]. The methods assign a score between zero and one to features based on ambiguity. A score close to one means that the feature is unambiguous and close to zero implies ambiguity of the feature.

Once we have the features, we can then further improve the quality of the data by reducing the redundant features present. Feature extraction enables us to reduce the dimensionality of features and we discuss it in more details in the next subsection [37].

6.5 Feature Extraction

Feature extraction is a type of dimensionality reduction used in the data mining process to extract a subset of new features from the original set of features using some functional mapping while

retaining as much information in the data as possible [86]. New variables derived from the combination of other variables are generated during feature extraction to reduce dimensionality.

Principal Component Analysis (PCA) is a feature extraction or dimensionality reduction technique used to extract more information from the given dataset. PCA is an orthogonal linear transformation. The main goal in PCA is to create lower dimensional sets of features from the original higher dimensional feature set [86]. It is important to determine what the number of principal components and this number should adequately represent the data adequately [86]. Cross-validation, broken stick model, Kaiser's criterion and cumulative percentage of variance are certain methods used to determine the optimum number of principal components.

Latent semantic indexing (LSI) is another feature extraction method for dimensionality reduction. It uses mathematical methods to determine the relationship between various terms and ideas in a document [86]. The LSI is used in indexing webpage. Other options for feature extraction include Autoencoders, Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF).

As earlier discussed, data mining is the discovery of unknown patterns from data. There are several tasks that can be performed in data mining such as classification, neural networks, regression, and clustering. In the subsequent subsections, we will be discussing other processes that make up the data mining lifecycle.

6.6 Machine learning methods

In data mining, there needs to some form of learning. Data mining learning methods can be classified in different ways. The classification method used depends largely depend on the kind of data being processed, the type of knowledge to be discovered and the algorithms utilized. Machine learning experts have divided these classification methods into two categories, the supervised and unsupervised methods which were discussed in section 4.4. In Chapter 8, we discuss how these machine learning methods were used for the prototype developed for the study.

6.7 Model Evaluation

One aspect that is highly important in implementing a data mining model is to critically evaluate and analyze the proposed model. The analysis of a system helps to highlight the strengths and weaknesses of the solution. We consider criteria that enable us to gain insights into the system. Some of the criteria considered are robustness, scalability, interpretability, algorithm efficiency, properties of data used, accuracy and error rate or misclassification rate. These criteria are unpacked below.

Robustness: Given noisy data or data with missing values, a robust system will still be able to make predictions correctly [19]. The data pre-processing step in the system design is responsible for cleaning up the data before machine learning methods can apply to it. A system that cannot handle noisy or incomplete data will not be suitable for deployment as data in the real world does not come in the required format and as a result, data can be incomplete, noisy and inconsistent.

Scalability: In developing a data mining system for healthcare claims fraud detection, the volume of available data needs to be considered. The systems need to be designed in such a way that given the large volume of data, they should be able to function effectively [24]. The scalability of the system is assessed using a series of data sets of increasing size.

Interpretability: The interpretability of the model refers to the level of insights and understanding that is generated by the model. Interpretability is subjective as it is difficult to assess. The more insights a system can provide, the more useful it is to the problem it is trying to solve.

Algorithm efficiency: The efficiency of an algorithm mainly depends on time and space usage. Algorithm efficiency relates to the number of computational resources used by the model. To maximize efficiency, we must minimize resource usage. The efficiency of the systems refers to the cost of computation incurred when generating and using the system [24]. An efficient data mining system will make use of minimal computational resources thereby increasing the speed of operations and reducing memory usage [19].

Properties of data used: Two very important properties of data we consider are variety and volume. These characteristics of data influence the results generated. The more data used the more

inferences can be made. The higher the variety of data the more the deductions can be generalized [19].

Accuracy: Accuracy is a measure of the overall correctness of the model. To compute the accuracy of the model, there are four additional terms that form the building block for calculating several evaluation measures: true positives (TP), true negatives (TN), false positives FP) and false negatives (FN) [24]. The TP refers to the positive tuples that were correctly identified by the classifier. The TN refers to the negative tuples that were correctly identified by the classifier. Accuracy is the ratio of the sum of the TP and TF to the total predictions. Accuracy is useful in inferring the correctness of the system [24]. Accuracy involves testing a generated model against an already calibrated data that contains output. The aim is to build models that have high accuracy.

$$Accuracy = (TP + TN) \div (TP + TN + FP + FN) \quad (6.7.1)$$

Error rate or misclassification rate: This is the ratio of the sum of false positives and false negatives to the total predictions. It can also be calculated by subtracting the accuracy from 1. A good model will have a very low error rate which implies a high accuracy [24].

6.8 Conclusion

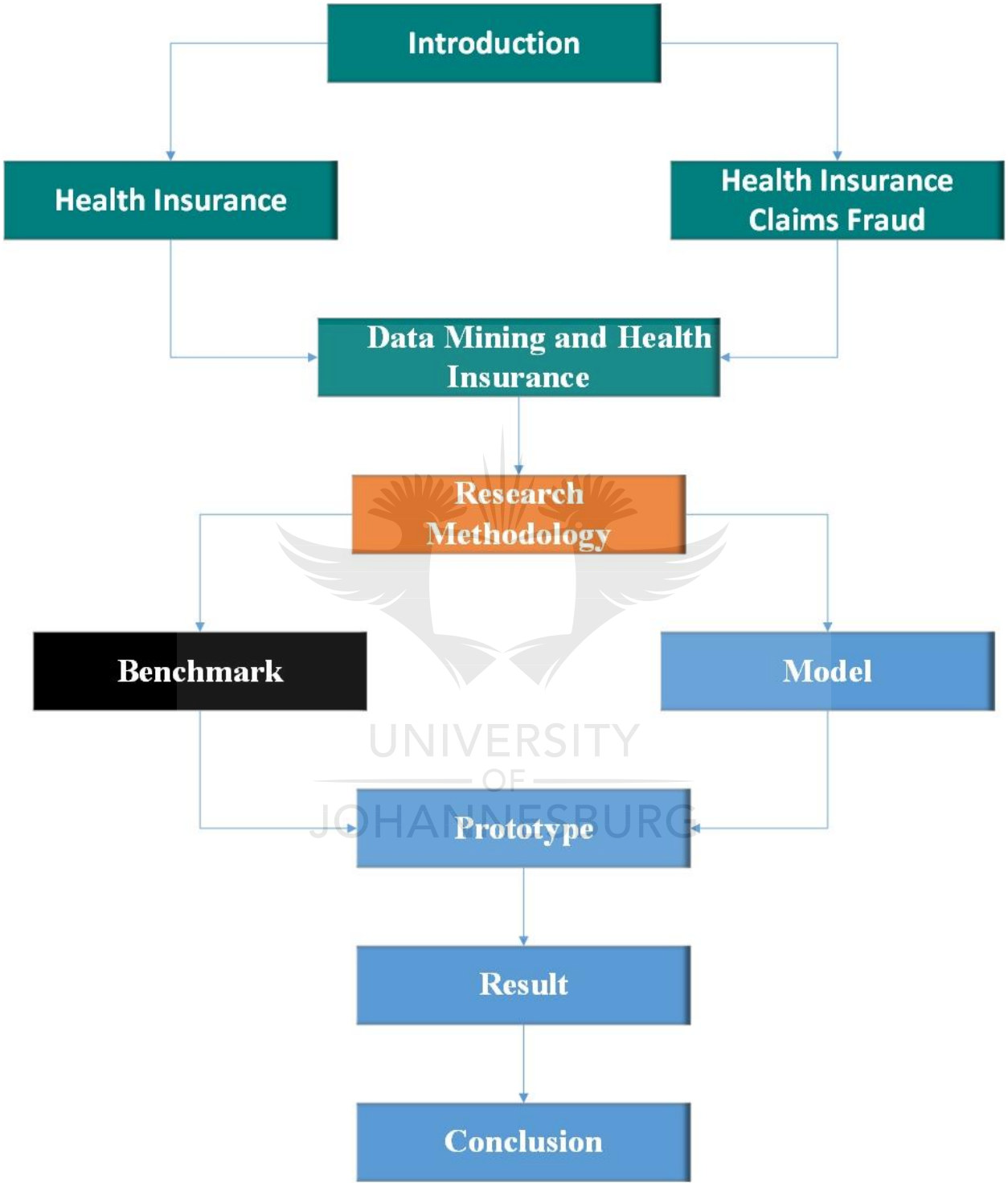
The chapter begins by describing the overall architecture of the proposed model as shown in figure 6.1 and then continues with an elaboration of each of the components making up the model. These components that make up the model are namely the:

1. Data collection process which is responsible for retrieving the data from the Medicare LEIE databases and then loading it into our environment for processing.
2. Data pre-processing and cleaning for handling dirty and incomplete data. The process is responsible for preparing the data for a state that is suitable for the model to use.
3. Feature selection a process of selecting the subset from the pre-existing feature set based on the importance of the features of the data.
4. Feature extraction a type of dimensionality reduction used in the data mining process to extract a subset of new features from the original set of features using some functional mapping while retaining as much information in the data as possible.

5. Application of machine learning methods to derive knowledge from data to classify healthcare claims.
6. The evaluation of a model to gain a better understanding of what its strengths and weaknesses are.

By carrying out these different processes, we can create an effective machine learning system that can detect possible fraud in healthcare claims. We also achieved an objective (**O2**) which is to build a model for applying the features to create a health insurance claim fraud detection system by defining the proposed model as shown in figure 6.1. Now we have discussed and gained a better understanding of the processes that make up the model, the next part of our study creates a benchmark for that will be used to evaluate the implementation of the proposed model.

Dissertation roadmap



Chapter 7 Benchmark

7.1 Introduction

Winston Churchill once said “Criticism may not be agreeable, but it is necessary. It fulfills the same function as pain in the human body. It calls attention to an unhealthy state of things.” Evaluating a model creates a way to analyze the strengths and weaknesses of the model. The weaknesses or limitations help determine the contributions to the field as well as reveals more future research opportunities. It is important when evaluating a model to be objective and to avoid any bias that can be introduced by the researcher. Being objective in evaluating models maximizes the insights that can be gained from the research.

In Chapter 7, we define the scope of the evaluation, the method of evaluation used as the evaluation metrics used. To gain a better understanding of the evaluation metrics used and how they apply to the model, we further unpack the evaluation criteria. Subsequently, we discuss each class of evaluation metrics and then a conclusion of the chapter.

7.1.1 Comprehensive Evaluation

Historically, to know how well a machine learning model performs, we evaluate the performance. The performance of a machine learning model determines the viability of the model. Certain performance metrics are examined to determine the viability or how well the model will operate in a real environment. The insight gained from the performance metrics is used to promote the use of the model as well as act as a yardstick for comparison with other similar systems. As discussed in section 6.7, there are different ways in which a machine learning model can be evaluated. We classify these evaluation criteria into two different core categories namely:

1. Performance Metrics
2. Resource Metrics

7.2 Performance Metrics

Once the usual data capturing, feature engineering, feature selection as well as the implementation of the machine learning model and receiving output in the form of probability scores. Next, we determine how effective the model is based on some test datasets. There are different performance metrics that can be used to evaluate a model which include metrics such as Confusion matrix, Precision, Recall, F1 Score, Specificity, Accuracy, AUC (Area under Curve), Root Mean Squared Error (RMSE), Equal Error Rate.

7.2.1 Confusion Matrix

The confusion matrix presents an intuitive and easy way to determine the correctness and accuracy of a model. A confusion matrix can be used to show the of a solution to a classification problem. The information relating to the actual and predicted classification is contained in the confusion matrix [88]. These derived metrics are described below:

1. True Positive is a measure of the cases that were predicted as true and their actual data class is true.
2. True Negative is the number of cases that were predicted as false and their actual data class is false.
3. False Positive is the number of cases that were predicted true when the actual data class is false.
4. False negative is the number of cases that were predicted as negative, but the actual data class is positive.

Using the metrics derived we can calculate the following values for the model:

Precision can be defined as a measure of the ratio between the number of true positives (TP) over the sum of the number of true positive (TP) and the number of false positives (FP) [88].

$$Precision = TP/(TP + FP) \quad (7.2.1)$$

Recall shows how much information is derived. It is a measure of the ratio between a number of true positives (TP) and the sum of the number of true positives (TP) and false negatives (FN) [88].

$$Recall = TP/(TP + FN) \quad (7.2.2)$$

F1 Score: The F-measure can be described as the harmonic mean of recall (R) and precision (P). Some literature refers to the F1 score as the F-Score or the F-Measure [88].

$$F1 = 2 * (Recall * Precision)/(Recall + Precision) \quad (7.2.3)$$

Specificity: The specificity tell what portion of the negatives were predicted correctly.

$$Specificity = TN/(TN + FP) \quad (7.2.4)$$

Accuracy: Accuracy for classification problems is the number of correct predictions i.e. the sum of TP and TN over the entire predictions made [88].

$$Accuracy = TP + TN/(TP + FP + TN + FN) \quad (7.2.5)$$

The ROC curve is calculated by plotting the values of the true positive (TP) rates on the Y- axis and on the x-axis, we plot the false positive rate (FP) [89]. Every point on the ROC curve is the TP rate(sensitivity)/FP rate (specificity) pair which corresponds to a particular decision threshold. Perfect discrimination is achieved in a test when there is no overlap in the two distributions) and has a ROC curve that goes through the upper left corner of the graph (100% sensitivity, 100% specificity). This implies that the closer the ROC curve is to the upper left corner, the higher the overall accuracy of the test [89].

7.2.2 Robustness

Real world data is noisy and can contain missing values, a machine learning system needs to be able to handle the data and still make predictions effectively. The machine learning system has a data pre-processing step, where data is cleaned up before the machine learning algorithms can be applied. With the presence of noisy data in the real world, a system that is not robust will not be suitable for deployment. So, the robustness is its ability to tolerate data of different qualities while producing consistent results.

7.2.3 Characteristics of features used

Features are data characteristics that define the data. Selecting the right features is an important task as the training time increases exponentially as the number of features increase. We also run the risk of over-fitting when the features are excessive. Overfitting is a problem that occurs when the machine learning model learns from trends and patterns which are present in the training data but do not reflect the data generating process [90]. Selecting features is not a trivial task and it involves the application of feature extraction and selection methods. We evaluate the machine learning model based on the quality and versatility of the features used. In detecting healthcare claim fraud, we need to make sure that the data we use covers a large aspect of the healthcare claims process.

7.3 Resource Metrics

Even with the abundance of resource in the present-day computing, there is a need to understand the evaluation metrics pertaining to the resource requirements in the design, implementation and the deployment of a machine learning system. Some resource metrics we have identified that need to be evaluated for a machine learning system are scalability and the efficiency of algorithms used. In the next subsection, we discuss these resource metrics as well as how the metrics can be measured.

7.3.1 Scalability of model

In building a data mining system, it is very important to have a scale in mind. The volume of data to be processed needs to be considered as well. The systems need to be designed in a manner that given additional data, it can continue to function normally. To determine the scalability of the machine learning model by using a series of data with increasing sizes.

7.3.2 Algorithm efficiency

The efficiency of the machine learning algorithm used can be determined based on time and space usage. The efficiency of an algorithm relies on the number of computational resources used up by the system. To improve algorithm efficiency, we must minimize the number of computational

resources consumed. The efficiency of a machine learning system is measured by the cost of computation incurred when generating and applying the system. An efficient data mining system will use minimal computational resource i.e. reduced time and space usage to achieve the expected results.

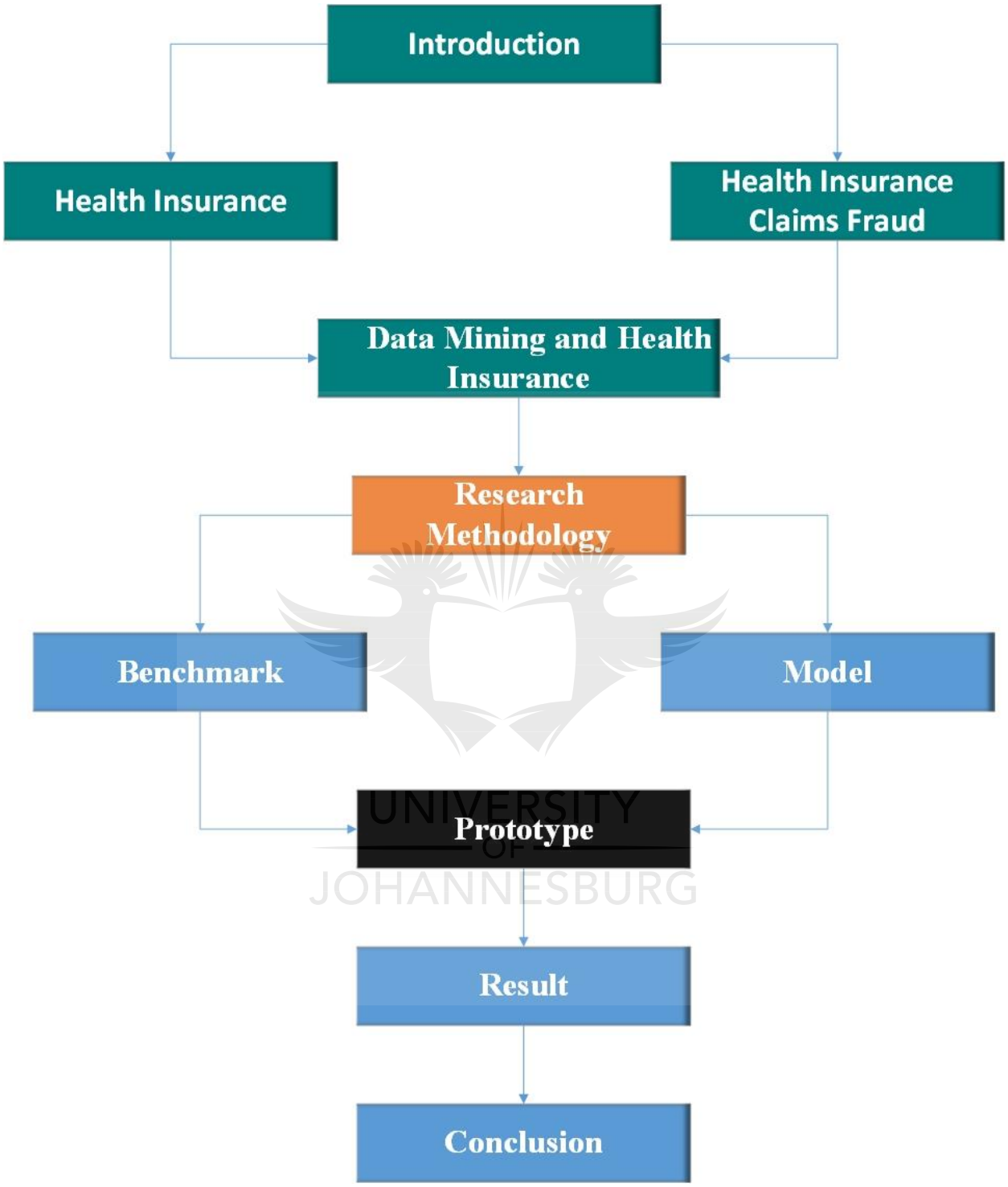
7.4 Conclusion

In the chapter, we created a benchmark which can be used to evaluate the implementation of the proposed model. We started with an overview of why benchmarking is needed. We further created metrics that will be used in the evaluation of models and these metrics were grouped into three different categories, namely:

1. Performance metrics, which analyses how well the module solves the problem and how effective the model is in differentiating the fraudulent and non-fraudulent healthcare claims.
2. Resource metrics, which addresses the need to build efficient models that solve the best problem in the most cost-effective manner. Resource metrics as part of the evaluation criteria is very important because even with the abundance of resources in modern day computing it is still important to build systems that deliver the best outcome in the most efficient manner.

Now that we have a firm understanding of the benchmark created for evaluating the implementation of the proposed model, a prototype that encapsulates it can then be implemented and benchmarked against the evaluation criteria set out. The next chapter elaborates on the model by implementing a prototype.

