BIG DATA ANALYTICS AND ENGINEERING FOR MEDICARE FRAUD DETECTION

by

Matthew Andrew Herland

A Dissertation Submitted to the Faculty of
The College of Engineering and Computer Science
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

Florida Atlantic University
Boca Raton, FL
May 2019

BIG DATA ANALYTICS AND ENGINEERING FOR MEDICARE FRAUD DETECTION

by

Matthew Andrew Herland

This dissertation was prepared under the direction of the candidate's dissertation advisor, Dr. Taghi M. Khoshgoftaar, Department of Computer and Electrical Engineering and Computer Science, and has been approved by the members of his supervisory committee. It was submitted to the faculty of the College of Engineering and Computer Science and was accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

SUPERVISORY COMMITTEE:

Taghi M. Khoshgoftaar, Ph.D.
Dissertation Advisor

Martin Solomon, Ph.D.

Hanqi Zhuang, Ph.D.

Xingquan Zhu, Ph.D.

Nurgun Erdol, Ph.D.
Chair, Department of Computer and Electrical Engineering and Computer Science

Stella N. Batalama, Ph.D.
Dean, The College of Engineering and Computer Science

Khaled Sobhan

Khaled Sobhan, Ph.D.
Interim Dean, Graduate College

April 1, 2019
Date

ACKNOWLEDGEMENTS

First, I would like to thank my graduate advisor Dr. Taghi M. Khoshgoftaar. Without his guidance and support I would not be reaching the end of this long journey and completing the doctorate program at Florida Atlantic University. I would like to further thank him for encouraging me to pursue my doctorate and his patience throughout this process and the fundamental role he played in my growth as a student and researcher. I would also like to extend my gratitude to my PhD supervisory committee members Dr. Martin Solomon, Dr. Hanqi Zhuang, and Dr. Xingquan Zhu.

Additionally, I would like to thank all members in the Data Mining and Machine learning Laboratory of Florida Atlantic University for all the time spent reviewing my research and all assistance provided. I would like to thank FAU Data Mining and Machine Learning Laboratory team members Richard Bauder and Joffrey Leevy for their help with the preparation and review of my dissertation. Also, a further thanks to Richard Bauder for his fundamental assistance throughout my research here at Florida Atlantic University. Finally, I would like to thank my parents, Richard and Linda Herland as well as my girlfriend, Fay Fantauzzi for their unending support, and encouragement throughout my academic journey.

I gratefully acknowledge partial support by the National Science Foundation, under grant number CNS-1427536. Any opinions, findings, and conclusions or recommendations expressed in this dissertation are those of the author and do not necessarily reflect the views of the National Science Foundation.

ABSTRACT

Author:	Matthew Andrew Herland
Title:	Big Data Analytics and Engineering for Medicare Fraud Detection
Institution:	Florida Atlantic University
Dissertation Advisor:	Dr. Taghi M. Khoshgoftaar
Degree:	Doctor of Philosophy
Year:	2019

The United States (U.S.) healthcare system produces an enormous volume of data with a vast number of financial transactions generated by physicians administering healthcare services. This makes healthcare fraud difficult to detect, especially when there are considerably less fraudulent transactions than non-fraudulent. Fraud is an extremely important issue for healthcare, as fraudulent activities within the U.S. healthcare system contribute to significant financial losses. In the U.S., the elderly population continues to rise, increasing the need for programs, such as Medicare, to help with associated medical expenses. Unfortunately, due to healthcare fraud, these programs are being adversely affected, draining resources and reducing the quality and accessibility of necessary healthcare services. In response, advanced data analytics have recently been explored to detect possible fraudulent activities. The Centers for Medicare and Medicaid Services (CMS) released several 'Big Data' Medicare claims datasets for different parts of their Medicare program to help facilitate this effort.

In this dissertation, we employ three CMS Medicare Big Data datasets to evaluate the fraud detection performance available using advanced data analytics techniques, specifically machine learning. We use two distinct approaches, designated as anomaly detection and traditional fraud detection, where each have very distinct data processing and feature engineering. Anomaly detection experiments classify by provider specialty, determining whether outlier physicians within the same specialty signal fraudulent behavior. Traditional fraud detection refers to the experiments directly classifying physicians as fraudulent or non-fraudulent, leveraging machine learning algorithms to discriminate between classes. We present our novel data engineering approaches for both anomaly detection and traditional fraud detection including data processing, fraud mapping, and the creation of a combined dataset consisting of all three Medicare parts. We incorporate the List of Excluded Individuals and Entities database to identify real-world fraudulent physicians for model evaluation. Regarding features, the final datasets for anomaly detection contain only claim counts for every procedure a physician submits while traditional fraud detection incorporates aggregated counts and payment information, specialty, and gender. Additionally, we compare cross-validation to the real-world application of building a model on a training dataset and evaluating on a separate test dataset for severe class imbalance and rarity.

BIG DATA ANALYTICS AND ENGINEERING FOR MEDICARE FRAUD DETECTION

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

CHAPTER 1
INTRODUCTION

1.1 MOTIVATION

The costs associated with delivering quality healthcare in the United States (U.S.) continue to increase while technology and medical knowledge advance. Programs such as the U.S. government funded Medicare [26, 32] are dedicated to reducing these costs and allowing more affordable access to healthcare and promoting beneficiary well-being [50]. Medicare provides health insurance and financial support for the elderly population, ages 65 and older, and other select groups of beneficiaries [38]. Quality and affordable healthcare is an important aspect in people's lives, particularly as they age. In the U.S., life expectancy and population size continue to increase [52, 75], especially in the elderly population, increasing 28% from 2004 to 2015 [1]. Couple this with the rise of chronic diseases, and there are many factors imparting additional, unavoidable strain on programs providing financial support for healthcare services. Healthcare costs are often beyond what a single patient or family can afford, premiums are increasing [74], and insurance programs can only provide a limited amount of funding across all its beneficiaries. Therefore, it is imperative to minimize avoidable or harmful expenditures wherever possible, such as fraud. Unfortunately, healthcare fraud contributes a substantial portion of needless financial loss, which negatively impacts the quality and affordability of healthcare available to U.S. citizens. The objective of the healthcare system should be to provide essential care to as many patients as possible, but with fraud prevalent in the healthcare system, this objective cannot be adequately fulfilled. Thus, finding ways to identify and thwart fraudulent activity is necessary. The focus of this dissertation is to assess data analytics and engineering processes using publicly available datasets to provide insight into improving current fraud detection procedures.

The identification and removal of fraud would have an enormous impact on healthcare, allowing a larger number of patients to receive the care and services they need. The number of physicians who commit fraud are far fewer than not, yet accumulate a significant percentage of financial loss. The Federal Bureau of Investigation (FBI) estimates that 3-10% of all healthcare expenditure is from fraud [104], which is further supported by [79] and [115]. The Coalition Against Insurance Fraud (CAIF) [42] notes that healthcare fraud is by far the largest type of insurance fraud in the U.S, and that there is no consensus on the exact amount lost from fraud. Thus, we present the following statistics to provide an understanding of the effects of fraud, including an estimated value lost from fraud. According to [35], total healthcare spending in the United States, in 2017, was approximately \$3.5 trillion, with Medicare contributing a significant portion with a total spending of \$702 billion (∼20%) [65]. Using the high-end estimate of fraud percentage presented by the FBI, the amount lost from fraudulent actions would be \$350 billion, with up to \$70 billion from Medicare alone. This is a large sum, that affects a large number of patients, rendering them unable to receive financial assistance due to fraudulent behavior. For perspective, the average healthcare cost in the United Sates, per person, per year, is about \$10,000 [126]. This translates to a potential 35 million (\$350 billion/\$10,000) more U.S. citizens able to receive full financial assistance per year, and 6 million from Medicare alone. Other authors offer estimates of potential savings if fraud were adequately handled, including Ko et al. [87] who estimate \$125 million from the field of Urology alone if appropriate regulations were put into place. Fox News released an article describing a single insurance fraud scheme totaling \$1 billion in Medicare costs [54] which was carried out by three individuals, demonstrating the financial burden of fraud by a single case. The Centers for Medicare and Medicaid Services (CMS) [37] note that healthcare spending is forecast to increase 5.5% per year through 2026, with a total of $5.7 trillion by 2026. Thus, the amount lost due to fraudulent behavior would also be expected to increase without intervention. These statistics demonstrate the potential impact of minimizing fraud, where if even a fraction was eliminated, there would be a significant impact for American citizens.

The detection of fraud is critical in identifying and, subsequently, stopping perpetrators. Fraudulent detection, in healthcare, has generally been conducted by auditors or investigators manually reviewing medical records and files to pinpoint possibly suspicious behavior that is potentially fraudulent [120]. Healthcare generates an enormous amount of data, and along with the complexity of financial transactions, effective fraud detection is an impossible task for manual reviewers, and is significantly less efficient compared to an automated data analytics approach, such as what data mining and machine learning could offer [92]. The volume of digital data, within healthcare, is continually increasing along with technology, such as the employment of Electronic Health Records (EHR), which allows for the storage and usage of Big Data. Even though the definition of Big Data continually changes, and has proven difficult throughout the literature to develop one that is agreed upon [71, 72, 78, 98, 100, 107], the amount of data within healthcare would qualify under any definition with an estimated 25,000 petabytes worldwide by 2020 [123, 151]. It is worth noting that Senthilkumar et al. [130] developed a definition, specifically for healthcare, classifying Big Data into six V's: Volume, Variety, Velocity, Veracity, Variability, and Value. Volume refers to vast quantities of data, Variety applies to high levels of complexity of data (i.e. incorporating data from different sources, mash-ups), Velocity represents the high frequency at which new data is generated/-collected, Veracity pertains to the correctness of the data, Variability refers to sizable fluctuations, or variation, in the data, and Value signifies significant data quality in reference to the intended results (i.e. fraud detection). The application of advanced data analytics, specifically machine learning methods and data mining strategies, can be leveraged to improve current fraud detection processes and reduce the resources needed to find and investigate possible fraudulent activities. The Medicare Fraud Strike Force [109], developed by the Office of Inspector General (OIG) [110], demonstrates the effectiveness of data analytics for identifying fraudulent activity with 2,498 indictments encompassing over $3 billion in total fraud between March 2007 through January 2018. Unfortunately, current methods of decreasing monetary losses facing the U.S. healthcare system are not significantly effective, especially compared to the statistics presented above. Therefore, there is a need for developing data processing and advanced data analytics methods that can more efficiently and effectively detect fraud and enhance current methods.

In response to fraud, which includes a recent policy declared by the U.S. Department of Health and Human Services [73], CMS has begun releasing datasets to help produce transparency in healthcare and to assist in identifying fraud within Medicare [49]. There are various datasets available from the Centers for Medicare and Medicaid Services website [33]. The three Public Use File (PUF) Big Data datasets we will be employing throughout this Dissertation are: 1.) Medicare Provider Utilization and Payment Data: Physician and Other Supplier (Part B) [27], 2.) Medicare Provider Utilization and Payment Data: Part D Prescriber (Part D) [29] and 3.) Medicare Provider Utilization and Payment Data: Referring Durable Medical Equipment, Prosthetics, Orthotics and Supplies (DMEPOS) [30]. Information provided in these datasets includes the average amount paid for services and other data points related to claims filed for procedures performed, drugs administered or medical equipment issued by U.S. healthcare providers and its commonwealths, providing a comprehensive view into physician behavior within Medicare. In order to determine whether fraud detection models work in the real-world, identifying physicians who have demonstrated fraudulent behavior is required for validating our detection results, and currently, CMS does not provide fraud labels with their Medicare datasets. However, there is one such dataset that contains real-world fraudulent physicians: the List of Excluded Individuals and Entities (LEIE) [91], established and maintained by the OIG. The LEIE contains physicians who have been found unsuited to practice medicine and thus excluded from practicing in the U.S. for a given period of time. We also searched and compared other real-world fraudulent physician repositories [143] for a number of states and found that either they did not contain the necessary information or the physicians therein were already listed in the LEIE. Other sources, such as the OIG's Criminal and Civil Enforcement website [111] or national and local news sources, do not clearly summarize provider fraud nor include detailed exclusion information in a single data source; thus, the LEIE remains the most reliable and encompassing source for determining real-world fraudulent physicians. Additional details concerning Medicare and Medicare fraud are provided in [15, 36, 26, 38]. Now that CMS and the OIG have publicly released these datasets, it is the responsibility of researchers to perform advanced data analytics and data engineering in order to extract useful information. There are numerous ways data can be engineered, processed and manipulated in order to provide analytical tools and algorithms, such as machine learning [105], the best opportunity to successfully achieve an outcome, such as fraud detection.

In the Medicare datasets we use throughout our experiments, we mapped fraud labels from the LEIE and found that there is a massively larger percentage of non-fraudulent physicians than fraudulent. In machine learning, when one class has a substantially larger number of instances (majority/negative) compared to the other (minority/positive), this is known as class imbalance [62, 85, 90, 127, 128]. In many real-world examples, the class of interest (positive class) is in the minority and Medicare fraud is no exception. The healthcare system in the U.S. contains an extremely large number of physicians, who perform numerous services for an even larger number of patients. Every day, there are a massive number of financial transactions generated by physicians administering healthcare services. The vast majority of these financial transactions are conducted without any fraudulent intent, but there is a very small minority of physicians who maliciously defraud the system for personal gain, translating to severe imbalance. Big Data can exacerbate class imbalance with an enormous over-representation of the negative class [82, 90]. Furthermore, we have found that the number of matching fraudulent physicians between the Medicare datasets and the LEIE has continually dropped since CMS began releasing their datasets in 2012. There are a number of possibilities for this, including insufficient documentation, fraud becoming less prevalent, fraud detection efforts minimizing the frequency of fraudulent behavior, and malicious actors hiding their activity better. Moreover, this decrease does not diminish the problem of needing fraud detection, especially since we have demonstrated that even a single case of fraud can inflict large monetary losses with, in theory, no limit.

In an effort to detect Medicare fraud by leveraging the CMS and LEIE data through machine learning, we employ two distinct approaches: anomaly detection and traditional fraud detection. The main difference between these two methods is that anomaly detection performs classification by physician specialty and traditional fraud detection classifies fraudulent physicians versus non-fraudulent. Furthermore, each approach has unique data processing and engineering steps, resulting in different features extracted from the CMS datasets and different classes for prediction. The other differences will be explicitly explained throughout this dissertation. The idea behind our anomaly detection approach is that if a physician is classified as another specialty (e.g. a Cardiologist classified as a Rheumatologist) based solely on the type and number of procedures performed, then a physician is acting outside the norm of their respective specialty. Note, the only features used in the anomaly detection datasets are the number of times a physician performs each possible procedure by HCPCS code [34], constructed by transforming the Part B dataset only. Based on this idea, misclassification could indicate fraud, misuse, lack of knowledge around billing procedures, or the treatment of mostly unique patients. There are a large number of fields in modern healthcare, resulting in a multi-class classification problem. This approach provides a model that can identify physicians who are potentially misusing insurance systems and require further investigation, which could minimize the time spent searching for fraudulent physicians by narrowing the window of possible suspects for law enforcement. We also propose a number of improvement strategies within our anomaly detection research, including feature selection [141] (to include adjusting for class imbalance through sampling), medical specialty grouping, and the removal of specific, overlapping specialties. For evaluating our anomaly detection approaches we employ exclusion testing, where the models are built using the non-fraudulent physicians and tested using only fraudulent physicians. These fraudulent physicians were identified by inclusion in the LEIE using their unique National Provider Identifier (NPI) [28]. Additionally, with our anomaly detection research, we also experimented with one traditional improvement strategy referred to as class isolation, where we isolate a single medical specialty and predict between fraudulent and non-fraudulent. Note, throughout this dissertation, the terms “field of expertise”, “provider type”, “specialty”, and “class” will be used interchangeably.

For our traditional fraud detection experiments, we employ a different dataset than used for anomaly detection, through our unique data processing and feature engineering built for all three Medicare parts. Our unique process includes choosing features, and transforming the original data from the procedure-level to the provider-level in order to match the LEIE to allow for fraud labeling for all three Medicare datasets. We also construct a new Combined dataset, which encompasses all information from the three CMS datasets. We match and assign fraud labels (fraud and non-fraud) using the NPI variable in the LEIE and each Medicare dataset. We also decided to keep count and payment variables, as well as gender and provider type. Furthermore, we calculate the exclusion period based on the LEIE, which determines the time period a physician's behavior is considered fraudulent. Among our traditional fraud detection research, we also conduct experiments comparing Cross-Validation (CV) and Train_Test for both severe class imbalance and class rarity. Both CV and Train_Test emulate data mining capabilities as they would be employed in real-world situations, providing an estimation of how a model will behave when implemented. CV allows for building and evaluating models on a single dataset, while Train_Test (hold-out) builds the model on past occurrences and evaluates using separate, new occurrences. Train_Test is a more accurate representation of real-world machine learning implementations, and we therefore evaluate the efficacy of CV in this domain. Furthering this argument, Rao et al. [119] note: "Any modeling decisions based upon experiments on the training set, even cross validation estimates, are suspect, until independently verified [by a completely new Test dataset]." Thus, for these experiments, we generated a training and test dataset based on calendar year, where CV uses only the training dataset. For CV, we employ stratified k-fold cross-validation. In comparing Train_Test and CV, we also study rarity, which is an extreme form of class imbalance, where the number of positive class instances (fraud) is so small that machine learning algorithms will have trouble discriminating its unique qualities and will be overwhelmed by the negative class instances (non-fraud). We perform our class rarity experiments by generating data subsets, where positive class instances are randomly removed, lowering the Positive Class Count (PCC) from severe class imbalance to class rarity. Studying class rarity is important in Medicare fraud detection. As mentioned, the number of documented real-world fraudulent physicians available are decreasing every year, moving towards rarity. Additionally, we apply Random Undersampling in order to assess potential improvements in fraud detection by reducing the adverse effects of class imbalance and rarity.

1.2 CONTRIBUTIONS

As discussed, the need for effective methods for identifying fraudulent behavior in healthcare is necessary. These methods would allow law enforcement to efficiently eliminate more fraudulent activity and preserve these funds for programs such as Medicare, which would translate into a better healthcare system and a healthier nation. Additionally, all information gathered from this Medicare research could provide insight into other healthcare domains, helping to minimize fraud throughout the entire healthcare system, through the concept of transfer learning [146]. Therefore, in this dissertation, we make the following contributions to detecting fraud through anomaly detection and traditional fraud detection approaches by leveraging available Medicare data and applying machine learning techniques.

- For anomaly detection:
  - Explore the efficacy of our proposed approach for detecting potentially anomalous providers.
  - Assess whether anomalies within a physician's behavior, based on provider specialty classification, indicate real-world fraudulent behavior through machine learning.
  - Investigate several improvement strategies implemented, through additional data manipulation, and assess the increase in fraud detection.
  - Describe the data processing steps used, including feature engineering for the Medicare Part B dataset for our baseline models and improvement strategies, and real-world fraud mapping with the LEIE.
  - Assess the best strategy for combining HCPCS codes when a physician submits claims from both an office and facility.

- For traditional fraud detection:

  - Compare our proposed anomaly detection approach with traditional fraud detection.
  - Provide a detailed description for our unique data processing of the Medicare Part B, Part D, and DMEPOS datasets including real-world fraud labeling with the LEIE.
  - Combine the three Medicare datasets into one Combined dataset, which encompasses claims from all three parts and assess any improvements in fraud detection performance.
  - Determine which of our four Medicare datasets have better fraud detection performance.
  - Perform experiments that determine whether results garnered through CV estimates are reliably accurate through comparing detection results with Train_Test.
  - Determine whether utilizing Medicare claims data, using our unique data processing methods, can satisfactorily be used to detect real-world healthcare fraud by evaluating results of the Train_Test experiments.
  - Determine the extent data sampling can help with mitigating the negative effects of severe class imbalance and rarity by studying the Train_Test and CV results.
  - Inspect the effects class rarity has on traditional fraud detection through the removal of positive class (fraudulent) instances, evaluating the results as the already very small PCC becomes smaller.

1.3 DISSERTATION STRUCTURE

This dissertation is organized as follows. Chapter 2 details the Medicare data, the LEIE and provides thorough detail for the data processing and feature engineering employed for anomaly and traditional fraud detection. Chapter 3 presents the experimental design, the machine learning algorithms, and the performance metrics used throughout this research. Chapter 4 covers our experiments conducted using anomaly detection. Chapters 5 and 6 cover traditional fraud detection, where the former displays the viability of using multiple data sources and the latter provides a comparison of Train_Test and CV. In Chapter 7, we conclude our study and present future work.

CHAPTER 2
DATA AND FEATURE ENGINEERING

Without a proper understanding of the datasets, available features, or which processing steps are needed, one cannot properly determine the full extent a dataset can be utilized. When data is used to its full potential, machine learning results become more reliable and can be leveraged to solve many problems, which in this case, leads to the most effective detection of Medicare fraud. Therefore, In this chapter we discuss the Medicare datasets, the List of Excluded Individuals and Entities, the data processing and feature engineering for our anomaly detection and traditional fraud detection experiments.

2.1 MEDICARE DATA

In this section, we describe the three CMS datasets we employ throughout this dissertation: Part B, Part D, and DMEPOS. These three datasets offer numerous attributes for provider claims, drugs and medical equipment. Our research efforts revolve around trying to find fraudulent behavior at the provider (or NPI level), which implies a single provider with a single procedure per Medicare claim. Before implementing the CMS data into research, it is important to understand each dataset and how to manipulate and leverage them in the most efficient and effective way. Since CMS records all claims information after payments are made [39, 40, 41], we assume the Medicare data is already cleansed and is correct. Note that NPI is not used in the data mining step, but rather for aggregation and identification. Additionally, for each dataset, we added a year variable which is also used for aggregation and identification.

The information within each dataset is based on CMS's administrative claims data for Medicare beneficiaries enrolled in the Fee-For-Service program. Note, this data does not take into account any claims submitted through the Medicare Advantage program [138].

2.1.1 Part B

The Part B dataset is currently available for the years 2012 through 2016 [27] and provides claims information for each procedure provided to Medicare beneficiaries by physicians and other healthcare professionals within a given year. Each physician is denoted by his or her NPI and each procedure is labeled by its Healthcare Common Procedure Coding System (HCPCS) code [34]. Other claims information includes average payments and charges, the number of procedures performed and medical specialty (also known as provider type). The Part B data is aggregated (grouped by) the following: 1) NPI of the performing provider, 2) HCPCS code for the procedure or service performed, and 3) the place of service which is either a facility (F) or non-facility (O), such as a hospital or office, respectively. Each row, in the dataset, includes a physician's NPI, provider type, one HCPCS code split by place of service along with specific information corresponding to this breakdown (i.e. claim counts) and other non-changing attributes (i.e. gender). Additionally, we have found that in practice, physicians perform the same procedure (HCPCS code) at both a facility and their office, as well as a few physicians that practice under multiple provider types (specialties) such as Internal Medicine and Cardiology. Therefore, for each physician, there are as many rows as unique combinations of NPI, Provider Type, HCPCS code, and place of service, thus the Part B data can be considered to provide procedure-level information. For example, if a physician (NPI = 1003000126) has claimed 20 different procedures and three of them were conducted at both an office and facility (while the other 17 were conducted at one place), there would be 23 rows for this physician (assuming this physician is labeled as only one provider type). Table 2.1 depicts an example of one physician (NPI = 1003000126) from the 2015 Part B dataset. This table shows a few selected attributes, but in the full dataset all variables are available for each NPI/Provider Type/HCPCS code/Place of Service combination.

Table 2.1: Sample of Part B Dataset

npi	nppes__ provider__ gender	provider__ type	place__ of__ service	hcpcs__ code	line__ srvc __cnt	bene__ unique__ count	average__ submitted__ chrg__ amt
1003000126	M	Internal Medicine	F	99217	23	23	328.00000
1003000126	M	Internal Medicine	F	99219	18	18	614.00000
1003000126	M	Internal Medicine	F	99221	59	58	333.28814
1003000126	M	Internal Medicine	F	99231	38	18	100.84211
1003000126	M	Internal Medicine	F	99232	1117	481	200.93196
1003000126	M	Internal Medicine	F	99291	21	13	633.80952.

To provide further understanding, we present two examples for interpreting the data from the 2014 Part B dataset. Example 1: provider "1003000126" performs the "99222" procedure a total of 357 times. From these 357 procedures performed, there were 341 distinct beneficiaries (or patients), where 357 unique services were performed for each unique patient to reduce double counting of services performed (equivalent to how many "office visits" were made for this procedure in 2014). Example 2: provider "1003000134" performs the "88305" procedure a total of 6,760 times. From these 6,760 procedures performed, there were 4,105 distinct beneficiaries (or patients), where 5,109 unique services were performed for each unique patient to reduce double counting of services performed (equivalent to how many "office visits" were made for this procedure in 2014).

2.1.2 Part D

The Part D dataset is currently available for the years 2013 through 2015 [29] and provides information pertaining to the prescription drugs physicians administer under the Medicare Part D Prescription Drug Program within a given year. Physicians are identified using their unique NPI while each drug is labeled by their brand and generic name. Other information includes average payments and charges, variables describing the drug quantity prescribed and medical specialty. The Part D data is aggregated (grouped by) the following: 1) the NPI of the prescriber, 2) the drug name (brand name in the case of trademarked drugs) and generic name (according to CMS documentation). Same as with Part B, we found a few physicians that practice under multiple specialties. Each row in the Part D dataset lists a physician's NPI, provider type and drug name along with specific information corresponding to this breakdown (i.e. claim counts) and other static attributes (i.e. gender). Therefore, for each physician, there are as many rows as unique combinations of NPI, Provider Type, drug name and generic name and thus, Part D data can be considered to provide procedure-level information. For example, if a physician (NPI = 1003000126) has prescribed 20 different drugs, there would be 20 rows for this physician (assuming this physician is labeled as one physician type). For each prescriber and drug, the dataset includes the total number of prescriptions that were dispensed (including original prescriptions and any refills), total 30-day standardized fill counts, total day's supply for these prescriptions, and the total drug cost. To protect the privacy of Medicare beneficiaries, any aggregated records, which are derived from 10 or fewer claims, are excluded from the Part D Prescriber PUF. The total drug cost includes the ingredient cost of the medication, dispensing fees, sales tax, and any applicable administration fees and is based on the amount paid by the Part D plan, Medicare beneficiary, government subsidies, and any other third-party payers. Table 2.2 shows an example of a select number of instances for one physician (NPI = 1003000126) as it is presented in the 2015 Part D dataset. This table shows a few chosen attributes, but in the dataset all of the variables listed, for Part D physicians and drugs, are available for each physician/drug combination. Note that some of these values are suppressed based on factors such as being below 11 (for example claim_count).

To provide further understanding, we again present two examples for interpreting

Table 2.2: Sample of Part D Dataset

npi	specialty__ description	drug__ name	total__ claim__ count	total__ day__ supply	total__ drug__ cost	total__ claim__ count__ ge65	ge65__ suppress__ flag
1003000126	Internal Medicine	AMLODIPINE BESYLATE	27	990	120.01	NA	#
1003000126	Internal Medicine	ATORVASTATIN CALCIUM	15	450	188.85	NA	*
1003000126	Internal Medicine	AZITHROMYCIN	16	87	139.24	NA	#
1003000126	Internal Medicine	CEPHALEXIN	12	96	76.09	NA	#
1003000126	Internal Medicine	CIPROFLOXACIN HCL	15	114	119.36	NA	#
1003000126	Internal Medicine	CLOPIDOGREL	24	780	205.46	NA	#
1003000126	Internal Medicine	FUROSEMIDE	12	360	34.83	NA	*
1003000126	Internal Medicine	HYDRALAZINE HCL	14	375	249.54	14	

the data from the 2014 Part D dataset. Provider "1003000126" prescribes "CEPHALEXIN" a total of 11 times. From these 11 prescriptions, there were 11 distinct beneficiaries (or patients) and a total of 84 days this drug was administered. This drug does not have a value presented for counting number of patients over the age of sixty-five and the corresponding flag is marked with a ‘*’ meaning total_claim_count_ge65 is between zero and ten. The total drug cost is $64.73. In another example, provider "1003000126" prescribes "LEVOFLOXACIN" a total of 59 times. From these 59 prescriptions, there were 57 distinct beneficiaries (or patients) and a total of 361 days this drug was administered. The number of patients over the age of sixty-five was 43 where 42 were unique. The total drug cost is $548.51.