Dissertation roadmap



Chapter 8 Prototype

8.1 Introduction

"Experience without theory is blind, but theory without experience is mere intellectual play." Immanuel Kant's words show that while it is important to create a theoretical base for a concept, it is even more important to have a practical implementation which shows the true value of the research. To gain the true value mentioned, we create a concrete implementation of the model discussed in chapter 6. The practical implementation provides valuable insights as well as the potential problems that may arise.

The prototype makes use of different methods at the different data mining stages to achieve its goal. The application of alternative algorithms allows for the comparison of these methods to determine the best suited for the data and task at hand.

In chapter 8, we start by analyzing the requirements for the prototype which ties to the initial objectives set in chapter 1. We then follow with a discussion of the implementation of the prototype as well as an in-depth discussion of the different processes that make up the implementation pipeline.

8.2 Prototype Requirements

In view of the original aims and objectives as defined in chapter 1, we set out to identify the features that can be used to differentiate between fraudulent and non-fraudulent healthcare claims (**O1**). We also set the objective of using these features to build a model which applies machine learning algorithms to the features to detect possible fraudulent healthcare claims (**O2**). A prototype is then created to validate the model (**O3**).

By creating a data mining pipeline that contains different processes that can allow for the identification of characteristics of healthcare claims that can be used to distinguish between fraudulent and non-fraudulent health care claims. These features are then fed into a machine learning model and the model is then used to determine if a claim is fraudulent or not.

In summary, the system should provide the following capabilities to achieve its objectives:

1. Collect and process healthcare claims data (**O1**).
2. Identify common features that can be used to differentiate between fraudulent and non-fraudulent healthcare claim (**O1**).
3. Apply different machine learning methods to these features to classify fraudulent and non-fraudulent healthcare claims (**O2**).
4. Determine which of the methods perform best given the same Medicare healthcare claims dataset (**O3**).

Now we have defined the different objectives, we need to achieve to ensure success, we then discuss how we implemented a prototype to achieve these objectives. The prototype achieves these objectives by implementing the model discussed in chapter 6. We discuss the implementation of the model in the subsequent sections.

8.3 Prototype Implementation

To implement the model described in chapter 6 for classifying healthcare claims data, there several processes which make up the data mining pipeline that needs to be completed. These processes are listed below:

1. Data collection process which allows for data to be loaded from the source into the system's environment for further processing.
2. As previously mentioned, data in the real world is inconsistent and dirty so we need a data processing stage that handles any possible incomplete and noisy data.
3. Data transformation is done to extract the features that can be used to distinguish between fraudulent and non-fraudulent healthcare claims.
4. We also create a machine learning stage where the feature extracted from the data transformation is passed to the machine learning algorithm to gain insight from the data.
5. The final stage of the prototype implementation derives metrics from the different machine learning models implemented.

The prototype was implemented using the Python programming language in the Apache Spark framework (PySpark version 2.3.2). The use of the Python programming language is because it's easy to learn and has very matured data science libraries (Numpy, Scipy, Pandas, and Scikit-Learn.), a very active developer community and the varied options for visualizations and graphics (Matplotlib and Seaborn). We used Spark to take advantage of the distributed data structure Spark provides as well as the machine learning library from Spark. The following subsections will be looking more in-depth at the different stages of the implementation pipeline.

8.3.1 Data Capturing

The first stage in the implementation pipeline is the collection of data from a source. The data we made use of in the research is from Medicare and Medical Services Centre USA. We specifically made use of the Physician and Other Supplier Data for the 2016 payment records. The dataset contains the description of payment and utilization claims data and the information on the services and the procedures rendered to the healthcare subscriber.

The Medicare dataset contains records with 30 features spanning over 91 different providers. We also made use of data from the List of Excluded Individuals or Entities (LEIE) database. The LEIE dataset describes the providers excluded from practicing due to offenses committed. We load the Medicare data as retrieved from the Centre for Medical Services (CMS) storage as well as the LEIE data into an SQL database for easy manipulation. Now that we have the data stored in a repository, the next step will be to prepare the data and make it more suitable to be passed to the machine learning model.

8.3.2 Data Pre-processing and transformation

The dataset from Medicare comprises values from payments recorded after payments for claims were made and the data is properly captured and cleansed by the Centre for Medical Service. The cleaned version of the dataset is in the correct format and does not have values that will adversely affect the model. The first step in the data pre-processing stage is filtering the data for only non-prescription data.

The next challenge was labeling the Medicare data as fraudulent or non-fraudulent. We made use of data from the LEIE database to match up practitioners that have been excluded due to corrupt activities to the claims they submitted and marked those claims as fraudulent. The National Provider Identifier (NPI) code is the unique identifier of practitioners in the Medicare data but unfortunately, the LEIE data does not contain the NPI code. To link the payments made to practitioners in the Medicare dataset to the LEIE data, we make use of a technique called fuzzy string matching. Fuzzy string matching is the technique used in finding strings that approximately match a pattern. For the prototype, we perform the fuzzy matching using the first name, last name and zip code of physician. The exclusions are grouped by different rule numbers. The numbers indicate severity as well as the length of time of each exclusion. We used the providers excluded for more serious reasons, as seen in Table 8.3.1. Applying the LEIE dataset to the Medicare dataset reduced the dataset we worked with to 311,521 records.

Rule Number	Description
128(a)(1)	Conviction of program-related crimes.
1128(a)(2)	Conviction relating to patient abuse or neglect.
1128(a)(3)	Felony conviction relating to health care fraud.
1128(b)(4)	License revocation or suspension.
1128(b)(7)	Fraud, kickbacks, and other prohibited activities.

Table 8-1 LEIE Exclusion Rules

We used the string indexer to convert string columns of labels to indices. For example, the column provider type (a, b, c) was converted to indices (1,2,3). We also applied the standard scaler to the entire dataset to ensure the entire dataset was on the same scale. The selected features used in the model was based on previous works done by [91] using the Medicare dataset as seen in table 8.1 We improved the quality of the features used by calculating the feature importance of each feature to the outcome of the model. Out of the 30 features, we focus on detecting fraud by only using the procedures performed, charges, and payments, with the additional features being used for filtering and identification purposes. We select the original features from categorical variables like gender and provider type, as well as numerical values, such as payments. The remaining excluded features are demographic in nature, such as address, or redundant features, such as the HCPCS code and description features. As part of the discussion of the outcome of the model in section 9.5, we show how much influence each of these features had in the final prediction made. The use of any

remaining variables, along with applying different feature engineering approaches, is left as future work. Table 8.3.2 describes the subset of features chosen for our study.

Feature	Description	
NPI	Unique provider identification number	Categorical
nppes_provider_gender	Provider's gender	Categorical
bene_day_srvc_cnt	Number of distinct Medicare beneficiary / per day services performed	Numerical
provider_type	Medical provider's specialty (or practice)	Categorical
line_srvc_cnt	Number of procedures performed per provider	Numerical
bene_unique_cnt	Number of distinct beneficiaries per day services	Numerical
average_submitted_chrg_amt	Average of the charges that the provider submitted	Numerical
average_medicare_payment_amt	Amount paid to the provider for services performed	Numerical
Exclusion label	Fraud labels from the LEIE database	Categorical

Table 8-2 Description for some of the Medicare Features

We used stratified 5-fold cross-validation to evaluate the performance of each of the learners. The reasoning behind the use of the 5-fold cross-validation is due to the extremely low percentage of fraud labels in the entire Medicare dataset. This reduces the likelihood that a fold has too few positive class instances and retains more equitable labeled data for fair evaluation. In order to further decrease bias due to bad random draws and for better representation of the claims data, we repeat the 5-fold cross-validation process 10 times and average the scores to get the final performance results.

8.3.3 Machine learning Process

Once we are done with processing the data and extracting the relevant features, the next step is to pass these features to a machine learning method to derive insight from the data and classify the dataset. We applied different machine learning methods to the dataset to determine the machine learning method that performs best with the dataset. The problem posed in the research is a binary

classification problem as there are only two possible output classes for the Medicare dataset which are fraud or no fraud. We performed a binary classification of the data using the following algorithms listed below:

1. The first machine learning method used was the Naïve Bayes classifier. The Bayesian classifier is a statistical classifier that assigns probabilities to member classes such that a given sample belongs to a class. The classifier assumes the effect a given attribute value has on a particular class is independent of the values of other attributes. We implemented a simple Bayesian classifier that is assigned a likely class based on the highest probability as described by the feature vectors. We used the Bayesian classifier to assign the dataset to either the fraudulent or non-fraudulent class. It makes these decisions based on prior knowledge and uses it to make future predictions.

   The Naïve Bayes model was simple to implement the model. We made use of the multinomial Naïve Bayes algorithm provided by Apache Spark ML which specifies that the distribution of features is multinomial rather than the other types of distribution and a default smoothing of 1.0 as recommended by Spark documentation.

2. Random Forest: The random forest machine learning algorithm used to solve regression and classification problems. The random forest classification method makes use of an ensemble of decision trees. They operate by constructing multiple decision trees during training and then output which class is the mode of all classes for classification problems. Random forest helps reduce the problem of overfitting of training set faced in the decision tree. Individual trees in the random forest are generated from a random vector that is constructed to grow the ensembles.

   The parameters for random forest classifier was tuned using the gridSearchCV library. We created a grid of parameters which was used with the cross-validation process. The process was resource intensive as will be seen in section 9.3. Due to resource constraints, we only experimented with include the following parameters as shown in table 8.3.

Table 8-3 An explanation of hyperparameters used in the random forest classifier

Hyperparameter	Explanation
Bootstrap	Method for sampling data points (with or without replacement)
Max_depth	Maximum number of levels in each decision tree
Max_feature	Maximum number of features considered for splitting a node
Min_samples_split	Minimum number of data points placed in a node before the node is split
Min_samples leaf	Minimum number of data points allowed in a leaf node
N_estimators	Number of trees in the forest

3. Logistic Regression is another machine learning algorithm we used in the prototype. Logistic regression is suitable for predicting categorical outcomes. Linear regression is a specialized form of the generalized linear models. It makes use of binomial logistic regression to predict a binary outcome and a multinomial logistic regression to predict outcomes with multiple classes. Linear regression does not perform well with a small training set, but logistic regression uses non-linear logistic functions that improve the performance.

   To implement the logistic regression classifier, we used a regular parameter ranging from 0.1 to 2.0 and an elastic net parameter of 0.1 to 0.2. We experimented these different hyperparameters by making use of the GridSearchCV library from Scikit learn for tuning the hyperparameters. The maximum iteration was then set to 10 also based on the recommendation by the documentation.

4. Gradient-boosted tree classifier is a powerful machine learning method which we also considered. The Gradient-boosted tree classifier makes use of ensembles of decision trees. The idea of boosting came from an attempt to modify weaker learner to become better. A weak learner in this instance is that which has a slightly better performance than random chance. Gradient-boosted tree classifier tries to produce a very accurate prediction rule by using a combination of inaccurate and suboptimal rules of thumb. It makes use of a loss

function, a weak learner (decision tree) and an additive model which adds up weak learners and reduces the loss function. For the prototype, we ran the model through a variety of iterations and achieved the results shown in section 9.2.5.

5. We also used Artificial Neural Networks which can be simply described as a network of interconnected nodes that gives the ability to perform regression and classification tasks. The design of the neural network is based on the functioning of the biological neurons in the brain. The nodes function as a group of linear function but do not contain any computations. An activation function is used to define the behavior of the individual nodes. The activation functions receive inputs from other connecting nodes and produce an out on reaching a certain threshold. A neural network comprises several nodes grouped in layers with each layer having multiple nodes. A simple artificial neural network will consist of three layers: the input, hidden and output layers although the artificial neural net can technically have multiple layers. The input layer collects the input required while the output serves as a presentation layer giving the result in the form of classification through one of the several output nodes.

   For this prototype, we made use of the Keras implementation of the TensorFlow python feedforward artificial neural network libraries for our implementation. We created a three-layer sequential neural network model. There was no noticeable increase in the accuracy of the model with an increase in the number of layers in the model which influenced the decision to create the model with 3 layers. The model took in six input dimensions in the first layer. In the second layer, we experimented with a suite of activation functions to determine the best performing function. The third and output layer had a single node and used the sigmoid activation function. The model made use of binary cross-entropy as its loss function due to the problem at hand is a binary classification problem. We also experimented with various optimizers and monitored the accuracy and root mean square error metrics. The model was run with varying epochs and in different batches and the results collected. More details about the hyperparameters are discussed in section 9.2.4.

8.3.4 Testing

Now we have trained our machine learning model to understand how to differentiate between fraudulent and non-fraudulent healthcare claim, we then evaluate how well the model performs by testing the model against the test dataset. We then use this evaluation to derive metrics that describe how well the model performs.

A confusion matrix is derived after applying the model to the test data to the trained model. We derive other evaluation metrics subsequently including specificity, sensitivity, and precision. We also plot the data and we are able to determine the area under the curve.

8.3.5 Prototype Development Environment

Upon the development of the prototype, we established specific requirements and made specific design decisions to address these issues accordingly. The implementation of the prototype took about two months from start to completion and used a wide variety of the latest tools and technologies.

One of the initial issues faced was the compatibility of the different Spark versions with Scala. It was also a choice between using SparkScala or PySpark and we ultimately chose PySpark (version 2.3.2) as PySpark not only gave us access to the Spark machine learning libraries, it also gave us access to the rich Python data processing and machine learning libraries. We also made use of Microsoft SQL Server 2017 as the storage database where the claims dataset is captured and pre-processed as the Microsoft SQL server provides integration tools that can be used in the data transformation process.

The PySpark and Microsoft SQL Server environment were created using Docker. The use of Docker was motivated by how easy it is to create, deploy, and run applications by using containers. These containers allow us to package up the application with all of the parts it needs, such as the python libraries and other dependencies, and ship it all out as one package independent of the development environment.

8.4 Conclusion

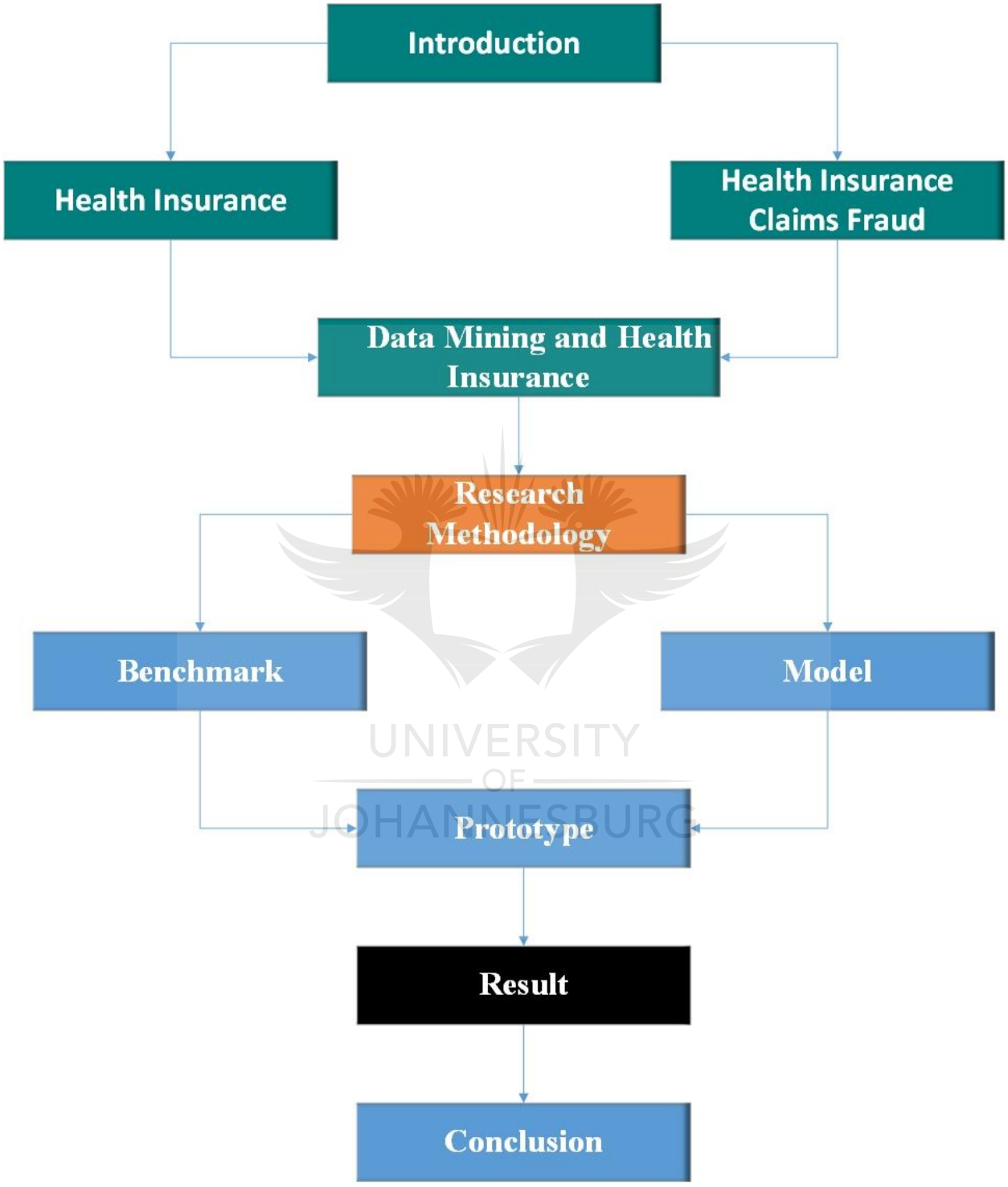
The prototype chapter starts by further unpacking the proposed model by creating a concrete implementation that outlines the main components of the pipeline as well as the guidelines and methods followed in the design and deployment of the system. We addressed the core methods used in each stage of the model and applied different machine learning methods to gain a firm understanding of which approach suits our problem the most.

The chapter shows how the objectives of the study are met as well as the decisions that were made and the reasoning behind these decisions during the implementation. The prototype was implemented using the Python programming language in the Apache Spark framework. Microsoft SQL server was used as the data storage. The components in the model have been implemented in the following manner:

1. A data collection process that loads data from the Medicare and LEIE database into the SQL database in our implementation environment.
2. A data pre-processing and transformation process that removes performs a fuzzy matchup between the practitioner in the Medicare database and the excluded practitioners on the LEIE database. The process also allows for identifying features that will be used in the machine learning model.
3. Application of different machine learning methods to the processed data. The stage consists of two phases, the testing and training phase.
4. We gained a better understanding of the different machine learning methods we applied by evaluating the model against some pre-determined test data.

The prototype we created and metrics derived after benchmarking the prototype with the dataset achieves the objective (**O3**) which is validating the model by implementing a prototype and benchmarking with a dataset. Now that we have a better understanding of the model implementation as well as the finer details of the implementation. We have established that valuable results and insights can be derived through the implementation of the system. The next chapter reveals the results and acts as a precursor to the conclusion of the study.

Dissertation roadmap



Chapter 9 Results

9.1 Introduction

The previous chapters gave sufficient background information, discussed the proposed model, followed by the benchmark and the implementation details that make up the model. The remaining part of the study is the discussion on the results as well as the conclusion of the research study. In this chapter, we discuss the results from the benchmark in chapter 7 applied to the prototype implemented.

The results that we present and analyze in chapter 9 will give a better and more firm understanding of the model discussed in chapter 6. It shows the validity and viability of the model as well as the appropriate conclusions that can be drawn on how successful the model will be in a real-world application.

This chapter begins by presenting the performance metrics results for each of the varying machine learning method used. An explanation of the implication of such a result is given as well. We then discuss the resource metrics and how the model faired in this regard. Finally, we analyzed the model to see how scalable it is as well as the quality of the data that was used.

9.2 Performance Metrics Result

As specified in the benchmark chapter, a comprehensive evaluation framework is proposed to be applied to the model created to gain a better understanding of the full potential. We apply all the evaluation criteria that we have prescribed to the different machine learning models we developed to gain a measure of the model's effectiveness.