2.1.3 DMEPOS

The DMEPOS (referring Durable Medical Equipment, Prosthetics, Orthotics and Supplies) dataset is currently available for the years 2013 through 2015 [30] and provides claims information about medical equipment, prosthetics, orthotics and supplies that physicians referred patients to either purchase or rent from a supplier within a given year. Physicians are identified using their unique NPI within the data while products are labeled by their HCPCS code. Note, this dataset is based on supplier's claims submitted to Medicare while the physician's role is referring the patient to the supplier. Other claims information includes average payments and charges, the number of services/products rented or sold and medical specialty (also known as provider type). The DMEPOS data is aggregated (grouped by) the following: 1) NPI of the performing provider, 2) HCPCS code for the procedure or service performed by the DMEPOS supplier, and 3) the supplier rental indicator (value of either 'Y' or 'N') derived from DMEPOS supplier claims (according to CMS documentation). Each row provides a physician's NPI, provider type, one HCPCS code split by rental or non-rental with specific information corresponding to this breakdown (i.e. number of supplier claims) and other non-changing attributes (i.e. gender). We have found that some physicians place referrals for the same DMEPOS equipment, or HCPCS code, as both rental and non-rental as well as a few physicians that practice under multiple specialties. The DMEPOS data, per year, is organized where each row contains the physician's NPI and provider type (along with all non-changing physician information) corresponding to one HCPCS code and further split by rental status (yes or no) and all the procedure information corresponding to these four attributes. Therefore, for each physician, there are as many rows as unique combinations of NPI, provider type, HCPCS code and rental status, and thus the DMEPOS data also can be considered to provide procedure-level information. As an example, if physician (NPI = 1003000126) has claimed 20 different procedures and three of them were issued as both a rental and non-rental (while the other 17 were issued as one), there would be 23 rows for this physician (assuming this physician is labeled as one physician type). The DMEPOS data has been aggregated based on the supplier rental indicator since separate fee schedules apply for rental versus purchase of products. To protect the privacy of Medicare beneficiaries, any aggregated records which are derived from 10 or fewer claims are excluded from the Referring Provider DMEPOS PUF. Table 2.3 shows select instances for one physician (NPI = 1003000126) from the 2015 DMEPOS dataset. This table shows a few chosen attributes, but in the complete DMEPOS dataset all of the variables are available for each physician/HCPCS code combination.

Table 2.3: Sample of DMEPOS

REFERRING_ NPI	REFERRING_ PROVIDER_ TYPE	HCPCS_ CODE	SUPPLIER_ RENTAL_ INDICATOR	NUMBER_ OF_ SUPPLIERS	NUMBER_ OF_ SUPPLIER _ CLAIMS	AVG_ SUPPLIER_ SUBMITTED_ CHARGE
1003000126	Internal Medicine	E0431	Y	6	51	48.8546154
1003000126	Internal Medicine	E1390	Y	6	85	251.0091861

Again, to provide further understanding, we present two examples for interpreting the data from the 2014 DMEPOS dataset. For the first example, provider "1003000126" referred patients to suppliers for the HCPCS code "E0431" a total of 86 times with this procedure/service being rendered 89 times. There were 13 distinct beneficiaries (or patients) associated with the claims, where 4 different suppliers are used by the referring provider. The rental indicator was marked Y. In the second example, provider "1003000522" referred patients to suppliers for the HCPCS code "A4259" a total of 76 times with the procedure/service being rendered 116 times. There were 29 distinct beneficiaries (or patients) associated with the claims, where 13 different suppliers are used by the referring provider. The rental indicator was marked N.

2.2 LIST OF EXCLUDED INDIVIDUALS AND ENTITIES

In order to accurately assess fraud detection performance as it appears in real-world practice, we require a data source that contains physicians that have committed real-world fraud. Therefore, throughout almost all experiments presented in this dissertation, we use the List of Excluded Individuals and Entities (LEIE) [91] which is currently the most useful and complete list of excluded providers. It contains the following information: reason for exclusion, date of exclusion and reinstate/waiver date for all current physicians found unsuited to practice medicine and thus excluded from practicing in the United States for a given period of time. This dataset was established and is maintained monthly by the Office of Inspector General (OIG)

[110] in accordance with sections 1128 and 1156 of the Social Security Act [108]. The OIG has authority to exclude individuals and entities from federally funded healthcare programs, such as Medicare. Unfortunately, the LEIE is not all-inclusive where 38% of providers with fraud convictions continue to practice medicine and 21% were not suspended from medical practice despite their convictions [114]. This lack of knowledge regarding all possible fraudulent providers could lead to predicting a provider as fraudulent when they are not, or vice versa, which may reduce the overall accuracy of a prediction model. Even so, fraud cases, like most criminal cases, are only known because those individuals were caught by law enforcement. There are many cases for which the perpetrators are never caught, thus we have no record of these activities.

Moreover, the LEIE dataset only contains the NPI values for a small percentage of physicians and entities. An example of four different physicians and how they are portrayed within the LEIE is shown in Table 2.4, where any physician without a listed NPI has a value of 0. The LEIE is aggregated at the provider-level and does not have specific information regarding procedures, drugs or equipment related to fraudulent activities. There are different categories of exclusions, based on severity of offense, described by various rule numbers. We do not use all exclusions, but rather filter the excluded providers by selected rules indicating fraud was committed [5]. Table 2.5 gives the codes that correspond to fraudulent provider exclusions and the length of mandatory exclusion.

Table 2.4: Sample of LEIE

SPECIALTY	NPI	EXCLTYPE	EXCLDATE
GENERAL PRACTICE/FP	0	1128b6	19770701
EMPLOYEE	0	1128b6	19780124
GENERAL PRACTICE	1003016742	1128a1	20170720
NURSE/NURSES AIDE	1003011644	1128b4	20091220

With other real-world fraudulent physician repositories (which can be found at [143]), we found that they either did not contain the necessary information or the physicians were already in the LEIE. In order to accurately assess fraud detection per-

Table 2.5: LEIE Rules Involving Fraud

Rule Number	Description	Min. Period
1128(a)(1)	Conviction of program-related crimes.	5 years
1128(a)(2)	Conviction relating to patient abuse or neglect.	5 years
1128(a)(3)	Felony conviction relating to health care fraud.	5 years
1128(b)(4)	License revocation or suspension.	5 years
1128(b)(7)	Fraud, kickbacks, and other prohibited activities.	5 years
1128(c)(3)(g)(i)	Conviction of two mandatory exclusion offenses	10 years
1128(c)(3)(g)(ii)	Conviction of 3 mandatory exclusion offenses	Indefinite

formance, we require a data source that contains fraud labels, like the LEIE database, with, at least, the following information: reason for exclusion, date of exclusion, and reinstate/waiver date. After reviewing each state's exclusion lists (if available), there were two states (Missouri [103] and Texas [134]) containing the necessary information, but when compared to the LEIE, we found that all members (with fraudulent reasons for exclusion) were already contained in the LEIE. The LEIE and the state-level exclusion lists are updated often and should be routinely checked for changes to these repositories.

2.3 DATA PROCESSING AND FEATURE ENGINEERING

This section provides an overview of the data processing and feature engineering strategies. Section 2.3.1 covers experiments in Chapter 4 including: processing, dataset information and fraud labeling. Section 2.3.2 covers Chapters 5 and 6 including processing, the Combined dataset, fraud labeling, one-hot encoding, and the training and test datasets.

2.3.1 Anomaly Detection

We employed only the Part B data for the anomaly detection experiments and we are only interested in the provider's specialty, procedures performed, number of each procedure performed, and place of service (office or facility). Using only the procedure codes and associated count of the procedures performed, we transformed each physician entry into a vector where the class label for each instance is the physician's specialty and the features are all available procedures, identified by unique HCPCS codes. The value for each feature is the number of times a given provider billed Medicare for the given procedure (even if the procedure is only done once by one physician in a given specialty). This results in a sparse vector, since most physicians only use a relatively fixed number of codes necessary for their own practice. The rest of the features are then zero for that physician. Table 2.1 shows a small sample of the original CMS data. Table 2.6 shows a small example of the sparse vector, where each line is a physician (NPIs are masked), specialty is the class attribute and every other attribute (codes 99222 through 64482 in this example) are the procedures, where for every instance, there is a value for the number of times the given physician performed that procedure for a given year.

Table 2.6: Sample of Dataset used for this Study

NPI	Specialty	99222	99223	88304	88305	...	62311	64483
001100110	Ophthalmology	0	59	0	9505	...	0	0
987321654	Cardiology	45	0	0	0	...	0	0
555888222	Anesthesiology	6	0	3	0	...	0	0

For experiments conducted in Sections 4.1 and 4.2 we employ only physicians who practice in Florida and from an office (not facility) to manually limit the dataset size, where Section 4.1 only uses the 2013 data and Section 4.2 use both the 2013 and 2014 datasets. These datasets are represented in Table 2.7, which shows the statistics for each year after being transformed into the sparse vector listing the number of physicians, number of procedures and the number of provider types. For Section 4.3 we use the Medicare Part B claims dataset which includes the 2012 through 2015 calendar years, which we combined into a Full dataset as shown in Table 2.8. We developed two schemes for handling the different locations office or facility. The problem with having different places of service is that a physician can bill the same procedure in either an office or facility. To account for these possible multiple entries, we devised two methods of handling this by either Combining Office and Facility (COF) or Separating Office and Facility (SOF). COF is where each procedure performed by each

Table 2.7: Dataset Summary

2013 Statistic	Florida Only
Number of Physicians	40,040
Number of Procedures	2,789
Provider Types (Specialties)	82

2014 Statistic	Florida Only
Number of Physicians	41,896
Number of Procedures	2,563
Provider Types (Specialties)	82

physician, per year, is summed, whether they are billed from an office or facility. For example, in a given year, physician X performed code A 100 times in their office and 500 times in a facility, such as a hospital. Thus, these counts would be summed together with physician X performing code A 600 times. The second method, SOF, is where the codes are treated as separate codes depending on where they were administered. Using the same example with physician X, for a given year, code A would be transformed into two different codes such that code A would be codified as code AO, being done 100 times, and code AF, being done 500 times. Therefore, COF leaves the data with the original number of procedure codes, while SOF increases the number of procedures as seen in Table 2.9.

Table 2.8: Sample of Dataset by Year

PROVIDER_TYPE	2012	2013	2014	2015	Full
Number of Physicians	874,743	909,606	938,147	947,824	1,120,904
Number of Procedures	5,949	5,983	5,973	5,983	7,023
Provider Types (Specialties)	89	90	90	91	89

Table 2.9: Sample of Full Dataset

PROVIDER_TYPE	COF	SOF
Number of Physicians	1,120,904	1,120,904
Number of Procedures	7,023	10,029
Provider Types (Specialties)	89	89

Fraud Mapping

For model testing and validation, we incorporate the LEIE data for determining real-world fraudulent physicians. For the experiments conducted in Sections 4.2 and 4.3, we decided to select physicians located in the LEIE that violated the codes described in Table 2.5 with the exception of 1128(c)(3)(g)(i) and 1128(c)(3)(g)(ii). Only physicians who violated these codes were used as fraudulent physicians in our study. We match physicians from our procedure code Medicare datasets with the LEIE data by NPI. Unfortunately, the LEIE database does not include NPI numbers for all physicians and after preliminary analysis, we found that combining first name, last name, and address is not 100% reliable in determining identity. Therefore, we used only those physicians with NPI numbers to identify matches for mapping fraud labels to the Part B data. Due to the small number of matching physicians in our datasets, we supplemented the matching fraudulent physicians with two other documented fraud cases [44, 113]. Note that the LEIE dataset is only used to identify fraudulent doctors corresponding to the information in the Medicare data. Only the Medicare dataset is used for model training and testing. One important item to mention is that because physicians can be added to the LEIE at any time of the year and some instances in the exclusion dataset may not have complete years (e.g. if they were put on the LEIE on February 1, 2015, they would only have submitted procedures for January 2015), we decided to retain all relevant and available instances. We generate two datasets (exclusion and non-exclusion) from our Medicare Part B datasets with the LEIE excluded physicians as identification as to which dataset they will have membership.

2.3.2 Traditional Fraud Detection

For each dataset (Part B, Part D and DMEPOS), we combined the information for all available calendar years [11]. For Part B and DMEPOS, the first step was removing all attributes not present in each available year. The Part D dataset had the same attributes in all available years. For Part B, we removed the standard deviation variables from 2012 and 2013 and standardized payment variables from 2014 and 2015 as they were not available in the other years. For DMEPOS, we removed a standard deviation variable from 2014 and 2015 as it was not available in 2013. For all three datasets, we removed all instances that either were missing both NPI and HCPCS/drug name values or had an invalid NPI (i.e. NPI = 0000000000). For Part B, we filtered out all instances with HCPCS codes referring to prescriptions. These prescription-related codes are not actual medical procedures, but instead are for specific services listed on the Medicare Part B Drug Average Sales Price file [26]. Keeping these instances would muddy the results as the line_srvc_cnt feature in these cases represents weight or volume of a drug, rather than simply quantifying procedure counts.

We are only interested in particular attributes from each dataset in order to provide a solid basis for our experiments and analyses. For the Part B dataset, we kept eight features while removing the other twenty-two. For the Part D dataset, we kept seven and removed the other fourteen. For the DMEPOS dataset we kept nine and removed the other nineteen. The excluded attributes provide no specific information on the claims, drugs administered, or referrals, but rather encompass provider-related information, such as location and name, as well as redundant variables like text descriptions which can be represented by using the variables containing the procedure or drug codes. For Part D, we also did not include variables that provided count and payment information for patients 65 or older as this information is encompassed in the kept variables. In this case, the claim count variable (total_claim_count) contains counts for all ages to include patients 65 or older. Table 2.10 details the features we chose from the datasets, including a description and feature type (numerical or categorical) along with the exclusion attribute (fraud label) derived from the LEIE.

The data processing steps are similar for Part B, Part D and DMEPOS. All three unaltered datasets are originally at the HCPCS or procedure level, meaning they were aggregated by NPI and HCPCS/drug. To meet our needs of mapping fraud labels using the LEIE, we reorient each dataset, aggregating to the provider-level where all information is grouped by and aggregated over each NPI (and other specific features). For Part B, the aggregating process consists of grouping the data by NPI, provider type, gender and year, aggregating over HCPCS and place of service. Part D was grouped by NPI, provider type and year aggregating over drugs. DMEPOS was grouped by NPI, provider type, gender and year, aggregating over HCPCS and rental status. For the Part D and DMEPOS datasets, their beneficiary counts are suppressed to a value of 0 if originally below 11, and in response we imputed the value of 5 as recommended by CMS.

In an effort to bypass information loss due to aggregating these datasets, we generated six numeric features for each chosen numeric feature for each dataset (Table 2.10). Therefore, for each numeric value, per year, in each dataset, we replace the original numeric variables with the aggregated mean, sum, median, standard deviation, minimum and maximum values, creating six new features for each original numeric feature. The resulting features are all complete except for standard deviation which contains NA values. These NA values are generated when a physician has performed/prescribed a HCPCS/drug once in a given year. Therefore, the population standard deviation for one unique instance is 0, and thus we replace all NA values with 0 representing that this single instance has no variability in that particular year. Two other features included are the categorical features: provider type and gender (Part D does not contain a gender variable).

We excluded repetitious features including physician names, addresses, or code descriptions as they provide no extra value. We also did not include several features containing missing or constant values. NPI was used for identification purposes but not for building the models, and other features, such as Medicare participation, were used for data filtering. Also features, like standardized payments and standard deviation values, are removed since they are not present in all of the Medicare years. Details on all of the available Medicare features can be found in the "Public Use File: A Methodological Overview" documents, for each respective dataset, available at [39, 40, 41].

Combined Dataset

The Combined dataset is created after processing Part B, Part D, and the DMEPOS datasets, containing all the attributes from each, along with the fraud labels derived from the LEIE as displayed in Table 2.10. The combining process involves a join operation on NPI, provider type, and year. Due to there not being a gender variable present in the Part D data, we did not include this variable in the join operation conditions and used the gender labels from Part B while removing the gender labels gathered from the DMEPOS dataset after joining. In combining these datasets, we are limited to those physicians who have participated in all three parts of Medicare. Even so, this Combined dataset has a larger and more encompassing base of attributes for applying data mining algorithms to detect fraudulent behavior, as demonstrated in our study.

Fraud labeling

For all four datasets, we use the LEIE dataset for generating fraud labels, where only physicians within are considered fraudulent, otherwise they are considered non-fraudulent. In order to obtain exact matches between the Medicare datasets and the LEIE, we determined that the NPI value is the only way to match physicians exactly, assuring our data the utmost reliability. The LEIE gives specific dates (month/-day/year) for when the exclusion starts and the length of the exclusion period, where we use only month/year (no rounding within a month, i.e. May 1st through 31st is

Table 2.10: Description of Features Chosen from the Medicare Datasets

Datasets		Feature	Description	Type
Combined	Part B	npi*	Unique provider identification number	Categorical
		provider_type	Medical provider's specialty (or practice)	Categorical
		nppes_provider_gender	Provider's gender	Categorical
		line_srvc_cnt	Number of procedures/services the provider performed	Numerical
		bene_unique_cnt	Number of distinct Medicare beneficiaries receiving the service	Numerical
		bene_day_srvc_cnt	Number of distinct Medicare beneficiary / per day services performed	Numerical
		average_submitted_chrg_amt	Average of the charges that the provider submitted for the service	Numerical
		average_medicare_payment_amt	Average payment made to a provider per claim for the service performed	Numerical
	Part D	npi*	Unique provider identification number	Categorical
		specialty_description	Medical provider's specialty (or practice)	Categorical
		bene_count	Number of distinct Medicare beneficiaries receiving the drug	Numerical
		total_claim_count	Number of drug the provider administered	Numerical
		total_30_day_fill_count	Number of standardized 30-day fills	Numerical
		total_day_supply	Number of day's supply	Numerical
		total_drug_cost	Cost paid for all associated claims	Numerical
	DMEPOS	referring_npi*	Unique provider identification number	Categorical
		referring_provider_type	Medical provider's specialty (or practice)	Categorical
		referring_provider_gender**	Provider's gender	Categorical
		number_of_suppliers	Number of suppliers used by provider	Numerical
		number_of_supplier_beneficiaries	Number of beneficiaries associated by the supplier	Numerical
		number_of_supplier_claims	Number of claims submitted by a supplier due to an order by a referring order	Numerical
		number_of_supplier_services	Number of services/products rendered by a supplier	Numerical
		avg_supplier_submitted_charge	Average payment submitted by a supplier	Numerical
		avg_supplier_medicare_pmt_amt	Average payment awarded to suppliers	Numerical
	All	exclusion	Fraud labels from the LEIE database	Categorical

*Not used for Data Mining
**Not used in the Combined Dataset

considered May). For example, if a provider breaks rule number 1128(a)(3) ('felony conviction due to healthcare fraud') carrying a minimum exclusion period of 5 years beginning February 2010, then the end of the exclusion period would be February 2015. Note that we used the earliest date between the exclusion end date (based on minimum exclusion period summed with start date), waiver, and reinstatement date. Therefore, continuing this example, if there is also a waiver date listed as October 2014 and a reinstatement date of December 2014, the exclusion period would be between February 2010 and October 2014. This accounts for providers that may still be in their exclusion period but received a waiver or reinstatement to use Medicare, thus being no longer considered fraudulent on or after this waiver or reinstatement date.

Contrary to the LEIE data, the Medicare datasets are released annually where all data is provided for each given year. In order to best handle the disparity between the annual and monthly dates, we round the new exclusion end date to the nearest year based on the month. If the end exclusion month is greater than 6 (majority of the year), then the exclusion end year is increased to the following year; otherwise, the current year is used. We do not want a physician to be considered fraudulent during a year unless more than half that year is before their exclusion end date. Continuing the above example, we determined that the end exclusion date was October 2014. Therefore, since October is the tenth month and 10 is greater than 6, the end exclusion year would be rounded up to 2015. Translating this to the Medicare data, any activity in 2014 or earlier would be considered fraudulent when creating fraud labels. For further clarification, if the waiver date would have been March 2014, the end exclusion year would be 2014 and only activity from 2013 or earlier would be labeled fraudulent.

The LEIE dataset is joined to all four datasets based on NPI. We create an exclusion feature which is the final categorical attribute discussed in previous subsections, which indicates either fraud or non-fraud instances. Any physician practicing within a year prior to their exclusion end year is labeled fraudulent. With an exclusion year of 2015, from the physician in our previous example, for Part B, the years 2012 through 2014 would be labeled fraudulent, while for Part D, DMEPOS, and the Combined datasets, 2013 and 2014 would be marked fraudulent (as 2012 is not available for these datasets). Through this process, we are accounting for two types of fraudulent behavior: 1) actual fraudulent behavior, and 2) payments made by Medicare based on submissions from excluded providers, where both drain funds from Medicare inappropriately. For the former, we assume any activity before being caught/excluded is fraudulent behavior. We also include the latter as fraud because, according to the False Claims Act (FCA), this is a form of fraudulent behavior [57]. The final four datasets include all known excluded providers marked via the categorical exclusion feature. Table 2.11 shows the distribution of fraud to non-fraud and number of features within all four datasets. All four datasets are considered highly imbalanced, ranging between 0.038% and 0.074% of instances being labeled as fraud.

Table 2.11: Summary of Final Datasets: 2015 and prior

Dataset	Features	Non-Fraudulent	Fraudulent	% Fraudulent
Part B	126	3,691,146	1,409	0.038%
Part D	126	2,098,715	1,018	0.048%
DMEPOS	145	862,792	635	0.074%
Combined	173	759,267	473	0.062%

One-hot encoding

In order to build our models with a combination of numerical and categorical features, we employ one-hot encoding, transforming the categorical features. For example, one-hot encoding gender would first consist of generating extra features equaling the number of options, in this case two (male and female). If the physician is male, the new male feature would be assigned a 1 and the female feature would be 0; while for female, the male would be assigned a 0 and the female assigned a 1. If the original gender feature is missing then both male and female are assigned a 0. This process is done for all four datasets for gender and provider type/specialty. Table 2.12 summarizes all four datasets after data processing and after the categorical features have been one-hot encoded. Note that NPI is not used for building models and is removed from each dataset after this step.

	Part B		Part D		DMEPOS		Combined	
	Instances	Features	Instances	Features	Instances	Features	Instances	Features
After Processing and Fraud Labeling	3,692,555	35	2,099,733	34	863,427	41	759,740	102
After One-hot encoding	3,692,555	126	2,099,733	126	863,427	145	759,740	173

Table 2.12: Summary of Medicare Datasets: Before and After One-Hot Encoding (2015 and prior)

The Training and Test Datasets

For Chapter 6, we also include the 2016 years as test datasets whereas all years prior are used as training datasets. The methods outlined in the previous subsection were applied to each training and test dataset. Table 2.11 summarizes each training dataset and Table 2.13 summarizes each test dataset used, listing the number of features, number of fraudulent and non-fraudulent instances, and the percentage of fraudulent cases after aggregation, one-hot-encoding, and fraud labeling. The main difference between the training and test datasets are in the provider type labels. There are several provider type labels within the 2016 CMS datasets that were either entered incorrectly, had slight variations, or changed completely. We edited as many of these provider type labels as possible in the 2016 datasets to match them to the training dataset labels. A few examples of these discrepancies are as follows: Allergy/ Immunology to Allergy/Immunology, Hematology-Oncology to Hematology/Oncology, and Audiologist to Audiologist (billing independently). There were also some provider types that were added or removed, which we were unable to match between training and test datasets. To the best of our knowledge, there is no documentation discussing differences in provider types between years. For the Train_Test evaluation method, we removed the non-matching physician types from both the training and test datasets (after one-hot encoding), such as Dentist which was added in 2016 and Optician which was removed in 2016. The removal of these non-matching physicians could affect the Train_Test results due to possible information loss, and we believe, due to the removal of such a small percentage of provider types, there is no significant impact on the model fraud detection results. All the non-matching provider types are documented in Table 2.14. It is important to note that there is no known documentation discussing differences in provider types across Medicare years.

Table 2.13: Summary of Final Datasets: Test (2016)

Dataset	Features	Non-Fraudulent	Fraudulent	% Fraudulent
Part B	126	999,815	99	0.010%
Part D	123	744,918	135	0.018%
DMEPOS	119	290,548	75	0.026%
Combined	171	256,529	55	0.021%

Table 2.14: Provider Type Labels Removed from Training and Test Datasets

Dataset	Not in Train	Not in Test
Part B	Hospitalist Dentist	Psychologist (billing independently) Pharmacy
Part D	Dentist Hospitalist Individual Certified Prosthetist-Orthotist	Medical Supply Company, Other All Other Suppliers Ambulance Service Supplier Pharmacy Voluntary Health or Charitable Agencies Centralized Flu
DMEPOS	Hospitalist	All Other Suppliers Ambulatory Surgical Center Anesthesiologist Assistants Audiologist (billing independently) Centralized Flu Clinical Laboratory HHA (Dmercs Only) Independent Diagnostic Testing Facility Individual Certified Orthotist Individual Certified Prosthetist Mass Immunization Roster Biller Medical Supply Company, Other Medical Supply With Certified Orthotist Medical Supply With Certified Prosthetist-Orthotist Medical Supply With Prosthetist Medical Supply With Resp. Therapist (Dmercs Only) Occupational therapist Ocularist Optician Pharmacy (Dmercs Only) Physical Therapist Public Health Welfare Agency Slide Preparation Facility SNF (Dmercs Only) Speech Language Pathologist Supplier of Oxygen and/or Oxygen Related Equip. Voluntary Health or Charitable Agency
Combined	Hospitalist	Clinical Psychologist Occupational therapist Physical Therapist

CHAPTER 3
THEORY AND METHODOLOGY

This chapter discusses the methodologies used throughout this dissertation, covering experimental design for both anomaly and traditional fraud detection research efforts. We provide descriptions for each machine learning algorithm and techniques employed in our research. Lastly, we describe our performance metrics used to evaluate the level of fraud detection our models were able to achieve, also split by anomaly and traditional fraud detection experiments.

3.1 EXPERIMENTAL DESIGN

In this section, we outline the methodology used for anomaly and traditional fraud detection. For anomaly detection, we discuss the hypothesis behind this line of research, how we test for fraudulent behavior and present the improvement strategies employed, where these experiments are presented in Chapter 4. The discussion for traditional fraud detection covers how we evaluate our models, providing insight into both CV and Train_Test as well as presenting the issue of severe class imbalance and rarity, and methods of mitigating the negative effects through data sampling. The traditional fraud detection experiments are carried out in Chapters 5 and 6.

3.1.1 Anomaly Detection

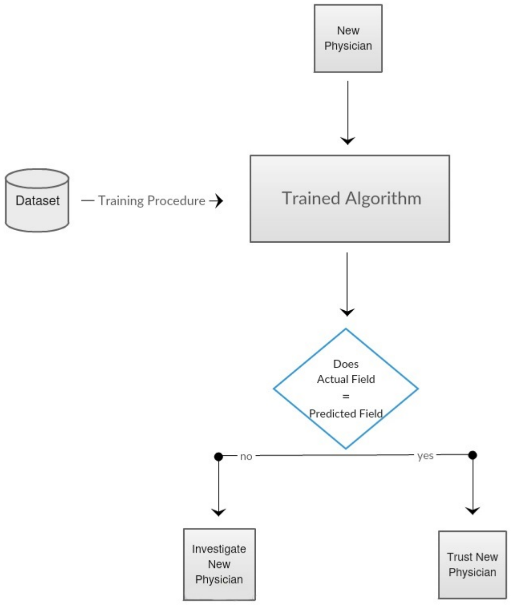
Figure 3.1 explains the basic workflow of our anomaly detection experiments, where the Part B procedure data is leveraged to create a model which takes the physician in question and determines whether they fit into the norm of their respective field. If a physician is determined to fit into the norm of their respective field, then they are considered trustworthy; if not, there may be reason to investigate the physician for possible fraud or misuse. We are determining whether or not it is possible to predict a physician's field of expertise based on procedure counts.

In order to assess the viability of anomaly detection, we need to validate, through the use of real-world fraudulent physicians, the hypothesis that by using only procedural data, a prediction model can successfully detect possibly fraudulent or wasteful behaviors. Therefore, if our model points to "Investigate New Physician", as shown in Figure 3.1, for a physician who is actually fraudulent, then our model is successful. The procedure used for testing models is exclusion testing, where we removed the fraudulent physicians from the training dataset (non-exclusion) and created a test dataset (exclusion) composed of all matching fraudulent physicians. The models are built using the non-exclusion dataset and evaluated on the exclusion dataset. In Section 4.2, the test evaluation was done by reviewing the resulting confusion matrix, where the number of instances predicted as a class other than their actual field are denoted as possibly fraudulent. Since the exclusion/test dataset contains only fraudulent physicians, every physician correctly identified as their actual specialty fails our hypothesis. For Section 4.3, due to the larger number of fraudulent physicians, when testing the models, the exclusion dataset is split into a number of smaller datasets, one for each physician type (specialty), and each of these datasets is used as the test set.

Improvement Strategies

We also experimented with a number of improvement strategies in Sections 4.2 and 4.3, to compare to our baseline models. The baseline model for Section 4.3 is determined between COF and SOF as discussed in Section 2.3.1 and the best learner combination, whereas for Section 4.2 the baseline model is the model examined in Section 4.1, which is also described in Section 2.3.1. Additionally in Section 4.3,

Figure 3.1: Machine Learning Workflow



we perform an experiment branching away from anomaly detection, but using the same dataset split by specialty, where the classes are fraudulent and non-fraudulent (combined the exclusion and non-exclusion subsets) in order to compare results with anomaly detection, referred to as class isolation. The COF or SOF method which gives better performance results is also used for the class grouping, removal, and class isolation method. In this subsection, we discuss the improvement strategies employed throughout our anomaly detection experiments to include feature selection and sampling, grouping of similar physician types, removing specific classes, and class isolation.

Feature Selection and Sampling This strategy is used only in Section 4.2 where the dataset was altered into two classes, or specialties prior to feature selection or sampling is performed. The specialty that we wish to focus on for classification was kept as a single class (the positive class), while each instance of the remaining fields were grouped into a single class (the negative class or *other class*), creating a one-versus-all scenario. For example, we can choose Podiatry as the positive class and then group all other classes in the *other class* (negative class) for binary classification. We use the Weka [60] platform to apply the Gain Ratio feature selection technique [58]. Gain Ratio returns the top n features (procedure codes in our case) to keep, while removing the rest. Through initial experimentation, we chose a range of procedures for n from all the procedures to 500 procedures, in increments of 500. The performance results will indicate the optimal number of procedures to keep, as well as any model improvements due to feature selection.

Gain Ratio is a slightly altered version of the information gain feature selection technique, born out of a weakness of the latter, where it tends to have bias toward attributes with many values [58]. In the Medicare claims data, each attribute can potentially be any value given that each attribute is a procedure, and a physician can perform a procedure any number of times. Therefore, using the Gain Ratio feature selection technique, that considers removing such bias, is beneficial to our anomaly detection research.

In addition to feature selection, we also experiment with undersampling and oversampling techniques to improve model performance. Sampling could be beneficial for feature selection since the field of interest (positive class) is relatively small compared to the other class (negative class). Thus, incorporating sampling techniques could help improve the overall prediction results. Undersampling is the preprocessing technique of removing instances from the majority class in order to balance the two classes. Oversampling is another method for balancing classes, but rather than removing instances from the majority class, oversampling adds instances to the minority class. For undersampling, we use the preprocessing algorithm SpreadSubsample [148] and for oversampling, we use the Synthetic Minority Oversampling Technique (SMOTE) [23]. Both sampling techniques are done in Weka, where we assume the default configurations unless specifically stated otherwise.

Removing Classes For this approach, we remove classes based on unique procedures that have both a high number of instances and poor classification performance. In Section 4.2, four classes were chosen for removal from the dataset as outlined as criteria 1 in Table 3.1 including: Family Practice, Nurse Practitioner, Internal Medicine, and Physician Assistant. Together, these four specialties make up a large portion of the dataset. The choice to remove these four classes was determined by the confusion matrix of the 2013 Florida data and confirming that all removed classes do indeed cause a relatively large number of misclassifications. A confusion matrix is a tabular representation comparing the predicted class membership against the actual class membership for each instance present in the dataset, denoting true positives, true negatives, false positives, and false negatives. Specialties with high misclassification rates would be considered generic, meaning that they most likely perform HCPCS codes that a number of other fields also perform (i.e. overlapping procedures). We repeated the procedure to validate the model (with removed classes) based on the known fraudulent physicians to compare performance to the original (baseline) model.

For Section 4.3, we have two different sets of specialties (criteria 1 and 2) for removal to test model performance. Table 3.1 shows the specialties for both criteria of class removal. The first criteria is the same as used in Section 4.2. The second criteria for removal includes those removed via the first criteria plus specialties with low scores (Precision and Recall) and containing the words 'medicine', 'general', or 'unknown' (i.e. Unknown Supplier/Provider), indicating less specific practices (e.g. family practice) or ambiguous and less defined specialties (via the 'unknown' word).

Grouping Specialties The class grouping strategy groups similar specialties into a single group to reduce redundant specialties and decrease model variance. Similar specialties, or classes, are ones that share a significant number of HCPCS codes, thus indicating overlapping procedures. Table 3.2 shows the total number of procedures performed and the overlap between Ophthalmologist and Optometrist, across five

Table 3.1: Specialties used in the class removal strategy

Class Removal Strategy	
Criteria 1: Four Classes	Criteria 2: Chosen Classes
Family Practice	Certified Clinical Nurse Specialist
Nurse Practitioner	General Surgery
Internal Medicine	General Practice
Physician Assistant	All Other Suppliers
	Unknown Physician Specialty Code
	Unknown Supplier/Provider
	Nuclear Medicine
	Osteopathic Manipulative Medicine
	Sports Medicine
	Geriatric Medicine
	Preventative Medicine
	Addiction Medicine
	Pediatric Medicine

randomly chosen codes.

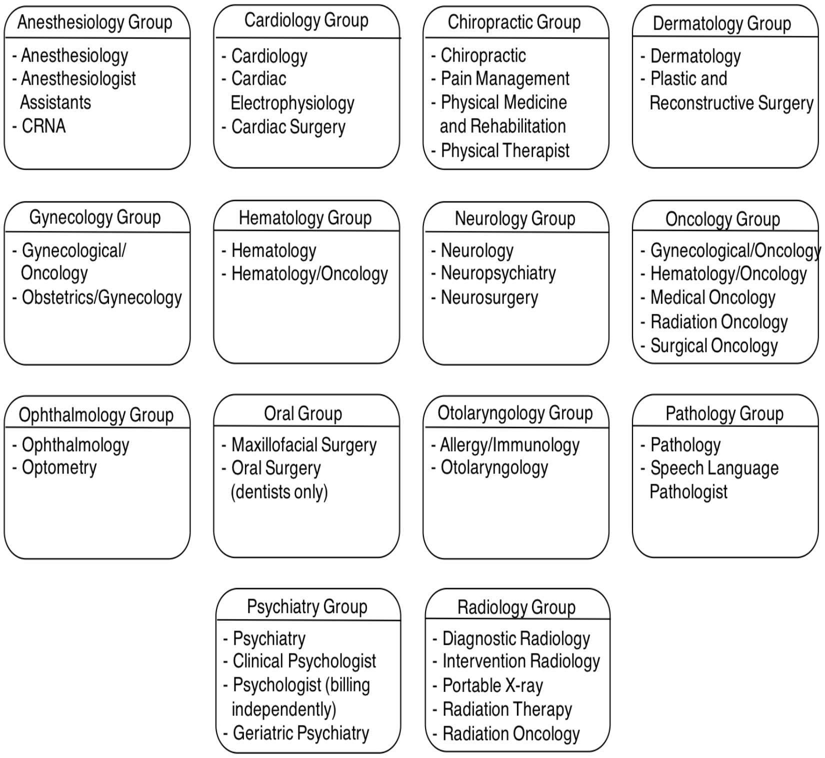
The process for determining similar classes in Section 4.2 was done by the confusion matrix and led to the following groupings: Otolaryngology, Cardiology, Dermatology and Ophthalmology. In Section 4.3, we chose additional groupings based on the assumption that if the specialties are similar then they share a significant number of HCPCS codes, thus indicating overlapping procedures, for a total of 14 groupings. The groupings are outlined in Figure 3.2. Note that classes such as Internal Medicine, which can be confused for many other classes, were not considered due to the large grouping that would be created (including too many other classes) and thus not useful for this experiment. A very large group that consists of many classes would defeat the purpose of grouping, as we want to find smaller groups of specialties where every class within this group is actually similar to each other in practice. For instance, Internal Medicine could share a number of procedures with both Cardiology and Optometry, but Cardiology and Optometry are not similar nor is Internal Medicine. Additional anecdotal confirmation of these similar class groupings was found based on whether these groups were reasonable or not. For example, Ophthalmology and

Optometry provide medical procedures focused on the eyes, and thus is a reasonable group. Classes were grouped manually on a specialty-by-specialty basis. Note that because these groups are manually formed, there could be other groupings, but our goal is to demonstrate the effectiveness of grouping for improving prediction and fraud detection performance and not specifically the formation of groups.

Table 3.2: Overlapping HCPCS Code Example

HCPCS Code	Ophthalmologist	Optometrist
92004	329	143
92012	904	423
92014	2090	354
92083	194	67
92250	151	186
Overlap	31%	88%

Figure 3.2: List of Groupings and Group Members



In order to evaluate any performance improvements via class grouping, we take a two stage approach in Section 4.3. The first stage consists of creating datasets for each of the 14 groupings. We evaluate the performance of each group, using Recall, Precision, F-score, G-measure, and Accuracy, versus the individual results for each of the members of the group, using the baseline model. In the second stage, we choose the groups that showed improvement over their individual members (per specialty) and conduct two tests determining improvement in fraud detection: 1) for each group individually and 2) all together (combined groups dataset).

Class Isolation The class isolation method improvement strategy is specific to Section 4.3. With this method, we split the Medicare Part B dataset and build fraud detection models for individual specialties. In the previous strategies, as with the baseline models, the result is a prediction of the physician's specialty from which we can validate fraud detection using the LEIE database. For these methods, to assess a correct prediction, the predicted specialty is compared to the actual specialty and if they differ, this could indicate possible fraud. This possible fraud result is compared to the LEIE to determine whether this prediction captured an actual fraudulent provider or not. For class isolation, however, the model results are not the physician's specialty but rather fraudulent and non-fraudulent derived from the mapped LEIE fraud labels, thus, qualifying as a traditional fraud detection approach. In order to provide both fraud and non-fraud labels, the exclusion and non-exclusion datasets are combined. For Section 4.3, we chose the following specialties, from the original dataset, based on having 50 or more available LEIE fraud labels: Chiropractic, Family Practice, General Practice, Internal Medicine, Physician Assistant, and Psychiatry. For each of these specialties, the class isolation method uses the SOF subset of data to build models, as chosen in the baseline model selection process, and can be compared to the other improvement strategies via the F-score measure.

For the first part of the class isolation method, we employ Random Under Sampling (RUS) with a 3:1 ratio (class distribution), resulting in 75% non-fraud and 25% fraud instances. RUS removes samples from the majority class (non-fraud), while keeping all of the minority class (fraud) observations. For example, the Chiropractic specialty had 86 excluded physician instances, which are retained, with 258 non-excluded Chiropractic instances chosen at random from the entire dataset. We decided to use a 3:1 ratio in order to retain a larger amount of non-fraud instances, thus retaining more information relative to the fraud instances. Based on our previous research in applying machine learning techniques with varying levels of imbalanced datasets [84, 140], we found that RUS is an effective method for mitigating the adverse effects associated with class imbalance, including higher levels of imbalance (i.e. a low number of fraud labels relative to non-fraud labels) such as in our study. Note that the goal in our study is not to find the best class distribution, but to determine the effectiveness of this particular class isolation approach in Medicare fraud detection. In order to reduce bias due to lucky or unlucky draws in the RUS process, we generated ten different datasets (RUS repeats) for the six specialties, each with the 3:1 ratio. For each of these ten datasets, we built each model, per specialty, using 10-fold cross-validation. The performance results were averaged over the 10 repeats for the final model evaluation. Type I and II error rates are used as the primary metrics, while F-score is used to validate the overall fraud detection performance. The second part of the class isolation method uses a cost-sensitive classifier [147]. This classifier is used to find the model with the lowest combination of Type I and II error rates, while minimizing the Type II error rate. To do this, only the cost of a Type II error is varied, which changes the ratio between Type I and II errors. In our case, the Type II error is more important as it can indicate money lost due to fraudulent activities. Thus, we increase the cost associated with this error to help the model better discriminate fraud instances.

3.1.2 Traditional Fraud Detection

This section presents the methods employed for traditional fraud detection. First we present the two methods used for evaluating our models: CV and Train_Test (hold-out). We also explore the concept of class imbalance, specifically severe class imbalance and rarity, as well as the strategies we employ for mitigating the negative effects of class imbalance to include data sampling.

Cross-validation and Train_Test

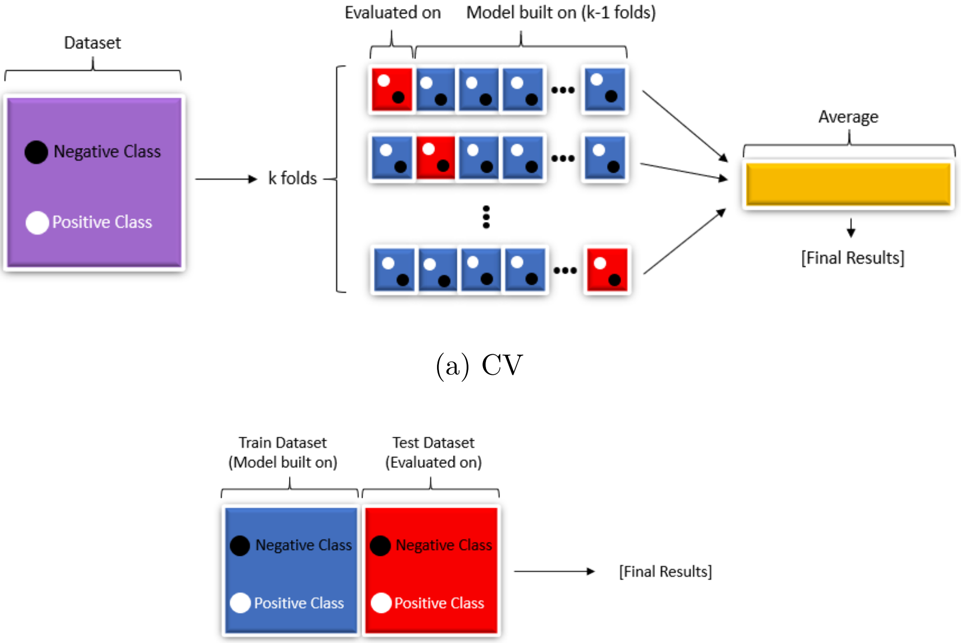
There are a number of different versions of CV, and throughout Chapters 5 and 6, we employ stratified k-fold cross-validation for evaluating our models, where k = 5. CV, as depicted in Figure 3.3a, allows a learner to evaluate a single dataset by evenly splitting the dataset into k-folds. A model is built on k-1 folds and evaluated on the remaining fold, repeated until each fold has been used for evaluation, with the final result being the average of the evaluation scores across all of the k-folds. Through this process, CV guarantees that each instance in the dataset is used in building and evaluation. Stratification ensures all folds have approximately the same ratio of class representation as the original data, which is important when dealing with highly imbalanced data, and even more so with rare data, which could result in a fold devoid of positive class instances. In order to validate the CV results and avoid any bias caused by bad random draws when creating folds, we repeat the CV process 10 times for each learner/dataset pair, where the final AUC score is the average over these repeats. The final detection performance score is the average of all 10 CV repeats.

CV is very popular among the Data Mining and Machine Learning community [17] as an evaluation method for prediction performance in almost every application domain, and can be useful when a researcher only has access to prior data. It should be noted that, Rao et al. [119] recommend validating CV results with a separate test dataset. They also mention that when a model is tuned by a test dataset, this is no longer an accurate simulation of the real-world event. A few other drawbacks of CV, as found in the literature, are CV can result in large errors using small sample sizes [142], the error introduced by bias or variance [80, 117], and CV being vulnerable to high levels of variability. Therefore in Chapter 6 we compare CV to Train_Test results and determine whether employing CV estimates are similar to results from the Train_Test evaluation method. Throughout Chapters 5 and 6 we use the dataset from 2015 and prior (Section 2.3.2), where in the latter Chapter this dataset is referred to as the training dataset. The dataset used with CV is unaltered to match the test dataset (Section 2.3.2) as CV does not employ the test dataset. Therefore, by employing the 2016 test dataset with Train_Test, we are evaluating the viability of CV for providing estimates that lead to model selection in Medicare fraud detection.

Train_Test is exclusively studied in Chapter 6, and uses a training dataset for building the model, and evaluates this model using a separate, distinct test dataset, as demonstrated in Figure 3.3b. Instances from the test dataset are completely new, and never used for model building, as opposed to CV. Examining the Train_Test method's performance is necessary for assessing whether, based on past occurrences, a model can accurately predict new occurrences. Through our experimentation, Train_Test will determine, based on prior Medicare data (years < 2016), whether a physician can be accurately classified as fraudulent or non-fraudulent given Medicare data from the most recently released 2016 datasets.

Severe Class Imbalance and Rarity

In addition to comparing Train_Test and CV, our experiments in Chapter 6 also explore the effects of severe class imbalance and rarity, as well as exploring techniques to curtail the negative effects of class imbalance. Having a large difference between the number of majority and minority class instances can create bias towards the majority class when building machine learning models. This is known as class imbalance [140],



(a) CV

(b) Train_Test

Figure 3.3: Flowcharts: Model Evaluation Methods

which presents issues for machine learning algorithms when attempting to discriminate, often complex, patterns between classes, particularly when applied to Big Data. Table 3.3a demonstrates the severe degree of imbalance present in each dataset, by year. One interesting observation is that the number and percentage of fraudulent instances matching between the LEIE and the Medicare datasets decreases every year, across each dataset. There are a few possibilities explaining this decrease, including the continued efforts to remove fraudulent physicians from practice, fraudulent physicians more efficiently avoiding detection, law enforcement shifting focus from physician fraud, or the deterrent effect of technological advances in fraud detection. We also note from the labeled Medicare datasets that each year, the non-fraudulent instances generally increase at a faster rate than the fraudulent cases are decreasing, furthering the imbalance. If this trend continues, machine learning for Medicare fraud detection could move from a severe class imbalance problem to a class rarity problem.

Rarity is an exceptionally severe form of class imbalance.

In real-world situations, when severe class imbalance and rarity are present, the minority class is generally the class of interest [127]. When employing Big Data in machine learning, severe class imbalance and rarity exhibit a large volume of majority class instances, increased variability, and disjuncts. Small disjuncts are associated with issues, such as between- and within-class imbalance [62, 76]. Generally, a learner will provide more accurate results for large disjuncts, which are created based on a large volume of instances. Large disjuncts can overshadow small disjuncts, leading to overfitting and misclassification of the minority class due to the under-representation of subconcepts [145]. Therefore, rarity is an important topic to study in Medicare fraud detection, and in order to study rarity, in Section 6.2, we generate additional training datasets as shown in Table 3.3b. All non-fraudulent instances are kept, while we remove a number of fraudulent instances, achieving further levels of severe class imbalance and rarity. The PCCs in these new generated datasets range from 1,000 to 100, based on original number of fraudulent instances. These PCCs were further chosen, based on preliminary results, which demonstrate that these adequately represent class rarity in Big Data. In order to get a thorough representation of fraudulent instances, we generate ten different datasets by re-sampling for each dataset/PCC pair. For example, with regard to the 200 PCC for the Part D, we randomly select 200 instances from the original 1,018, with this process repeated ten times. The final result for each dataset/PCC pair is the average score across all ten generated rarity subsets.

Data sampling is used to minimize the effects caused by severe class imbalance and rarity in Chapter 6. The two main branches of data sampling are oversampling and undersampling. Oversampling generates new minority instances and undersampling removes majority instances, both with the goal of achieving a given ratio between the majority and minority representation. Oversampling has a few disadvantages,

Table 3.3: Summary of Medicare datasets

Year	Part B		Part D		DMEPOS		Combined	
	Fraud	%Fraud	Fraud	%Fraud	Fraud	%Fraud	Fraud	%Fraud
2012	546	0.062%	-	-	-	-	-	-
2013	403	0.044%	465	0.069%	323	0.110%	229	0.090%
2014	285	0.030%	329	0.047%	193	0.068%	154	0.061%
2015	175	0.018%	224	0.031%	119	0.041%	90	0.035%
2016	99	0.010%	135	0.018%	75	0.026%	55	0.021%

(a) By Year

PCC	Part B		Part D		DMEPOS		Combined	
	Fraud	%Fraud	Fraud	%Fraud	Fraud	%Fraud	Fraud	%Fraud
All	1,409	0.038%	1,018	0.048%	635	0.074%	437	0.062%
1000	1000	0.027%	-	-	-	-	-	-
400	400	0.011%	400	0.019%	400	0.046%	-	-
200	200	0.005%	200	0.010%	200	0.023%	200	0.026%
100	-	-	100	0.005%	100	0.012%	100	0.013%

(b) Class Imbalance and Rarity

including decreased model generalization provoked by the duplication of existing minority class instances [22] and the increased processing time due to these additional instances. For these reasons and based on our prior research where oversampling has been shown to decrease fraud detection [12], we select Random UnderSampling (RUS). The main drawback of RUS is the potential removal of useful information, but it is beneficial when applied to Big Data as removing instances decreases both required computing resources and build time, as well as being supported by [61, 84, 140]. As we employ RUS, our goal is to incur minimal information loss while simultaneously removing the maximum number of majority instances (i.e. determine which ratio delivers the best fraud detection). Therefore, we chose the following class ratios: 1:99, 10:90, 25:75, 35:65, and 50:50 (minority:majority), including the full, non-sampled datasets as the baseline (labeled as all:all or Full). In applying these ratios, we generate ten datasets for each original and generated training dataset, to reduce bias due to poor random draws. These ratios were chosen because they provide a good distribution, ranging from balanced 50:50 to highly imbalanced 1:99 [10]. Note, for Train_Test, when applying RUS or creating the severe class imbalance and rare subsets, only the training datasets are sampled, as they build the model, while test datasets are kept unaltered for model evaluation.

3.2 MACHINE LEARNING ALGORITHMS

In this section, we discuss the machine learning models used in our research. For running and validating models, we use either Weka [60] or Apache Spark [152, 153], depending primarily on dataset size. Apache Spark is a unified analytics engine for large-scale data processing [131] and is capable of handling Big Data, offering dramatically quicker data processing over traditional methods or other approaches using MapReduce. Spark is used on top of a Hadoop [3] YARN cluster and can effectively handle the large datasets presented by Medicare. Weka is an application providing a suite of machine learning algorithms that can be applied either via a graphical interface or the command line, and is suitable for smaller datasets. Therefore, when the datasets are small we employ Weka due to the larger library of available machine learning algorithms and techniques, but when using full datasets, due to their Big Data, we employ the Apache Spark Machine Learning Library (MLlib) [102][1]. The MLlib provided by Apache is a scalable machine learning library built on top of Spark.

3.2.1 Multinomial Naïve Bayes

Multinomial Naïve Bayes (MNB) classifies new instances by finding the posterior probabilities of class membership based on each feature value, which is learned from a set of labeled training instances. The approximation is done using Bayes' rule by assuming conditional independence. Conditional independence is the idea that each feature in the dataset is independent from one another which is rarely true in practice. However, the model is very effective and is used extensively in the field of data mining and machine learning [48, 150].

[1] https://spark.apache.org/docs/latest/ml-classification-regression.html

3.2.2 Logistic Regression

Logistic Regression (LR) predicts probabilities for which class a categorically distributed dependent variable belongs to by using a set of independent variables employing a logistic function. LR uses a sigmoidal (logistic) function to generate values between [0,1], that can be interpreted as class probabilities. LR is similar to linear regression but uses a different hypothesis class to predict class membership [48, 89, 150]. Unlike Naïve Bayes, there is no requirement for statistical independence between independent variables, though there is an assumption of collinearity. Multinomial Logistic Regression (MLR) is an extension of binomial Logistic Regression that allows for more than two categories of the dependent variable.

Specifically for Section 4.3, we use the the Limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm (LBFGS) version of LR to improve memory usage [97], and employ both Binomial and Multinomial Logistic Regression as necessary. For Chapters 5 and 6, we set the bound matrix to match the shape of the data (number of classes and features) so the algorithm knows the number of classes and features the dataset contains. The bound vector size is equal to 1 for binomial regression, and no thresholds are set for binary classification.

3.2.3 Support Vector Machine

Support Vector Machine (SVM) models create a space consisting of training instances portraying them as points, mapping them in a way that best creates linearly separable categories, where the goal is to have the largest gap between them. The specific implementation of SVM used in Weka is called Sequential Minimal Optimization (SMO), which uses this optimization algorithm as the training method for Support Vector Classification [48, 116, 150].

3.2.4 Random Forest

Random Forest (RF) is an ensemble learning method that generates a large number of trees, employing sampling with replacement, creating a number of randomized datasets to build each tree, where features are selected automatically, at each node, through entropy and information gain. Each tree within the forest is dependent upon the values dictated by a random vector that is independently sampled and where each tree is equally distributed among the forest. The generation of random datasets minimizes overfitting. The class value that appears most often among these trees (mode) is the class predicted as output from the model. As an ensemble learning method, RF is an aggregation of various tree predictors. Each tree within the forest is dependent upon the values dictated by a random vector that is independently sampled and where each tree is equally distributed among the forest [19, 48, 150]. The RF ensemble inserts randomness into the training process which can minimize overfitting and is fairly robust to imbalanced data [10, 83].

Specifically for Chapters 5 and 6 we build each RF learner with 100 trees as our research group has found little to no benefit using more trees. The parameter that caches node IDs for each instance, was set to true and the maximum memory parameter was set to 1024 MB in order to minimize training time. The setting that manipulates the number of features to consider for splits at each tree node was set to one-third, since this setting provided better results upon initial investigation. The maximum bins parameter, which is the max number of bins for discretizing continuous features, is set to 2 because we no longer have categorical features since they were converted using one-hot encoding.

3.2.5 Gradient Boosted Trees

Gradient Boosted Trees (GBT) [102] is an ensemble of decision trees which trains each decision tree individually in order to minimize loss determined by the algorithm's loss function. During each iteration, the current ensemble is used to predict the class for each instance in the training data. The predicted values are evaluated with the actual values allowing the algorithm to pinpoint and correct previously mislabeled instances. The parameter that caches node IDs for each instance, was set to true and the maximum memory parameter was set to 1024 MB to minimize training time.

3.3 PERFORMANCE METRICS

This section provides definitions for the performance metrics used for both anomaly detection and traditional fraud detection. These metrics are used to evaluate our models and determine the level and significance of our fraudulent detection efforts using Medicare claims data.

3.3.1 Anomaly Detection

For Sections 4.1 and 4.2, the calculations for model performance are derived from the resulting confusion matrices, from a single run of 5-fold cross-validation, where we took the average over each fold. We use the one-versus-all approach for calculating the error rates, which considers the class in question as the positive class then considers the rest of the classes as being in the same negative class. True positive, true negative, false positive and false negative (*tp*, *tn*, *fp* and *fn* in the subsequent equations) are described in Table 3.4. We determined, through preliminary experimentation, that Precision, Recall and F-score are the most suitable metrics, because the dataset contains numerous classes where a metric such as Accuracy would not be useful due to the data sparseness and multi-class imbalances.

For Section 4.3, in order to select the best baseline model and office or facility method, we use Precision, Recall, F-score, G-measure, and overall Accuracy as metrics for performance evaluation. We also use the so-called Inverse Overall Accuracy (IOA), for each specialty, and the overall weighted average IOA (owaIOA) to assess performance with the best models having the highest IOA per specialty. For these experiments, we are interested in the incorrectly classified (as we consider these fraud) which is given by the IOA and owaIOA, which is normalized by the number of fraudulent instances for each specialty. Additionally, for the class isolation method, we use Type I (false positive) and Type II (false negative) error rates to assess the predictive capabilities of the models for detecting Medicare Part B fraud.

Recall

Recall measures the ability of a classifier to determine the rate of positively marked instances that are in fact positive; therefore, for this dataset, Recall is showing the proportion that a given physician is labeled correctly.

$$Recall = \frac{tp}{tp + fn} \tag{3.1}$$

Precision

Precision indicates how well a classifier has predicted a class by finding the ratio of actually positive instances from the pool of instances that it has marked as part of the positive class; therefore Precision here is showing the proportion that a given physician is marked correctly against the amount of physicians, from the other class(es), marked also as the class in question.

$$Precision = \frac{tp}{tp + fp} \tag{3.2}$$

F-score

F-Score (also known as F1-Score or F-measure) is the harmonic mean of both Precision and Recall, and is used to organize the model performance results into one concise metric for performance comparisons, generating a number between 0 and 1, where values closer to one indicate better performance. F-Score is reasonably robust to imbalanced data and we are primarily interested in the prediction of true positives (i.e. the prediction of actual fraudulent behaviors). F-Score is used as a gauge to assess any performance improvements due to our proposed improvement strategies. For our experiments, we assume equal weighting between Precision and Recall, with $\beta = 1$, as seen in Equation 3.3.

$$F_1 = (1 + \beta^2) \times \frac{Recall \times Precision}{(\beta^2 \times Recall) + Precision} \tag{3.3}$$

G-measure

G-measure, also known as the Fowlkes-Mallows index, gives the geometric mean of Precision and Recall giving the central point between the values as seen in Equation 3.4.

$$\text{G-measure} = \sqrt{Recall \times Precision} \tag{3.4}$$

Type I and II Error Rates

Type I error rate (false positive rate) is the percentage of instances that are actually non-fraud but marked as fraud, in relation to the number of actual non-fraud instances. A fire alarm going off indicating a fire when in fact there is no fire would be an example of this kind of error. Type II error rate (false negative rate) is the percentage of instances that are actually fraud but marked as non-fraud, in relation to the actual number of actual fraud instances. As an example, a fire breaking out and the fire alarm does not ring would be considered a false negative. Note that in binary classification, finding a balance between the error rates, while minimizing the Type II error rate, is generally preferred. Recall measures the ability of a classifier to determine the rate of positively marked instances that are in fact positive; therefore, in our study, Recall is the fraction of physicians labeled correctly and not as any of the other specialties. Precision indicates how well a classifier has predicted a class by finding the ratio of actually positive instances from the pool of instances that it has marked as part of the positive class; therefore, Precision shows the fraction of physicians marked correctly against the number of physicians, from any of the other specialties, also marked as the class in question.