Table 9-1 Result for the performance metric of the different machine learning methods

ML Model	Weighted Precision	Weighted Recall	AUC	Test Error	Sensitivity	Specificity	F1
Naïve Bayes	82.8	90.8	54.0	9.1	0.0	99.7	86.6
GBT Classifier	92.8	93.5	93.0	6.4	34.9	99.2	92.3
Random Forest	94.9	95.0	84.0	5.0	47.3	99.7	94.2
Logistic Regression	82.9	91.0	63.5	9.0	0.0	100	86.8
Neural Network	89	91	91.5	10.0	34.8	99.2	91.9

9.2.1 Naïve Bayes

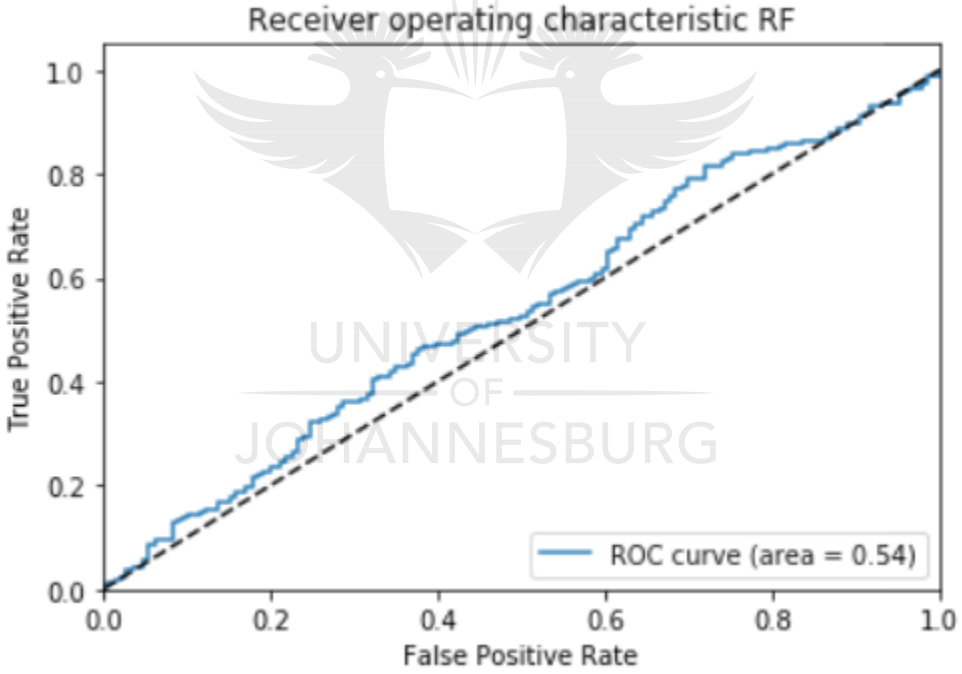


Figure 9.2-1 The ROC curve for the Naive Bayes classification model

In the Naïve Bayes model, we achieved a weighted precision of 82.8%, weighted recall of 90.8%, Test error of 9.1%, sensitivity of 0.0, a very high specificity of 99.7% and finally an F1 Score of 86.6% One noticeably low metric is AUC with a value of 54.0% meaning the predictions by the Naïve Bayes model were just as good as random guesses. The ROC curve also gives more insight into the randomness of the prediction. The result shows specificity of 0 and the implication of a

zero specificity shows how poorly the model performs when trying to identify fraudulent claims. Based on the ROC curve and specificity, the Naïve Bayes model would not be recommended for building the claims fraud detection system using the Medicare dataset.

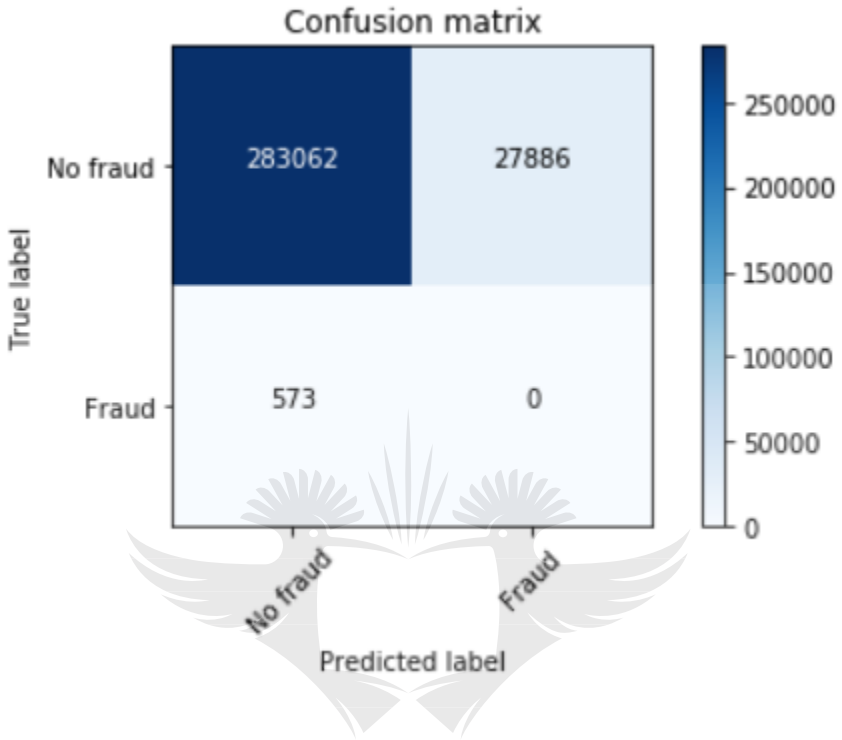


Figure 9.2-2 The Confusion Matrix for the Naive Bayes classification model

9.2.2 Logistic Regression

The logistic regression model used was a slight improvement to the Naïve Bayes model. The model achieved an accuracy of 91.0% which is slightly better than the Naïve Bayes model. There was a slight improvement in the weighted precision and the weighted recall when compared to the Naïve Bayes model with a score of 82.9% and 91.0 for both metrics. The sensitivity score was 0.0% while specificity was 100.0% and F1 score 86.8%.

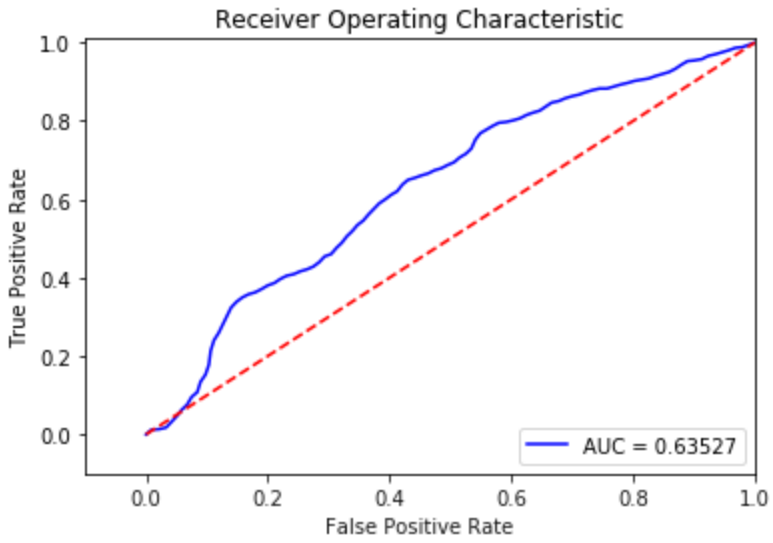


Figure 9.2-3 ROC curve for the logistic regression classification model

The logistic regression model also had a poor ROC curve with an AUC of 63.5% Although the logistic regression model presented a slightly improved performance metrics, the AUC, ROC curve and the specificity show it is not well suited for the problem we aim to solve as the low AUC score indicates the model is slightly better than random guessing. It cannot identify fraudulent claims as shown by the specificity.

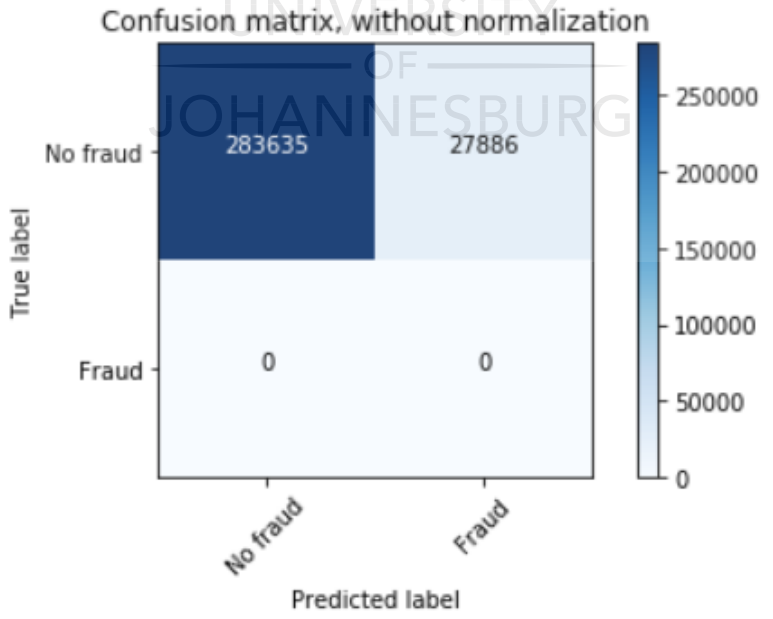


Figure 9.2-4 The Confusion Matrix for the logistic regression classification model

9.2.3 Random Forest Classifier

As explained in chapter 8, the random forest is an ensemble of multiple decision trees. We created a random forest model with a suite of parameters as shown in table 9.2. For the bootstrap, we experimented with both true and false. We also experimented with a maximum depth ranging from 10 to 140 in increment of 10. For the maximum number of features, we tried 1 to 5 features and we used 3,4 and 5 minimum sample leaves. The last two hyperparameters which we experimented with is min_sample_split (8,10,12,14) and number of estimators (100,200,300,400,800). At the end of the experiment, the results derived suggests that for the Medicare dataset, the random forest classifier performs best with a combination of the parameters and their respective values as shown in table 9.2.

Table 9-2 Hyper parameter tuning outcome for Random forest classifier

Hyperparameter	Value
Bootstrap	True
Max_depth	80
Max_feature	3
Min_samples_split	12
Min_samples leaf	5
N_estimators	100

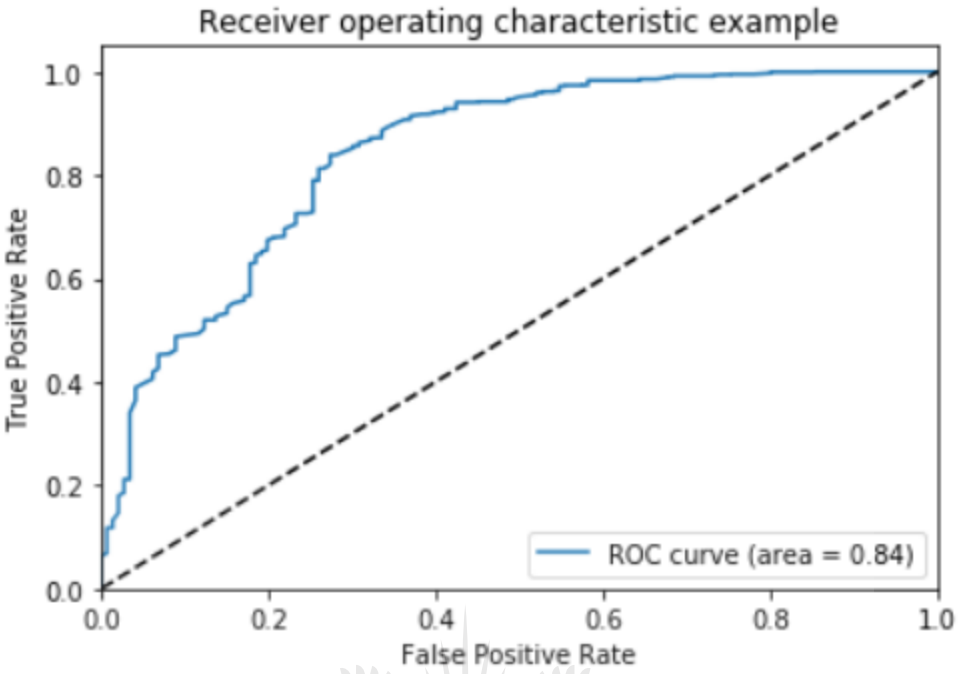


Figure 9.2-5 The ROC curve for the random forest classification model

The random forest offered a great improvement to the logistic regression with an AUC of 84.0% after tuning the hyperparameters of both models. The random forest classifier had a sensitivity of 47.3%, a specificity score of 99.7%, and an f1 score of 94.2 and an accuracy of 95%. The model recorded a weighted precision of 94.9% and a weighted recall of 95.0%.

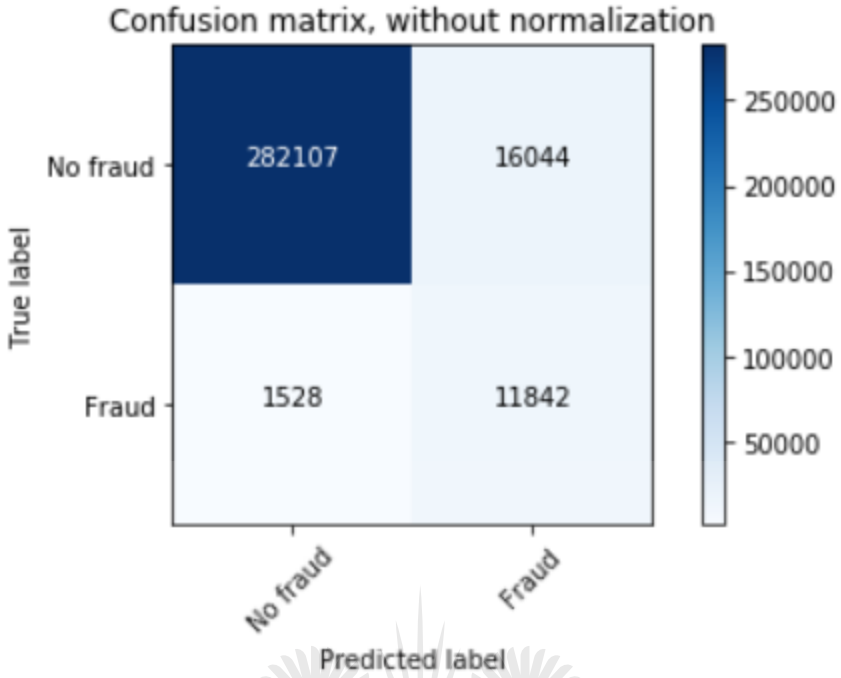


Figure 9.2-6 The Confusion Matrix for the random forest classification model

The metrics derived from the random forest classifier suggests that the random forest classifier is more suitable for classifying the Medicare dataset as the score implies a great improvement to random guess noticed in the previous models. The AUC score of 84% shows that the model actually learned how to classify the data rather than random guessing.

9.2.4 Artificial Neural Net

The TensorFlow feedforward artificial neural network (Basic multi-perceptron feed forward) slightly underperformed against the gradient boosted tree classifier. The artificial neural network was run through varying epochs. The model made use of the binary cross-entropy as the loss function. We made use of the GridSearchCV library provided by Keras to tune and optimize the hyperparameters. To determine the best combination of batch size and epochs.

We evaluated a suite of different mini batch sizes from 10 to 100 in increments of 20. We also evaluated the model with epoch sizes ranging from 10 to 100 in increments of 10. The batch size specifies the number of patterns shown to the network before the weights are updated. The number of epochs defines how many times the entire training dataset is shown to the network while

training. The results in table 9.3 show that the best result (accuracy 89) is achieved with a combination of 100 epochs and batch size of 20.

Table 9-3 Result for the hyperparameter tuning of the epoch and batch size from the ANN model

Epochs	Batch Size	Accuracy
10	10	0.648958
50	10	0.648958
100	10	0.766146
10	20	0.847135
50	20	0.860156
100	20	0.891198
10	40	0.789583
50	40	0.352344
100	40	0.954948
10	60	0.518229
50	60	0.605469
100	60	0.665365
10	80	0.537760
50	80	0.591146
100	80	0.658854
10	100	0.402344
50	100	0.652344
100	100	0.542969

We tuned the activation function used in the hidden layer. The activation function is important as it controls when individual neurons fire as well as their non-linearity. From the result, in table 9.4 we see that we achieved the best accuracy (90) when we used the relu activation function. The ReLu activation function has become very popular and according to [92], it was proved that the ReLu function had 6 times improvement in convergence than the Tanh function.

Table 9-4 Result for the hyperparameter tuning of the activation function from the ANN model

Activation	Accuracy
softmax	0.848957
softplus	0.863482
softsign	0.871384
relu	0.900317
tanh	0.886427
Sigmoid	0.822292
Hard_sigmoid	0.873125
linear	0.892543

The choice of optimization algorithm we choose for our model can determine the difference between good results in minutes, hours, and days. For the model optimizer, we used the following optimizers in the parameter grid: SGD, RMSprop, Adagrad, Adadelta, Adam, Adamax and Nadam. The result would suggest that with an accuracy of 90, the Adam optimizer is the best for the problem we are trying to solve. Another reference point for the choice of Adam as the optimization algorithm comes from the Stanford course on deep learning for computer vision titled "CS231n: Convolutional Neural Networks for Visual Recognition" delivered by Andrej Karpathy, he suggested that the Adam algorithm is used as the default optimization method for deep ANN applications.

Table 9-5 Result for the hyperparameter tuning of the neurons in the hidden layer from the ANN model

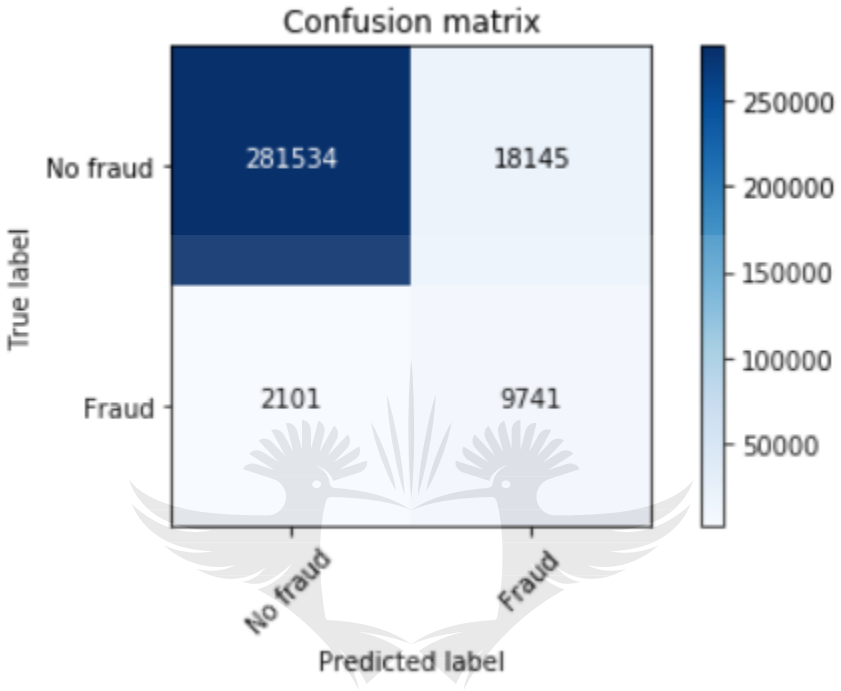
Neuron	Accuracy
1	0.848957
6	0.863482
12	0.900317
18	0.800317
24	0.816427
30	0.822292
36	0.823125

Table 9-6 Result for the hyperparameter tuning of the optimizer from the ANN model

Optimizer	Accuracy
SGD	0.348958
RMsprop	0.348958
Adagrad	0.471354
Adadelta	0.669271
Adam	0.904427
Adamax	0.682292
Nadam	0.703125

From table 9.2.3 we can see that the best accuracy score was derived when the combination of 100 epochs and batch size of 20. Table 9.2.4 shows the outcome of the hyperparameter tuning for the activation function. The result suggests that the relu is best suited for the Medicare dataset. Experimenting with a different number of neurons in the hidden layer presents the results as can be seen in table 9.2.5 that suggest 12 neurons in the hidden layer is the most ideal. Finally, Adam

optimizer slightly outperformed the other optimizers as seen in table 9.2.6. The results we achieved using the accuracy as a benchmark suggests that a combination of 100 epochs with a batch size of 20, relu activation function and the Adam Optimization algorithm with 12 neurons in the hidden layer produces the best outcome.