Inverse Overall Accuracy

The Inverse Overall Accuracy is specific to Section 4.3 and to use this metric we manipulated the datasets, by filtering out certain specialties only, so that we could test one fraudulent specialty at a time (based on the model predicting the physician's specialty). As mentioned in Section 2.3.1, the mapped exclusion datasets contain only one specialty at a time, and the overall Accuracy would be the percentage of real-world fraudulent physicians that are considered not fraudulent. In order to capture the the percentages of classes that are labeled as another class, we incorporate the Inverse Overall Accuracy (IOA) performance measure. IOA, where IOA $= 1 -$ overall Accuracy, is the percentage of fraudulent physicians marked as fraudulent for a given specialty. As shown in Equation 3.5, to calculate the model's overall weighted average IOA (owaIOA), we take the IOA for a specialty, the number of fraudulent instances (n) for that specialty and the total number of instances (N), and sum over the total number of fraudulent instances between all specialties with F-score of 0.75 or above (NoS).

$$owaIOA = \sum_{i=1}^{i=NoS} \frac{n_i}{N} \times IOA_i \tag{3.5}$$

3.3.2 Traditional Fraud Detection

For the traditional fraud detection experiments we employ a single, encompassing performance metric: Area Under the ROC Curve (AUC) [16]. In order to provide more rigor to our AUC results, we also perform ANalysis Of VAriance (ANOVA) [55] and Tukey's HSD tests [137].

Area Under the Receiver Operating Characteristic Curve

We use the AUC as our evaluation metric for traditional fraud detection experiments to demonstrate the fraud detection capabilities of each model. AUC has been found to be an effective metric for quantifying results for studies employing datasets with class imbalance [77]. AUC illustrates performance over all decision thresholds as a single value ranging from 0 to 1, where a perfect classifier will receive an AUC of 1, while an AUC of 0.5 is equivalent to random guessing and less than 0.5 can indicate bias towards a given class. The ROC curve illustrates a learner's capability to discriminate between both classes and is the comparison between false positive (fall-out or 1-specificity) and true positive (Recall) rates, where fall-out is calculated by $\frac{FP}{FP+TN}$ and Recall is calculated as described in equation 3.1. True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) are described in Table 3.4.

Table 3.4: Confusion Matrix Calculations

Element	How to Calculate
True Positive	Number of actual positive instances correctly predicted as positive
True Negative	Number of actual negative instances correctly predicted as negative
False Positive	Number of negative instances incorrectly classified as positive
False Negative	Number of positive instances incorrectly assigned as negative

Significance testing

We perform hypothesis testing to demonstrate the statistical significance around our AUC results through ANalysis Of VAriance (ANOVA) [55] and Tukey's HSD tests [137]. ANOVA is a statistical test determining whether the means of several groups (or factors) are equal. Tukey's HSD test determines factor means that are significantly different from each other. This test compares all possible pairs of means using a method similar to a t-test, where statistically significant differences are grouped by assigning different letter combinations (e.g. group 'a' is significantly better than group 'b' in correlation to the issue).

CHAPTER 4
DETECTING FRAUD THROUGH ANOMALY DETECTION

4.1 PREDICTING MEDICAL PROVIDER SPECIALTIES TO DETECT ANOMALOUS INSURANCE CLAIMS

The human body is a complex system; therefore, it is necessary to have specialized physicians trained to diagnose and treat diseases in different parts of the body. This leads to different types of treatment plans and procedures that doctors perform for patients in various fields. The goal of a healthcare system is to effectively treat as many patients as possible, but there is cost associated with every treatment on various levels. Physicians, drug manufacturers, and medical staff must be paid for their time and expertise, in addition to various medical equipment and facilities. As these costs are frequently much more than an individual patient can afford, insurance plans are used to distribute costs across all patients in the network and pay for the necessary personnel and equipment. As with any insurance system, there is potential for misuse or fraudulent activities. The Federal Bureau of Investigation estimate that fraud accounts for 3-10% of all billings [104]. Clearly, healthcare fraud continues to be a major problem for the government and taxpayers, with a need for more effective detection methods. Given this need, the detection of potential fraudulent behaviors, such as physician referrals or upcoding [15], is the purpose of our research. We attempt to identify such activity within the healthcare system by examining claims for medical procedures performed by healthcare providers.

One such insurance provider is Medicare, the government-run organization that provides health insurance for seniors (as well other select groups). CMS, in response to a new policy declared by the U.S. Department of Health and Human Services [73], recently started releasing datasets in an attempt to assist in identifying fraud, waste, and abuse within Medicare [49]. One such dataset outlines all procedure claims made by each health provider in the U.S., and the average amount paid for these services, among other data points.

Malicious or wasteful use of any medical financial system makes healthcare inefficient, potentially leaving patients without the treatment they need. Ko et al. [87] estimate a 9% savings just in the field of Urology (over $125 million) by regulating the use of Medicare. In order to both set and control regulations, there needs to be a process to determine when regulations are broken. Anomaly detection, as part of a machine learning process, can determine when certain physicians do not practice in a similar manner as their peers [122]. While this model would not be able to guarantee that a physician is practicing maliciously, it can help insurance systems flag outliers that would require further investigation. To achieve this, we explore the use of a machine learning model to predict the specialty of a physician, based solely on the procedures that he or she performs, as depicted in Figure 3.1. If the predicted specialty matches the provider's actual specialty, then the assumption is the provider is practicing within the norm of his or her field. If not, the provider could have very unique patients or could be sending wasteful or malicious claims. For the latter case, these providers exhibit aberrant, and possibly fraudulent, behaviors, warranting additional scrutiny into their practicing habits. To the best of our knowledge, we are the first to propose such a method.

There are a large number of fields in modern healthcare, resulting in a multi-class classification problem. In this section, the terms "field of expertise", "provider type", "specialty", and "class" will be used interchangeably. For prediction using multiple classes, we build a MNB classifier evaluated using 5-fold cross-validation and 3 performance metrics: Precision, Recall and F-score (all averaged over the 5 folds). The inputs to the model are physician specialties and the number of times each provider performs a particular procedure.

Our work demonstrates that machine learning can be used to indicate possible fraudulent behaviors, and that further research may provide better models as well as a better understanding of provider practices [14]. This is clearly indicated by our results, as there are a satisfactory number of specialties that had high or mediocre prediction results. Currently, only the classes that are predicted well can be used as part of an anomaly detection framework; therefore, the providers in these classes can be adequately flagged for further investigations. The classes with mediocre, or bad, results open up opportunities for future research, such as comparing various learners, using feature selection methods, and adding different types of data. This research focuses on exploring the efficacy of our proposed approach in detecting potentially anomalous providers. Due to the novelty of this research problem, comparisons to other studies are not currently available.

The rest of this section is organized as follows. Section 4.1.1 discusses works related to the current research in this domain. In Section 4.1.2, the experimental methods used are detailed to include the dataset, learner and performance metrics. Section 4.1.3 presents the results of our experiment. Finally, Section 4.1.4 outlines our conclusions and ideas for future work.

4.1.1 Related Works

The data that the Centers for Medicare and Medicaid Services (CMS) has released, at the point of this experiment was conducted, is only for 2012 and 2013 and were released in 2014 and 2015, respectively. Therefore, all research done using this data, up to that point, was in the preliminary stages with additional future work needed for finding misuse in medical insurance utilization. One such research effort, which uses the 2012 data, looked into how a given physician's past schooling determines the way he or she practices [49]. They compare medical school charges, procedures, and payments as well as look to find possible anomalies in the data by presenting a geographical analysis with the national distribution of school procedure payments and charges. By following this line of research, the authors attempt to find correlations between educational backgrounds and the practices and procedures physicians perform to help pinpoint those physicians who are misusing or inefficiently using medical insurance systems.

Another study that used the 2012 CMS data specifically looked at one field: Urology [87]. The authors analyze variability among Urologists within the field's service utilization and payment and determine an estimated savings from a standardized service utilization. They found that there is a strong correlation between the number of patient visits and reimbursement from Medicare. They also found, in terms of services per visit, that there was a high utilization variability and a possible 9% savings within the field of Urology. This research can lead to finding rules for service utilization.

Though not only using CMS data, a general coverage paper by Chandola et al. [21] assesses healthcare fraud using data with labels for fraudulent providers, primarily from the Texas Office of Inspector General's exclusion database. The authors employ several techniques including social network analysis, text mining, and temporal analysis in order to translate the problem of healthcare data analysis into some well-known data mining methods. More specifically, Chandola et al. [21] discuss the use of typical treatment profiles, i.e. procedures performed, in order to compare among providers and spot possible misuses or abuses in procedures to treat particular ailments.

4.1.2 Methodology

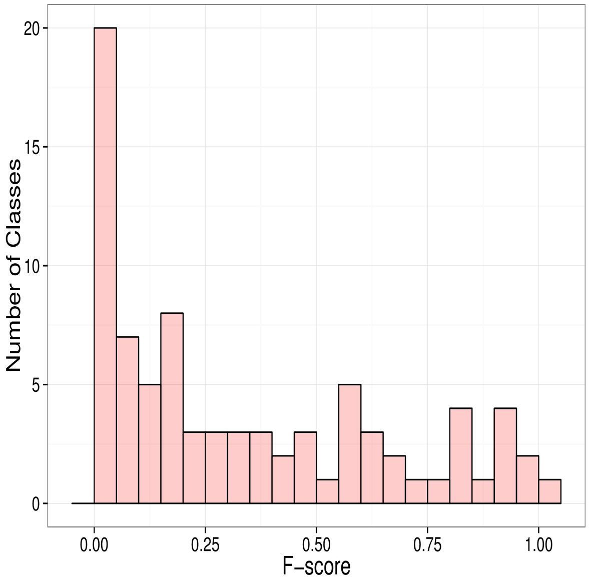
This subsection discusses the dataset, learner and metrics used in this experiment. Section 3.1.1 discusses the experimental design for our proposed method. Figure

3.1 explains the workflow of our proposed model, where the CMS procedure data is leveraged to create a model which takes the physician in question and determines whether they fit into the norm of their respective field. For this section, we decided to use the Part B dataset for the 2013 calendar year, found at CMS.gov [27], to determine whether or not it is possible to predict a physician's field of expertise based on the procedures they perform. Our research could lead to assisting other researchers, and eventually the government, in finding discrepancies in the everyday dealings of physicians who are abusing the system, committing insurance fraud or wasting funds through the detection and flagging of outliers by finding physicians that do not fit into the norm of their respective field. This line of research can help determine which doctors should be investigated and assist insurance systems (such as Medicare) with setting up rules and regulations for physicians to run more cost-effective practices free of abuse and mistreatment.

The 2013 Part B dataset represents 5,983 distinct types of procedures done by 909,606 physicians throughout the United States. Due to the large size of the dataset, we decided to only use data from office clinics in Florida (as opposed to larger facilities, such as hospitals and academic institutions). This resulted in a subset of the dataset composed of 82 classes of physicians, 40,940 instances (individual physicians) and 2,789 distinct procedures. Due to Florida's unique demographic, the use of this subset is not necessarily representative of the entire US population. Even so, Florida is a good candidate for testing our method for several reasons to include having the second highest number of Medicare beneficiaries and being second in total Medicare spending [63]. Table 2.7 illustrates the size of the original CMS dataset in comparison to our subset, in terms of classes, instances, and procedures, in order to demonstrate the proportion of the original used to validate this work. We provide a description of our unique data processing and engineering approach we applied to the Part B data in Section 2.3.1. Table 2.1 shows a small sample of the the original CMS data and for comparison, Table 2.6 shows a sample of the dataset after it was converted to the sparse vector (NPIs are masked).

For our multi-class classification experiment, we use the MNB classifier. This model is very effective and is used extensively in the field of data mining and machine learning [48, 150]. We used the WEKA [60] machine learning toolkit to perform the experiment, with 5-fold cross-validation for model evaluation. In order to evaluate model performance, we employ the following metrics: Recall, Precision, and F-score. In the calculation of these metrics, we took the average over each fold of the 5-fold cross-validation for each metric. We use the one-vs-all approach for calculating the error rates, which considers the class in question as the positive class then considers the rest of the classes as being in the same negative class.

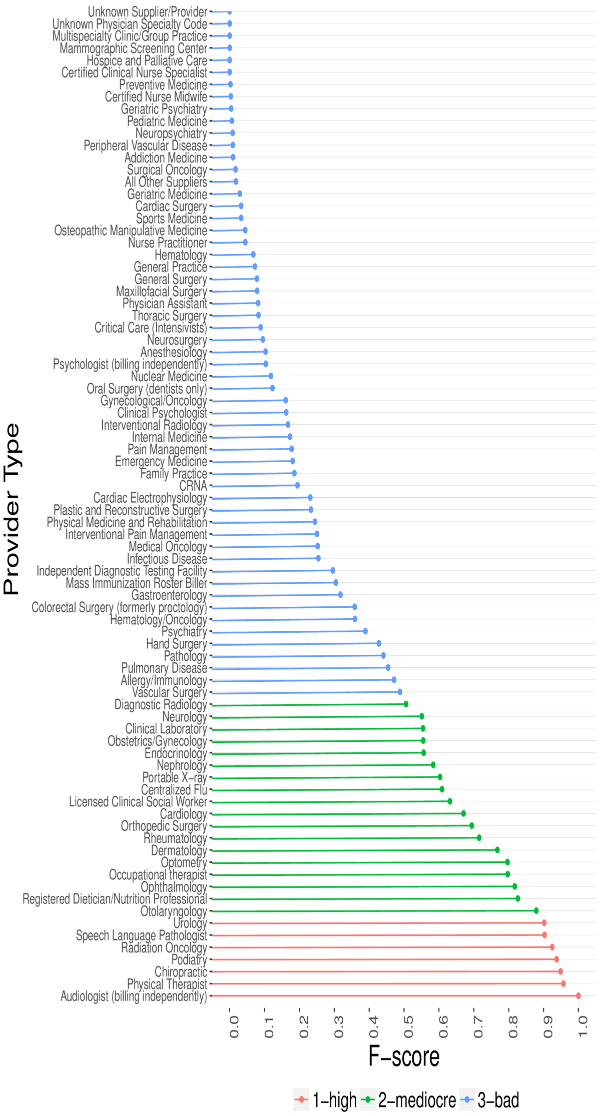
Figure 4.1: Histogram of F-Scores



4.1.3 Results and Discussions

Figure 4.1 presents the distribution of F-score values throughout the dataset indicating visible groups that classes of physicians fall into, informing the categories with which our results are organized and discussed. Figure 4.2 shows the model F-score

Figure 4.2: F-Score value per Provider Type



values per provider type, indicating the chosen groupings by F-score. In order to find physicians that are misusing insurance, we need to determine the conditions that achieve the best possible prediction for each field of expertise. Along with the performance metrics outlined above, two other explanatory values are shown for each provider type, which are the number of instances and number of codes.

The results are organized in three groups, as they have similar needs, based on Figure 4.1 and Figure 4.2: the group that scored very high in terms of F-score (above

0.90), the group that scored mediocre but needs work (between 0.5 and 0.90) and the group that had bad results (between 0.0 and 0.50). There are 7 classes that scored very high, 18 that scored mediocre and 57 that scored unfavorable. In an attempt to not clutter the results with 82 classes, the tables below will only display a sample of the classes within each breakdown, with the exception of the high scoring partition.

The fields that have F-scores over 90%, shown in Table 4.1, do not need much more focused work as multinomial Naïve Bayes is a simple learner, compounded with the fact that there are numerous instances showing that most likely any method of prediction, under most circumstances, will garner good results. Even with this simple prediction model, these classes are, for the most part, ready to be used in determining whether or not a particular physician is exhibiting anomalous behavior.

Table 4.1: Results for .90+ F-score

Provider Type	# of Instances	# of Codes	Recall	Precision	**F-score**
Audiologist (billing independently)	306	29	1.00	1.00	**1.00**
Physical Therapist	1525	41	0.93	0.99	**0.96**
Chiropractic	2010	3	1.00	0.91	**0.95**
Podiatry	1132	228	0.92	0.96	**0.94**
Radiation Oncology	224	202	0.92	0.94	**0.92**
Speech Language Pathologist	19	6	0.83	1.00	**0.90**
Urology	606	298	0.87	0.94	**0.90**

Table 4.2: Count of 6 unique procedures by provider type

Provider Type	# of Times Procedures Performed
Speech Language Pathologist	13801
Otolaryngology	3943
Neurology	1325
Internal Medicine	942
General Surgery	300
Nurse Practitioner	277
Diagnostic Radiology	220
Gastroenterology	109
Family Practice	94
Allergy/Immunology	47

Most of the classes within this partition have an adequate number of instances (more than 100) but there is one standout, Speech Language Pathologist, that has a very small number of instances (19). Speech Language Pathologists use 6 unique procedures codes (92506, 92507, 92526, 92610, 92611, 92612), but these same procedures overlap with 9 other specialties. As seen in Table 4.2, Speech Language Pathologists performed about 50% of the procedures, across these 6 unique procedure codes, with the other 9 providers accounting for the rest, indicating overlap in the procedures performed.

Even so, the F-score is still quite high indicating that the model picked up on some pattern in the procedure distribution resulting in this high value. Interestingly, both Speech Language Pathology and Otolaryngology have similar, high F-scores and also have similar procedure distributions. Future work will involve further testing on a dataset containing more instances to confirm these results and/or produce additional distinctions between these two specialties.

It seems intuitive that in order to receive a high F-score, especially under these conditions, there needs to be little or no overlap between procedures done in their field compared to procedures done by any of the other 81 fields. For example, we take Chiropractic (having an F-score of 0.95), which only has 3 codes (98940, 98941, 98942), all referring to variations of Chiropractic manipulative treatment. We performed further testing and found that only five doctors, other than Chiropractors, billed one of these codes from only two other physician fields. These two fields, Physical Therapist and Physical Medicine and Rehabilitation, indicate overlapping procedures but account for less than 0.5% of all procedures performed; therefore, the Chiropractic specialty has minimal overlapping procedures and high model classification performance, as indicated by the F-score.

We also looked at Family Practice which had a low F-score of 0.17 (shown in Table 4.3) and 651 procedures, where many of these procedures are shared with various other classes. Therefore, it would appear that getting good results, when using codes to predict a field of expertise, depends upon the number of other similar classes that use the procedures in that field. Similar classes would be any class that shares at least one procedure. In the future, more classes will need to be tested to further support this theory and methods should be implemented to appropriately handle overlapping procedures to garner better overall model performance.

Table 4.3 shows the 18 classes with mediocre results, as previously defined. The classes that fall into this category are of the most interest for future research, because they have the best chance of being improved through various data mining techniques and the general number of instances are high. Even with the simple multinomial Naïve Bayes learner and the inclusion of the other classes' procedures in the dataset, the model still produced a result better than 0.50. Only Portable X-ray and Occupational Therapist have a small number of instances; therefore, the results from these fields could benefit more complex learners and/or additional data. There are a few fields that have a large difference in their Recall and Precision (leading to a low F-score), but a high number of instances. These fields could use some extra attention such as Cardiology, Licensed Social Worker, Centralized Flu, and Diagnostic Radiology. These can benefit from further research using techniques that will increase classifier performance, such as feature selection or ensemble techniques, or future methods to address procedure overlap.

Table 4.3: Sample of results .50 - .90 F-score

Provider Type	# of Instances	# of Codes	Recall	Precision	**F-score**
Otolaryngology	475	2453	0.84	0.92	**0.88**
Registered Dietician/Nutrition Professional	69	6	1.00	0.80	**0.83**
Opthamology	1138	227	0.75	0.90	**0.82**
Occupational Therapist	211	37	0.96	0.69	**0.80**
Optometry	1364	74	0.81	0.78	**0.80**
Dermatology	864	276	0.73	0.81	**0.77**
Rheumatology	267	315	0.65	0.80	**0.72**
Orthopedic Surgery	1182	394	0.59	0.84	**0.69**
Cardiology	1528	525	0.53	0.90	**0.67**
Licensed Clinical Social Worker	560	8	0.98	0.47	**0.63**
Centralized Flu	713	12	0.90	0.50	**0.61**
Portable X-ray	27	79	0.81	0.50	**0.60**
Nephrology	447	255	0.67	0.52	**0.58**
Endocrinology	284	283	0.55	0.58	**0.56**
Obstetrics/ Gynecology	1283	285	0.45	0.73	**0.56**
Clinical Laboratory	174	773	0.42	0.86	**0.55**
Neurology	748	339	0.42	0.81	**0.55**
Diagnostic Radiology	950	499	0.36	0.86	**0.51**

Table 4.4 shows a sample of the fields (48/57) that did not have good results under these basic conditions but could still possibly benefit from testing other learners as well as applying various other methodologies in order to improve the results [43]. The results with F-scores between 0.25 and 0.50 contain only three classes (Colorectal Surgery, Hand Surgery and Medical Oncology) with a small number of instances; thus,

the results for that range are fairly dependable for each class. As with the previous group, there are a few classes that have a high Precision or Recall but low F-score values for classes, such as Allergy/Immunology, Hand Surgery, Hematology/Oncology and Mass Immunization Roster Biller. For these fields, classifiers have a greater chance of being improved over their counterparts with evenly low Precision and Recall.

In general, the classes that received an F-score below 0.25 have a low number of instances making it indiscernible whether their low results were due to the lack of representation in the dataset or difficulty with distinguishing each class. There are four standouts in this groups that have a quite large number of instances: Family Practice, Internal Medicine, Physician Assistant and Nurse Practitioner which, on their own, make up a large percent of the dataset. Since these four classes have so many instances and each have over 500 performed procedures but low F-score results, it would lead one to believe that there are many overlapping procedure codes between other classes. This is inferring that physicians in Internal Medicine, for example, perform a lot of the same procedures done by the other 81 fields. Several other fields have an adequate amount of instances within this group, such as Clinical Psychologist and General Surgery, but still exhibit overlapping procedures amongst providers. None of the classes that scored an F-score of 0.0 are shown here as all of them have very few instances.

All of the classes that had a low number of instances within the dataset, whether they received a high F-score or not, need further study using a dataset that contains more instances of these said classes, either by using a different region (state), more regions, or the entire original dataset. For the classes with a reasonable or large number of instances, future work should look to data mining techniques and methodologies such as testing the dataset with different learners, adding other data, determining which procedures to remove from each class (feature selection), overlap handling, or any other various data mining techniques. There is great importance in determining which procedures to keep and which to remove for each class (compared to the binary inclusion method as discussed in Section 2.3.1) as we found a very small percentage of codes that were unique to a given class. Our method reveals that providers with unique procedure distributions and/or codes can be classified successfully by using only the number of procedures performed. As seen, overlapping procedures, though, inhibit the successful classification by blurring the distinct patterns in the procedures performed amongst some provider types, making it difficult for a machine learning algorithm to adequately classify a provider.

Table 4.4: Sample of < .50 Average F-score

Provider Type	# of Instances	# of Codes	Recall	Precision	**F-score**
Vascular Surgery	155	180	0.56	0.44	**0.49**
Allergy/ Immunology	188	169	0.84	0.33	**0.47**
Pulmonary Disease	522	297	0.44	0.48	**0.45**
Pathology	235	233	0.56	0.37	**0.44**
Hand Surgery	84	127	0.77	0.30	**0.43**
Psychiatry	743	121	0.39	0.40	**0.39**
Hematology/ Oncology	428	301	0.24	0.74	**0.36**
Colorectal Surgery (formerly proctology)	89	48	0.77	0.24	**0.36**
Gastroenterology	843	206	0.33	0.31	**0.32**
Mass Immunization Roster Biller	1915	15	0.26	0.85	**0.30**
Independent Diagnostic Testing Facility	273	283	0.34	0.26	**0.30**
Infectious Disease	284	132	0.31	0.23	**0.25**
Medical Oncology	96	202	0.48	0.17	**0.25**
Interventional Pain Management	209	318	0.19	0.37	**0.25**
Physical Medicine and Rehabilitation	293	324	0.19	0.35	**0.24**
Plastic and Reconstructive Surgery	276	221	0.21	0.29	**0.23**
Cardiac Electrophysiology	85	97	0.73	0.14	**0.23**
CRNA	34	21	0.53	0.18	**0.19**
Family Practice	3806	790	0.12	0.45	**0.19**
Emergency Medicine	270	416	0.19	0.18	**0.18**
Pain Management	103	225	0.30	0.13	**0.18**
Internal Medicine	4216	961	0.10	0.61	**0.17**
Interventional Radiology	41	222	0.37	0.11	**0.17**
Clinical Psychologist	733	20	0.09	0.84	**0.16**
Gynecological/ Oncology	58	98	0.22	0.14	**0.16**
Oral Surgery (dentist only)	84	44	0.23	0.09	**0.12**
Nuclear Medicine	16	136	0.30	0.07	**0.12**
Psychologist (billing independently)	16	15	0.13	0.09	**0.10**
Anesthesiology	197	309	0.07	0.23	**0.10**
Neurosurgery	243	130	0.08	0.12	**0.10**
Critical Care (Intensivists)	60	77	0.38	0.05	**0.09**
Thoracic Surgery	103	41	0.12	0.07	**0.08**
Physician Assistant	1479	555	0.05	0.29	**0.08**
Maxillofacial Surgery	49	20	0.53	0.04	**0.08**
General Surgery	845	386	0.04	0.31	**0.08**
General Practice	615	464	0.07	0.08	**0.07**
Hematology	49	94	0.23	0.04	**0.07**
Nurse Practitioner	2462	589	0.02	0.36	**0.04**
Osteopathic Manipulative Medicine	96	113	0.25	0.02	**0.04**
Sports Medicine	34	57	0.27	0.02	**0.03**
Cardiac Surgery	58	124	0.08	0.02	**0.05**
Geriatric Medicine	89	86	0.32	0.02	**0.03**
All Other Suppliers	25	12	0.27	0.01	**0.02**
Surgical Oncology	25	22	0.52	0.01	**0.02**
Addiction Medicine	41	12	0.20	0.01	**0.01**
Peripheral Vascular Disease	60	60	0.20	0.00	**0.01**
Neuropsychiatry	58	44	0.10	0.00	**0.01**
Pediatric Medicine	57	185	0.03	0.00	**0.01**

As noted previously, one pattern worth mentioning is when there is a relatively large difference in a given classes' Recall and Precision. The Recall is the higher value when the class has a low number of instances and Precision is the higher value when the class has a large number of instances. These groups show the most promise for favorable results in the future and one technique that might help with this could be to create a similar dataset that balances the number of instances accounting for some of the data imbalance.

4.1.4 Section Summary

The misuse of medical insurance, whether malicious or not, can lead to many undesirable outcomes such as patients not getting the funding they need or physicians not getting reimbursed for their time. This is unacceptable for the field of healthcare and there needs to be misuse and fraud-detection rules that are actionable. The purpose of this section is to effectively use machine learning to determine whether or not using procedure data could accurately predict a physicians' field of practice. This research explores the possibility of creating a machine learning model for assessing fraudulent provider behaviors based on their medical procedure history. The process, as seen in Figure 3.1, provides validation that a provider is performing normally within their specialty, or indicates possibly aberrant medical practices when the model classifies that provider into another specialty for which they do not practice. The results show that there is certainly promise for this research, as results were good for several specialties, even with using a relatively simple learner and a dataset with a large number of classes.

For the group of results with an F-score of 0.90 or above, it would appear that they most likely will have good results in any situation. Of course, additional testing will be performed to further solidify this claim as well as find the optimal conditions for aberrant procedure classification. Even so, given these positive results, we recommend using our model to predict and flag possible fraudulent practices for the providers found in Table 4.1. For example, the procedures (submitted claims) from a local Physical Therapist can be input into our model to detect any instances of anomalous behavior, that were not classified as Physical Therapist, for further investigation. These in-depth inquiries into the flagged provider's procedure practices can reveal possible abuse or fraudulent activities.