The artificial neural network slightly underperformed against the gradient boosted tree classifier as can be seen in the metrics in table 2. The model made use of the binary cross-entropy as the loss function. The model had precision and a recall of 89% and 90% respectively. It also had an accuracy of 90% with specificity and sensitivity measuring 71.2% and 94.8% respectively. The F1 score was 90% and it also had an AUC of 93.8%.

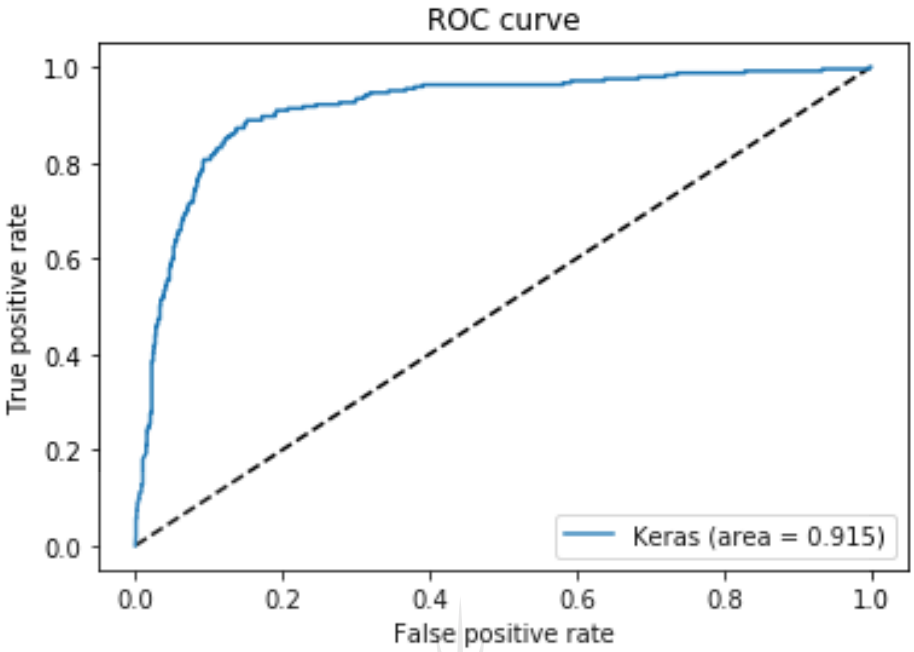


Figure 9.2-7 ROC curve for the Artificial Neural Network classification model

9.2.5 Gradient Boosted Tree Classifier

The gradient boosted tree classifier proved to be a very powerful machine learning method as seen in the results it produced. A simple to implement model yet combines the strength of weaker learners to produce a very powerful learner. We ran the gradient boosted tree classifier for ten iterations.

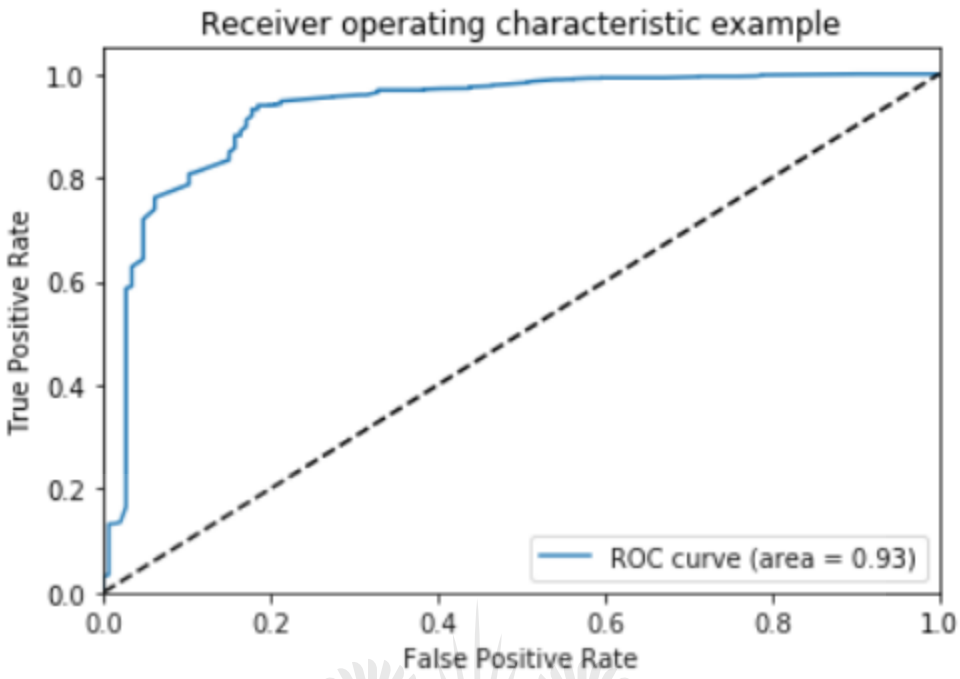


Figure 9.2-8 3 The ROC curve for the Gradient Boosted Tree classification model

The gradient boosted tree classifier model presented a great improvement to the previous models. The model was run through 10 iterations. The gradient boosted tree classifier had a weighed precision and recall of 92.8 and 93.5% respectively. It had an accuracy of 93.6% and an F1 score of 92.3 %. It recorded a sensitivity score of 34.9% and a specificity of 99.2%. The most important improvement was the noticeable increase in the AUC score of 93.0%. Based on the results the gradient boosted tree classifier will be ideal for detecting possible fraud in healthcare claim using the Medicare data. The high performance of the gradient boosted tree classifier can be attributed to the boosting model. Boosting models perform well because they intelligently give more weight to the observations that are hard to classify making it more efficient in classifying datasets.

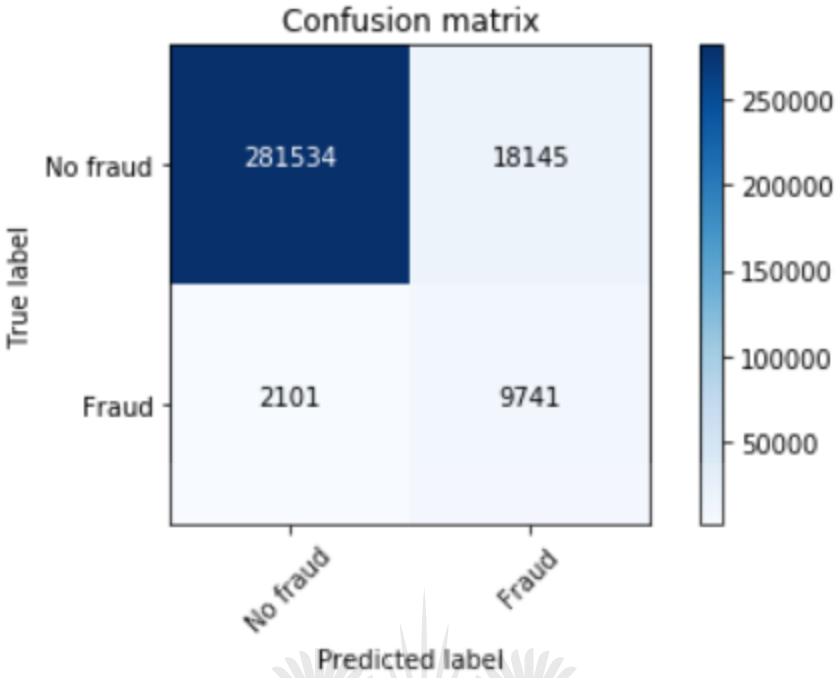


Figure 9.2-9 The Confusion Matrix for the Gradient Boosted Tree classification model

The result shows that the models have F1 scores ranging from 86% to 94%. The result means that we have low false positives and low false negatives, so we are correctly identifying real threats and we are not disturbed by false alarms. The results obtained from the performance of the models also show that most of the models had a low specificity score due to the class imbalance of the Medicare dataset. The Medicare data contained more non-fraudulent claims than fraudulent claims as can be seen from the confusion matrixes above. The low specificity among the models shows that all the models struggled to classify the negative class of the dataset. There was an increase in the specificity score with the random forest classifier, gradient boosted tree classifier and artificial neural net. Although the increase was noticeable, more work still needs to be carried out to resolve the issue of class imbalance in order to reduce the bias of the models towards the non-fraudulent cases. We have enlisted the resolution of class imbalance as part of the future work to be done.

9.3 Resource Metrics

In creating the model, we kept in mind the need to have a solution that is scalable as well uses efficient algorithms. The choice of PySpark as the development was to create a more scalable system that can be deployed to a Hadoop cluster and using the distributed computing, we can

perform computations faster. In terms of algorithm efficiency, we implemented the model by applying good programming practices. The model made minimal use of memory with the highest memory usage seen in the artificial neural nets. The model successfully ran on a machine of 16gb of RAM.

We also measure how long it took to train each model. We did this by using the timit function. Table 9.3.1 below presents shows the duration elapsed by each model while training.

Table 9-7 Model training duration

Model	Duration (in fractional seconds)
GBT	0.21821822
NV	0.11381394
Logistic Regression	0.12121312
Random Forest Classifier	57.02785873
Artificial Neural Net	2303. 1218587

The time statics shows that while the gradient boosted tree classifier, Naïve Bayes and logistic regression models took a fraction of a second, the random forest model classifier took almost a minute to train. The artificial neural net also showed a high time statistic. From the time statistics, we can see that at a fraction of the time used by the random forest and artificial neural net classifier, the gradient boosted classifier achieved similar results as can be seen in table 9.3.1. Hence, we can recommend the gradient boosted tree classifier as an ideal classifier since it performed better than the other models in resource usage as well as the performance metrics when there is a limited supply of computational resource, but the higher sensitivity of the random forest classifier makes it also ideal even though it takes more computational resource to train.

9.4 Robustness

The model implemented ensured it catered to the possibilities of malformed data by creating a pipeline that incorporates a well-established data pre-processing and transformation process. We created a data cleaning process that effectively handles incorrect data. We created a very tolerant system that can handle data of varying quality due to the well-established feature engineering process of the model. The feature engineering process ensures that the model produces a consistent result with different types of data quality.

9.5 Characteristics of data used

The characteristics of data used metrics define how good the data used is. The measure of the quality of the data used can depend on several factors such as the volume of data and the population of the data. Typically, the more data the better the model. For our prototype, the Medicare data used contained 311,521 records which were good enough for the learning and testing of the model. An area of improvement is the population of the data as we experienced issues of class imbalance. The class imbalance was as a result of the Medicare data containing a significantly higher number of non-fraudulent claims than fraudulent claims which can make the model biased towards the non-fraudulent claims.

We also collected metric result for the feature importance of each model. The table below shows the measurement of how much each feature contributed to the outcome of the model.

Table 9-8 Feature Importance for the features and corresponding machine learning models

	Naïve Bayes	Logistic regression	Random forest	Gradient boosted tree
NPI	0.2068	**0.3023**	0.3606	0.2138
nppes_provider_gender	0.1300	0.1345	0.0198	0.1198
bene_day_srvc_cnt	0.0220	0.0245	0.0138	0.0702
provider_type	0.0312	0.0287	**0.4659**	**0.4106**
line_srvc_cnt	0.1173	0.0023	0.0486	0.0466
bene_unique_cnt	0.1150	0.2300	0.0703	0.0300
average_submitted_chrg_amt	0.0327	0.0327	0.0123	0.0120
average_medicare_payment_amt	**0.3450**	0.2450	0.0087	0.0970

The feature importance for the random forest and gradient boosted tree classifiers were derived from the feature importance method in PySpark while the Naïve Bayes and the logistic regression models made use of the theta function to retrieve the feature importance since the models do not directly expose the feature importance metrics. The closer the score is to 1 the more influence the variable has on the outcome of the prediction and the closer the value is to 0 the most likely it is to have some marginal impact on the outcome of the response variable. For example, in the gradient boosted tree classifier, we can see that the NPI has the highest impact on the outcome while the feature average_submitted_chrg_amt with a score of 0.0120 has the least impact on the outcome of the model.

The function for determining relative importance for the artificial neural net model is the permutation importance module from the ELI5 package. Although it most easily works with a scikit-learn model the wrapper for sequential models which Keras provides helped in understanding the relative importance of the features in the artificial neural net. The score shows to what extent the model performance decreased with a random shuffling (in our case, it uses "accuracy" as the performance metric)

Table 9-9 Feature Importance for ANN model

Variable	Score
NPI	1.8246
nppes_provider_gender	0.1198
bene_day_srvc_cnt	1.3138
provider_type	0.7002
line_srvc_cnt	1.0466
bene_unique_cnt	1.0250
average_submitted_chrg_amt	1.1013
average_medicare_payment_amt	0.0167

Table 9.5-2 shows us that the variables provider_type and NPI have the strongest relationships, respectively with the response variable. Similarly, variables that have relative importance close to zero, such as average_medicare_payment_amt, do not have any substantial importance for the outcome of the fraud label. It is worth noting that these scores show relative importance to an outcome such that a feature with a score of zero or less will most likely have some marginal impact on the outcome of the response variable.

9.6 Algorithm Comparison

We have seen how each of the models performed against the benchmark used. In this section, we compare how these machine learning methods compare to each other. Figure 9.6 shows a comparison of all the models based on their ROC score. The box plot represents the distribution of the AUC scores for all the iterations of the cross-validation for each model. It shows that the gradient boosted tree classification model, the random forest classifier and the artificial neural network models have similar performance in terms of AUC scores which range from about 0.75 to 0.85. The logistic regression model recorded a poor AUC score which ranged from about 0.30

to 0.65 with a median score of 0.55. The Bayesian model had the worst AUC score with the median AUC score falling just under 0.40 and the best AUC score at about 0.54.

The result shows that while the gradient boosted tree classification model, the random forest classifier and the artificial neural network models can be used to classify the Medicare data, the performance of the logistic regression and Bayesian models makes it unsuitable for classifying the Medicare data set.

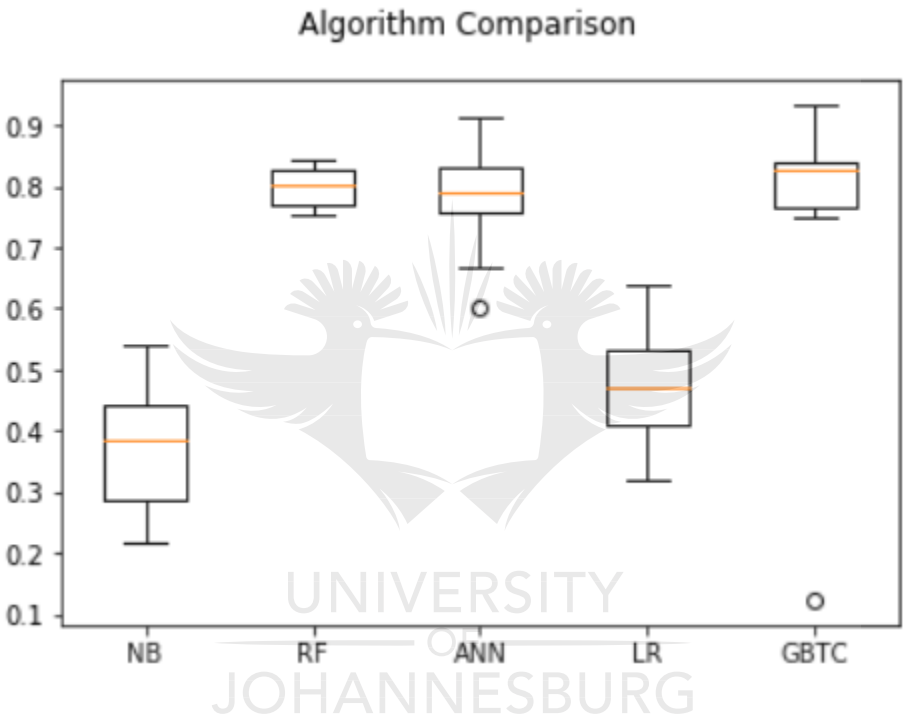


Figure 9.6 A comparison of all the models based on their ROC curve.

The models struggled to classify the fraudulent claims. The reason for the misclassification by the models can be largely attributed to the class imbalance in the data set. The fraudulent cases in the dataset were only11842 record compared to the 299679 records for non-fraudulent instances. The class imbalance introduced a bias towards the non-fraudulent cases in the model.

9.7 Error Analysis

We have the results of the performance of the metrics and we have seen how the gradient boosted classifier performs best when compared with the other models. In this section, we perform an error

analysis of the gradient boosted classifier to understand the possible cause of the misclassified predictions in the cross-validated dataset.

The process of error analysis involves manually examining the examples of the misclassified data in the cross-validation dataset to understand why the algorithm misclassified the data. For the gradient boosted tree classifier model, we see that out of the 311,521 records, 20,246 were misclassified.

We manually examine these 20,246 to see the relationship between the records. A look at the feature importance on the gradient boosted tree classifier in table 9.8 shows that provider type had the highest impact on the outcome of the predictions. We then categorize the misclassified data based on the type of providers. We make the following deductions.

Table 9-10 Error analysis for gradient boosted tree classifier.

Provider type	Percentage Error
General Surgery	14.7
Pain Management	12.5
Internal Medicine	13.5
Family Practice	49.3
General Practice	4.3
Otolaryngology	2.1
Nurse Practitioner	3.6

From table 9.10, we can see that almost 50% of the misclassified data comes from the family practice provider type. The misclassification of internal medicine type showed that the model had difficulty in understanding the dataset with regards to this provider type. That means more attention needs to be paid to why records belonging to the family practice provider type were misclassified.

A look at the entire dataset shows that about 23% of the fraudulent records come from the family practice provider type which represent a large portion of the entire dataset consisting of 91 provider types in total. The model misclassified the records as even though the NPI indicate that they are non-fraudulent, the high feature importance of the provider type in the gradient boosted tree classifier means that more priority will be given to the NPI belonging to the family practice provider type which influences the outcome. In this case even though the practitioner's claim was

non-fraudulent, the provider type which it falls under influenced the model to classify the record as fraudulent.

We can see that the feature importance, the label from the LEIE database as well as the type of provider being analysed are very important in the final outcome of the prediction. Therefore, future work on this prototype will involve more attention to the feature engineering process to see how we can derive more feature from the provider type and the NPI.

9.8 Conclusion

The chapter created a better understanding of the strengths and weaknesses of each system by presenting facts and figures to show the viability of the solution. The chapter started by discussing the performance metrics and how each of the machine models performs based on the metrics. The metrics used in the chapter were based on the defined metrics in the benchmark.