The classes with lower than 0.90 F-score need additional research work. One of the noted reasons for lower F-scores, across several provider types, is the overlap in the procedures performed. Even with overlapping procedures, as discussed in Section 4.1.3, the model does not always produce low F-scores, but this overlap could make it difficult to detect fraudulent behaviors across providers with very similar procedure profiles. Additionally, classes with a low number of instances, with the majority having low F-scores, make determining their results unverifiable. In order to alleviate this situation, tests need to be performed on a dataset with more instances of these classes, which can be done using a larger subset or the full dataset released by CMS. Furthermore, for classes in the second partition (0.50-0.90), which generally have a high number of instances per provider, more sophisticated machine learning techniques, such as feature selection or clustering [95], could be used to improve classification results.

Throughout this section, we have discussed possible avenues for future work. With regards to our fraud detection method, given the novelty of our research, there is still work yet to be done before all physicians can be accurately classified into their respective fields in a reliable manner. This research, along with other research done using publicly-available Medicare datasets [26], can lead to a more cost effective healthcare system by detecting and flagging potentially fraudulent behaviors exhibited by healthcare providers. The hope for this line of research is to develop a generalizable model that finds physicians that work outside the norm of their fields through the successful application of anomaly detection methods. One noted potential threat to this line of research is the fact that many different types of physicians can perform the same procedures. Future models will account for this procedure overlap, as this does not represent an anomaly or fraud, merely a physician practicing everyday medicine. As no standard dataset for insurance fraud exists, there needs to be a set of rules and regulations that establish a baseline behavior for physicians in terms of insurance utilization for each specialty. We believe that by continuing the research started herein, we can help determine these baselines for providers through the detection of anomalous medical procedures. Therefore, the next step is to incorporate the discussed future work and find ways of improving upon the results shown by our model to better detect fraudulent provider behaviors.

4.2 MEDICAL PROVIDER SPECIALTY PREDICTIONS FOR THE DETECTION OF ANOMALOUS MEDICARE INSURANCE CLAIMS

As both healthcare costs and the average life spans continue to increase, more financial tension is applied to public and private institutions which assist in funding doctor visits and treatments [52, 56]. Therefore, the goal of the healthcare system should be offering sufficient and necessary treatment to as many patients as possible, at fair cost to both patients and providers. One such healthcare system is Medicare, which is a government-run program that specifically provides financial support to seniors (and other select groups) [38]. Similar to other insurance-related programs each covered procedure is submitted with a specific procedure code. In general, a physician performs a series of procedures and then submits a claim to Medicare for payment. We do not detail Medicare or the Medicare claims process in this section, but the interested reader can find additional information in [26, 32].

Insurance systems, such as Medicare, are set in place to to keep medical costs reasonable, as they are generally not affordable for the average patient. One way to help keep costs more sensible is to curtail fraud, waste, and abuse (FWA) in medical practices and claims. Malicious or wasteful use of any medical financial system makes healthcare inefficient, potentially leaving patients without the treatment they need. FWA and other unwanted behaviors can be discovered using anomaly detection methods, allowing more patients to get treatment at lower and more reasonable costs. The methods and improvements proposed in this section could be used to help in not only monitoring Medicare claims data, but in guiding regulations attempting to pinpoint fraudulent behavior, such as by requiring a physician to act similar to their peers or have a justifiable reason for the departure in practice.

In this section, we use the 2013 Part B data released by CMS [27, 26]. In response to a new policy declared by the U.S. Department of Health and Human Services [73], CMS has begun releasing datasets in an attempt to assist in identifying fraud, waste, and abuse within Medicare [49]. One such dataset outlines every Medicare claim made by healthcare providers throughout the U.S., the average amount paid for these services and several other data points related to the specific procedures. In Section 4.1, we built an anomaly detection model to flag outliers using only the physician's procedures performed found in the Medicare claims data. Our model detects possibly fraudulent behavior by predicting a provider's specialty based on the number of procedures performed. If the physician is predicted into a different medical field (e.g. Chiropractor classified as a Rheumatologist), then the physician in question either has mostly unique patients or is exhibiting possibly fraudulent behaviors. In this section, we do not consider different types of healthcare fraud, such as upcoding or self-referrals, but the reader is referred to [15] for additional information. We use the 2013 and 2014 Medicare Part B claims data for testing and validation, and subset the data to include only the Florida-based claims performed within an office. Additionally, we use data released by the OIG from the LEIE [91] data repository, which includes all current physicians who have been found unfit to practice thus excluded from practicing in the U.S. for a given period of time. We employ the LEIE information to identify fraudulent physicians within the Medicare dataset for model testing and validation. In addition to the LEIE database, we found two documented fraud cases: Michael Burgos, who was indicted on charges of Medicare fraud constituting $13.8 million [44] and Salomon Melgen, an eye doctor who was charged with scamming Medicare out of $105 million [113].

Our contribution includes two related areas. We first expand upon our previous work from Section 4.1 [14], also referred to as the original model, by testing our model against a dataset of real-world fraudulent physicians. The model performance is determined by how many of physicians are classified into a specialty other than their actual practicing specialty and compared to those known fraudulent physicians found in the LEIE database. We then use our proposed strategies for improving the original model's performance. The first strategy employs feature selection and sampling to account for class imbalance. The second strategy is the removal of a selection of low scoring medical fields with a large number of instances (presumably caused by procedural overlap) in order to help boost model performance. The final strategy is grouping of similar medical specialties that have a relatively large overlapping of procedures, as confirmed through the confusion matrix. Results are shown using the 2013 Medicare data for comparative purposes, with the 2014 data used to validate those results. Note the 2014 data is not used as a test set, per say, but as validation of the 2013 model results to confirm the results are similar (for both 2013 and 2014). We employ the MNB classifier evaluated using 5-fold cross-validation and three performance metrics. The only attribute used for each model is the number of times each provider performs a particular procedure.

Out of the 18 physicians found to be fraudulent, 12 were correctly detected using our model with a 67% detection rate [68]. Our findings indicate that the performance of our strategies varies depending on the medical specialty. For instance, grouping worked well for Ophthalmology, while removing classes worked well for Cardiology. Our experiments suggest that improvements can be made by using one or more of our proposed strategies. Specifically, we found that different techniques provide different results across the specialties, possibly depending on the characteristics of a given specialty or the extent of overlapping procedures.

The rest of this section is organized as follows. Section 4.2.1 discusses works related to the current research in this domain. Section 4.2.2 details the experimental methods used, including the dataset, learner and performance metrics. Section 4.2.3 presents the results of our experiment. Finally, Section 4.2.4 outlines our conclusions and ideas for future work.

4.2.1 Related Works

The data released by the Centers for Medicare and Medicaid Services, at the point this experiment was conducted, was for 2012, 2013 and 2014. Therefore, all research done using this data was still relatively new, with additional future work needed for finding misuse in medical insurance utilization. One such effort by Feldman et al. [49] looked into how a given physician's past schooling determines the way he or she practices, from the 2012 Medicare data. The authors compared medical school charges, procedures, and payments as well as looked to find possible anomalies in the data by presenting a geographical analysis with the national distribution of school procedure payments and charges. The authors attempted to find correlations between educational backgrounds and the practices and procedures physicians perform to help pinpoint those physicians who are misusing or inefficiently using medical insurance systems.

Another study by Ko et al. [87] used the 2012 data for the urology specialty only. The authors analyzed the variability among urologists within the field's service utilization and payment, and determined an estimated savings from a standardized service utilization. They found that the number of patient visits had a strong correlation with reimbursement from Medicare. They also found, in terms of services per visit, there was high utilization variability and a possible 9% savings within the field of urology. This research could culminate in finding rules for better service utilization.

Though CMS was not the only data used, a general coverage paper by Chandola et al. [21] assessed healthcare fraud using data with labels for fraudulent providers, primarily from the Texas Office of Inspector General's exclusion database. The authors employed several techniques including social network analysis, text mining, and temporal analysis in order to translate the problem of healthcare data analysis into some well-known data mining methods. More specifically, the authors discussed the use of typical treatment profiles, i.e. procedures performed, in order to compare among providers and spot possible issues or abuses in procedures to treat particular ailments.

There are additional, related studies utilizing data from sources other than the CMS data. Pawar et al. [115] wrote a survey giving a small overview of fraud detection using publicly available data, specifically through the use of geo-location clustering along with various other clustering algorithms. Joudaki et al. [79] collected claims data, and other information, for general physicians specifically in the area of drug prescription. They attempted to find indicators of fraud within the healthcare system using Iranian-based data. They employed a total of thirteen indicators in determining abuse and fraud in physician behavior using clustering and discriminate analysis with satisfactory results. In the study [139], Van et al. used Medicaid data for 369 dentists with a total of 650,000 claims and similar to our work employed outlier detection in order to determine fraudulent physicians. In their work they used unsupervised techniques with 14 metrics (which they decided after review of additional sources such as FBI reports) and determined that 17 out of the 369 dentists should be investigated for fraud. These results were then discussed with professionals within the field of dentistry and 12 out of these 17 physicians were determined worthy of further inspection giving them a 71% detection rate.

In Section 4.1 [14], we investigated whether or not it is possible to predict a physician's field of expertise based on only the procedures they perform. The idea was that if we could perform this prediction accurately (determined by F-Score), then we could find potentially anomalous behavior in a particular physician's procedures (compared to the norm of other physicians in their field). If they are anomalous then they are more likely to be fraudulent, wasteful or abusive. Even though the initial results were satisfactory, considering the large number of fields present in the dataset, the model could be improved to better detect possibly fraudulent behaviors. This study was a proof-of-concept endeavor and did not confirm whether or not the model can actually predict a physician's fraudulent behavior. As such, we intend to improve upon this prior study by evaluating the performance of our original model, using real-word known fraud cases, and assess the improvements made through our three proposed strategies.

4.2.2 Methodology

In this section we use the Part B 2013 and 2014 dataset [26]. Due to the large size of the dataset, we decided to only use data from office clinics in Florida (as opposed to larger facilities, such as hospitals and academic institutions). Table 2.7 summarizes the 2013 and 2014 Medicare data, nationally and for Florida only. We provide a detailed discussion, outlining our unique data processing and feature engineering we applied to the Part B dataset in Section 2.3.1. Table 2.6 shows a small example of the sparse vector where each line is a physician, PROVIDER_TYPE is the class attribute and every other attribute (codes 99222 through 64482 in this example) are the procedures. For every instance, there is a value for the number of times the given physician performed that procedure. Additionally, for model testing and validation, we incorporate the LEIE. Section 2.3.1 discusses the fraud mapping procedure we employed for this section. We matched physicians from our 2013 procedure code dataset with the LEIE data. The LEIE is constantly changing as physicians are added or removed; therefore, we accessed the LEIE database twice during this experiment, for 2016 and 2017, and found 16 physicians in the Florida Medicare data. We supplemented these 16 physicians with two other documented fraud cases [44, 113], found in the 2013 data, giving us 18 fraudulent physicians. Note that the LEIE dataset is only used to identify fraudulent doctors corresponding to the information in the Medicare data. Only the Medicare dataset is used for model training and testing.

For this section, we chose MNB to build each model. Specifically, we used the implementation in the Weka machine learning toolkit, with a single run of 5-fold cross-validation for model evaluation. The calculations for model performance are from the confusion matrices. We use a one-vs-all approach for calculating the error rates, which considers the class in question as the positive class (true positive) and the rest of the classes as being in the same negative class (true negative). We leverage Recall, Precision and F-Score to assess model performance, with F-Score being the primary metric for comparison.

Original Model Testing and Validation

In order to create and assess improvements to our original model, we need to validate, through the use of real-world fraudulent physicians, the hypothesis that by using only procedural data, a prediction model can successfully detect possibly fraudulent or wasteful behaviors. For this, we use the 2013 Medicare claims data with labels consisting of known fraudulent physicians from the LEIE database and the two documented cases. The procedure used for testing our original model is exclusion testing, where we removed the fraudulent physicians from the training dataset and created a test dataset composed of the 18 known fraudulent physicians. The model was built using MNB, in Weka [60], from the modified training data (after the 18 were removed) and evaluated on the test dataset (including only the 18 removed instances). The test evaluation was done by reviewing the resulting confusion matrix, where the number of instances predicted into a class other than their actual field are denoted as possibly fraudulent.

Improvement Strategies

From preliminary analysis, we found that specialties with comparable characteristics are similarly improved by a particular strategy, and thus we chose a small, representative number of specialties to focus on in this section. It is important to note that the Medicare 2013 dataset is identical to that used in Section 4.1. All data manipulation was performed using the R language [136], or Weka. There are three improvement strategies used in this section:

- Feature Selection and Sampling
- Removing Classes
- Grouping Specialties.

Each improvement strategy is explained in Section 3.1.1.

4.2.3 Results and Discussion

This subsection presents the results from testing our model using the fraudulent dataset, and the implementation of the three improvement strategies. In order for our original hypothesis to be valid, a specialty must be able to be confused (i.e. labeled as a different specialty) via two primary factors: 1) the physician performs procedures in a way that is significantly different from their peers, or 2) the classification model is sub-optimal. Otherwise, our prior hypothesis is not true. We endeavor to validate our previous model's performance, using physician's procedures, and provide meaningful improvements through our proposed strategies. Therefore, in this section, we aim to provide research to find strategies that will yield the largest improvements in model performance, while not compromising fraud detection capabilities.

The original model is validated against a real-world exclusion dataset to confirm the successful detection of fraudulent behaviors. Due to the limited number of physicians in the LEIE corresponding to the Florida-only Medicare data, testing the strategies would not be reliable; therefore, with these strategies, we focus on improving F-Score results over the original model results found in [14]. To assess performance changes, we chose to focus on a few select fields from which we found that different strategies improve model performance differently based on the specialty. As mentioned, the three improvement strategies are 1) feature selection and sampling, 2) removing classes, and 3) grouping similar classes.

Five specialties, also known as classes, were chosen to assess changes in model performance based on the proposed strategies. These specialties included: Otolaryngology, Dermatology, Ophthalmology, Cardiology and Internal Medicine. It is important to note that the class removal strategy does not include Internal Medicine. These five specialties were chosen to adequately capture the variability in F-Score results from the original model (which had scores ranging from 0.91 to 0.33). As seen in Table 4.6, Otolaryngology has a high F-Score with a large number of different procedures, whereas, Internal Medicine has a low F-Score, an above average number of procedure codes, and a large number of instances. The remaining specialties are consistent with above average F-Scores, average number of procedure codes, and similar numbers of instances.

Original Model Testing and Validation

The results from the exclusion testing, where we tested our original model against a list of known fraud cases, showed the model correctly classified 12 out of 18 physicians (67%) whom exhibited fraudulent behavior, based on the violation of specific LEIE rules listed in Table 2.5, minus rules 1128(c)(3)(g)(i) and 1128(c)(3)(g)(ii). Table 4.5 depicts the 18 excluded physicians to include their class, the number of instances in each physician's class, and their predicted class. The "Classified As" column name indicates which class(es) the fraudulent physicians were confused for. For example, General Practice finds three matching fraudulent cases, with two of them labeled as different classes (general surgery and emergency medicine) and one instance labeled as its actual class (General Practice). Each fraudulent instance labeled as a different class is considered a possibly fraudulent behavior by our model. Not all classifications should be considered possible fraud, with some specialties simply being confused with other similar classes, such as Ophthalmology and Optometry (both specializing in the eyes), ergo the implementation of the class grouping strategy to improve specialty classification.

Table 4.5: Classification of Fraudulent Physicians

Actual Class	Matching Instances	Classified As
Cardiology	1	-
Family Practice	3	Psychiatry Internal Medicine Gastroenterology
General Practice	3	General Surgery Emergency Medicine
Gynecological/ Oncology	1	Obstetrics/ Gynecology
Hematology/ Oncology	1	-
Internal Medicine	3	Gastroenterology Family Practice Geriatric Psychiatry
Ophthalmology	2	Optometry
Otolaryngology	1	Dermatology
Podiatry	2	-
Psychiatry	1	Nurse Practitioner

In contrast to Ophthalmology and Optometry, Otolaryngology is classified as Dermatology which is not a similar specialty. Even so, they both have fairly high F-Score values, as shown in Table 4.6, which seems to indicate that this particular Otolaryngologist is not performing procedures similarly to their peer group and possibly acting in a fraudulent or wasteful manner. Overall, 67% of the real-world fraudulent physicians were classified as something other than their actual field (i.e. 67% accuracy) and, under these basic conditions with no proposed improvements implemented, appears to be quite promising. With that, classification performance has room for improvement in detecting actual fraudulent behavior versus normal behaviors. The remaining experiments involve the implementation and testing of our proposed strategies for improving classification results as seen in changes to F-Scores.

Table 4.6: Chosen Fields for 2013 CMS Data

Specialty	Original F-Score	# of Codes	# of Instances
Otolaryngology	0.91	2453	477
Cardiology	0.82	525	1540
Dermatology	0.80	276	866
Ophthalmology	0.73	227	1139
Internal Medicine	0.33	961	4243

Strategies

In this subsection, we present the results of our three strategies, assess improvements made, and discuss any caveats and limitations regarding the application of these strategies per specialty. Additionally, we provide comparative testing values with the 2014 Medicare data to validate and confirm the results and improvements seen using the 2013 Medicare data. Note, results presented are for the 2013 data unless specified as 2014.

Feature Selection and Sampling In this subsection, feature selection (or HCPCS procedure code selection) results are presented, across a range of procedures from the full feature set down to the 500 top features. Table 4.7 shows the results of the Dermatology class versus the *other class*, in order provide an example of both class results and the average F-Scores. We selected Dermatology because feature selection provided neither a substantial increase or decrease in performance over the original model. It is necessary for the F-Scores of both the positive and negative classes to be high in order to have a reliable model, which implies the ability to correctly classify both as true positives and true negatives. The *other class* makes up the majority of the dataset, thus the generally high negative class F-Score across the chosen specialties. With that, the class in question is of higher importance, thus the weights will be set to 50% (equal weighting) for the average F-Score values.

Table 4.7: Dermatology Feature Selection Results

# of Procedures	Provider Type	F-score	Average F-score
Full (2,789)	Dermatology	0.39	0.68
	Other Class	0.97	
2500	Dermatology	0.39	0.68
	Other Class	0.97	
2000	Dermatology	0.40	0.68
	Other Class	0.96	
1500	Dermatology	0.42	0.69
	Other Class	0.96	
1000	Dermatology	0.42	0.70
	Other Class	0.97	
500	Dermatology	0.43	0.70
	Other Class	0.97	

We are interested in comparing performance versus the original model, thus the remaining F-Scores, by specialty, are listed in Table 4.8. This table shows the number of procedures, the class in question (positive class), and the weighted average with each class receiving equal weighting. From these results, the higher F-Scores indicate better model performance. The average scores indicate that the F-Scores marginally improve when the model uses less procedures. The best scores are seen when using 1,000 or 1,500 procedures, or less than 50% of the full feature set. So even given only this marginal improvement, the reduction in the number of features needed can reduce computational complexity.

Unfortunately, feature selection alone does not produce sufficiently improved performance results. Ophthalmology and Internal Medicine show increased performance over the original model, as seen in Table 4.6, but the others have lower average F-

Table 4.8: Feature Selection Results for the Remaining Specialties

	Average F-Score of the Positive and Negative Class			
# of Procedures	Otolaryngology	Cardiology	Ophthalmology	Internal Medicine
Full (2,789)	0.53	0.55	0.80	0.50
2500	0.52	0.57	0.79	0.50
2000	0.54	0.57	0.79	0.50
1500	0.54	0.60	0.80	0.51
1000	0.53	0.60	0.79	0.51
500	0.53	0.60	0.80	0.48

Scores. The lack of improvement is most likely due to the large difference between the size of the classes, which, conversely, is why Internal Medicine performs better as it has a high number of instances relative to the entire dataset versus the other tested specialties. Since class imbalance is seen to be an issue, balancing the class sizes through sampling could be used to further increase model performance.

In order to balance the classes, we used both undersampling and oversampling techniques. Undersampling, reduces the number of the majority class by some percentage in order to reduce overall class imbalance. In Weka, the reduce setting is the "distibutionSpread" value that was varied to adjust the number of instances for each class. Conversely, oversampling adds more instances to the minority class in order to better balance the classes. With oversampling, the addition setting in Weka is the "percentage" value with which we use to vary the increase in the minority class. The results are presented in Table 4.9 indicating minimal improvements using under-sampling versus feature selection alone, but increased performance across all specialties when employing the SMOTE oversampling method. The highest gains with oversampling are seen with Internal Medicine, Dermatology, and Ophthalmology.

Removing Classes In Table 4.10, we present the results based on the removal of specific classes. In this table, we show the original model's 2013 F-Scores, update scores after class removal, the 2013 change in F-Scores, and the changes in scores using

Table 4.9: Results of Random Undersampling and Oversampling

		Undersampling		Oversampling	
Specialty	# of Procedures	Reduction Setting	Average F-Score	Addition Setting	Average F-Score
Ophthalmology	1500	20	0.86	1800	0.97
Dermatology	1000	20	0.79	2200	0.96
Cardiology	1500	10	0.73	1200	0.91
Otolaryngology	1500	30	0.65	4000	0.90
Internal Medicine	1500	4	0.50	450	0.77

the 2014 data for confirmation and validation of model improvements. To reiterate, the 2014 data was not used as a test set but rather to replicate the procedure we used on the 2013 data, creating a new model, in order to show similar results between years. We do not show the original and post class removal F-Scores for 2014 due to space, but the scores are very similar the 2013 data results. Additionally, Fig 4.3 shows the results for the remaining 78 specialties.

After removing the four classes, as described in Section 3.1.1, we notice that there are large improvement in the low scoring specialties, as well in Otolaryngology and Dermatology, with Ophthalmology being the only class to have a decrease in F-Score. From these results, the classes with relatively large improvements most likely have procedures that are easily confused with those in the removed classes, i.e. a large number of overlapping procedures. The classes with little to no improvements would indicate more specialized services with minimal overlapping procedures, with regards to the removed classes. Furthermore, we tested the class removal strategy after removing the same four specialties.

Once the classes were removed, 12 of the 18 fraudulent physicians remained with 5 of these 12 (42%) labeled as possibly fraudulent by the model. Our original model's fraudulent physician detection results are actually the same if we were to remove these four classes. Because there is no change, the physicians not removed were not affected by the removal of these four classes, so any improvements via this strategy in detecting fraudulent behaviors is inconclusive requiring additional work. Even so, for some individual specialties, this strategy appears promising and future work with class removal could improve performance across multiple classes. It is important to note that the overly liberal removal of classes could reduce the detection of real fraudulent behaviors, by removing the potential classes of interest.

Table 4.10: Improvements from Removing Classes

Specialty	Original 2013 F-Score	Updated 2013 F-Score	2013 Gain	2014 Gain
Otolaryngology	0.91	0.95	0.04	0.04
Cardiology	0.82	0.90	0.08	0.08
Dermatology	0.80	0.97	0.17	0.17
Pathology	0.79	0.79	0.00	0.00
Orthopedic Surgery	0.74	0.81	0.07	0.08
Ophthalmology	0.73	0.72	-0.01	0.00
Psychiatry	0.51	0.71	0.20	0.20
Emergency Medicine	0.20	0.47	0.27	0.29
General Practice	0.13	0.35	0.22	0.22

Grouping Specialties The results for the grouping of specialties are outlined in Tables 4.11 and 4.12. These results summarize model performance (Recall, Precision, and F-Score) for the individual specialties and the grouping of the specialties which have similar practices. The group score is the weighted average of the individual performance results. Classes were grouped manually on a class-by-class basis, where, for example, Otolaryngology shows the results for Otolaryngology and its similar classes. As in Section 4.2.3, we used the 2014 data in order to corroborate and validate the 2013 model results. The original models and the grouped models were built using Multinomial Naïve Bayes, to fairly compare the results and show any improvements due to our specialty grouping strategy.

F-Scores improved for each grouping, with Cardiology and Ophthalmology having the most significant improvement. These improvements appear to be heavily dependent on the specialty for which the grouping is applied; therefore, this strategy will be effective for some classes but less effective for others. By grouping Ophthalmology with Optometry, the F-Score increased over the individual scores resulting in a very high F-Score of 0.96. The Cardiology grouping shows two of the individual specialties with F-Scores below 0.5 and Cardiology only with a score of 0.82. Even with two

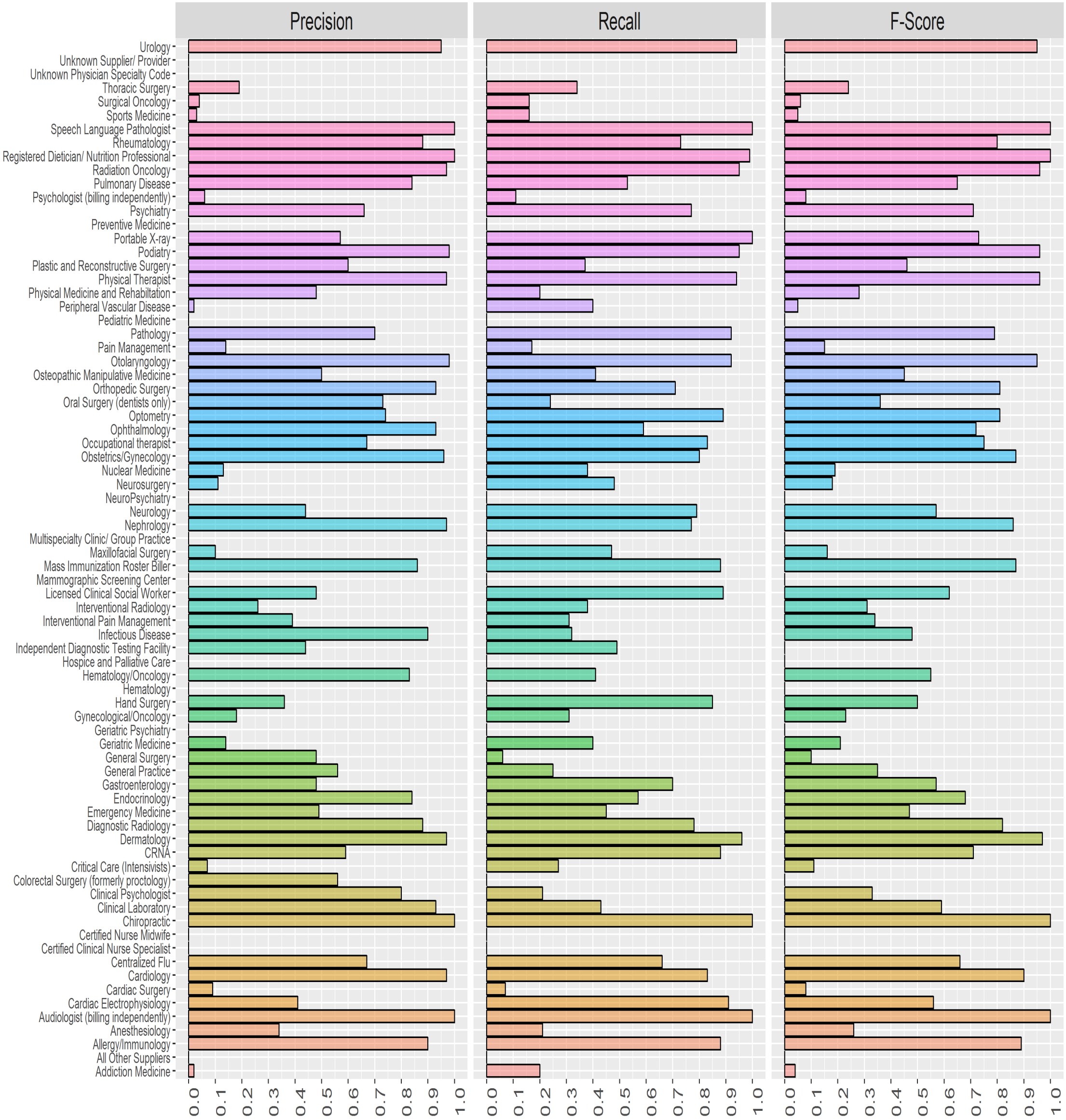


Figure 4.3: Class Removal for All Physician Types (2013)

of the group members having low individual F-Scores, the Cardiology group had a good F-Score of 0.9. Dermatology indicated negligible improvement with grouping, while the Otolaryngology group had a slight decrease in performance. The improvements shown by grouping specialties may not actually increase the effectiveness of our model to detect fraudulent and wasteful behavior. For instance, the fraudulent Ophthalmology case in our test dataset was confused due to the inclusion of the Optometry specialty. This particular grouping would decrease the 67% fraud detection rate of our model. Thus, additional experimentation is needed to confirm whether this strategy, not only improves F-Scores, but also improves real-world detection. In particular, we would need more real-world fraudulent Ophthalmology cases to determine if they are commonly confused for Optometry, thus demonstrating that some groupings can mask possible fraud activities.

Table 4.11: Class Grouping Results

Group	Specialty	Recall	Precision	F-Score
Otolaryngology	Otolaryngology	0.90	0.92	0.91
	Allergy / Immunology	0.77	0.88	0.83
Cardiology	Cardiology	0.83	0.81	0.82
	Cardiac Electrophysiology	0.32	0.91	0.48
	Cardiac Surgery	0.04	0.04	0.04
Dermatology	Dermatology	0.70	0.93	0.80
	Plastic Surgery	0.51	0.36	0.42
Ophthalmology	Ophthalmology	0.93	0.60	0.73
	Optometry	0.74	0.89	0.81

Table 4.12: Improvements from Grouping Classes

Group	Group F-Score	2013 F-Score Gain	2014 F-Score Gain
Otolaryngology	0.90	-0.01	0.01
Cardiology	0.90	0.14	0.06
Dermatology	0.73	0.02	0.04
Ophthalmology	0.96	0.19	0.21

4.2.4 Section Summary

Our intent for this line of research is to develop a system that successfully detects physicians who work outside the norm of their field, through the successful application of anomaly detection. We continue our previous research and expand upon this original model in order to increase fraud detection capabilities. In this section, we tested and validated our original model against known fraudulent physicians which resulted in 12 of the 18 (67%) physicians being successfully labeled as fraudulent. Our model hypothesis is if a physician is not submitting procedures in a similar manner compared to their peers, which can be seen as abnormal, then that physician may be committing fraudulent or wasteful behavior. The 67% detection rate of our original model is quite good, even considering that a number of specialties garnered low or very low F-Scores. Even so, the high detection rate could be due to inadequacies in the original model to include too many classes (82) being uniquely evaluated and a large number of low scoring specialties. In order to address these concerns and increase the performance of our original model, we proposed three improvement strategies that include: feature selection and sampling, removing specialties with a large number of overlapping procedures, and grouping similar specialties.

Table 4.13: Summary of F-Score Results

Specialty	Original	Feature Selection	Removing Classes	Grouping
Otolaryngology	0.91	0.15	0.95	0.90
Cardiology	0.82	0.30	0.90	0.90
Dermatology	0.80	0.43	0.97	0.73
Ophthalmology	0.73	0.61	0.72	0.96
Internal Medicine	0.33	0.32	-	-

Table 4.14: Best Strategies for Classes

Strategy	Specialty	Reason
Feature Selection	Internal Medicine Dermatology	Large number of instances
Over Sampling	Ophthalmology Internal Medicine	Works well with a range of class
Removing Classes	Dermatology Cardiology Otolaryngology	Works well with classes that have overlapping procedures with "generic" specialties
Grouping	Ophthalmology Cardiology	Classes that are very similar with other specialties

In Table 4.13, we summarize the F-Score results from the original model and all proposed improvement strategies. With feature selection only, we found that using 1,000 to 1,500 procedures gave the best F-Scores across the classes, but none of the classes demonstrated improvement over the original model. In addition to feature selection, we used under- and over-sampling to improve model performance. The results showed that under-sampling did not positively change the F-Score, whereas over-sampling improved results across the board. Our second strategy, class removal, showed large improvements for a few fields (e.g. Dermatology) implying this method could improve model performance relative to certain specialties. As mentioned, there is a concern with removing too many specialties reducing potential fraud detection capabilities. Grouping of similar specialties is our third strategy which showed some significant improvements over individual specialty results, such as the combination of Ophthalmology and Optometry, but less noticeable changes to other groupings. Overall, our results indicate that different strategies can improve model performance depending on the selected specialties; therefore, the choice of strategy is, in large part, determined by the specialty. With that, we provide specific characteristics or reasons why a specialty would improve given one of the three proposed strategies in Table 4.14. As the LEIE dataset contains very few NPIs, with only 18 were available for Florida, future work will include adding other states so that more comparisons of real-world fraudulent cases can be done using the LEIE database. Additionally, using different machine learning methods and performance metrics will be pursued.

4.3 APPROACHES FOR IDENTIFYING U.S. MEDICARE FRAUD IN PROVIDER CLAIMS DATA

Healthcare is a vital component in the daily life of most citizens in the United States. Even given healthcare's prominence in society, costs and premiums continue to increase placing additional strain on those in need of medical services. In particular, the generally increasing life expectancy as well as increasing population size, with the rise of chronic diseases, puts additional burdens on programs focused on the health of the elderly [52, 74, 75]. As such, it is imperative that healthcare costs are kept fair and reasonable for quality medical services [50]. Medicare is a U.S. government program that provides healthcare insurance and financial support for the elderly population, ages 65 and older, and other select groups of beneficiaries [38]. Within the Medicare program, each covered medical procedure is codified for claims and payment purposes. The basic claims process entails a physician performing one or more procedures and then submitting a claim to Medicare for payment, rather than directly billing the patient, thus making Medicare a type of "middle man" in this process. A claim is defined as a request for payment for benefits or services received by a beneficiary. In this section, we use the terms providers and other entities [31] interchangeably with physicians, and note any specific differences as needed. Additional information on the Medicare claims process can be found in [25].

As in Sections 4.1 [14] and 4.2 [68], we use data released by CMS [11, 26, 27]. In contrast to our other works that use smaller subsets of Medicare data, we use the complete 2012 to 2015 *Medicare Provider Utilization and Payment Data: Physician and Other Supplier* dataset, also known as Medicare Part B, which includes provider Medicare claims information for the U.S. and its commonwealths. Because the Medicare Part B dataset does not include labels indicating provider fraud, we use the LEIE [91] database to generate fraud class labels (i.e. fraud or no fraud) for each provider to assess fraud detection capabilities of our baseline model and proposed improvement strategies. The LEIE contains all physicians who are excluded from practicing medicine for federally funded programs, such as Medicare. It is important to note that in this section, fraud detection is flagging possible or suspected fraudulent activities (based on known LEIE exclusions) and any non-fraud provider claims should be considered either non-confirmed fraud or indicative of exhibiting no fraud. We use fraud and non-fraud labels interchangeably with possible/suspected fraud and non-confirmed fraud designations. The effective detection of fraud can help to reduce costs related to time and resources needed for further investigation, by focusing the efforts on candidates most likely to exhibit fraudulent behaviors. Our baseline model, closely based on the original model from Sections 4.1 [14] and 4.2 [68], predicts the expected medical specialty of a physician based on the type and count of procedures performed. For example, if our model predicts a physician as a Dermatologist but the actual specialty is Optometrist, then this physician could be performing procedures indicating possible fraud. This difference in expected and actual specialties, for a particular provider, could indicate possible fraud to include possible upcoding [15] or coding incorrect procedures. It is important to note that issues related to medical coding do not necessarily imply fraud, but are still problematic in submitting, accepting, and paying on claims. A medical coder may not have enough information (e.g. poor documentation or lack of easy access to the provider) regarding the procedure to document the coding at the highest level of specificity, or may be using incorrect medical coding code sets [101].

In this section, our primary contributions include determining the best baseline model and assessing our proposed improvement strategies which include: class grouping, class removal, and class isolation [66]. The baseline model is based on our prior research in predicting provider specialties but takes into account providers practicing in offices and/or facilities, such as a hospital, since services can be offered at one or both locations. Class grouping takes similar specialties and combines them into a single class, whereas class removal entails removing a selection of low scoring specialties based on two specified criteria. Class isolation is a different approach than the previous two improvement strategies that randomly sample a percentage of non-fraud class labels, retaining all fraud labels, per specialty, building models for each specialty to predict fraudulent providers. Overall, the class grouping and removal strategies had inconsistent results, with limited improvements and some cases of worsening fraud detection performance versus the baseline model. The class isolation method, however, indicated good improvements in overall fraud detection.

The rest of this section is organized as follows. Section 4.3.1 discusses works related to current research in this area. Section 4.3.2 details the methodology used, including the data, learners, performance metrics, and baseline model selection and improvement strategies. Section 4.3.3 discusses the results of our baseline model and strategies, as well as some possible research limitations. In Section 4.3.4, we conclude and present ideas for future work.

4.3.1 Related Works

CMS has been releasing a given year's data on average two and a half years after the end of a particular year. Therefore, research employing this data is relatively new with no comprehensive studies leveraging all the publicly available Medicare data. All related works use varying subsets of the full Medicare dataset with most providing preliminary assessments and results, with or without machine learning approaches. Two related works that use Medicare data employing more typical data analysis to include descriptive statistics and regression are by Feldman et al. [49] and Ko et al. [87]. Using 2012 Medicare data, Feldman et al. try to determine if there is any correlation between a physician's schooling and the way he or she practices. They compared medical school charges, procedures, and payments for a given physician, researching whether they could identify possible anomalies in the data. With this information the authors could flag physicians that are at risk of performing fraudulent activities at the beginning of their careers. Another study by Ko et al. focused on the field of Urology. The authors analyze variability among Urologists within the field's service utilization and payment data (2012 Medicare) to determine the estimated savings from a standardized service utilization. They determined there was a strong correlation between the number of patient visits and reimbursement received from Medicare. They establish that in the specialty of Urology alone there could potentially be a $125 million savings, or about 9% of the total expenditure within the field. Neither of these studies though employs more advanced analytics or machine learning to predict fraudulent providers or behaviors, nor do they leverage the full scope of available Medicare data.

The idea of looking for deviations from normal or expected patterns is part of a study by Bauder et al. [7]. The authors built a multivariate regression model for each Medicare specialty, such as Cardiology. From this model, the studentized residuals were generated and used as inputs into a Bayesian probability model in order to produce the probability of an instance being an outlier, which indicates the likelihood of fraud. Sadiq et al. [124] used the 2014 CMS dataset in order to find anomalies that possibly point to fraudulent or other interesting behavior. The framework they employ is the Patient Rule Induction Method based bump hunting method attempting to determine peak anomalies by spotting spaces of higher modes and masses within the dataset. They explained that by applying their framework they could characterize the attribute space of the data helping uncover the events provoking financial loss.

None of the previously discussed works, whether using descriptive statistics or machine learning, provide fraud detection performance validation using known fraudulent providers. In a general coverage paper by Chandola et al. [21], they used the CMS Medicare among other datasets to assess healthcare fraud using labeled data for fraudulent providers, primarily from the Texas Office of Inspector General's exclusion database. The authors employed several techniques including social network analysis, text mining, and temporal analysis. In particular, the authors discussed the use of typical treatment profiles, i.e. procedures performed. The idea is leveraging these profiles as the normal activity of physicians by comparing them with any provider in question which will determine possible issues or abuses in procedures. This is akin to predicting expected values and comparing these predictions to the associated actual values. In [18], Branting et al. use the 2012, 2013, and 2014 Medicare Part B and Part D data with the LEIE. They present a method for pinpointing fraudulent behavior by determining the fraud risk through the application of network algorithms from graphs. One algorithm, which they denote as Behavior-Vector Similarity, determines similarity in behavior for real-world fraudulent and non-fraudulent physicians using nominal values such as drug prescriptions and medical procedures. A group of algorithms makes up their Risk Propagation, which uses geospatial co-location (such as location of practice) in order to estimate the propagation of risk from fraudulent healthcare providers.

In our previous papers [14] and [68] (Sections 4.1 and 4.2), we experimented with whether or not it is possible to predict a physician's field of expertise based on only the procedures they perform. We found that there were a small percentage of specialties that could be accurately predicted. The overall goal of these papers and this current work is to better pinpoint fraudulent behavior. In essence, if we can predict a physician's specialty accurately (determined by F-score), then we could potentially find anomalous behaviors in a physician's procedures if they are predicted to be a specialty other than their own (e.g. Dermatologist as a Rheumatologist). The idea is that if a physician is predicted as having a specialty other than their actual specialty, they could be behaving in fraudulent, wasteful or abusive ways. The strategies we constructed were successful in increasing the prediction capabilities for many specialties, using 2013 Florida-only Medicare data. Our results showed we were able to detect 12 of the 18 known fraudulent physicians. These studies are limited in the location and amount of claims data used, and do not encompass the full scope of Medicare claims for the United States. Our previous works applied a different approach in trying to detect Medicare provider claims fraud, versus other related works, predicting a provider's specialty using the procedure type and utilization. We extend and improve upon our prior research considering procedures performed in an office and/or facility, and use data for the entire U.S. (not just Florida) over all available years. In this section, we evaluate fraud detection performance of our proposed improvement strategies by leveraging the LEIE excluded providers as fraud or no-fraud labels.

4.3.2 Methodology

Data

In this section, we use the Medicare Part B claims dataset which includes the 2012 to 2015 calendar years [27]. We combined the four available years of Part B data into a single dataset. We discuss our unique data processing and feature engineering steps in Section 2.3.1. Table 2.8 shows a small example of the sparse matrix where each line is a physician, indicated by *provider_type* (i.e. specialty), with the remaining attributes (codes 99222 through 64482 in this example) being the procedures. For every instance in this sparse vector, there is a value for the number of times the given physician performed that procedure for that given year. Table 2.6 shows a small example of the sparse matrix where each line is a physician, indicated by provider_type (i.e. specialty), with the remaining attributes (codes 99222 through 64482 in this example) being the procedures. For every instance, there is a value for the number of times the given physician performed that procedure per year. In order to validate fraud detection performance, we need labels indicating fraudulent provider claims. The Medicare Part B dataset does not include fraud labels; thus, we incorporate the LEIE [91]. We discuss our fraud mapping process between the Medicare Part B dataset and the LEIE in Section 2.3.1. From the mapping with the Full dataset, we ended up with a total of 1,310 physicians found as fraudulent. We supplemented these 1,310 physicians with two other documented fraud cases, found during our previous research [68], giving us 1,312 fraudulent physicians. We generate two datasets (exclusion and non-exclusion) from the combined Medicare Part B data with the LEIE excluded physicians as fraud labels with a more detailed discussion in Section 3.1.1. Table 4.15 provides a high-level summary of each dataset.

Table 4.15: Dataset Descriptions

Dataset	Description
Part B	- Released by CMS. - Provides claims information for each procedure a physician/provider performs within a given year. - Oriented by: 1) NPI, 2) HCPCS, and 3) Place of Service
LEIE	- Released by OIG. - Contains physicians/providers that have committed real-world fraud.

Learners

We perform the data manipulations in order to generate each of the datasets using the R [136] and Python [118] programming languages. For building and testing our models, we use either Weka or PySpark depending on the improvement approach and the size of the dataset. The PySpark library in Python is used to interface with and leverage the capabilities of Spark [152, 153], which is a unified analytics engine for large-scale data processing [131]. Weka is an application providing a suite of machine learning algorithms that can be applied either via a graphical interface or the command line, and is suitable for smaller datasets. More specifically, for the class grouping and removal strategies, we use PySpark for building models through the Apache Spark Machine Learning Library (MLlib) [102], due to the large size of the full Medicare Part B dataset. The class isolation method, however, generates much smaller datasets and we can leverage the larger variety of tools for machine learning in Weka. In this section, we use four machine learning models: MNB, LR, SVM, and RF. We used Logistic Regression in both Weka and PySpark. For Pyspark, Logistic Regression with the Limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm (LBFGS) is the version used to improve memory usage [97]. In this research, we will be using data for both binomial and multinomial LR. The specific implementation of SVM used in Weka is called Sequential Minimal Optimization (SMO), which uses this optimization algorithm as the training method for Support Vector Classification. MNB, LR are used with both PySpark and Weka while SVM and RF are used in Weka only.

Performance Metrics

In order to provide a robust experiment on fraud detection, we use several different metrics to assess model and improvement strategy performance. We leverage Recall, Precision, F-score, G-measure, and overall accuracy to assess model performance, with F-score being the primary metric for general performance comparisons. We also use the so-called Inverse Overall Accuracy (IOA), for each specialty, and the overall weighted average IOA (owaIOA), which determine the fraud detection ability of our models. For this section, we are interested in the incorrectly classified (as we consider these fraud) which is given by the IOA and owaIOA, which is normalized by the number of fraudulent instances for each specialty. Additionally, for the class isolation method, we use Type I (false positive) and Type II (false negative) error rates to assess the predictive capabilities of the models for detecting Medicare Part B fraud. These performance metrics are defined in Section 3.3.1

Baseline model selection and improvement strategies

First, we determine the best model to be used as our baseline model for assessing the change in fraud detection performance using the class grouping and removal strategies. Recall that the baseline model is based on our prior research in producing models that predict provider specialties to assess possible fraudulent behaviors. There are a couple of requirements for selecting a baseline model: 1) able to support multi-class classification, and 2) able to output the Precision, Recall, F-score and overall accuracy in order to make fair comparisons. Due to the large data size and through preliminary analysis, we found the PySpark versions of MNB and LR met these criteria.

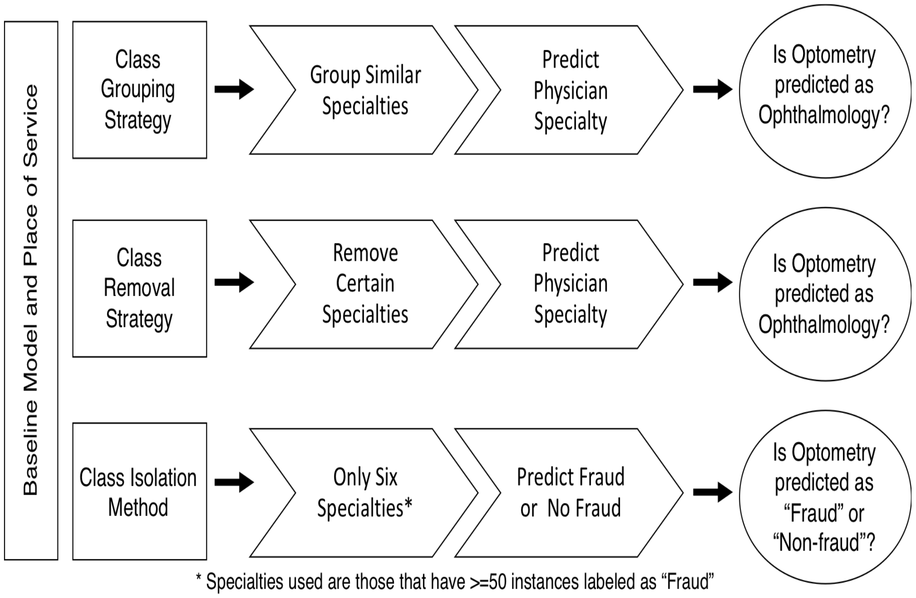
The methods for handling different locations, office or facility, from which physicians perform medical procedures is also tested in order to get the best overall baseline model using either office- or facility-based procedures. The problem with having different places of service is that a physician can bill the same procedure in either an office or facility. To account for these possible multiple entries, we devised two methods of handling this by either Combining Office and Facility (COF) or Separating Office and Facility (SOF). Section 2.3.1 provides details and examples for COF and SOF. COF leaves the data with the original number of procedure codes, while SOF increases the number of procedures as seen in Table 2.9. The best performing PySpark model using either COF or SOF is to be used as the baseline model for the class grouping and removal strategies. Additionally, the COF or SOF method which gives better performance results is also used with the Weka models for the class isolation method. In order to select the best baseline model and office or facility method, we use Precision, Recall, F-score, G-measure, and overall accuracy as metrics for performance evaluation. For this experiment, we use the non-exclusion dataset for training and the exclusion dataset for testing, in order to validate the models and assess the detection of fraud instances. When testing the models, the exclusion dataset is split into a number of smaller datasets, one for each physician type (specialty), and each of these datasets is used as the test set. We use IOA to assess fraud detection performance with the best models having the highest IOA per specialty. The baseline model is thus the best combination of MNB or LR and COF or SOF. With this baseline model, we experiment with the class grouping and removal strategies, as well as the combination of both strategies. Lastly, the class isolation method also compares performance against that of the baseline model.

The class grouping strategy groups similar specialties into a single group to reduce redundant specialties and decrease model variance. This is based on research in our previous work [68], where we found that specialties that practice on similar parts of the body or have similar medical descriptions provide improvements to prediction results. The process used for creating the groupings is discussed in Section 3.1.1. We decided to test fourteen different groups as shown in Figure 3.2. In order to evaluate any performance improvements via class grouping, we take a two stage approach discussed in Section 3.1.1. The class removal strategy removes certain specialties from the dataset prior to building a model. We have two different sets of classes (specialties) for removal to test model performance. The first, referred to as the 'original four classes', is from our previous work and was based on unique procedures that have both a high number of instances and poor classification performance. The second criteria for removal includes Table 3.1 and shows the specialties for both criteria of class removal. The class removal strategy is discussed in Section 3.1.1. The class isolation method improvement strategy differs from the other two approaches, qualifying as a traditional fraud detection approach. This method employs data sampling and adjusting the costs associated with Type I and Type II errors. The class isolation method is described in Section 3.1.1.

4.3.3 Results and discussion

In this subsection, we present the fraud detection results for the baseline model and each of the improvement strategies. For ease of presentation, we organize this subsection into three parts corresponding to the experimental flow and design as outlined in Section 4.3.2. The three parts include: 1) baseline model selection with SOF and COF, 2) class grouping and removal strategies with combinations, and 3) class isolation method. For each of these parts, we present the salient results with discussions pertaining to the Medicare Part B fraud detection performance and validity. Figure 4.4 depicts a high-level flow of these parts with inputs, descriptions, and possible output behaviors. Note that these output behaviors are examples of a possible predicted specialty versus actual specialty.

Figure 4.4: Improvement Approaches


Baseline Model and Place of Service
Class Grouping Strategy
Group Similar Specialties
Predict Physician Specialty
Is Optometry predicted as Ophthalmology?
Class Removal Strategy
Remove Certain Specialties
Predict Physician Specialty
Is Optometry predicted as Ophthalmology?
Class Isolation Method
Only Six Specialties*
Predict Fraud or No Fraud
Is Optometry predicted as "Fraud" or "Non-fraud"?
* Specialties used are those that have >=50 instances labeled as "Fraud"

Baseline model selection with place of service

During the baseline selection process, we select the best model using the SOF or COF configurations[1] between MNB and LR. Table 4.16 shows the average scores for each performance metric by learner and SOF/COF method. These results indicate that LR outscores MNB for all but Recall. Since LR, in general, is shown to perform better than MNB, we focus on the SOF and COF configuration results for LR only for which SOF has higher scores (possibly due to the larger number of features to draw from). Given these results, we selected LR and SOF as the baseline model for the class grouping and removal strategies, and the SOF configuration for the class isolation method.

We evaluate the baseline model's fraud detection performance for each specialty (IOA) and overall (owaIOA). Table 4.17 shows the results for the baseline model, per specialty, with F-scores at or above 0.75 indicating good detection performance. In our previous work [68], we group different scores into performance groups and from these, we selected the 0.75 minimum threshold indicating moderate to high prediction

[1] Separate Office or Facility or Combined Office or Facility

Table 4.16: Average Metric Scores

	Naïve Bayes		Logistic Regression	
	COF	SOF	COF	SOF
Precision	0.442	0.447	0.512	0.514
Recall	0.546	0.550	0.499	0.510
F-score	0.407	0.410	0.487	0.497
G-Measure	0.436	0.440	0.495	0.504
Accuracy	0.517	0.520	0.679	0.682

performance. Note that only specialties with moderate to high prediction performance can accurately determine our model's ability to detect fraud given our hypothesis of "if mislabeled, then fraudulent", and thus only tested on these specialties. In order to summarize the by-specialty results into a single metric representing the overall fraud detection performance of the model, we calculate the owaIOA over the IOA values seen in Table 4.17. The owaIOA for the baseline model, which is the model created to compare our improvement strategies, is 0.231 indicating 23.1% of the instances were correctly labeled as fraud.

Class grouping and removal strategies

Given the baseline model, in this subsection, we present the results of the class grouping and removal strategies and discuss any areas of improvement to this baseline. Unfortunately, both strategies produce mixed results and do not demonstrate general improvements on the baseline model. In Table 4.18, we present the results for the class grouping strategy. In this table, the Group Name is the name we denoted to this selected grouping and the Group Members are the specialties that compose this grouping, with changes in performance shown in the 'Overall Accuracy (Improvement)' column. The results are heavily influenced by the physician specialties in each of the groupings, with some groups having better performing individual specialties, thus better group performance and vice versa. Therefore, the class grouping strategy could be effective with some classes and boost fraud detection performance. In gen-

Table 4.17: Baseline Fraudulent Detection Results

PROVIDER_TYPE	F-score	Instances	IOA
Ambulance Service Supplier	1.00	2	0.00
Chiropractic	1.00	86	0.00
Audiologist (billing independently)	0.99	8	0.00
Physical Therapist	0.99	17	0.00
Speech Language Pathologist	0.97	3	0.33
Pathology	0.95	4	0.00
Diagnostic Radiology	0.94	4	0.25
Occupational therapist	0.92	6	0.17
Podiatry	0.90	47	0.09
Urology	0.90	13	0.23
Gastroenterology	0.88	7	0.29
Cardiology	0.85	13	0.46
Ophthalmology	0.84	20	0.25
Optometry	0.84	8	0.25
Dermatology	0.83	11	0.10
Otolaryngology	0.82	13	0.62
Emergency Medicine	0.81	32	0.53
Obstetrics/ Gynecology	0.80	38	0.58
Allergy/Immunology	0.79	4	0.00
Independent Diagnostic Testing Facility	0.78	5	0.00
Nephrology	0.78	1	0.00
CRNA	0.76	12	0.17
Orthopedic Surgery	0.76	11	0.82

eral, we do not discriminate between different provider types (specialties) with the assumption that each provider (with a specific provider type) has legitimate claims (payment and utilization information). Furthermore, each class, or specialty, can be tied back to an NPI for a particular claim for further investigations.

There were improvements from the baseline model compared to individual members within each group, with good group improvements seen for Anesthesiology, Cardiology, Gynecology, Ophthalmology, Oral, and Otolaryngology. Each of these groups showed good overall improvements, the F-scores are considerably higher than most of the individual member's results indicating significant improvements due to class grouping for these specialties. These particular specialties show promise when employing a grouping strategy. The Chiropractic, Dermatology, Hematology, Neurology, Psychiatry, and Radiology groups showed moderate improvements from the

Table 4.18: Grouping Strategy Performance Results

Group Name	Group Members	Recall	Precision	F-score	G-measure	Overall Accuracy (Improvement)
Anesthesiology	Anesthesiology	0.75	0.65	0.70	0.70	0.701 (0.19)
	Anesthesiologist Assistants	0.05	0.01	0.02	0.02	
	CRNA	0.71	0.82	0.76	0.76	
	Grouped Values	0.98	0.97	0.97	0.97	
Cardiology	Cardiology	0.86	0.85	0.85	0.86	0.683 (0.01)
	Cardiac Electrophysiology	0.48	0.52	0.50	0.50	
	Cardiac Surgery	0.39	0.34	0.36	0.36	
	Grouped Values	0.86	0.86	0.86	0.86	
Chiropractic	Chiropractic	1.00	1.00	1.00	1.00	0.682 (0.00)
	Pain Management	0.24	0.15	0.18	0.19	
	Physical Medicine and Rehabilitation	0.50	0.48	0.48	0.49	
	Physical Therapist	0.98	0.99	0.99	0.99	
	Grouped Values	0.94	0.95	0.94	0.94	
Dermatology	Dermatology	0.78	0.88	0.83	0.83	0.682 (0.00)
	Plastic and Reconstructive Surgery	0.49	0.39	0.42	0.43	
	Grouped Values	0.76	0.75	0.75	0.75	
Gynecology	Gynecological/ Oncology	0.30	0.32	0.31	0.31	0.683 (0.01)
	Obstetrics/ Gynecology	0.86	0.76	0.80	0.81	
	Grouped Value	0.87	0.76	0.81	0.81	
Hematology	Hematology	0.03	0.03	0.03	0.03	0.682 (0.00)
	Hematology/ Oncology	0.46	0.62	0.53	0.53	
	Grouped Values	0.44	0.67	0.53	0.54	
Neurology	Neurology	0.69	0.71	0.70	0.70	0.682 (0.00)
	Neuropsychiatry	0.02	0.01	0.01	0.01	
	Neurosurgery	0.50	0.57	0.53	0.53	
	Grouped Values	0.61	0.71	0.66	0.66	
Oncology	Gynecological/ Oncology	0.30	0.32	0.31	0.31	0.684 (0.02)
	Hematology/ Oncology	0.46	0.62	0.53	0.53	
	Medical Oncology	0.22	0.12	0.15	0.17	
	Radiation Oncology	0.96	0.94	0.95	0.95	
	Surgical Oncology	0.34	0.16	0.21	0.23	
	Grouped Values	0.62	0.76	0.68	0.69	
Ophthalmology	Ophthalmology	0.85	0.83	0.84	0.84	0.688 (0.06)
	Optometry	0.85	0.84	0.84	0.84	
	Grouped Values	0.97	0.95	0.96	0.96	
Oral	Maxillofacial Surgery	0.48	0.14	0.21	0.26	0.682 (0.00)
	Oral Surgery (dentists only)	0.56	0.40	0.47	0.47	
	Grouped Values	0.69	0.64	0.67	0.67	
Otolaryngology	Allergy/ Immunology	0.79	0.79	0.79	0.79	0.682 (0.00)
	Otolaryngology	0.76	0.89	0.82	0.82	
	Grouped Values	0.86	0.87	0.87	0.87	
Pathology	Pathology	0.96	0.93	0.95	0.95	0.682 (0.00)
	Speech Language Pathologist	0.98	0.96	0.97	0.97	
	Grouped Values	0.96	0.93	0.95	0.95	
Psychiatry	Psychiatry	0.59	0.66	0.62	0.62	0.683 (0.01)
	Clinical Psychologist	0.62	0.51	0.56	0.56	
	Psychologist (billing independently)	0.08	0.00	0.00	0.01	
	Geriatric Psychiatry	0.00	0.00	0.00	0.00	
	Grouped Values	0.65	0.60	0.62	0.62	
Radiology	Diagnostic Radiology	0.94	0.94	0.94	0.94	0.683 (0.01)
	Interventional Radiology	0.29	0.18	0.22	0.23	
	Portable X-ray	0.90	0.93	0.92	0.92	
	Radiation Therapy	0.71	0.84	0.77	0.77	
	Radiation Oncology	0.96	0.94	0.95	0.95	
	Grouped Values	0.96	0.96	0.96	0.96	

class grouping strategy. These groups had members with both high and low F-scores, with the group score being either slightly above or slightly below the group's highest scoring member. The remaining groups, Pathology and Oncology, showed no useful improvements. The Pathology group F-score was quite high at 0.95, but the individual members had F-scores of 0.95 and 0.97; thus, no improvement was found via grouping. The Oncology group had good results, but this group has a lot of overlapping specialties. Specifically, the Oncology group overlaps classes with the Gynecological, Hematology, and Radiology groups. But by removing the overlapping groups, we would be removing the three highest scoring classes within this group leaving only the two lowest scoring specialties (Medical Oncology and Surgical Oncology), thus no notable improvements. Even with improved F-scores, class grouping does not seem to improve overall detection results, as indicated by the overall accuracy. Only two groups show good improvement, Anesthesiology and Ophthalmology, with improvements over the baseline model of 0.19 and 0.06, respectively. Otherwise, the accuracy remained the same or showed negligible improvement. Therefore, the improvements shown by only grouping may not actually increase the overall effectiveness of our baseline model.