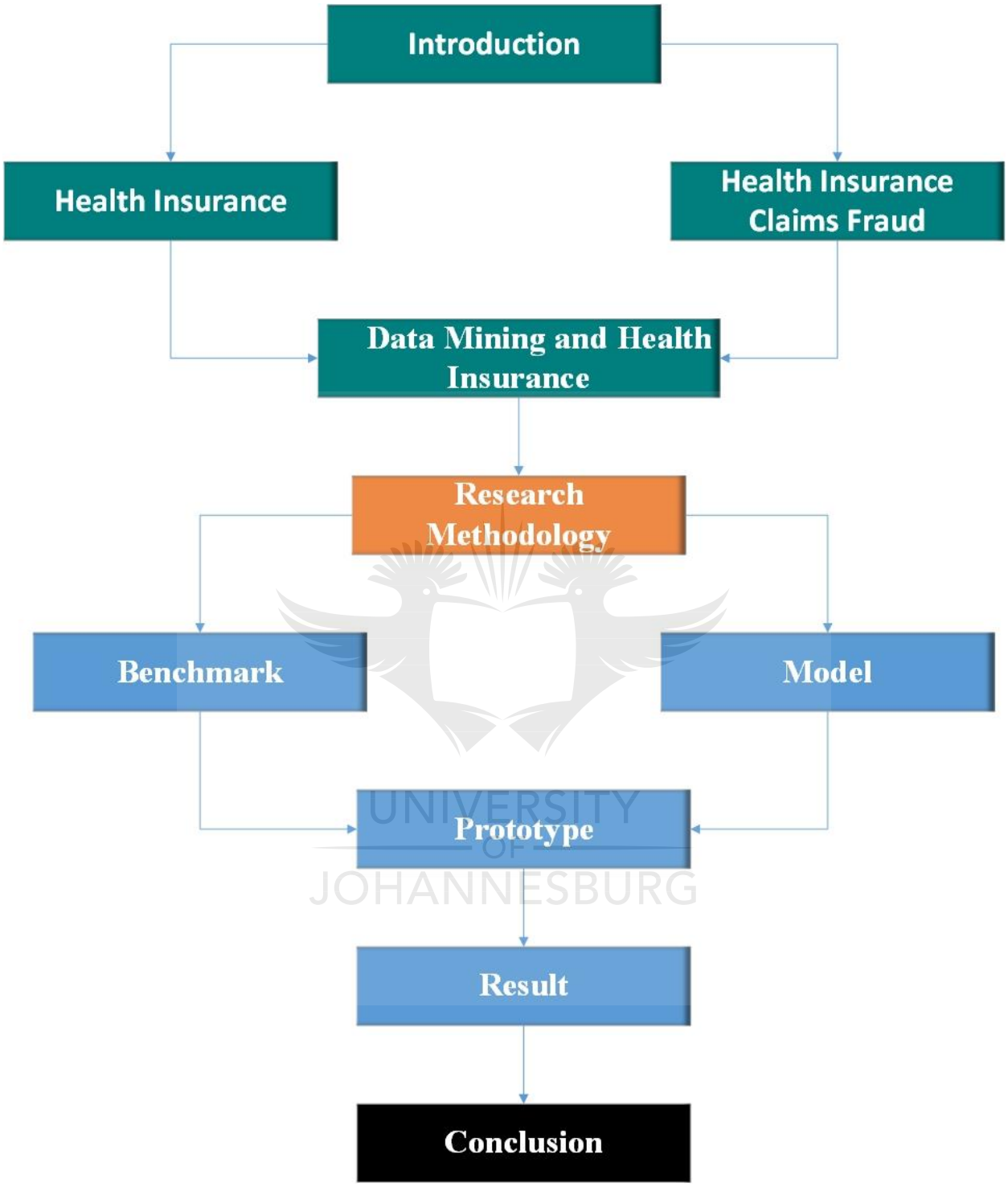
For each of the machine learning models implemented, we applied the defined metrics to benchmark the performance. The Naïve Bayes model performed poorly, with an improvement in the logistic regression model. The best performing models were the gradient boosted tree classification model, the random forest classifier and the artificial neural network model. We showed how each of the models performed against the metrics as well as an illustration of the ROC curves for each model.

We also analyzed the models in terms of robustness, characteristics of data and resources used. The presence of data processing methods in the model makes it robust enough to handle different kinds of data. The models made use of minimal resources and also used relevant data for classification. Finally, a comparison of the different models used was performed and the results suggests that the gradient boosted tree classifier was best for detecting possible healthcare fraud using the Medicare dataset. We then analysed the misclassification that occurred using the gradient boosted tree classifier to understand what led to the errors in the model.

The completion of the discussion of the results leads to the conclusion of the study. In conclusion, we unpack how the model can be applied practically as well as how the model

achieved the objectives set out for the study. A critique of the model, as well as support, is given followed by an overall conclusion.

Dissertation roadmap



Chapter 10 Conclusion

10.1 Introduction

The previous chapter presented the results of the model by evaluating the implemented prototype based on the benchmark criteria defined in chapter 7, thereby allowing the final conclusions of the study to be drawn. The potential of creating a data mining model to detect possible healthcare claims fraud is established and further be analyzed and unpacked for further insights.

The initial objectives set at the beginning of the study can be aligned with the results generated and the advantages and disadvantages of the model can be analyzed and further unpacked. The insights gained from conducting this study can also be presented as future work in the healthcare fraud detection field. By leveraging the study findings and the insights gained, an overall assessment and conclusion can be made in the study.

The conclusion chapter begins by performing hypothesis testing to determine the underlying hypothesis holds true and then discussing the potential applications of the model in the healthcare claims fraud detection problem space, along with its deployment potential. Next, we perform a critique of the research study and then arguments that support the study. Based on the lesson learned from the study, we discuss possible future work that can be carried out in the research problem space and finally an overall conclusion of the study is provided.

10.2 Hypothesis testing

The hypothesis H1 which states that machine learning methods can be used to effectively detect possible fraud in the health insurance claims process can be proved by the process of hypothesis testing. Hypothesis testing can be defined as a scientific assertion that can be tested by observing a process which is modeled with a set of random variables [93]. Hypothesis testing makes use of a null hypothesis H_0 which corresponds to the default natural state. It also has the alternative hypothesis H_a which is the opposite of the null hypothesis.

There are two kinds of statistical hypothesis error testing, the "Type I and Type II errors" which relies on the hypothesis which has been identified. An error occurs when after a test is conducted, the result of the test does not match the actual state of the condition. Type I error which is analogous to False positive from the confusion matrix occurs when the statistical test performed falsely states otherwise when the true state exists [93]. The Type II error which is analogous to the false negative occurs when the test carried out fails to identify false state existence. Type I error is when a null hypothesis is rejected when it is true and Type II error is when a null hypothesis is accepted when it is actually false.

The Neyman-Pearson paradigm suggests that Type I error occur with a probability expressed as α which is termed the significance level of the statistical test and the Type II error occurs with a probability of β [93]. To statistically test the hypothesis, we do the following below:

H_0: Machine learning cannot be used to effectively used to detect possible fraud in the health insurance claims process.

H_a: Machine learning methods can be used to effectively detect possible fraud in the health insurance claims process.

		Actual	
		Positive	*Negative*
Predicted	*Positive*	True Positive	False Positive (Type I error)
	Negative	False Negative (Type II Error)	True Negative

Table 10-1 Confusion matrix showing statistical error types as visualized by author

From the analysis of the confusion matrix in Chapter 9 and Table 10.2, we can see that the Type I Error represents the false positive and Type II error represents false negative. A perfect model which is difficult to attain is a diagonal matrix with only true positive and true negative values. As

this situation most likely doesn't occur, the decision to make what type of error should we favor between the two errors.

The answer to the question depends on the implications of the prediction. In the case of detecting healthcare claims fraud, we want the lowest possible false negative rate (you want to avoid letting fraudulent claims go undetected). All the models recorded very high true positive rate as can be seen in chapter 9. The highest sensitivity recorded was 47% by the random forest classifier. We established the reason for the low sensitivity score to be the class imbalance in the dataset.

Statistically, to prove the hypothesis we analyse the prediction probability. The prediction probability is a vector of two probabilities as seen in figure 10.2.1. The first probability for the case when the record is non-fraudulent and the second shows when the record is fraudulent. The sum of these probabilities equal to 1. The predicted class is determined based on whichever probability is higher.

We calculate the p-value of the distribution to determine whether to accept or reject the null hypothesis. The p-value is the probability that our model would be inconsistent with the hypothesis, thereby assuming the null hypothesis is true. The test we perform to determine the p-value is the one sample t-test.

We calculate the mean of the distribution and get the value 0.875 and a standard deviation of 0.13. With the use of the python ttest_1samp library, we get a t statistic of 117.3. We calculate the p-value to be 0.0 which is less than the chosen significant level 0.05. Therefore, we reject the null hypothesis and the hypothesis that states that that machine learning methods can be used to effectively detect possible fraud in the health insurance claims process stands.

	prediction	label	probability
0	0.0	0.0	[0.9232881876321789, 0.0767118123678211]
1086	0.0	0.0	[0.9233030265638716, 0.07669697343612836]
1076	0.0	0.0	[0.8900421801997246, 0.10995781980027541]
1085	0.0	0.0	[0.9233030265638716, 0.07669697343612836]
1084	0.0	0.0	[0.9337993917284765, 0.06620060827152352]
1083	0.0	0.0	[0.9337993917284765, 0.06620060827152352]
1082	0.0	0.0	[0.9375252491303431, 0.062474750869656925]
1081	0.0	0.0	[0.9344298048071163, 0.06557019519288365]
1080	0.0	0.0	[0.9303110417499213, 0.0696889582500787]
1079	0.0	0.0	[0.9303110417499213, 0.0696889582500787]

Figure 10.2-1 Prediction and Prediction probabilities for the gradient boosted tree classifier model

10.3 The Potential Applicability

The healthcare industry has become a very important pillar in modern society but has witnessed an increase in fraudulent activities. The increased difficulty in the task of detecting possible healthcare claims has created the need for measures that can effectively identify these claims. Unfortunately, the traditional methods put in place to detect these fraudulent claims do not suffice anymore due to human dependence as well as the required subject matter expert knowledge. Furthermore, more novel healthcare claims fraud detection systems have used machine learning to detect these fraudulent claims. However, there is still a need to further explore the problem space to address some of the objectives identified in section 1.3.

The model we proposed attempts to address these problems by creating very robust and effective data pre-processing and transformation process in the implementation pipeline to identify and select these features that differentiate between fraudulent and non-fraudulent healthcare claims (to address objective **O1**), consolidating those identified features and passing them to the classification model. We also defined and created a model for the detection of possible healthcare claims fraud using data mining methods in chapter 6 (to address objective O**2**). Finally, an implementation of

the model was done to validate the model created in chapter 6 as well as understand the advantages and disadvantages of the methods applied (to address **O3**).

Overall, as can be seen by the results derived from the implementations, the detection of possible fraudulent healthcare claims is viable using machine learning methods, therefore fulfilling the primary objective of the study outlined in section 1.2. The metric result shown in chapter 9, acts as validation and to show there is value implementing a machine learning model to detect possible healthcare fraud.

10.4 Critique

So far, we have shown the functionalities of the proposed model as well as the results achieved. Despite the applicability of the model to facilitate the identification of possible fraudulent healthcare claims, it is not without its caveats or limitations as well. The limitations can become a hindrance in the implementation of the proposed model for the healthcare industry. These limitations are to be considered when making decisions when implementing the proposed model. The areas that these limitations are predominant include the sample data, feature selection, and deployment.

10.4.1 Sample Dataset

When choosing a sample dataset that will be used to build the data mining model, it is important to ensure that the dataset covers a wide variety of features and also a large population. Data plays a very important role in the data mining process and the more variety your data has more insights that can be derived from the data. The data that was used for the model only considered fraud by the medical practitioner. Although the majority of the fraud occurs through the medical practitioner that does not necessarily mean that possible fraud relating to the health insurance subscriber should not be considered. The data used in the study restricts the scope of the problem of fraud related to medical practitioners.

10.4.2 Lack of labeled data

One of the main challenges that were encountered when implementing the model was the lack of label data to be used for supervised learning. We could not find labeled health insurance data that was publicly available at the time of the study. The Medicare data that was used was not labeled, hence the need for consolidation with the LEIE data to identify potentially fraudulent claims. The use of an already labeled data will be a better means of validating the model.

10.5 Support

Although from the above critique, we have seen that there are some limitations to the proposed model, there are also some redeeming features that encourage the use of the model over the traditional healthcare fraud detection methods. These advantages should serve to encourage the use of machine learning methods embedded in the data mining pipeline to detect possible fraud. These beneficial aspects of the model that make it an attractive option as discussed below.

10.5.1 A comparison of different machine learning methods

Choosing the right machine learning method to apply on a dataset and problem space is not a trivial task. The approach we took in the study presented several machine learning methods applied to the Medicare dataset. Using multiple machine learning methods allowed for the comparison of the different performances and therefore presented a clearer view of which model best classifies the data. Through the arguments presented for each of the machine learning methods, one can effectively analyze the advantages and disadvantages of each model before deciding on which method to implement.

10.5.2 Flexibility

As a result of the structure of the model, it allows for flexibility in the implementation of a healthcare fraud detection system. The components of the model were designed in a modular nature to allow for easy replacement of a component with another and switch between different methods in the different stages of the pipeline with more efficient algorithms. The modular design of the model reduces the burden of upgrades to the system that may be necessary for the future.

The flexibility in the design of the model makes with rapidly changing trends and formats in crime in healthcare fraud. With the critique and support of the study discussed, we proceed to discuss several insights that can be drawn from the study.

10.6 Lessons Learnt

Through the course of the study, there were several insights gained that further support the hypothesis of our study. We can further explore the insights gained, and lessons learned from the study to see its potential in the implemented model as well as other related fields.

The importance of data in the 21st century cannot be over-estimated as seen in the study so far. From the implementation of the model, we can see the choice and quality of data used plays a major role in the outcome of the model. The data mining system works based on the garbage in garbage out principle, with better outcome and performance coming from data of higher quality. From the study, it can be deduced that the data pre-processing and transformation processes are as important as the machine learning process. If the data is not processed in the right manner, there will be a negative impact on the results and performance of the machine learning process. Poorly engineered data can lead to increased computation cost as some of the machine learning methods will take an extended time to reach the decision boundary as well as reducing the probability of the algorithm converges.

During the training stage of the classification model, there's a great deal of time and computational resource used. Persisting the model becomes very important to reduce the computational cost used up during testing. Persisting the model allows for reusing the model to make classifications rather than the need for retraining. Persisting the model also allows for comparison between the old and previously trained model with newer models and determining the model degradation.

10.7 Future Work

The study has revealed areas of research that can possibly be explored to further improve the model we implemented. The lack of labeled data was a limitation in building the healthcare fraud detection model. Finding health insurance data with explicit ground truth label is a difficult task hence restricting the use of supervised machine learning methods. Due to the lack of labeled

healthcare payment data, an area of research that can be explored is using Semi-supervised online machine learning methods. The aim will be to create an adaptive model for healthcare fraud detection using computationally efficient online Semi-Supervised Learning methods. The proposed model will offer the potential to detect known and unknown attacks as well as the capability to be updated according to the novel trends of data discovered by the domain experts in a cost-effective manner.

In the study, we explored the conventional machine learning methods as well as the artificial neural network. Further work can be done in applying the machine methods based on the deep learning architecture to the problem of detecting fraud in the healthcare claims process. With the growth in popularity of the application deep learning methods to other domain areas, more insight can be gained on how deep learning performs with the detection of healthcare fraud.

10.8 Overall Conclusion

Enabling the identification and detection of possible fraudulent healthcare claims has become a very important yet difficult task due to the growing complexity of the medical billing process. The constant evolution in the fraud patterns, as well as the increase in the volume of healthcare, claims to be processed has led to the use of machine learning methods as a means of detecting these fraudulent claims. Machine learning methods can be seen as an improvement to traditional healthcare claims fraud detection.

The study explores the use of machine learning methods to effectively identify possible fraudulent healthcare claims by designing a model that specifies a machine learning fraud detection model pipeline (thereby solving **O2**). The model comprised of the different processes that need to be carried out to detect fraudulent claims. The study made use of data from the Medicare database for healthcare payments as well as data from the LEIE database for ground truth labels. The model was then implemented, a feature extraction phase in the implementation process helped identify the features that can be used to distinguish between fraudulent and non-fraudulent healthcare claims (thereby solving **O1**). Different machine learning methods were applied to the dataset and the results derived as well as the hypothesis testing performed helped achieve **O3**. These results proved that the initial hypothesis that machine learning methods can be used to detect possible healthcare fraud is valid. Therefore, the primary objectives of the study were achieved.

Through the insights gained from the study, we were able to identify some limitations in the study, which should be taken into consideration when implementing healthcare fraud detection systems. The use of the Medicare dataset limits the model to detecting fraud that only occurs through the medical practitioners and not fraud that may occur through the insurance subscriber and the insurance service provider. The difficulty in finding sample healthcare claims data that has been labeled and classified was also a limiting factor in the study.

However, the redeeming qualities that the model has, shows that it can be a viable solution to detecting possible healthcare fraud. The comparison of different machine learning methods helps for the analysis of the advantages and disadvantages of the different approaches to solving the problem. The modular implementation of the model makes it easy for the component to be easily replaced as well as upgraded as need be.

The study was aimed at building an automated model that can detect healthcare claims fraud. With the use of the data from Medicare and LEIE, we achieved the aim we set for the study to a plausible degree. We achieved the aim by implementing an automated data mining model consisting of different machine learning methods which allow us to determine which methods perform best in identifying fraud in the medical billing process. The knowledge derived from the dissertation will serve as guidance on how to build healthcare fraud detection systems using the Medicare dataset.

References

[1] J. Gee and M. Bu, “The Financial Cost of Healthcare Fraud 2014,” Centre for Counter Fraud Studies, Portsmouth, 2014.

[2] J. Broomberg, “Current fraud trends within the healthcare industry,” 2012. [Online]. Available: https://www.fanews.co.za/article/fraud-crime/5/general/1094/current-fraud-trends-within-the-healthcare-industry/12293. [Accessed 13 April 2017].

[3] S. Rajasekar, P. Philominathan and V. Chinnathamb, “RESEARCH METHODOLOGY,” Cornell University, New York, 2013.

[4] J. W. CRESWELL, Research Design: Quantitative, Qualitative and Mixed Methods Approaches, London: SAGEPublicatíons, Inc, 2009.

[5] O. Werby, “Health, Human Rights, and Maslow’s hierarchy of needs,” 2017. [Online]. Available: http://www.interfaces.com/blog/2013/09/health-human-rights-and-maslows-hierarchy-of-needs/. [Accessed 15 06 2017].

[6] H. Yu and A. W. Dick, “Impacts of Rising Health Care Costs on Families with Employment-Based Private Insurance: A National Analysis with State Fixed Effects,” *Health Service Research,* vol. 47, no. 5, 2012.

[7] B. Bemisaal, Handbook on Health Insurance, Mumbai: Insurance Regulatory and Development Authority, 2016.

[8] L. Gorman, “THE HISTORY OF HEALTH CARE COSTS AND HEALTH INSURANCE,” Wisconsin Policy Research Institute, Inc, Wisconsin , 2006.

[9] M. Thomasson, “Health Insurance in the United States,” Health US, Miami, 2010.

[10] Health24, “Medical schemes - a history,” 2016. [Online]. Available: http://www.health24.com/Medical-schemes/About-medical-schemes/Medical-schemes-a-history-20120721. [Accessed 12 June 2017].

[11] M. Billing, “Medical Billing Insurance Claims Process,” 2017. [Online]. Available: http://www.medicalbillingandcodingonline.com/medical-billing-claims-process/. [Accessed 28 March 2017].

[12] Z. Deng, W. Huang and L. Chuangxue, “A real-time medical assistance billing system,” *International Conference on Intelligent Computing and Integrated Systems,* pp. 757-760, 2010.

[13] K. Ngoepe, “Private health care costs have increased by 300% in 10 years,” 17 february 2016. [Online]. Available: http://www.news24.com/SouthAfrica/News/private-health-care-costs-have-increased-by-300-in-10-years-motsoaledi-20160217. [Accessed 18 August 2017].

[14] D. P. Ferenc, Understanding hospital billing and coding, 2nd ed., Chicago: St. Louis, Mo. : Elsevier Saunders, 2011.

[15] D. B. Evans, R. Elovainio and G. Humphreys, “The world health report financing for universal coverage,” World Health Organization, Geneva, 2010.

[16] H. P. Healthcare, “Payment Policies,” *Fraud, Waste and Abuse,* pp. 1-2, 03 15 2017.

[17] DEPARTMENT OF HEALTH AND HUMAN SERVICES Centers for Medicare & Medicaid Services, “Medicare Fraud & Abuse:Prevention, Detection, and Reporting,” Medicare Learning Network, Florida, 2016.

[18] M. Kirlidoga and C. Asuk, “A fraud detection approach with data mining in health insurance,” *SciVerse ScienceDirect,* vol. 62, pp. 989-994, 2012.

[19] S. Viaene and G. Dedene, “Insurance Fraud: Issues and Challenges,” *The Geneva Papers on Risk and Insurance,* vol. 29, no. 2, pp. 313-333, 2004.

[20] C. Piper, “10 popular health care provider fraud schemes,” February 2013. [Online]. Available: https://www.acfe.com/article.aspx?id=4294976280. [Accessed 7 August 2017].