As with the baseline model, the class grouping strategy's fraud detection performance is assessed with the IOA and owaIOA metrics. To better evaluate possible improvements made by this strategy, we look at two different grouping tests. The first group test is used to assess performance of each group separately, keeping those groups exhibiting improvements and testing only those with an F-score at or above 0.75. These results are shown in Table 4.19, where the number of total number of instances is calculated by taking the sum of instances from each group member. There are 293 instances between these groups with an owaIOA across the eight groups of 25.2% (a 2% increase over the baseline model). In general, some groups show noticeable improvements whereas others do not. The Cardiology and Chiropractic groups showed improvements due to class grouping, demonstrating that this strategy can improve fraud detection for certain groupings of specialties and groups. With that said, there were two groups, Otolaryngology and Ophthalmology, that showed noticeable decreases in performance because of class grouping. One possible reason for the negative results after grouping could be that the members within these groups are extremely similar and, with the baseline model, these classes were being labeled as another class within the group.

The second test, with results shown in Table 4.20, contains all specialties and groups with an F-score at or above 0.75. This includes the groups from the group one combinations plus the remaining specialties not included in a group. There are 444 instances across these specialties with an owaIOA of 26.1%. Thus, there is an increase of 3% over the baseline model using all the groups that had individually good or mediocre results. In particular, two specialties showed improvement for this grouping: Independent Diagnostic Testing Facility and Orthopedic Surgery. Otolaryngology showed an improvement in IOA compared to when grouping was done individually. Conversely, the Chiropractic group had a slight decrease in IOA. Based on our results for class grouping, in order to detect fraudulent behavior for specialties within groups, these specialties would need to be contained within a group (e.g. radiation therapy should be grouped in the Radiation group). The exception could be the Otolaryngology group. The reason is that when using only the group containing the specialty of the physician in question, there are fewer procedure codes available with which to be confused. Future research would involve improved grouping methods to maximize procedure code similarities among grouped specialties. The best situation, for class grouping, is to retain the largest number of specialties with the best fraudulent detection results.

Table 4.19: Fraud Detection Results with Groups Only (Group Test One)

Group Name	Group F-score	Instances	IOA
Anesthesiology Group	0.97	44	0.36
Cardiology Group	0.86	13	0.54
Chiropractic Group	0.94	148	0.19
Dermatology Group	0.75	11	0.09
Gynecology Group	0.81	40	0.55
Ophthalmology Group	0.96	28	0.04
Otolaryngology Group	0.87	17	0.18
Radiology Group	0.96	4	0.00

Table 4.21 presents the averaged results based on the removal of specific classes

Table 4.20: Fraud Detection Results with Groups and Non-grouped Specialties (Group Test Two)

PROVIDER_TYPE	F-score	Instances	IOA
Ambulance Service Supplier	1.00	2	0.00
Audiologist (billing independently)	0.99	8	0.00
Anesthesiology Group	0.97	44	0.36
Speech Language Pathologist	0.97	3	0.33
Radiology Group	0.96	4	0.00
Pathology	0.95	4	0.00
Ophthalmology Group	0.94	28	0.04
Chiropractic Group	0.94	148	0.18
Occupational therapist	0.92	6	0.17
Podiatry	0.90	47	0.09
Urology	0.90	13	0.23
Otolaryngology Group	0.88	17	0.24
Gastroenterology	0.88	7	0.29
Cardiology Group	0.86	13	0.54
Emergency Medicine	0.81	32	0.47
Gynecology Group	0.81	40	0.55
Nephrology	0.78	1	0.00
Independent Diagnostic Testing Facility	0.78	5	0.40
Dermatology Group	0.76	11	0.09
Orthopedic Surgery	0.76	11	0.91

for each of the criteria[2] compared to the baseline model. Using the original four-class criteria, we notice that there is decent improvement in the overall performance, especially regarding accuracy which was improved by 0.136 over the baseline model. Further investigation into these results show the baseline model had 15 classes with an F-score greater than 0.90, with the original four-class criteria improving the baseline result to 21 classes having over a 0.90 F-score. The chosen classes criteria demonstrates a larger performance improvement over the removal of the original four classes only. This improvement is most likely due to fact that some of the classes removed had very low individual performance scores. Thus, these low-scoring removed classes had no effect on the remaining high-scoring classes. Even with the demonstrated improvements using the chosen classes criteria, there was no noticeable improvement

[2] Original four classes are from our previous work and based on unique procedures that have both a high number of instances and poor classification performance, and chosen classes removes the original four classes plus twelve additional specialties

beyond the original four classes criteria. The classes with relatively large improvements most likely have procedures that are easily confused with those in the removed classes, whereas the classes with little to no improvement indicate that more specialized services are not well represented in the removed classes.

Table 4.21: Class removal results and improvements by criteria

	Original Four	Chosen	Original Four Improvement	Chosen Improvement
Precision	0.572	0.641	0.058	0.127
Recall	0.546	0.623	0.036	0.113
F-score	0.544	0.621	0.047	0.124
G-measure	0.551	0.626	0.047	0.122
Accuracy	0.818	0.843	0.136	0.161

The detailed performance results for the original four classes and chosen classes removal criteria are listed in Tables 4.22 and 4.23. Similar to the previously discussed class grouping strategy, we only test specialties with an F-score of 0.75 or above for evaluating fraud detection results. The baseline model results included 365 instances with an owaIOA of 23.1%, with the removal of the original four classes having 378 instances and a lower owaIOA of 15.9% and the chosen classes criteria, with 397 instances, having a 14.1% owaIOA. For both criteria, we found the IOA for each individual specialty was either decreased or stayed the same when compared to the baseline model. The reason that fraud detection performance decreased using the class removal strategy is because class confusion, related to each specialty's procedures performed, is removed from the dataset giving less specialties for classification. In other words, the only reason specialties were mislabeled was due to the confusion of the low scoring fields and not because these specialties were actually performing as any of the removed classes. Therefore, given our need to predict a physician's specialty, the class removal strategy does not provide any meaningful improvements.

In order to understand the impacts for class removal and grouping, we leverage the best results from both and evaluate the fraud detection performance. Table 4.24 shows the overall performance when mixing the combined groups with the original

Table 4.22: Class Removal (Original Four Classes) Fraud Detection Results

PROVIDER_TYPE	F-score	Instances	IOA
Ambulance Service Supplier	1.00	2	0.00
Chiropractic	1.00	86	0.00
Audiologist (billing independently)	0.99	8	0.00
Physical Therapist	0.99	17	0.00
Speech Language Pathologist	0.97	3	0.33
Dermatology	0.96	11	0.00
Urology	0.96	13	0.08
Pathology	0.95	4	0.00
Diagnostic Radiology	0.94	4	0.25
Otolaryngology	0.94	13	0.23
Cardiology	0.93	13	0.31
Podiatry	0.93	47	0.02
Emergency Medicine	0.93	32	0.50
Gastroenterology	0.92	7	0.14
Occupational therapist	0.92	6	0.17
Nephrology	0.91	1	0.00
Orthopedic Surgery	0.87	11	0.55
Obstetrics/Gynecology	0.86	38	0.42
Allergy/Immunology	0.85	4	0.00
Optometry	0.82	8	0.25
Ophthalmology	0.82	20	0.15
Pulmonary Disease	0.81	10	0.20
Independent Diagnostic Testing Facility	0.78	5	0.00
Clinical Laboratory	0.77	3	0.00
CRNA	0.76	12	0.17

four and the chosen classes removal criteria. We notice that for both there is a greater prediction improvement over the summation of each individual strategy. However, the owaIOA for the combined groups and original four classes is 20.8% with 457 instances, and the combined groups and chosen classes owaIOA is 21.1% with 484 instances. Both results indicate a decrease in performance when compared to the baseline model in terms of owaIOA. Thus, mixing class grouping and removal strategies does not have a positive performance impact.

Table 4.23: Class Removal (Chosen Classes) Fraud Detection Results

PROVIDER_TYPE	F-score	Instances	IOA
Ambulance Service Supplier	1.00	2	0.00
Chiropractic	1.00	86	0.00
Audiologist (billing independently)	0.99	8	0.00
Physical Therapist	0.99	17	0.00
Speech Language Pathologist	0.97	3	0.33
Urology	0.96	13	0.00
Dermatology	0.96	11	0.00
Gastroenterology	0.95	7	0.14
Diagnostic Radiology	0.94	4	0.25
Podiatry	0.94	47	0.02
Pathology	0.94	4	0.00
Emergency Medicine	0.94	32	0.34
Otolaryngology	0.94	13	0.23
Cardiology	0.93	13	0.23
Occupational therapist	0.92	6	0.17
Nephrology	0.91	1	0.00
Orthopedic Surgery	0.89	11	0.45
Allergy/Immunology	0.86	4	0.00
Obstetrics/Gynecology	0.86	38	0.37
Optometry	0.84	8	0.25
Ophthalmology	0.83	20	0.15
Pulmonary Disease	0.81	10	0.30
Clinical Laboratory	0.79	3	0.00
Independent Diagnostic Testing Facility	0.79	5	0.00
CRNA	0.76	12	0.17
Neurology	0.75	19	0.26

Table 4.24: Combined Removal and Grouping Strategy Performance Results

	Combined Averages		Improvement	
	Original Four	Chosen	Original Four	Chosen
Precision	0.599	0.706	0.085	0.192
Recall	0.579	0.693	0.069	0.183
F-score	0.573	0.688	0.076	0.192
G-Measure	0.580	0.694	0.092	0.19
Accuracy	0.864	0.891	0.182	0.209

Class Isolation Method

When employing the class isolation method, we build a model per physician specialty (Chiropractic, Family Practice, General Practice, Internal Medicine, Physician Assistant, and Psychiatry), each with a 3:1 RUS class distribution. Table 4.25 focuses on the results employing data sampling showing the Type I and Type II error rates for all six classes (specialties) separated by learners, with the best performers denoted in boldface. Family Practice using MNB is the only class that fits our previously mentioned goal of balancing error rates while minimizing Type II error, with a Type I error rate of 0.284 and Type II error rate of 0.256. No other learner was able to create a comparable model.

Table 4.25: Summary of Class Isolation Results with Data Sampling

Provider Type	MNB		LR		RF		SVM	
	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
Chiropractic	**0.363**	**0.639**	0.022	0.914	0.174	0.696	0.052	0.992
Family Practice	**0.284**	**0.256**	0.235	0.491	0.174	0.696	0.032	0.759
General Practice	0.137	0.601	**0.339**	**0.583**	0.183	0.520	0.059	0.747
Internal Medicine	0.056	0.569	0.196	0.444	0.101	0.589	**0.351**	**0.667**
Physician Assistant	**0.235**	**0.349**	0.250	0.498	0.052	0.689	0.059	0.633
Psychiatry	**0.220**	**0.485**	0.133	0.649	0.078	0.717	0.023	0.847

In addition to using data sampling, via RUS, to improve fraud detection performance, we use a cost sensitive classifier[3]. The purpose of a cost sensitive classifier is to determine the optimal model by finding the best learner and cost ratio combination for each specialty. Again, we want to have the lowest possible Type I and Type II error rates, thus minimizing both while treating the Type II error as the more important metric. In order to find the best model for each specialty, we first determined the intersection of the Type I and Type II errors for each of the four tested models. We found the intersections for each learner and specialty by randomly choosing one of the ten datasets for a given class. The variation between the ten datasets for each learner was small, thus this process is considered reliable. We found that Random Forest had the best results for all six specialties, and therefore, we use RF to find the optimal cost ratio. After determining where the Type I and Type II error rates intersected using a single dataset from a given specialty, we ran that same cost ratio over all ten datasets, checking for an optimal model. If that particular cost ratio was found to not produce an optimal model, we either increased or decreased the cost for the Type II error only until we reached the optimal cost. Table 4.26

[3] A model built by adjusting the costs associated with Type I and Type II errors

presents the final cost ratios and the F-scores for the optimal models, per specialty. The results for Chiropractic, Physician Assistant, and Psychiatry were similar with error rates around 0.300 and 0.400. Family Practice and General Practice had better results with error rates around 0.200. Internal Medicine had very good results with the lowest error rates. Overall the results of the class isolation method demonstrate strong fraud detection performance. It is particularly notable when compared to the F-score results of either the class removal or grouping strategies.

Table 4.26: Class Isolation Cost Sensitive Classifier Results with RF

Provider Type	Cost		Error		F-Score	
	Non-fraud	Fraud	Type I	Type II	Non-fraud	Fraud
Chiropractic	1	8	0.397	0.405		
	1	8.3	0.403	0.394	0.69	0.43
Family Practice	1	8	0.212	0.216		
	1	8.35	0.222	0.206	0.82	0.68
	1	8.5	0.215	0.223		
General Practice	1	4	0.236	0.293		
	1	5	0.274	0.238	0.8	0.59
Internal Medicine	1	8	0.193	0.126	0.85	0.71
Physician Assistant	1	10	0.262	0.324		
	1	13.5	0.361	0.291	0.74	0.51
	1	14.5	0.379	0.284		
Psychiatry	1	8	0.336	0.297	0.75	0.52

Summary of improvements

A summary of the fraud detection performance results comparing each strategy and the baseline model is shown in Tables 4.27 and 4.28. The results shown in Table 4.27 are for the class grouping and removal strategies with baseline model comparisons only. Table 4.28 summarizes the class isolation method results. With regards to improving performance, the class grouping strategy has produced results indicating improved detection. In class grouping, there were some large improvements over individual specialty results, such as the Ophthalmology and Optometry, and Anesthesiology and Anesthesiologist Assistants groups, while the other groups had moderate to

Table 4.27: Summary of Fraud Detection Results for Baseline and Improvement Strategies

Experiment	Description	owaIOA
Baseline	The learner used is Logistic Regression. The O/F method used was separate (SOF). Only classes that had a F-score above 0.75 were tested for fraudulent detection.	23.1%
Groups	Similar specialties were grouped together. All fourteen groups were tested separately. In this experiment only classes within groups were tested for fraudulent to see how the grouping affected the member within the groupings. Only groups with an F-score above 0.75 were chosen for testing fraudulent detection.	25.2%
Combining Groups	All groups found to yield improved prediction results in grouping were used in one dataset in order to determine the overall effects of grouping. All classes and groupings with an F-score above 0.75 are used for testing fraudulent detection.	26.1%
Class Removal (Original Four)	Classes were removed that were considered to have high overlap of procedures and low F-score. Only classes that had a F-score above 0.75 were tested for fraudulent detection.	15.9%
Class Removal (Chosen)	Along with the O4, classes were removed that created ambiguity, scored low in Precision and Recall, contained medicine, general or unknown. Only classes that had a F-score above 0.75 were tested for fraudulent detection.	14.1%
Mixed (Original Four)	This experimented combined the class removal O4 and combined grouping. Only classes that had a F-score above 0.75 were tested for fraudulent detection.	20.8%
Mixed (Chosen)	This experimented combined the class removal chosen and combined grouping. Only classes that had a F-score above 0.75 were tested for fraudulent detection.	21.1%

minimal improvements. The high scoring groups, with only the group in the dataset, had an owaIOA of 25.2% that was slightly higher than the baseline model result. When the eight specialties were in the dataset for our combined class grouping strategy, which combined groups with moderate to high improvements, we found similar results to grouping separately and produced minimal overall improvement, with a 26.1% owaIOA. The class removal strategy, with the two criteria, showed a significant decline in fraud detection capability compared to the baseline with an owaIOA of 15.9% for the original four classes and 14.1% for the chosen classes criteria. Mixing both combined grouping and class removal showed that the predictive results were increased more than the sum of each strategy alone. The fraud detection results, however, both decreased in comparison to the baseline model with owaIOA scores of 20.8% and 21.1%. However, the class isolation method, summarized in Table 4.28, shows good performance, especially for Internal Medicine with a Type I error of 0.193 and a low Type II error at 0.126. The IOA of fraudulent instances for Internal Medicine is 87.4% indicating that percentage of fraudulent behaviors would be correctly identified as fraudulent, exhibiting promising detection performance. Note that these lower error rates are only achieved using a cost sensitive classifier with an RF model.

Table 4.28: Isolation Method Error Rates

Specialty	Type I	Type II
Chiropractic	0.403	0.394
Family Practice	0.215	0.223
General Practice	0.274	0.238
Internal Medicine	0.193	0.126
Physician Assistant	0.361	0.291
Psychiatry	0.336	0.297

Research limitations

Given the difficulty in detecting Medicare fraud, we briefly summarize some possible limitations of our models and proposed improvement strategies. Overall, this section focuses only on claims information in the Medicare Part B dataset. Also, our fraud detection results depend on possible fraudulent activities being represented by a physician/provider showing claims patterns (from procedures performed) outside of their primary specialty. This, however, does not specifically account for or consider other possible confounding variables that could impact service utilization. For the baseline, grouping, and class removal strategies, we can only detect certain types of fraud where procedure codes are used that do not align with a primary specialty. In grouping similar classes, we could potentially remove a specialty with which another specialty could be classified as, such as Optometrist and Ophthalmology. Finally, class isolation only uses a subset of the entire Part B data, limited by physician type and number of non-fraudulent instances.

4.3.4 Section Summary

Medicare fraud continues to be a problem for the U.S. government and its beneficiaries. Reducing the impact of fraud is critical in helping to reduce costs and provide high quality of service. In this section, we demonstrate, through the use of data mining and machine learning, the successful detection of Part B provider fraud for different medical specialties. In our previous research, we created a unique model to detect fraud by predicting a physician's specialty. If this predicted (or expected) specialty differs from that physician's actual specialty, as listed in the Medicare Part B data, this could be indicative of possible fraud. The reason is that this misclassified physician is not performing procedures in a manner similar to their peers, which is considered to be anomalous. For instance, a physician whose expected specialty differs from their actual specialty could be performing fraudulent acts such as double billing, upcoding [15], or otherwise purposefully coding incorrect procedures. There are many examples of these types of fraudulent behaviors, and the interested reader can find a sample real-world Medicare conviction for upcoding at [45].

Due to the small size of the dataset used in our previous research, we were limited in performing robust fraud detection validation of the models and unable to assess any proposed improvement strategies. With that said, our current research incorporates the full Medicare Part B data and thus has both baseline model and strategy assessments and validation. Because our experiments used Big Data, we employed different methods in building and testing baseline models. We used PySpark to build the Multinomial Naive Bayes and Logistic Regression models, where each of these models can effectively perform multi-class classification on the 89 different specialties. Along with building a baseline model, we also addressed the concern regarding a physician's place of service. We tested both office and facility by either combining (COF) or separating (SOF) the procedures and split the results based on where the procedures were performed. From the model selection process, accounting for the place of service, the Logistic Regression model using the Separate Office or Facility (SOF) was chosen as the best baseline model. From this baseline model configuration, we tested two strategies specifically to assess improvement to the baseline model. Specifically, we tested two strategies in order to increase the performance of the baseline model. Grouping similar specialties and removing specialties with a large number of possibly overlapping procedures, as well as the combination of grouping and removal strategies, produced mixed results. Even so, class grouping showed some improvements but only for certain physician specialties. In addition to the aforementioned improvement strategies, we assessed the class isolation method, using the SOF selected dataset. This method builds separate models per specialty by either using 3:1 (non-fraud to fraud instances) sampled datasets or via cost sensitive classification. Both of these methods perform well and show improved fraud detection performance over the baseline model and associated improvement strategies.

Our proposed fraud detection method and improvement strategies can be used to flag possibly fraudulent physicians based on the procedures performed. These flagged providers are then the focus of further investigation to determine any actual fraudulent behaviors and legal culpability. It is important to note that our detection approach reduces the efforts and resources required to investigate possible fraud by limiting the number of fraud instances to a small subset of all possible providers. Even so, further scrutiny is still recommended to confirm and document any fraud. Through our research, we found that the improvements from our class grouping and class removal strategies were minimal, except for a few of the specialties, and thus not a viable approach for continued research in this area. The class isolation method, however, is promising and applying new approaches and data mining techniques to this method could yield increased fraud detection performance. These improvements to class isolation, such as the use of grouping, are left as future work. Other areas of future work involve adding additional features, as well as increasing the pool of known fraudulent providers.

4.4 CHAPTER SUMMARY

Overall, we have determined that through our anomaly detection approach, it is possible to classify a number medical specialties with sufficiently high accuracy, where a number of specialties had F-scores above 0.90. The hypothesis behind our unique anomaly detection approach was that if a model could be designed which predicts a physician specialty accurately enough, then a physician belonging to that specialty was misclassified, potentially pointing to fraudulent behavior. The datasets used in this Chapter are the Part B datasets which are transformed into a sparse vector where the attributes are the counts that each physician has performed for all available HCPCS codes. This dataset provides the behavior of each physician throughout entire calendar years. This anomaly approach is limited to physician specialties that can be predicted with very high accuracy, increasing the probability that misclassifications are due to behavioral differences and not due to the difficulty of predicting the specialty in question. Therefore, in order to increase the number of specialties with high prediction scores, we experimented with a number of anomaly detection improvement strategies including: feature selection and sampling, grouping specialties, class removal. We found that prediction scores were improved when applying these improvement strategies. After testing our anomaly detection approach on those specialties with high prediction scores, we found that the number of misclassified physicians who were confirmed to be actually fraudulent via the LEIE was unfortunately, at best 26.1%, when applied to the full Part B dataset from 2012 through 2015. Additionally, using this dataset, we also performed an experiment with the class isolation method, which is a traditional fraud detection technique. This technique performed well with one specialty scoring a Type I error rate of 0.193 and a Type II error rate of 0.126. Therefore, in this dissertation, we shift our focus to experiments employing traditional fraud detection.

CHAPTER 5
TRADITIONAL FRAUD DETECTION USING MULTIPLE MEDICARE BIG DATA SOURCES

5.1 BACKGROUND AND MOTIVATION

Healthcare in the United States (U.S.) is important in the lives of many citizens, but unfortunately the high costs of health-related services leave many patients with limited medical care. In response, the U.S. government has established and funded programs, such as Medicare [32], that provide financial assistance for qualifying people to receive needed medical services [50]. There are a number of issues facing healthcare and medical insurance systems, such as a growing population or bad actors (i.e. fraudulent or potentially fraudulent physicians/providers), which reduces allocated funds for these programs. In this chapter, we focus on fraud, and use the word “fraud" to include the terms “waste" and “abuse". Medicare is a federally subsidized medical insurance, and therefore is not a functioning health insurance market in the same way as private healthcare insurance companies [32]. There are two payment systems available through Medicare: Fee-For-Service and Medicare Advantage. For this chapter, we focus on data within the Fee-For-Service system of Medicare where the basic claims process consists of a physician (or other healthcare provider) performing one or more procedures and then submitting a claim to Medicare for payment, rather than directly billing the patient. The second payment system, Medicare Advantage, is obtained through a private company contracted with Medicare, where the private company manages the claims and payment processes [64]. Additional information on the Medicare process and Medicare fraud is provided within [15, 26, 38, 32].

The detection of fraud within healthcare is primarily found through manual effort by auditors or investigators searching through numerous records to find possibly suspicious or fraudulent behaviors [120]. This manual process, with massive amounts of data to sieve through, can be tedious and very inefficient compared to more automated data mining and machine learning approaches for detecting fraud [92, 125]. The volume of information within healthcare continues to increase due to technological advances allowing for the storage of high-volume information, such as in Electronic Health Records (EHR), enabling the use of Big Data. As technology advances and its use increases, so does the ability to perform data mining and machine learning on Big Data, which can improve the state of healthcare and medical insurance programs for patients to receive quality medical care. The Centers for Medicare and Medicaid Services (CMS) joined in this effort by releasing Big Data Medicare datasets to assist in identifying Fraud, Waste and Abuse within Medicare [49]. CMS released a statement that “those intent on abusing Federal health care programs can cost taxpayers billions of dollars while putting beneficiaries’ health and welfare at risk. The impact of these losses and risks magnifies as Medicare continues to serve a growing number of people [36].” There are several datasets available at the Centers for Medicare and Medicaid Services website [33].

In this chapter, we use three Public Use File (PUF) datasets: 1.Part B, 2. Part D, and 3. DMEPOS. We chose these parts of Medicare because they cover a wide range of possible provider claims, the information is presented in similar formats, and they are publicly available. Furthermore, the Part B, Part D, and DMEPOS dataset comprise key components of the Medicare program and by incorporating all three aspects of Medicare for fraud detection, this chapter provides a comprehensive view of fraud in the Medicare program. Information provided in these datasets includes the average amount paid for these services and other data points related to procedures performed, drugs administered, or supplies issued. We also create a dataset combining all three of these Medicare datasets, which we refer to as the Combined dataset. The last dataset examined is the LEIE [91], which contains real-world fraudulent physicians and entities.

These datasets released by CMS exhibit many ‘Big Data’ qualities. The datasets qualify as Big Volume since they contain annual claim records for all physicians submitting to Medicare within the entire United States. Every year, CMS releases the data for a previous year, increasing the Big Volume of available data. The datasets contain around 30 attributes each, ranging from provider demographics and the types of procedures to payment amounts and the number of services performed, thus qualifying as Big Variety. Additionally, the Combined dataset inherently provides Big Variety data, because it combines the three key (but different) Medicare data sources. As CMS is a government program with transparent quality controls and detailed documentation for each dataset, we believe that these datasets are dependable, valid, and representative of all known Medicare provider claims indicating Big Veracity. Through research conducted by our research group and others, it is evident that this data can be used to detect fraudulent behavior giving it Big Value. Furthermore, the LEIE dataset could also be considered as Big Value since it contains the largest known repository of real-world fraudulent medical providers in the United States.