[21] E. Ngai, Y. Hu, Y. Wong, Y. Chen and X. Sun, “The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature,” *Decision Support Systems,* vol. 50, pp. 559-569, 2011.

[22] Magellan Health Services,, “Fraud, Waste and Abuse Training for Medicare and Medicaid Providers,” Magellan Health Services Inc, Ohio, 2014.

[23] J. Gee and M. Bu, “THE FINANCIAL COST OF HEALTHCARE FRAUD 2014. What data from around the world shows,” Center for Counter Fraud Studies. University of Portsmouth, Portsmouth, 2014.

[24] L. E. Davis, “Growing health care fraud drastically affects all of us,” 22 October 2012. [Online]. Available: http://www.acfe.com/article.aspx?id=4294974475. [Accessed 16 August 2017].

[25] M. Cohen, “Human Impact of Healthcare Fraud,” 24 May 2016. [Online]. Available: https://www.medicfp.com/mfpnew2/human-impact-healthcare-fraud/. [Accessed 08 August 2017].

[26] Ponemon Institute, “Third Annual Survey on Medical Identity Theft,” Ponemon Institute, Chicago, 2014.

[27] M. B. &. Coding, “All About Medical Billing & Coding,” 20 September 2014. [Online]. Available: http://medicalbillingcodingworld.com/2014/09/technologys-impact-on-the-medical-billing-and-coding-field/. [Accessed 21 06 2017].

[28] A. Sheshasayee and S. S. Thomas, “Implementation of Data Mining Techniques in Upcoding Fraud Detection in the Monetary Domains,” in *International Conference on Innovative Mechanisms for Industry Applications*, India, 2017.

[29] M. Kirlidog and C. Asuk, "A fraud detection approach with data mining in health insurance," *Procedia - Social and Behavioral Sciences,* vol. 62, pp. 989-994, 2012.

[30] S. S and S. S.N, Introduction to data mining and its application, Heidelberg: Springer-Verlag Berlin , 2006.

[31] U. Fayyad, G. Piatetsky-Shapiro and a. P. Smyth, "From Data Mining to Knowledge Discovery in Databases," *American Association for Artificial Intelligence,* vol. 17, no. 3, pp. 37-54, 1996.

[32] R. Bhowmik, "Data Mining Techniques in Fraud Detection," *Journal of Digital Forensics, Security and Law,* vol. 3, no. 2, pp. 1-20, 2014.

[33] Ó. Marbán1, G. Mariscal and a. J. Segovia, Data Mining and Knowledge Discovery in Real Life Application, Rijeka: I-Tech Education and Publishing, 2009.

[34] R. Sint, S. Schaffert and S. Stroka, "Combining Unstructured, Fully Structured and Semi-Structured Information in Semantic Wikis," in *Conference: 4th Semantic Wiki Workshop (SemWiki 2009) at the 6th European Semantic Web Conference (ESWC 2009) Proceedings*, Hersonissos, 2009.

[35] J. S. Malik, P. Goyal and A. K. Sharma, "A Comprehensive Approach Towards Data Preprocessing Techniques & Association Rules," (BVICAM)/news/Indicompapers, Rajendra Nagar Indore, 2009.

[36] M. Pechenizkiy, S. Puuronen and a. A. Tsymbal, Feature Extraction for Classification in Knowledge Discovery Systems, Heidelberg: Springer, Berlin, 2003.

[37] Aparna.U.R and S. Paul, "Feature Selection and Extraction in Data mining," in *Online International Conference on Green Engineering and Technologies,* Kerala, 2016.

[38] F. P. Shah and V. Patel, "A Review on Feature Selection and Feature Extraction for Text Classification," in *International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET),* Tamilnadu, 2016.

[39] "Implementation of Data Mining Techniques in Upcoding Fraud Detection in the Monetary Domains," in *International Conference on Innovative Mechanisms for Industry Applications,* Bengaluru, 2017.

[40] B. Ramageri, "DATA MINING TECHNIQUES AND APPLICATIONS," *Indian Journal of Computer Science and Engineering,* vol. 1, no. 4, pp. 301-305, 2014.

[41] S. C. Pandey, "Data Mining Techniques for Medical Data: A Review," in *International conference on Signal Processing, Communication, Power and Embedded System*, Mumbai, 2016.

[42] J. Han, M. Kamber and J. Pei, Data mining Concepts and techniques, Waltham: Morgan Kaufmann Publishers, 2012.

[43] M. Gera and S. Goel, "Data Mining - Techniques, Methods and Algorithms: A Review on Tools and their Validity," *International Journal of Computer Applications,* vol. 113, no. 18, pp. 975-8837, 2015.

[44] "DATA MINING TECHNIQUES AND APPLICATIONS," *Indian Journal of Computer Science and Engineering,* vol. 1, no. 4, pp. 301-305.

[45] T. R and S. Kalaiselvi, "An Overview of Data Mining Techniques and Applications," *International Journal of Science and Research,* vol. 2, no. 2, pp. 506-509, 2013.

[46] J. Deogun and L. Jiang, "Prediction Mining – An Approach to Mining Association Rules for Prediction Rules for Prediction," *International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing,* vol. 3642, pp. 98-108, 2005.

[47] HelgeLangseth and LuigiPortinale, "Bayesian networks in reliability," *Reliability Engineering & System Safety,* vol. 92, no. 1, pp. 92-108, 2007.

[48] S. Sperandei, "Understanding logistic regression analysis," *Biochemia Medica,* vol. 24, no. 1, pp. 12-18, 2014.

[49] R. A. Bauder and T. M. Khoshgoftaar, “Medicare Fraud Detection using Machine Learning Methods,” in *16th IEEE International Conference on Machine Learning and Applications*, Cancun, 2017.

[50] A. K. Mishra and B. K. Ratha, “Study of Random Tree and Random Forest Data Mining Algorithms for Microarray Data Analysis,” *International Journal on Advanced Electrical and Computer Engineering,* vol. 3, no. 4, 2016.

[51] N. d. Freitas, D. Matheson and M. Denil, “Narrowing the Gap: Random Forests In Theory and In Practice,” *International Conference on Machine Learning,* vol. 32, no. 1, 2014.

[52] G. Biau, “Analysis of a Random Forests Model,” *Journal of Machine Learning Research,* vol. 13, pp. 1063-1095, 2012.

[53] Y. Freund and R. E. Schapire, “A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting,” *Journal of computer and system sciences,* vol. 55, pp. 119-139, 1997.

[54] J. H. Friedman, “Greedy Function Approximation: A Gradient Boosting Machine,” *The Annals of Statistics,,* vol. 25, no. 5, p. 2001, 1189-1232.

[55] A. L. Buczak and E. Guven, “A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection,” *IEEE COMMUNICATIONS SURVEYS & TUTORIALS,* vol. 18, no. 2, pp. 1153-1174, 2016.

[56] N. G. Paterakis, E. Mocanu, M. Gibescu, B. Stappers and W. v. Alst, “Deep Learning Versus Traditional Machine Learning Methods for Aggregated Energy Demand Prediction,” in *IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe),* Torino, 2017 .

[57] B. M. Ramageri, “DATA MINING TECHNIQUES AND APPLICATIONS,” *Indian Journal of Computer Science and Engineering,* vol. 1, no. 4, pp. 301-305.

[58] N. Joshi, A. Kumar, P. Chakraborty and R. Kala, "Speech Controlled Robotics using Artificial Neural Network," *International Conference on Image Information Processing,* vol. 15, 2015.

[59] V. Rawte and G. Anuradha, "Fraud Detection in Health Insurance using Data Mining Techniques," *International Conference on Communication, Information & Computing Technology (ICCICT),,* vol. 1, no. 1, pp. 16-17, 2015.

[60] Y. Liu, "A hybrid neural network learning system," *The Fourth International Conference on Computer and Information Technology,* pp. 1016-1021, 2004.

[61] W.-S. Yang and S.-Y. Hwang, "A process-mining framework for the detection of healthcare fraud and abuse," *Expert Systems with Applications,* vol. 31, pp. 56-68, 2006.

[62] F.-M. Liou, Y.-C. Tang and J.-Y. Chen, "Detecting hospital fraud and claim abuse through diabetic outpatient services," *Health Care Manage Science,* vol. 11, pp. 353-358, 2008.

[63] H. Shin, H. Park, J. L. a and W. C. Jhee, "A scoring model to detect abusive billing patterns in health insurance claims," *Expert Systems with Applications,* vol. 39, pp. 7441-7450, 2012.

[64] L. K. Branting, F. Reeder, J. Gold and T. Champney, "Graph Analytics for Healthcare Fraud Risk Estimation," *International Conference on Advances in Social Networks Analysis and Mining (ASONAM),* vol. 7, no. 16, 2016.

[65] A. Bayerstadler, L. v. Dijk and F. Winter, "Bayesian multinomial latent variable modeling for fraud and abuse detection in health insurance," *Insurance: Mathematics and Economics,* vol. 71, pp. 244-252, 2016.

[66] K. Yamanishi, J.-i. Takeuchi and G. Williams, "On-Line Unsupervised Outlier Detection Using Finite Mixtures with Discounting Learning Algorithms," *Data Mining and Knowledge Discovery,* vol. 8, no. 3, pp. 275-300, 2004.

[67] F. Lu, J. E. Boritz and D. Covvey, "Adaptive Fraud Detection Using Benford's Law," in *Advances in Artificial Intelligence: 19th Conference of the Canadian Society for Computational Studies of Intelligence, Canadian AI 2006,* Quebec, 2006.

[68] Y. Shan, D. Jeacocke, D. W. Murray and A. Sutinen, "Mining medical specialist billing patterns for health service management," in *Proc. of the 8th Australasian Data Mining Conference*, Tuggeranong, 2008.

[69] C. Lina, C.-M. Lin, S.-T. Li and S.-C. Kuo, "Intelligent physician segmentation and management based on KDD approach," *Expert Systems with Applications,* vol. 34, no. 3, pp. 1963-1973, 2008.

[70] "Discovering inappropriate billings with local density based outlier detection method," in *Proc. of the 8th Australasian Data Mining Conference (AusDM'09),* Tuggeranong, 2009.

[71] R. A. Bauder and T. M. Khoshgoftaar, "Multivariate outlier detection in medicare claims payments applying probabilistic programming methods," *Health Services and Outcomes Research Methodology,* vol. 17, no. 3-4, pp. 256-289, 2017.

[72] R. A. Bauder and T. M. Khoshgoftaar, "A Novel Method for Fraudulent Medicare Claims Detection from Expected Payment Deviations," *IEEE 17th International Conference on Information Reuse and Integration,* vol. V, no. 16, 2016.

[73] D. R. Riedinger and J. A. Major, "EFD: A HYBRID KNOWLEDGE/STATISTICAL-BASED SYSTEM," *The Journal of Risk and Insurance,* vol. 69, no. 3, pp. 309-324, 2002.

[74] V. Rawte and A. Srinivas, "Fraud Detection in Health Insurance using Data Mining Techniques," *International Conference on Communication, Information & Computing Technology,* pp. 1-5, 2015.

[75] C. Kothari, Research Methodology: Methods and Techniques, New Delhi: NEW AGE INTERNATIONAL (P) LIMITED, PUBLISHERS, 2004.

[76] S. Langkos, "RESEARCH METHODOLOGY: Data collection method and Research tools," Annals of Emerging Technologies in Computing (AETiC), Athens, 2014.

[77] N. Walliman, Research Methods: The Basics, New York: Routledge, 2011.

[78] C. Pilkington and L. Pretorius, "A Conceptual Model of the Research Methodology Domain With a Focus on Computing Fields of Study," *7th International Conference on Knowledge Engineering and Ontology Development,* vol. 2, pp. 96-107, 2015.

[79] S. K. Chang, Handbook of software engineering & knowledge engineering, Singapore: World Scientific, 2001.

[80] K. J. Cios, R. W. Swiniarski, W. Pedrycz and L. A. Kurgan, Data Mining, A knowledge Discovery Approach, SpringerLink, 2007.

[81] D. Pyle, Data Preparation for Data Mining, San Franccisco: Morgan Kaufmann Publishers, Inc., 1999.

[82] S. ZHANG, C. ZHANG and Q. YANG, "DATA PREPARATION FOR DATA MINING," *Applied Artificial Intelligence,* vol. 17, pp. 375-381, 2003.

[83] K. C. Davis, K. Janakiraman, A. Minai and R. B. Davis, "Data Integration Using Data Mining Techniques," in *Fifteenth International Florida Artificial Intelligence Research Society* , Florida, 2002.

[84] V. Palade, R. J. Howlett and L. Jain, "Feature Extraction for Classification in Knowledge Discovery Systems," in *7th International Conference, KES* , Oxford, 2003.

[85] F. P. Shah and V. Patel, "A Review on Feature Selection and Feature Extraction for Text Classification," in *IEEE WiSPNET,* Chennai, 2016.

[86] Aparna.U.R and S. Paul, "Feature Selection and Extraction in Data mining," in *Online International Conference on Green Engineering and Technologies,* Coimbatore, 2016.

[87] R.Kavitha and E.Kannan, "An Efficient Framework for Heart Disease Classification using Feature Extraction and Feature Selection Technique in Data Mining," in *International Conference on Emerging Trends in Engineering, Technology and Science*, Pudukkottai, 2016 .

[88] N. Jiang and H. Liu, "Understand System's Relative Effectiveness Using Adapted Confusion Matrix," in *International Conference of Design, User Experience, and Usability*, Berlin, 2013.

[89] N. L. A. Ghani, S. Z. Z. Abidin and N. E. A. Khalid, "Accuracy Assessment of Urban Growth Pattern Classification Methods Using Confusion Matrix and ROC Analysis," in *International Conference on Soft Computing in Data Science*, Putrajaya, 2015.

[90] P. Domingos, "A Few Useful Things to Know about Machine Learning," *Communications of the ACM,* vol. 55, no. 10, pp. 78-87, 2012.

[91] R. A. Bauder and T. M. Khoshgoftaar, "Medicare Fraud Detection using Machine Learning Methods," in *16th IEEE International Conference on Machine Learning and Applications*, California, 2017.

[92] P. Ramachandran, B. Zoph and Q. V., "Searching for Activation Functions," Cornell University, New York, 2017.

[93] E. D.-G. J. Owusu-Ansah, A. Sampson, S. K. Amponsah and R. C. Abaidoo, "Sensitivity and Specificity Analysis Relation to Statistical Hypothesis Testing and Its Errors: Application to Cryptosporidium Detection Techniques," *Open Journal of Applied Sciences* , vol. 6, no. 4, pp. 209-216, 2016.

[94] HDFC Life, "Health Insurance - Definition & Meaning," 2015. [Online]. Available: https://www.hdfclife.com/insurance-knowledge-centre/about-life-insurance/what-is-health-insurance. [Accessed 03 June 2017].

[95] "Application of Bayesian Methods in Detection of Healthcare Fraud," *The Italian Association of Chemical Engineering,* vol. 33, pp. 151-156, 2013.

[96] NHIS, "TOWARDS UNIVERSAL HEALTH COVERAGE," NATIONAL HEALTH INSURANCE FOR SOUTH AFRICA, Cape Town, 2015.

[97] R. Bhowmik, "Data Mining Techniques in Fraud Detection," *Digital Forensics, Security and Law,* vol. 3, no. 2, 2014.

[98] P. L. Brockett, X. Xia and R. A. Derrig, "Using Kohonen's Self-Organizing Feature Map to Uncover Automobile Bodily Injury Claims Fraud," *The Journal of Risk and Insurance,* vol. 65, no. 1998, pp. 245-274, 1998.

[99] E. Cox, "A Fuzzy System for Detecting Anomalous Behaviors in Healthcare Provider Claims," in *Intelligent Systems for Finance and Business,* Goonatilake, John Wiley and Sons Ltd, 1995, pp. 111-134.

[100] H. He, G. Warwick and Y. Xin, "application of Genetic Algorithms and k-Nearest Neighbour method in real world medical fraud detection problem," *Springer,* pp. 74-81, 1999.

[101] Health24, "Fraud in SA healthcare system," 2013. [Online]. Available: http://www.health24.com/Medical-schemes/General-info/Fraud-in-SA-healthcare-system-20130319. [Accessed 15 April 2017].

[102] IBM, "Using Data Mining to Detect Insurance Fraud," New York, 2011.

[103] J. Leskovec, A. Rajaraman and J. D. Ullman, Mining Of Massive Datasets, 2nd ed., California: Stanford Publisher, 2010.

[104] Q. Liu and M. Vasarhelyi, "Healthcare fraud detection: A survey and a clustering model incorporating Geo-location information," *WORLD CONTINUOUS AUDITING AND REPORTING SYMPOSIUM,* vol. 29, no. 1, pp. 1-10, 2013.

[105] M. Marketing, "Life insurance fraud and dishonesty hits record high," 2014. [Online]. Available: https://www.moneymarketing.co.za/life-insurance-fraud-and-dishonesty-hits-record-high/. [Accessed 16 April 2017].

[106] K. Melih and A. Cuneyt , "A fraud detection approach with data mining in health insurance," *Social and Behavioral Sciences,* p. 989 – 994, 2012.

[107] V. Mutyambizi, "Health Insurance in South Africa," UCT Press, Cape Town, 2015.

[108] M. B. a. C. Online, "Medical Billing Insurance Claims Process," 2017. [Online]. Available: http://www.medicalbillingandcodingonline.com/medical-billing-claims-process/. [Accessed 1 April 2017].

[109] P. A. Ortega, C. J. Figueroa and G. A. Ruz, "A Medical Claim Fraud/Abuse Detection System based on Data Mining: A Case Study in Chile," *Conference on Data Mining,* pp. 224-230, 2014.

[110] C. Phua, V. Lee, K. Smith and R. Gayler, "A Comprehensive Survey of Data Mining-based Fraud Detection Research," Baycorp Advantage, Melbourne, 2013.

[111] N. K. Sekhri, "Managed care: the US experience," *Bulletin of the World Health Organization,* vol. 78, no. 6, pp. 1-15, 2000.

[112] D. Thornton, G. v. Capelleveen, M. Poel, J. v. Hillegersberg and a. R. M. Mueller, "Outlier-based Health Insurance Fraud Detection for U.S. Medicaid Data," San Diego, 2014.

[113] K. Yamanishi, J.-i. Takeuchi and G. Williams, "Algorithms, Online Unsupervised Outlier Detection Using Finite Mixtures with Discounting Learning," *Discovery,* vol. 8, p. 275, 2004.

[114] M. Zareapoor and S. K. R, "Feature Extraction or Feature Selection for Text Classification: A Case Study on Phishing Email Detection," *I.J. Information Engineering and Electronic Business,* vol. 2, pp. 60-65, 2015.

[115] H. Joudaki, A. Rashidian, B. Minaei-Bidgoli, M. Mahmoodi, B. Geraili, Mahdi and M. Arab, "Using Data Mining to Detect Health Care Fraud and Abuse: A Review of Literature," *Global Journal of Health Science,* vol. 7, no. 1, 2015.

[116] Š. Furlan and M. Bajec, "Holistic Approach to Fraud Management in Health Insurance," *JIOS,* vol. 32, no. 2, 2008.

[117] S. p. A.Vinothini, "Survey of Machine Learning Methods for Big Data Applications," in *International Conference on Computational Intelligence in Data Science(ICCIDS)*, Chennai, 2017.

[118] D. Wang, X. Liu and M. Wang, "A DT-SVM Strategy for Stock Futures Prediction with Big Data," in *16th IEEE International Conference on Computational Science and Engineering*, Sydney, 2013.

A comparison of machine learning methods applicable to healthcare claims fraud detection

Some name and Some name

Some University

Abstract. The healthcare industry has become a very important pillar in the modern society but has witnessed an increase in fraudulent activities. Traditional fraud detection methods have been used to detect potential fraud, but for certain cases they have been insufficient and time consuming. Data mining which has emerged as a very important process in knowledge discovery has been successfully applied in the health insurance claims fraud detection. We implemented a prototype that comprised different methods and a comparison of each of the methods was carried out to determine which method is most suited for the Medicare dataset. We found that while ensemble methods and neural net performed, the logistic regression and the naive bayes model did not perform well as depeicted in the result.

Keywords: Healthcare, Fraud Detection, Neural nets, Gradient boosted tree classifier, Medicare

1 Introduction

The health insurance industry, a pillar in the modern-day society, that serves the purpose of providing affordable healthcare to individuals. Healthcare has become a necessity for households, hence the cost of healthcare forms a part of household expenditure. The increased cost of healthcare has made it a luxury rather than a basic need [1]. One of the reasons for the increased cost of health insurance can be attributed to the money lost through fraud in the healthcare system [2].

Fraud has impacted several aspects of life and the healthcare industry is no exception. Fraud occurs in the medical billing process leading to loss of funds by the insurance company which leads to the insurance company charging higher premiums to make up for the lost funds. The impact of fraud in healthcare has other implications aside from the monetary implications such as health risks that can arise from the altering of a patient's record by the physician [3].

Traditional fraud detection methods such as rule based statistical methods have been applied to detect possible fraud in the healthcare claims process, but these methods no longer suffice due to the large number of claims to be processed and the variety of patterns these fraudulent activities take. Machine learning methods

have been applied to other fraud detection problems such as credit card fraud detection and has also been applied to fraud detection in healthcare claims [4].

In the research study, we unpack the problem of healthcare claims fraud detection as well as the impacts it has . We then analyse how machine learning has been applied to the healthcare fraud claims detection problem by discussing the similar systems. With an understanding of the current research being done in the area, we create a data mining model that takes an exploratory approach to solving the problem of healthcare claims fraud detection by comparing and analysing different methods. We implement these methods in a prototype and then analyse the results to see which methods performed best with the Medicare dataset.

2 Problem Background

Abraham Maslow states the physiological needs of any individual are the most basic innate human needs that need to be satisfied and has the highest priority [5]. Maintaining a healthy body condition, eating, basic security is tantamount to satisfying the safety needs of an individual. To have good health, one needs access to adequate and affordable healthcare. Unfortunately, the cost of healthcare has been on the rise, making healthcare more of a luxury than a basic need [1]. One of the factors that have contributed to the increased cost of healthcare is the impact of the funds lost to fraudsters through healthcare claims fraud.

Before going deeper into the depth of health insurance, a definition of health insurance is needed. Health insurance represents a contract that a person pays an agreed premium to an insurance provider for a designated healthcare cover. The health insurance industry involves the transfer of funds and has been affected by fraudulent activities perpetrated by individuals that seek to gain illegal access to these funds.

Health insurance waste in healthcare is most times unrelated to fraud as it mainly the provision of unnecessary health services. Health insurance waste can only be seen as fraud and abuse when the act is intentional. Waste can occur when services are over utilized and then results in unnecessary expenditure [6].

Health insurance abuse is the billings of practices that either directly or indirectly is not consistent with the goals of providing patients with services that are medically necessary and these practices meet professionally recognized standards as well as being fairly priced [6].

Health insurance fraud is purposely billing for services that were never performed and or supplies not provided, medically unnecessary services and altering claims to receive higher reimbursement than the service produced [6].

To tackle the problem of fraud in the medical billing process, health insurance companies make use of traditional rule-base models, but these models do not suffice anymore due to several factors such as the large volume of claims to be processed which makes the medical billing process prone to error, slow and sometimes inefficient. Machine learning methods can be used to improve the detection of possible healthcare claims fraud. The next section discusses the related works on the applications of machine learning methods in the detection of possible fraudulent healthcare claims.

3 Review of related work

Machine learning methods have been effective in automatically extracting patterns from data to derive knowledge which yields meaningful results such as detecting which submitted claims are likely fraudulent. The first work we consider is the Outlier-based health insurance fraud detection for US Medicaid data presented by Thornton et al. [7]. They made use of Medicaid data for dental services which is a healthcare provider in the US that caters to low income people. Their model made use of 3 different univariate machine learning methods which are the linear regression, time series plot as well as box plot. They also used a multivariate method through clustering to detect possible health insurance fraud. The dataset used contained a case study for 500 dentists and they successfully identified 17 activities that can be deemed fraudulent among the 360 records analysed.

We also reviewed "Graph Analytics for Healthcare Fraud Risk Estimation" by Branting et al. which made use of a graph to link providers, drug prescriptions and the procedures [8]. They used two algorithms where the first algorithm was used to calculate the similarity to predetermined fraudulent and non-fraudulent providers while the second algorithm calculates the estimated fraud risk through location of practitioners. They achieved an F-score of 0.919 and an impressive AUC of 0.960.

Bauder et al. also carried out several works in the area of detecting fraud in the health insurance process. One of the systems, a multivariate outlier detection in Medicare claims payments applying probabilistic programming methods [9]. They created a base for what the expected Medicare payments should look like for each type of provider. Outliers were then identified by comparing payment amounts with the normative case and the deviations are categorised as outliers.

4 Experimental Setup

To establish a way to conduct the research, we define a research methodology that form the guide to solving the problem at hand. The quantitative research approach was chosen as it allows for the statistical analysis of the data and maintains an objective standpoint.

The data that was used for the study is the Medicare payments data between 2012-2015. The dataset contained payments and utilization healthcare claims data as well as the details about the procedures rendered to individuals. The data from the List of Excluded Individual or Entities (LEIE) was used to create the ground truth labels.

5 Model

Applying a machine learning algorithm to derive knowledge is just a piece in the puzzle of creating an effective data mining model. The data mining model consists of different processes and each of these processes play a major role in deriving knowledge from the data. In this section, we unpack these individual processes as well as the methods that were used for each individual process.

5.1 The data collection phase

Data is the raw material to be processed in a data mining system. Data can be structured, unstructured or semi-structured. The Medicare dataset used, was in a structured format. Both datasets were loaded and stored in a database for further analysis.

5.2 The data pre-processing and transformation phase

Normally, data in the real world is dirty, contains missing values and can be incorrect. Therefore, a lot of work to be done cleaning up the data to get it to the form that will be suitable for use. The Medicare data used was already pre-processed by the Centre for Medical Services (CMS). The work first task we performed was filtering the Medicare dataset for only non-prescription data. The Medicare dataset did not contain any label to be used to differentiate between fraudulent and non-fraudulent claims, therefore we used the LEIE dataset to flag the claims that were detected as fraudulent. We made use of fuzzy matching on the practitioners first name, last name and ZIP code to link the Medicare payment data to a practitioner in the LEIE dataset, as there was no explicit join between the two datasets.

After labelling the dataset and identifying ground truth labels, the next task performed was indexing categorical string features. We based the initial selection of features on the work done by [10] using the same Medicare dataset as shown in table 1. We also used feedback gotten from the model as calculated by the feature importance to adjust the features used for the model. Finally, we applied a 70:30 split to the dataset. 70 % of the data was used for training the machine learning model while the remaining 30 % was used for testing the outcome of the model.

Table 1. Description of Medicare Features.

Feature	Description

NPI	Unique provider identification number
last name	Providers last name
First name	Providers first name
Zip	Providers 5-digit zip code
provider type	Medical providers specialty (or practice)
line srvc cnt	Number of procedures performed per provider
bene unique cnt	Number of distinct beneficiaries per day services
average submitted chrg amt	Average of the charges that the provider submitted
average medicare payment amt	Amount paid to the provider for services performed

5.3 Machine learning application

Now that we have completed the pre-processing and transformation of the data, we apply machine learning algorithm to derive insights from the data. We used an exploratory approach in creating the machine learning model as several methods were used and results were collected on the performance of the different methods used.

1. The Bayesian classifier is a simple classifier based on statistical methods andit assigns probabilities to each member of the class in such a way that a given sample can be categorised into one class. It makes a strong assumption on the independence of features.
2. Random forest classifier, an ensemble learning method made up of a sequential construct of several decision trees in the training phase and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.
3. Logistic regression is another machine learning method based on statisticalprinciples. It is highly suited for predicting categorical features. In predicting a binary outcome, the logistic regression model makes use of the binomial logistic regression and for multiple outcomes it makes use of multinomial logistic regression.
4. Gradient Boosted tree classifier is yet another powerful classification method.The classification method uses ensembles of decision trees and applies the technique known as boosting to improve performance. The idea of boosting emanates from the attempt to combine weaker learners to become better learners. The gradient boosted tree classifier is a combination of a loss function, a weak learner and the additive function responsible for combining the weak learners and reducing the loss function.
5. The artificial neural network is a machine learning method based off thefunctioning of the neurons in the brain of a biological system. It is made up of a network of interconnected nodes. The nodes do not contain any computation but rather, they function as a group of linear functions. The nodes in the neural network are grouped into layers. The behaviour of each node is defined by an activation function.

6 Results

Once the data has been passed through to the machine learning process, we evaluate how well the model performed by using the following pre-determined benchmarks. We used the following metrics to evaluate the different machine learning models created: Weighted Precision, Weighted Recall, AUC, Test Error, Sensitivity, Specificity, F1. In this section, we unpack how each of the implementations performed against the defined metrics.

Table 2. Result for the performance metric of the different machine learning methods.

ML Model	Weighted Precision	Weighted Recall	AUC	Test Error	Sensitivity	Specificity	F1
Naive Bayes	82.8	90.8	54.0	9.1	0	99.7	86.6
GBT Classifier	**92.8**	**93.5**	**93.0**	6.4	**34.9**	99.2	**92.3**
Random Forest Classifier	94.9	95.0	84	8.0	47.3	98.8	86.9
Logistic Regression	82.9	91.0	63.5	9.0	0.0	**100**	86.8
Neural Net	89	91	91.5	**10**	34.8	99.2	91.9

6.1 Naive Bayes

We implemented the Bayesian model using the Apache SparkML library using multinomial distribution of features and a smoothing of 1.0. The model performed poorly with a ROC curve of 54.0 as seen in the figure below, which entails that the predictions of the model was just as good as random guesses. The Naive Bayes model scored well in the other metrics and also recorded a low-test error of 9.1%.

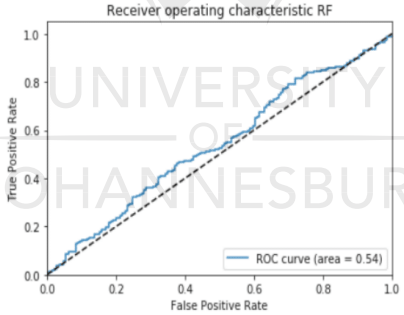


Fig.1. ROC curve for the Naive Bayes classification model

.

6.2 Logistic regressions

The logistic regression model was a slight improvement to the Naive Bayes model. The logistic regression model was created with a regularization parameter of 0.3 and the elastic net parameter was set at 0.8 based on the recommendation of the SparkML documentation. The logistic regression model was run over several iterations and achieved the best result at 10 iterations. An improved AUC of 63.5% was achieved as seen in figure 2, which is better than the Bayesian model but not good enough for predictions. The model achieved an accuracy of 91.0%.

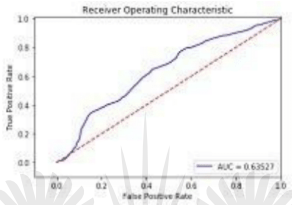


Fig.2. ROC curve for the logistic regression classification model

.

6.3 Random Forest Classifier

The random forest offered a great improvement to the logistic regression with an AUC of 84.0%. The random forest classifier had a specificity score of 99.7% and an accuracy of 95.0%. The AUC score of 84.0% makes the random forest classifier suitable for the Medicare dataset as the score implies a great improvement to random guess noticed in the previous models.

6.4 Gradient Boosted Tree Classifier

The gradient boosted tree classifier model presented a great improvement to the previous models. The model was run through 10 iterations. The gradient boosted tree classifier had a weighed precision and recall of 93% each. It had an accuracy of 93.3% and an F1 score of 92.3 %. It recorded a sensitivity score of 47.3% and a specificity of 98%. The most important improvement was the noticeable increase in the AUC score of 97.0%. Based on the results the gradient boosted tree classifier will be ideal for detecting possible fraud in healthcare claim using the Medicare data.

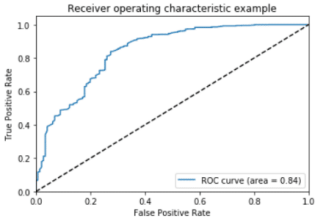


Fig.3. ROC curve for the random forest classification model

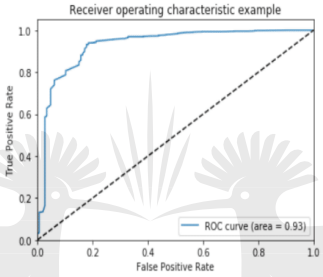


Fig.4. ROC curve for the Gradient Boosted Tree classification model

6.5 Artificial Neural Net

The artificial neural network slightly underperformed the gradient boosted tree classifier as can be seen in the metrics in table 2. The model made use of the binary cross-entropy as the loss function. The model had a weighted precision and a weighted recall of 89% and 90% respectively. It also had an accuracy of 90% with specificity and sensitivity measuring 99.2% and 3408% respectively. The F1 score was 92% and it also had an AUC of 91.5%.

7 Discussion

We have shown how the proposed model functions as well the results that were derived from the model. There are several limitations associated with the model

which can act as a hindrance in the implementation of the model. The following limitations were evident through the implementation of the model.

7.1 Critique

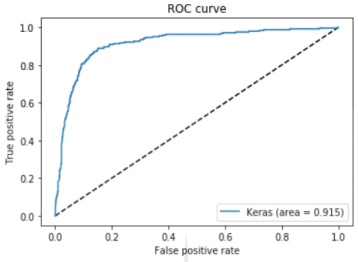


Fig.5. ROC curve for the Artificial Neural Network classification model

Lack of labelled data There was difficulty in obtaining openly available data with ground truth label. We made use of data from the LEIE database for ground truth label. The availability and use of a dataset that has been labelled initially would be a better validation of the model.

Sample dataset The data that was used in the model was only concerned with the healthcare provider. Although fraud mainly occurs through the healthcare provider, it doesnt imply that fraud cannot occur through the other entities in the health insurance ecosystem. The dataset we used constrain the research to only consider fraud from the medical practitioner.

7.2 Support

Applications of alternative machine learning methods The choice of the machine learning methods to be used in a data mining system is important. The approach taken in the study is an exploratory approach, applying different types of machine learning methods and the assessing them against the benchmark. The use of multiple machine learning methods allowed for a better analysis of how each method performs with the Medicare dataset as well as the different advantages and disadvantages of each model.

Flexibility of solution The model is structured in a modular manner which allows for a flexible implementation of a healthcare claims fraud detection system. The modular implementation of the system means that the different components of the system can be easily replaced thereby reducing the burdens of future system upgrades.

8 Conclusion

The task of building a system that enables the identification of possible fraudulent healthcare claims has become very important as the demand for healthcare increases. The medical billing process, due to its complexities and the volume of claims to be processed has been exploited by fraudsters looking to gain illegal advantage from the system. Machine learning methods have enabled the identification of these fraudulent claims and have improved the effectiveness of the medical billing process.

The study explored the application of different machine learning methods to detect fraudulent claims in the medical billing process. The result generated from the application of these machine learning methods were collected. The analysis of the result showed that the ensemble methods and the artificial neural network performed best with the Medicare dataset.

Through the insights gained from the study, we were able to identify the strengths and weaknesses of the model. The Medicare dataset limited the system to only identify fraud that occurs through the medical practitioner as the Medicare payment data only contained data regarding the medical practitioner. Notwithstanding the limitations, the success of the model in identifying the fraudulent healthcare claims as well as the explorative approach taken to determine the which machine learning method makes this a viable solution.

References

1. Yu, H.: Impacts of rising health care costs on families with employment-basedprivate insurance: A national analysis with state fixed effects. Health Services Research (2012)
2. Singh, A.: Fraud in insurance on rise. Technical report, Ernst & Young (2011)
3. Davis, L.E.: Growing health care fraud drastically affects all of us (October 2017)
4. Rabiul, J., Nabeel, M., Ahsan, H., Sifat, M.: An evaluation of data processing solutions considering preprocessing and special features. 11th International Conference on Signal-Image Technology & Internet-Based Systems (2015)
5. McLeod, S.: Maslow's hierarchy of needs. Simply Psychology **1** (2007)
6. Coustasse, J.B.S.T.V.: Medicare fraud, waste and abuse. In: Business and HealthAdministration Association Annual Conference. (2017)

7. Dallas, T., Guido, v.C., Mannes, P., van Hillegersberg, J., Roland, M.M.: Outlierbased health insurance fraud detection for u.s. medicaid data. 16th International Conference on Enterprise Information Systems (2014)
8. Branting, L.K., Reeder, F., Gold, J., Champney, T.: Graph analytics for healthcare fraud risk estimation. Advances in Social Networks Analysis and Mining (ASONAM) (2016)
9. Bauder, R.A., Khoshgoftaar, T.M.: A probabilistic programming approach for outlier detection in healthcare claims. 15th IEEE International Conference on Machine Learning and Applications (ICMLA) (2016)
10. Bauder, R.A., Khoshgoftaar, T.M.: Medicare fraud detection using machine learning methods. 16th IEEE International Conference on Machine Learning and Applications (2017)

References

1. McLeod, S.: Maslow's hierarchy of needs. Simply Psychology **1** (2007)
2. Jinhee Kim, B.B., Williams, A.D.: Understanding health insurance literacy:a literature review. Family and Consumer Sciences Research **42**(1) (2013) 3–13
3. for Medicare & Medicaid Services, C., et al.: Medicare fraud & abuse: Prevention, detection, and reporting (2015)
4. Ferenc, D.P.: 4. In: Understanding Hospital Billing and Coding, 3rd Edition. Elsevier (2014) 88–95
5. Coustasse, J.B.S.T.V.: Medicare fraud, waste and abuse. In: Business and Health Administration Association Annual Conference. (2017)
6. Sadoughi, M.F.A.S.F.: Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision (icd-10). ELSEVIER **30** (2010) 78–84
7. Kirlidog, M., Asuk, C.: A fraud detection approach with data mining in health insurance. Procedia-Social and Behavioral Sciences **62** (2012) 989–994
8. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery in databases. AI magazine **17**(3) (1996) 37
9. Bhowmik, R.: Data mining techniques in fraud detection. The Journal of Digital Forensics, Security and Law: JDFSL **3**(2) (2008) 35
10. Karahoca, J.P.: 1. In: Data Mining and Knowledge Discovery in Real Life Applications. In-Teh (20009) 15–20
11. Rawte, V., Anuradha, G.: Fraud detection in health insurance using data mining techniques. In: Communication, Information & Computing Technology (ICCICT), 2015 International Conference on, IEEE (2015) 1–5
12. Yang, W.S., Hwang, S.Y.: A process-mining framework for the detection of healthcare fraud and abuse. Expert Systems with Applications **31**(1) (2006) 56–68
13. Liou, F.M., Tang, Y.C., Chen, J.Y.: Detecting hospital fraud and claim abuse through diabetic outpatient services. Health care management science **11**(4) (2008) 353–358
14. Shin, H., Park, H., Lee, J., Jhee, W.C.: A scoring model to detect abusive billing patterns in health insurance claims. Expert Systems with Applications **39**(8) (2012) 7441–7450
15. Lu, F., Boritz, J.E., Covvey, D.: Adaptive fraud detection using benfords law. In: Conference of the Canadian Society for Computational Studies of Intelligence, Springer (2006) 347–358
16. Lin, C., Lin, C.M., Li, S.T., Kuo, S.C.: Intelligent physician segmentation and management based on kdd approach. Expert Systems with Applications **34**(3) (2008) 1963–1973
17. Sint, R., Stroka, S., Schaffert, S., Ferstl, R.: Combining unstructured, fully structured and semi-structured information in semantic wikis. SemWiki **464** (2009)
18. Han, J., Pei, J., Kamber, M.: Data mining: concepts and techniques. Elsevier (2011)
19. Shan, Y., Jeacocke, D., Murray, D.W., Sutinen, A.: Mining medical specialist billing patterns for health service management. In: Proceedings of the 7th Australasian Data Mining Conference-Volume 87, Australian Computer Society, Inc. (2008) 105–110

A Critical Analysis of the Application of Data Mining Methods to Detect Healthcare Claim Fraud in the Medical Billing Process

Obodoekwe Nnaemeka and Dustin van der Haar

University of Johannesburg

Abstract. The healthcare industry has become a very important pillar in the modern society but has witnessed an increase in fraudulent activities. Traditional fraud detection methods have been used to detect potential fraud, but for certain cases they have been insufficient and time consuming. Data mining which has emerged as a very important process in knowledge discovery has been successfully applied in the health insurance claims fraud detection. We performed an analysis of studies that used data mining techniques for detecting healthcare fraud and abuse using the supervised and unsupervised data mining methods. Each of these methods have their own strengths and weaknesses. This article attempts to highlight these areas, along with trends and propose recommendations relevant for deployment. We identified the need for the use of more computationally efficient models that can easily adapt and identify the novel fraud patterns generated by the perpetrators of healthcare claims fraud.

Keywords: Healthcare, Fraud Detection, Assessment, Supervised learning, Unsupervised leanring

1 Introduction

According to Maslows hierarchy of needs, the most basic need of every individual is the physiological need. The physiological needs of individuals consist of basic needs such as food, water and good health [1]. For the full functioning of any individual, the individual first needs to be in good health. To be in good health one needs to have access to the adequate medical treatment when needed. The continuous rise in the cost of the medical care, makes it more of a luxury than a basic need in recent times. Health insurance was introduced to help manage this cost and make healthcare more affordable for everyone. Health insurance is a contract between a group of individuals or person with an insurer stating that an individual pays an agreed premium for a specified health insurance cover [2]. Health insurance relates to an insurance type that covers medical and surgical expenses of a member.

The health insurance system has been impacted by fraudulent activities with

several parties involved trying to gain illegal benefits [3]. The impact of fraud in the health insurance system results in loss of revenue, excessive time spent on reviewing these claims by the insurance service provider thereby leading to a delayed feedback on reimbursements. We start by analyzing the medical billing process in section 2 to gain a better understanding of problem background. In section 3, we see how data mining can be applied to healthcare claims fraud and current work. Section 4 defines an assessment framework which we apply to the current work discussed in section 3.2 to derive the results in section 5. Based on the results, we make recommendations for further improvements in section 6 and conclude in section 7.

2 The medical billing process

In this section, we analyze the healthcare system to understand how it functions and the different role players involved. An understanding of the healthcare process forms a base for the further exploration of how fraud can occur in the healthcare claims process.

Fig. 1. A summary of the medical billing process flow (as visualized by the author)

The purpose of the health insurance claim process is for service providers to receive reimbursements for the services they provided. The reimbursements

can be paid to patients or the insurance providers. Figure 2 shows the flow of activities in the health insurance claim process.

There are several steps involved in processing a health insurance claim. A brief review of the steps involved in the medical insurance claim process involves a patient receiving care from the licensed practitioner. The practitioner records the services provided to the patient and the relevant International Classification of Diseases (ICD) codes if diagnoses were made or Current Procedural Terminology (CPT) codes if the patient was treated [4]. The patients data along with the patients insurance information are also captured and added to the bill for claim processing.

Now the claim has been sent through, the next step is processing the claim. Processing these claims involves taking the details on the claim form and then lining it up to a policy and then ensuring that these claims correspond to a rule set out in the policy. Technologies such as Optical Character Recognition have been employed to improve the speed and accuracy of claim processing. Software systems have been employed to capture health insurance claim, thereby reducing the possibility of unreadable information and reduce the risk of error in retrieving the information on the claim form. Optical Character Recognition equipment has also been used to process hard copy claims for efficiency and more accuracy [4].

Now, these claims have been processed, the next phase is the clearinghouses which serve as the third-party or intermediate between the healthcare providers and the insurance providers. The clearinghouses act as the central hub where all the insurance claims are brought to be sorted and sent through to the various insurance carriers. The clearinghouses are necessary because the number of claims that provider needs to submit daily can be quite enormous and all these claims go to different carrier. The clearinghouses are also susceptible to error or fraudulent activities, so the process can go wrong at this stage [4]. The entire process described above from the practitioner filling out the claim form to the insurer receiving the claim can be exposed to various fraudulent activities and can lead to loss of revenue from either the insurer or the healthcare provider.

2.1 The definition of health insurance fraud

Fraud has affected many facets of life and from the discussion above, the Healthcare sector is no stranger. To gain a clearer understanding of what healthcare fraud is, a distinctive comparison is given between the terms fraud, waste and abuse as these terms are sometimes used interchangeably.

Health insurance waste in healthcare is most times unrelated to fraud as it mainly the provision of unnecessary health services [3]. Waste can only be fraud and abuse when the act is intentional. Waste can occur when services are over-utilized and then results in unnecessary expenditure [5].

Health insurance abuse describes the billings of practices that either directly or indirectly, is not consistent with the goals of providing patients with services that are medically necessary, meet professionally recognized standards, and are fairly priced [5]. Abuse occurs when the practices of the service provider are not in line with sound business practices.

Health insurance fraud is purposely billing for services that were never performed and or supplies not provided, medically unnecessary services and altering claims to receive higher reimbursement than the service produced [5]. Fraud is when healthcare is paid for by the insurance subscriber but not provided or a situation whereby reimbursements are paid to the service provider while no such services were provided.

Now we have discussed the differences between waste, abuse, and fraud. The next section discusses the role data mining plays in reducing or eliminating fraud in the healthcare billing process. We look at the drivers for fraud in the healthcare industry and how these factors have affected the health insurance industry.

3 Data mining in health insurance

Technology has played a major role in the Health insurance industry. The impact of advances in technology can be seen in the various aspects of health insurance and the medical billing process is no exception. In the past, the medical billing and coding process was carried out with a paper and pen, then submitted via mail but today the professionals in the insurance field have found their workspace in the virtual and electronic world [6]].

Data mining techniques have in recent times been applied in the health insurance domain to detect these fraudulent claims. Data mining involves extracting, discovering or mining knowledge from a large amount of data. The availability of data and the advancement in technology allows for the design of data mining systems that can extract previously unknown knowledge and insight from the available data [7].

3.1 Data Mining (DM), Knowledge Discovery from Databases (KDD)

The terms data mining and knowledge discovery are used interchangeably. It is the nontrivial extraction of implicit, previously unknown and potentially useful information from data in a repository [8]. Data mining enables us to filter through immense volume of data to find unknown or hidden patterns that can give new perceptions [9]. Although the terms data mining and knowledge discovery in the databases are synonyms, data mining forms a part of the knowledge discovery

system as can be seen from the figure below. In the context of this research, we are going to use the terms knowledge discovery in database and data mining interchangeably.

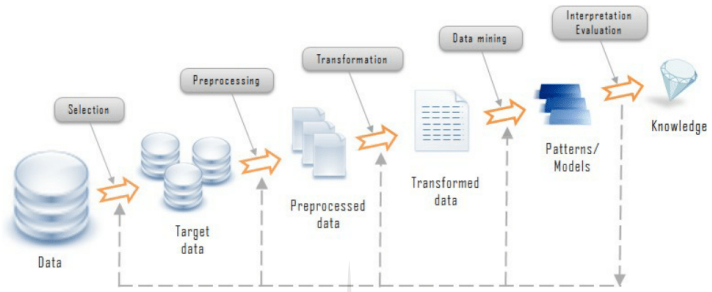


Fig. 2. Diagram illustrating the processes in the data mining system adapted from [10] .

3.2 Current Work

Electronic fraud detection is a relatively new field and grew in popularity with the advent of larger databases. Data mining, used in electronic fraud detection became a reality now as cloud services, data warehouse, and big data have now become a commonplace. Advances in the field of information technology, digitalization of the medical billing process and the vast amount of research on healthcare fraud have created an avenue for the use of data mining and machine learning to fight fraudulent activities [11]. Data mining has gained popularity in researchers as a potential tool to combat health care fraud.

There are several works that have been done in using data mining for fraud detection and this section would be exploring some of those systems. Literature has categorized data mining and machine learning techniques into supervised, unsupervised or hybrid methods. The supervised methods require labelling of data to build a training set, the unsupervised methods, on the other hand, deals with data statistical detection of outlier behaviour. In the following subsections, we discuss several systems that make use of data mining to detect health insurance fraud.

3.3 Systems based on supervised learning

In this subsection, we discuss the systems that make use of the supervised data mining learning algorithms to solve the problem of health insurance fraud. The

systems are reviewed in a chronological order according to the date of publication to see the progression of methods used.

In the National Health Institute (NHI) Taiwan, Wan-Shiou, and San-Yih developed a process- mining framework which detected healthcare fraud [12]. Data from the from a regional service provider for NHI. They created two datasets by filtering out noisy data, identifying medical activities using domain knowledge, identifying fraudulent instances and non-fraudulent instances with the help of domain experts. They made use of the filter model for feature selection as the filter model algorithm does not need to search through the entire space for feature subsets. It is very efficient for domains with many features such as the health insurance domain.

Liou et al. in Taiwan made use of supervised data mining methods to analyze claims for diabetic outpatients that were submitted to National Health Insurance Taiwan [13]. The data used in validating the model was from a random sample of diabetes patients health insurance claims. The fraudulent claims in the dataset were identified by the termination of claim contract. The fraud detection model was built by selecting nine expense related variables and a comparison of these variables was done in two groups of fraudulent and non-fraudulent claims. The expense related input variables used include average drug cost, average drug cost per day, average consultation and treatment fee, average diagnosis fee, average days of drug dispense, average claim amount, average medical expenditure per day and average dispensing service fee.

Shin et al. created a scoring model to detect abusive patterns in health insurance using the 3705 Korean internal medical clinics [14]. They made use of data from the Health Insurance Review and Assessment Services (HIRA). They examined the relationship between the intervention listed by HIRA and the indicating factors to get the data to a manageable size. They extracted 38 indicators which were further validated by HIRA domain experts. The 38 features extracted were used for identifying fraudulent claims using a simple definition of anomaly score.

Now we have seen the systems that used supervised learning algorithms in the data mining process to detect fraud in healthcare. The next sets of systems we look at are the systems that use unsupervised learning methods.

3.4 Systems based on unsupervised learning

Supervised methods work well when used with labelled data. Sometimes the healthcare claim dataset does not have a clear distinction between the records that are fraudulent and those that are not. In this subsection, we discuss the several approaches that used unsupervised learning to detect fraud and abuse in healthcare.

In Canada, Fletcher Lu et al. built an adaptive fraud detection system using

Benfords law. Adaptive Benfords law specifies the probabilistic distribution of digits for many repeating phenomena, notwithstanding incomplete records [15]. The data used was retrieved from Ernst and Young which contained already audited claims data. Their approach created a new fraud discovery approach by combining digital analysis with reinforcement learning technique.

Lin et al. in the work they did on detecting fraud in the Nation health insurance Taiwan general physicians practice data, they made use of unsupervised clustering methods [16]. The applied PCA feature compression for dimensionality reduction of the feature space. They made use of 10 features in the clustering of physicians data. The indicators used include number of cases, average treatment fee per case, average fee per case, amount of fee, average consultation fee per case, percentage of antibiotic prescriptions, number of visits per case, percentage of injection prescriptions, average drug fee per case, amount of prescription days and percentage of injection prescriptions.

4 Assessment Framework

One aspect that is highly important in implementing a data mining model is to critically evaluate and analyze the proposed model. The analysis of a system helps highlight the strengths and the weaknesses of the solution. To compare the similar systems discussed in section 3.2, we defined criteria for the evaluation of the system.

We consider criteria that enable us to gain insights into the system. Some of the criteria considered are highlighted below:

Robustness: Given noisy data or data with missing values, a robust system will still be able to make predictions correctly [17]. The data preprocessing step in the system design is responsible for cleaning up the data before machine learning methods can applied to it. A system that cannot handle noisy or incomplete data will not be suitable for deployment as data in the real world is dirty as a result data can be incomplete, noisy and inconsistent.

Scalability: In developing a data mining system for healthcare claims fraud detection, the volume of available data needs to be considered. The systems need to be designed in such a way that given the large volume of data, they should be able to function effectively [18]. The scalability of the system is assessed using a series of data sets of increasing size. Interpretability: The interpretability of the model refers to the level of insights and understanding that is generated by the model. Interpretability is subjective, hence not so easy to assess. The more insights a system can provide, the more useful it is to the problem it is trying to solve.

Algorithm efficiency : The efficiency of an algorithm mainly depends on the time and space usage. Algorithm efficiency relates to the number of com-

putational resources used by the model. To maximize efficiency, we must minimize resource usage. The efficiency of the systems refers to the cost of computation incurred when generating and using the system [18]. An efficient data mining system will make use of minimal computational resources thereby increasing the speed of operations and reducing the memory usage [17].

Properties of data used: Two very important properties of data we consider are variety and volume. These characteristics of data influence the results generated. The more data used the more inferences can be made. The higher the variety of data the more the deductions can be generalized [17].

Accuracy: Accuracy is a measure of the overall correctness of the model. To compute the accuracy of model, there are four additional terms that form the building block for calculating several evaluation measures: true positives (TP), true negatives (TN), false positives FP) and false negatives (FN) [18]. The TP refer to the positive tuples that were correctly identified by the classifier. The TN refer to the negative tuples that were correctly identified by the classifier. Accuracy is the ratio of the sum of the TP and TF to the total predictions. Accuracy is useful in inferring the correctness of the system [18]. Accuracy involves testing a generated model against an already calibrated data that contains output. The aim is to build models that have high accuracy.

Error rate or misclassification rate: This is the ratio of the sum of false positives and false negatives to the total predictions. It can also be calculated by subtraction the accuracy from 1. A good model will have a very low error rate which implies a high accuracy [18].

5 Results

In this section, we discuss how the different criteria in previous section applies to the systems considered in section 3.2. The systems are analyzed from a structural point of view and then a discussion on the methods used by the system is given. The end goal of the systems discussed is to detect fraudulent claims. We discuss how the systems achieved this goal and to what extent did they solve the intended problem.

The healthcare fraud detection process mining framework defined by Wan-Shiou and San-Yih made use of clinical pathways which are defined as multidisciplinary care plan which diagnosis and therapeutic intervention are performed. In simple terms, it means the order which a physician is supposed to take when carrying out a treatment. They mined frequent patterns from clinical instances and systematically identified practices that deviated from the pathway and flagged them as fraudulent. The approach was evaluated with a real-world dataset retrieved from NHI Taiwan. The results showed that the proposed model could capture and identify several fraudulent and abusive cases that cannot be detected manually [12].

Liou et al. used three data mining techniques including logistic regression, neural networks and classification tree for the fraud detection model [13]. They compared the three data mining techniques and discovered that all three were accurate, but the classification tree method had the best performance as it recorded a 99% accuracy in the overall identification rate. The model had a limitation in the sample data used. The sample data used only consisted three fraudulent service providers whose contracts were cancelled by BNHI. The dataset did not include fraudulent service providers that padded claims but didnt face any penalty such as contract termination. This caused some data limitation and hence result cannot be generalized [13].

The model created by Shin et al. comprised two aspects: scoring to rate the degree of abusiveness and a second part which is identifying the problematic providers by performing segmentation and finding similar utilization patterns. They made use of a decision tree in classifying the providers.

The model performed well when presented with different payment arrangements in detecting abusive patterns. The system made use of the scoring model to alert payers of a potential fraudulent billing pattern [14]. The scoring model provides information on the attributes most dissimilar from the norm. Fraud keeps growing in sophistication and the patterns identified for fraudulent and non-fraudulent behaviour quickly become outdated. The model proposed by Shin et al. is scalable, flexible, easy to use and update. The use of decision trees in the model improved the level of complexity in creating the model as preparing decision trees especially the ones with numerous branches can be difficult and time-consuming.

The adaptive fraud detection system built by Fletcher Fu et al. made use of the Benfords distribution to benchmark the unsupervised machine learning method used to discover new cases of fraudulent activities [15]. When this technique was applied to several records of naturally occurring events, the fraud detection system finds the deviations from expected Brenfords law distributions showing an anomaly in the behaviour indicating a strong possibility of fraud. The system then searches for the root cause of the anomalous behaviour by identifying the underlying attributes causing the anomaly. The model proposed by Shan et al. made use of association rules to mine medical specialist billing pattern. Association rule can be described as statements in the form of antecedents and consequences [19]. For example, if a patient is diagnosed with A then the physician would prescribe drug B and C with a likelihood of 95%. 215 of the association rules were identified Using these predetermined association rules, the model could pick out the physicians who broke these rules and were flagged with a high likelihood of fraud. The study introduced the mining of negative and positive association rules and not just positive rules only, as found in previous models.

Lin et al. used expert opinions to determine the impact of some of these features on the health expenditure. The opinions of the experts were then used to identify and rank critical clusters. The model successfully integrated the data mining process with the segmentation of the GPs practice patterns [16]. The GPs practice pattern detected by applying clustering methods using the features of expenditures of the GP. The final step was then to illustrate managerial guidance based on these expert opinions. The model was benchmarked against real-world data and the results show that the model can effectively and accurately identify fraudulent abuse and behaviours in healthcare [16].

6 Recommendations

The works we have examined in the review paper clearly demonstrates that machine learning and data mining methods can be applied successfully in the health insurance claims fraud detection. However, the review of this literature also shows that there are several gaps that need to be addressed in applying data mining and machine learning methods to effectively detect healthcare claim fraud. Section 7 highlights these gaps and suggests possible ways to address these gaps.

An observation that can be made from the systems reviewed is that less attention was paid to the efficiency of the algorithms and methods used. Even with the abundance of computational resource, we need still need to build systems that not only solve the problem at hand, but we must ensure that we make use of computationally efficient methods.

The practical implementation parameters and the deployment of the proposed model were not discussed by most literature. More research needs to include the deployment and practical implementation of the proposed system in the actual environment to understand what the deployment constraints are and any anomalies that may occur.

The healthcare ecosystem consists of several role players who can be involved in fraud. More research needs to be done to detect fraud that can arise from other roles in the medical billing process. Studies need to be carried out on the potentials of applying data mining methods to detect the insurer fraud.

An interesting observation is the unavailability of literature for the application of data mining methods to detect healthcare fraud different contexts, such as third world countries. The lack of electronic systems for data capturing and auditing could be the major cause. However, where data is available the use of machine learning methods can be used to detect fraudulent activities that may be going

unnoticed in the healthcare process.

Finally, the perpetrators that carry out health insurance claims fraud always find new ways to circumvent measures in place to detect fraudulent claims, therefore making it very important for the fraud detection systems to keep up with the novel forms of fraud. The process of making improvements to systems, such as retraining classification models or gearing implementation parameters can be an expensive exercise especially in the supervised methods where dataset needs to be labelled. One area that can be explored is the use of semi-supervised online machine learning model that can regularly update itself with the new data and then be able to detect novel fraudulent claims pattern leveraging the strengths of unsupervised machine learning.

7 Conclusion

This paper presents an analysis of the literature on the application of data mining methods to health insurance claims fraud detection. Several of the literature reviewed show that the healthcare claims system is highly prone to fraudulent activities which can occur at the different stages of the medical billing process. The adoption of electronic health systems and availability of medical claims data have created an avenue for the application of data mining methods to detect fraud in the healthcare claims process. Data mining methods have been successfully applied to detect these fraudulent activities.

However, each of these data mining methods has their own strengths and weaknesses. In this article, we highlighted the areas, along with trends and propose recommendations relevant for deployment. We identified that there is a need for attention to be paid to the computational efficiency of methods used even with the abundance of the resource. The review of the current data mining solutions to healthcare fraud also highlights the need for data of large volume and variety to be able to improve accuracy and generalize findings.

The development of practical implementation guides that contain details about deployment as well as implementation parameters may improve the adoption and usage of data mining methods to prevent possible claim fraud and misuse of techniques.

Finally, both supervised and unsupervised techniques have important merits in discovering different fraud strategies and schemes. However, to keep the model updated with the latest trends and patterns used by the perpetrators in healthcare claims fraud we suggest exploring a cost-saving model which uses semi-supervised learning to reduce the cost of retraining classification models and subsequently improves the level of accuracy and currency of the model.