In this chapter, we provide analyses to show the best learners and datasets for detecting Medicare provider claims fraud [70]. Our unique data processing steps consist of data imputation, determining which variables (dataset features) to keep, transforming the data from the procedure-level to the provider-level through aggregation to match the level of the LEIE dataset for fraud label mapping, and creating the Combined dataset. Note that the fraud labels are used to assess fraud leveraging historical exclusion information, as well as payments made by Medicare to currently excluded providers. The resulting processed datasets are considered Big Data and thus, for our fraud detection experiments, we employ Apache Spark [4] on top of a

Hadoop [3] YARN cluster which can effectively handle these large dataset sizes. For our experiments, the four Medicare datasets were trained and validated using 5-fold cross-validation, and the process was repeated ten times. From the Apache Spark [152, 153] Machine Learning Library, we build the RF, GTB and LR models, and use the AUC metric to gauge fraud detection performance. We chose these learners, as they are commonly used and provide reasonably good performance, for our exploratory analysis to assess fraud detection performance using Big Data in Medicare. In order to add robustness around the results, we estimate statistical significance with the ANalysis Of VAriance (ANOVA) [55] and Tukey's Honest Significant Difference (HSD) tests [137]. Our results indicate that the Combined dataset with LR resulted in the highest overall AUC with 0.816, while the Part B dataset with LR was the next best with 0.805. Additionally, the Part B dataset had the best results for GBT and RF with both resulting in a 0.796 AUC. The worst fraud detection results were attributed to the DMEPOS dataset, with RF having the lowest overall AUC of 0.708. The results for the Combined dataset using LR, indicate better performance than any individual Medicare dataset; thus, the whole in this case is better than the sum of its parts. This, however, is not the case for RF or GBT with Part B having the highest average AUC. Even so, the Combined dataset showed no statistical difference when compared to the Part B dataset results. Therefore, the high fraud detection results, paired with our assumption that Medicare fraud can be committed in any or all parts of Medicare, demonstrates the potential in using the Combined dataset to successfully detect provider claims fraud across learners. To summarize, the unique contributions of this chapter are as follows:

- Experimenting with our unique Medicare Part B, Part D, and DMEPOS data processing and real-world fraud label mapping

- Combining the three Medicare big datasets into one Combined dataset in order to demonstrate high fraud detection performance that takes into account the different key parts of Medicare

- Exploring fraud detection performance and learner behavior for each of the four big datasets

The rest of this chapter is organized as follows. Section 5.2 covers related works, focusing on works employing multiple CMS branches of Medicare. Section 5.3 discusses the different Medicare datasets used, how the data is processed, and the fraud label mapping approach. Section 5.4 details the methods used including the learners, performance metric, and hypothesis testing. Section 5.5 discusses the results of our experiment. Finally, we conclude and discuss future work in Section 5.6.

5.2 RELATED WORKS

There have been a number of studies conducted, by our research group and others, using Public Use Files (PUF) data from CMS in assessing potential fraudulent activities through data mining and other analytics methods. The vast majority of these studies use only Part B data [5, 6, 8, 14, 21, 49, 68, 86], neglecting to account for other parts of Medicare when detecting fraudulent behavior. Within the healthcare system, anywhere money is being exchanged, there is an opportunity for a bad actor to manipulate the process and siphon funds, affecting the efficiency and effectiveness of the Medicare healthcare process. There is limited prior information as to where (in the Medicare system) a physician will commit fraud, so choosing a single part of Medicare could miss fraud committed elsewhere. In this chapter, we focus on evaluating fraud detection performance for multiple Medicare datasets. Therefore, we generally limit our discussion in this section to the small body of works attempting to identify fraudulent behavior using multiple CMS datasets. We only found two works [18] and [124] that fall under that category.

In [18], Branting et al. use the Part B (2012 - 2014), Part D (2013) and LEIE dataset. They do not specifically mention how they preprocess the data or combine Part B and Part D, but they do take attributes from both Part B and Part D datasets, treating drugs and HCPCS codes in the same way. They matched 12,153 fraudulent physicians using the National Provider Identifier (NPI) [28] with their unique identity-matching algorithm. They decided against distinguishing between LEIE exclusion rules/codes and instead used every listed physician. It is unclear whether the authors accounted for waivers, exclusion start dates or the length of the associated exclusion during their fraud label mapping process. These details are important in reducing redundant and overlapping exclusion labels and for assessing accurate fraud detection performance. Therefore, due to this lack of clarity in the exclusion labeling methodology, the results from their study cannot be reliably reproduced and can be difficult to compare to other research. They developed a method for pinpointing fraudulent behavior by determining the fraud risk through the application of network algorithms from graphs. Due to the highly imbalanced nature of the data, the authors used a 50:50 class distribution, retaining 12,000 excluded providers while randomly selecting 12,000 non-excluded providers. They put forth a few groups of algorithms and determined their fraud detection results based on the real-world fraudulent physicians found in the LEIE dataset. One set of algorithms, which they denote as Behavior-Vector Similarity, determines similarity in behavior for real-world fraudulent and non-fraudulent physicians using nominal values such as drug prescriptions and medical procedures. Another group of algorithms makes up their Risk Propagation, which uses geospatial co-location (such as location of practice) in order to estimate the propagation of risk from fraudulent healthcare providers. An ablation analysis [24] showed that most of this predictive accuracy was the result of features that measure risk propagation through geospatial collocation.

Sadiq et al. [124] use the 2014 CMS Part B, Part D and DMEPOS datasets (using only the provider claims from Florida) in order to find anomalies that possibly point to fraudulent or other interesting behavior. The authors do not go into detail on how they preprocessed the data between these datasets. From their study, we can assume the authors use, at minimum, the following features: NPI, gender, location (state, city, address etc.), type, service number, average submitted charge amount, the average allowed amount in Medicare and the average standard amount in Medicare. It is also unclear as to whether they used the datasets together or separately or which attributes were used and which were not, making the reproduction of these experiments difficult. The authors determine that when dealing with payment variables, it is best to go state-by-state as each state's data can vary. However, in this chapter, we found that good results can be achieved by using Medicare data encompassing the entire U.S. The framework they employ is the Patient Rule Induction Method based bump hunting method, which is an unsupervised approach attempting to determine peak anomalies by spotting spaces of higher modes and masses within the dataset. They explain that by applying their framework, they can characterize the attribute space of the CMS datasets helping to uncover the events provoking financial loss.

We note a number of differences from these two studies [18] and [124], including data processing methods, the process for data combining and comparisons made between the three Medicare datasets both individually and combined. In this chapter we compare fraud detection within three different Medicare big datasets, as well as a Combined version of the three primary Medicare datasets. Additionally, we incorporate all available years in each CMS dataset covering the entire United States, requiring us to incorporate software which can handle Big Data.

5.3 DATA

In this chapter, we employ three CMS datasets Part B, Part D, and DMEPOS. The Part B dataset contains information for calendar years 2012 through 2015, while Part D and DMEPOS contain information for 2013 through 2015. Each part was processed individually and combined using our unique processing methodology described in Section 2.3.2. This process includes processing, fraud label mapping between the Medicare datasets and the LEIE, the design of our Combined dataset, and one-hot encoding for categorical variables. The details for the four final datasets used in this chapter are presented in Table 2.11, showing number of features after one-hot encoding, number of non-fraudulent and fraudulent physicians as well as the percentage of fraudulent physicians. All four datasets are severely imbalanced, ranging between 0.038% (Part B) and 0.074% (Combined) of instances being labeled as fraud.

5.4 METHODS

For running and validating models, we used Apache Spark on top of a Hadoop Yarn cluster due to the Big Volume of the datasets. We used three classification models available in the Apache Spark Machine Learning Library: Logistic Regression, Gradient Boosted Trees and Random Forest. In Section 3.2, we describe each learner and note any configuration changes that differ from the default settings.

In assessing Medicare fraud, we are presented with a two-class classification problem where a physician is either fraudulent or non-fraudulent (traditional fraud detection). In this chapter, the positive class, or class of interest, is fraud and the negative class is non-fraud. Spark presented us with a confusion matrix for each model and is commonly used to assess the performance of learners. Confusion matrices provide counts comparing actual counts against predicted counts. From the resultant matrices, we employ AUC [16, 129] to measure fraud detection performance. In order to provide additional rigor around our AUC performance results, we use hypothesis testing to show the statistical significance of the Medicare fraud detection results. Both ANOVA and post hoc analysis via Tukey's HSD tests are used. Detailed descriptions of AUC, ANOVA and Tukey's HSD tests are provided in Section 3.3.2.

We employ stratified k-fold cross-validation for evaluating our models, where k =

5, and is discussed in Section 3.1.2. Spark will automatically create different folds each time the learner is run, and to validate our results we ran each model 10 times for each learner/dataset pair to reduce bias due to bad random draws when creating the folds. The final performance for every presented result is the average over all 10 repeats.

5.5 RESULTS AND DISCUSSION

This section discusses the results for this chapter, assessing dataset and learner performance for Medicare fraud detection. The practices of individual physicians are unique, where a given physician might only submit claims to Medicare through Part B, Part D, DMEPOS, or to all three. Therefore, we show learner performance in relation to each of the Medicare datasets to establish the best fraud detection combinations. In Table 5.1, we show the AUC results for each dataset and learner combination. The boldfaced values depict the highest AUC scores per dataset, whereas the underlined values are the highest per learner. LR produces the two highest overall AUC scores, with 0.816 for the Combined dataset and 0.805 for Part B. The Combined dataset has the best overall AUC, but the Part B dataset shows the lowest variation in fraud detection performance across learners, which includes having the highest AUC scores for GBT and RF. The Part D and DMEPOS datasets have the lowest AUC values for all three learners, but show improvement when using LR and GBT compared to RF.

Dataset	Logistic Regression	Gradient Boosted Trees	Random Forest
Combined	**0.81554**	0.79047	0.79383
Part B	**0.80516**	0.79569	0.79604
Part D	**0.78164**	0.74851	0.70888
DMEPOS	**0.74063**	0.73129	0.70756

Table 5.1: Learner AUC results by Dataset

The favorable results using LR with each of the datasets may be due to the squared-error loss function with the application of L2 regularization, also known as

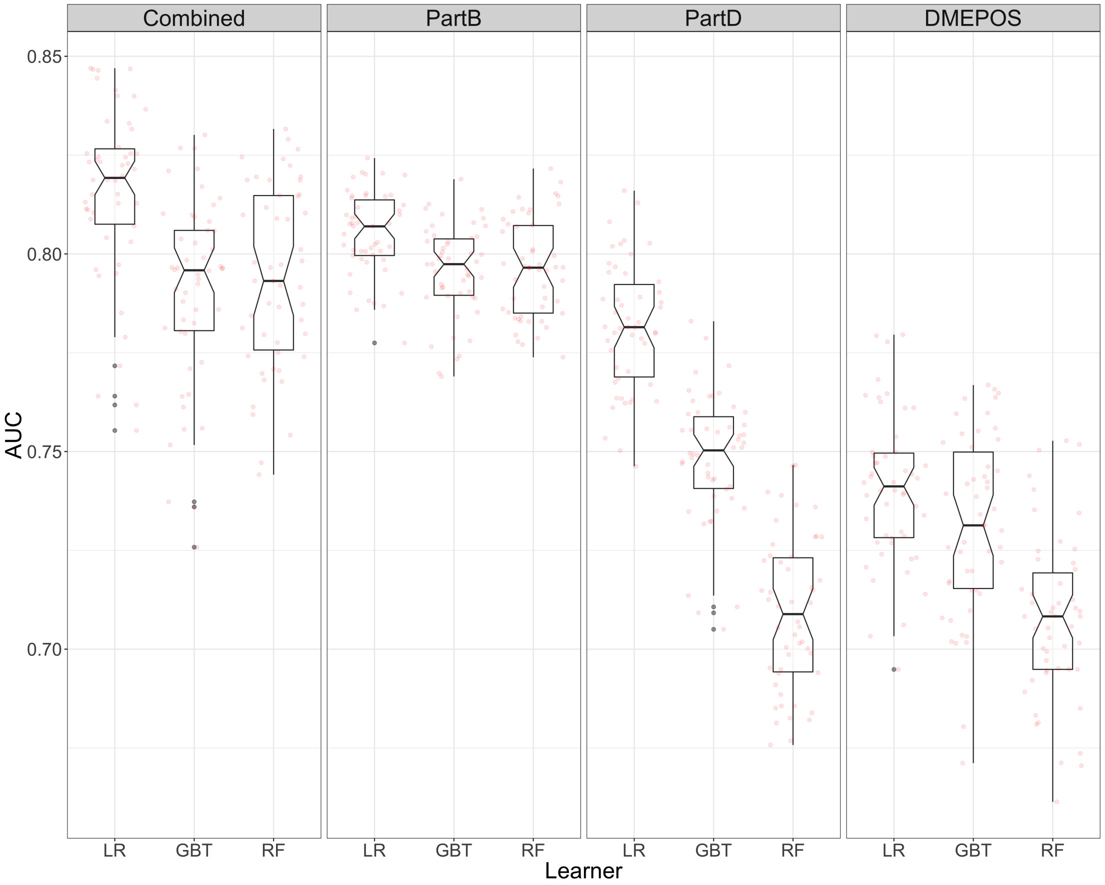


Figure 5.1: AUC values for 10 runs of 5-fold cross-validation for each learner and dataset combination

Ridge Regression, penalizing large coefficients and improving the generalization performance. This makes LR fairly robust to noise and overfitting. Even though LR performs well on the Part B and Combined datasets, additional testing is required to determine whether the Part D and DMEPOS datasets have particular characteristics contributing to their lower fraud detection performance. The poor performance of the tree-based methods, particularly RF, may be due to the lack of independence between individual trees or the high cardinality of the categorical variables. The Combined dataset contains features across the three parts of Medicare, creating a robust pool of attributes, presumably allowing for better model generalization and overall fraud detection performance. In particular, the Combined dataset using LR has the highest AUC with better performance versus each of its individual Medicare parts. This is not the case with RF or GBT, with Part B indicating the highest AUC scores. Interestingly, the Part B dataset has the lowest variability across the learners and within each individual learner, which could be due, in part, to having the largest number of fraud labels. The Part D and DMEPOS datasets not only show poor learner performance, but exhibit generally higher AUC variability across individual learners. This could indicate possible adverse effects of high class imbalance or less discriminatory power in the selected features. With regards to our above discussions, Figure 5.1 shows a box plot of our experimental results over all 50 AUC values from the ten runs of 5-fold cross-validation for each dataset/learner pair.

Table 5.2 presents the results for the two-factor ANOVA test over each Dataset and Learner, as well as their interaction (Dataset:Learner). The ANOVA test shows that these factors and their interactions are statistically significant at a 95% confidence interval. In order to determine statistical groupings, we perform a Tukey's HSD test on the results for the Medicare datasets, which corroborates the high performance of the LR learner and the Combined dataset for Medicare fraud detection (as seen in Table 5.1).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Dataset	3	0.6257	0.20855	594.15	<2e-16
Learner	2	0.1174	0.05868	167.17	<2e-16
Dataset:Learner	6	0.0658	0.01097	31.26	<2e-16
Residuals	588	0.2064	0.00035	-	-

Table 5.2: Two-factor ANOVA Test Results

In Table 5.3, the results for each learner across all datasets show that LR is significantly better than GBT and RF. Moreover, LR and GBT have similar AUC variability, but LR has the highest minimum and maximum AUC scores which, again, substantiate the good performance of LR for each dataset. Table 5.4 summarizes the significance of dataset performance across each learner. We notice that the Combined and Part B datasets show significantly better performance than either the Part D or DMEPOS datasets, and that the DMEPOS dataset is significantly worse than the

Part D dataset. Since the Part B and Combined results are not significantly different, we consider the Combined dataset preferable for general fraud detection since we do not necessarily know beforehand exactly which part of the Medicare system a physician/provider will target with their fraudulent behavior (e.g. medical procedures/services, drug submissions, or prosthetic rental). With the Combined dataset, we have a larger web for monitoring fraudulent behavior as opposed to monitoring only one part of Medicare for a given healthcare provider. Additionally, the Combined dataset with LR provides the only results where the Combined dataset produces the best performance, greater than the results for the individual Medicare datasets. Therefore, based on these exploratory performance results, we demonstrate that when a physician has participated in Part B, Part D, and DMEPOS, the Combined dataset, using LR, indicates the best overall fraud detection performance.

Learner	Group	AUC	sd	r	Min	Max
Logistic Regression	a	0.78574	0.03369	200	0.69487	0.847
Gradient Boosted Trees	b	0.76649	0.03343	200	0.67119	0.83013
Random Forest	c	0.75158	0.04753	200	0.66138	0.83161

Table 5.3: Two-factor Tukey's HSD learner results over all datasets

Dataset	Group	AUC	sd	r	Min	Max
Combined	a	0.79995	0.02549	150	0.7258	0.847
Part B	a	0.79896	0.0123	150	0.769	0.82425
Part D	b	0.74634	0.03443	150	0.67576	0.81602
DMEPOS	c	0.72649	0.02506	150	0.66138	0.77957

Table 5.4: Two-factor Tukey's HSD dataset results over all learners

5.6 CHAPTER SUMMARY

The importance of reducing Medicare fraud, in particular for individuals 65 and older, is paramount in the United States as the elderly population continues to grow. Medicare is necessary for many citizens, and therefore, quality research into fraud detection to keep healthcare costs fair and reasonable is very important. CMS has made available several Big Data Medicare claims datasets for public use over an ever-increasing number of years. Throughout this chapter, we use our unique approach (combining multiple Medicare datasets and leverage state-of-the-art Big Data processing and machine learning approaches) for determining the fraud detection capabilities of three Medicare datasets, individually and combined. We employ three learners, against real-world fraudulent physicians and other medical providers taken from the LEIE dataset.

We ran experiments on all four datasets: Part B, Part D, DMEPOS, and Combined. Each dataset was considered Big Data, requiring us to employ Spark on top of a Hadoop YARN cluster for running and validating our models. Each dataset was trained and evaluated using three learners: Random Forest, Gradient Boosted Trees and Logistic Regression. The Combined dataset had the best overall fraud detection performance with an AUC of 0.816 using LR, indicating better performance than each of its individual Medicare parts, and scored similarly to Part B with no significant difference in average AUC. The DMEPOS dataset had the lowest overall results for all learners. Therefore, from these experimental findings and observations, coupled with the notion that a physician/provider can commit fraud using any part of Medicare, we show that using the Combined dataset with LR provides the best overall fraud detection performance.

CHAPTER 6
COMPARING MODEL EVALUATION METHODS WITH TRADITIONAL FRAUD DETECTION

6.1 THE EVALUATION OF MEDICARE FRAUD PREDICTIVE MODELS

The quality of life for many individuals in the United States (U.S.) depends on access to quality healthcare but for many patients, associated costs, such as exorbitantly priced procedures or medication, can interfere. Programs, such as the U.S. government established Medicare program [32], assist in alleviating financial burden for its beneficiaries, but issues like fraud reduce available funding. Unfortunately, current methods are not significantly decreasing monetary losses facing the U.S. healthcare system. In an effort to detect potential healthcare fraud, we use three publicly available Medicare datasets released by CMS: Part B, Part D and DMEPOS. These datasets comprise key components of Medicare, containing a wide array of claims information providing a comprehensive view of Medicare fraud and delivering an encompassing perspective into physician behavior. All three were organized in a similar manner, and therefore, used in this section, as we also create a Combined dataset. The information available within these datasets includes payment information for claims submitted to Medicare (i.e. submitted charges), payments made by Medicare (i.e. average payments), and other data points related to procedures performed, drugs administered, or supplies issued. Currently, CMS does not provide fraud labels with their Medicare datasets. To solve this, we utilize the LEIE [108] to generate fraud labels. Additional details concerning Medicare and Medicare fraud are provided in

[15, 26, 32, 36, 38].

Big Data provides a major challenge for identifying fraud, especially when traditionally, this process was carried out by a limited number of investigators, manually inspecting massive amounts of claims, allowing only enough time to identify suspicious behaviors with very specific patterns. In other words, successfully detecting fraud with such copious amounts of data manually is impractical. Therefore, we employ machine learning methods for detecting fraud [81]. The LEIE contains the largest known repository of real-world fraudulent medical providers in the United States. The Medicare datasets we use, after the addition of the fraud labels via the LEIE, are considered severely imbalanced. Class imbalance can be defined as a dataset where class labels have an unequal ratio [62, 127], where for this section, our class labels are fraud (positive) and non-fraud (negative). Severe imbalance occurs when this ratio, between positive and negative classes, becomes exceedingly one-sided. Commonly, in real-world data sources, the class comprising the minority, of a dataset, is the class of interest (positive) and this holds true for Medicare fraud as most claims are not considered to be fraudulent. When machine learning is applied to datasets with class imbalance, there is a difficulty in discriminating positive class members due to minimal data representation. Big Data can exacerbate class imbalance with an enormous over-representation of the negative class [82].

Many practitioners and researchers use various evaluation methods in order to estimate generalization error and to determine the best performing model/learner [17]. Two prevalent evaluation methods for validating machine leaning models include: 1) building the model on a training dataset and evaluating on a separate, distinct test dataset containing a full year of data (Train_Test), and 2) CV which subsets a single dataset into smaller training and test datasets for building and evaluation, allowing for the assessment of prediction performance without a separate test dataset. There are various versions of CV, such as k-fold, stratified k-fold, leave-p-out, and holdout

[117]. In this section, due to its usefulness and ubiquity in machine learning, we employ stratified k-fold cross-validation. Technically, Train_Test can be seen as a hold-out CV. However, for Train_Test, we separate and use both the training datasets (prior to 2016) and test datasets (2016 only) for evaluation, whereas CV only uses the training datasets. Train_Test and CV emulate the process real-world practitioners would employ when using machine learning to detect fraud, by learning on known, historical data (training) in order to provide a model that can classify new, unknown data (test). CV can be useful when a researcher only has access to prior data. This, however, poses the question as to how accurate are CV predictions compared to Train_Test. CV may not be as reliable as Train_Test, with Train_Test generally providing more accurate results over CV [80, 99]. Furthermore, CV is susceptible to overfitting (variance), due to each instance, at some point, being used both for model building and evaluation. This can be lessened when using larger data, which allows for larger training and test datasets, but class imbalance can negate this benefit due to the reduced number of positive class instances in each training dataset. Also, there is the possibility of bias being introduced into CV, as learners are built on a reduced training dataset, which limits the model's discriminatory ability in finding meaningful trends and patterns [117]. Rao et al. [119] provide the argument that, "Any modeling decisions based upon experiments on the training set, even cross validation estimates, are suspect, until independently verified [by a completely new Test dataset]." Having a separate test dataset for validation (Train_Test) is important and may produce different results over CV, thus our focus on CV and test dataset model evaluations.

We assess overall fraud prediction performance using our big, severely imbalanced Medicare datasets to compare Train_Test and CV. We provide our recommendations based on the results of these experiments, over three different learners, across all Medicare parts and a Combined dataset. The results of these comparisons are evaluated using the AUC, with differences assessed via significance testing. Our results demonstrate that the Train_Test evaluation method outperforms CV for the majority of dataset/learner pairs. We determine that CV, although slightly conservative, is comparable to Train_Test. Overall, the evaluation of predictive models indicates good fraud detection performance on the latest 2016 test dataset. Additionally, to reduce the negative effects of class imbalance, we perform data sampling, using the higher scoring Train_Test models, and generate five datasets with various class ratios ranging from no sampling to fully balanced (i.e. equal class distributions), and provide recommendations based on this experiment. In general, sampling exhibits significantly better performance across the dataset/learner pairs compared to non-sampling, but results start to decrease as the datasets become more balanced. This section is Medicare fraud domain-focused, incorporating three big, severely imbalanced datasets to evaluate the ability of a classifier built with training data and evaluated on new, distinct test data, and to compare these Train_Test results to CV estimates [69]. The main contributions of this section are:

- Establish whether CV can be reliably used to evaluate fraud detection performance using Medicare claims datasets, by comparing results with Train_Test
- Determine the degree for which data sampling can help mitigate the negative effects of class imbalance and improve fraud detection performance in the Train_Test experiments

The rest of this section is organized as follows. Section 6.1.1 presents related works, discussing studies on Medicare fraud detection and problems related to CV. We discuss the Medicare and LEIE datasets, and summarize our data processing and fraud label mapping in Section 6.1.3. Section 6.1.4 outlines the learners and performance evaluation methods used, as well as class imbalance and the CV and Train_Test experiments. Section 6.1.5 presents the results of our experiments. Conclusions and future work are presented in Section 6.1.6.

6.1.1 Related Works

Medicare Fraud Detection

Since CMS began releasing their big Medicare datasets, there have been a number of studies pursuing the detection of fraudulent activity employing data mining, machine learning, and other analytical methods [132, 144], where a considerable majority of these studies use only the Part B data [6, 8, 14, 21, 49, 68, 86]. Utilizing only one part of Medicare can limit a study's comprehensiveness and the overall applicability of the results. Therefore, we limit our discussion to studies that employ more than one Medicare big dataset.

Branting et al. [18] conducted a study utilizing the Part B (2012 - 2014) and Part D (2013) datasets, while generating fraud labels from the LEIE. Through their identity-matching algorithm centering around a physician's National Provider Identifier (NPI) [28], they matched over 12,000 fraudulent physicians. Due to the large size and significant imbalance, they limit their dataset(s), through sampling, to achieve a 50:50 balanced class ratio. The authors did not experiment with any further ratios, which may have produced better results. They developed a unique method for discriminating fraudulent behavior by determining the fraud risk through the application of algorithms containing graph-based features in conjunction with a decision tree learner. Their overall results demonstrate good fraud detection results using Medicare claims data through machine learning. Sadiq et al. [124] use 2014 Florida only data from the CMS Part B, Part D and DMEPOS datasets in an effort to find anomalies that possibly indicate fraudulent or other interesting behavior. The authors utilize the Patient Rule Induction Method based bump hunting method, which is an unsupervised approach attempting to determine peak anomalies by spotting spaces of higher modes and masses within the dataset. They determine that this method can accurately characterize the attribute space of CMS datasets and can uncover the events contributing to financial loss. In [70], we present a preliminary study, not focused on class imbalance, to determine which part of Medicare better predicts fraudulent behavior where fraud labels are determined by the LEIE, where Part B and the Combined dataset were superior. In contrast to these related works, we incorporate three big Medicare datasets, as well as a Combined dataset, using three different learners to compare Train_Test and CV. We also experiment with various sampling ratios, thoroughly assessing the capabilities of the Train_Test method.

6.1.2 Cross-validation

CV is very popular among the Data Mining and Machine Learning community [17] as an evaluation method for predicting performance in nearly every application domain. We focus on a selection of studies that evaluate the viability of CV as a reliable model evaluation method and specifically outline shortcomings of CV, including instructor notes [80].

Domke [80] discusses the inverse relationship between bias and variance experienced when employing CV. Bias occurs when the model built lacks discrimination between classes and does not fully recognize the full complexities of the data. Variance occurs when the model built is overly discriminatory becoming overfit. Optimal CV models will balance this trade-off in order to minimize error. In [119], Rao et al. provide experimentation demonstrating the impact of large sets of algorithms and increasingly large data dimensionality on Leave One Out Cross-Validation (LOOCV) results, specifically on its ability to measure generalization of performance. LOOCV is the same process as k-fold CV, where the number of folds is one (k=n). The authors utilize an array of data including a synthetically generated, standardized benchmark dataset (from the UCI repository) and a real-world dataset dealing with clinical diagnosis based on virtual colonoscopy. They determine that as sample size decreases and the number of algorithms and data dimensionality increase, the effectiveness of CV to estimate generalization becomes less reliable. They provide both recommendations and warnings when employing CV for machine learning. They recommend validating CV results with a separate test set. They warn that when a model is tuned based upon performance on a test set, this is no longer viable for simulating a real-world scenario, and the test set should only be used for evaluation. In this section, we use the 2016 CMS datasets as the test sets in order to evaluate the viability of CV. The authors do not consider Big Data or class imbalance, which could provide more insight into CV.

Varoquaux [142], conduct a study using brain image analysis to promote awareness for the shortcomings when applying CV, specifically for LOOCV and 80/20 Train/Test splits with 50 repeats. They determine that for small sample sizes, CV results in large errors. In [88], Kodovsky performs a study in the field of steganalysis, presenting the risks involved in utilizing CV, specifically in the JPEG domain. They demonstrate that k-fold CV results are inadequate with a significant difference between predicted error and real testing error. Bengio et al. [17] note k-fold CV is susceptible to large degrees of variability, potentially misleading a researcher's decision during model selection. They demonstrate that determining the level of variance in k-fold CV is challenging and that there exists no variance estimation technique without bias. Their results show, in very simple cases, the bias centered around ignoring the dependencies between test errors will be relatively equal to the quantity of variance. Forman et al. [53] bring forth the issue that there is no universal calculation method for performance metrics such as accuracy, F-measure, and AUC in correlation to CV. This disparity causes deviation in results among studies. Their results determine that generally the measurements introduce bias, particularly when applied to high class imbalance. Westerhuis et al. [149] use permutation testing and CV in the metabolomics domain using Partial Least Squares Discriminant Analysis (PLSDA). They determine that, if applied improperly, CV results become overly optimistic. None of these studies assess or use Big Data in Medicare fraud detection, but if these issues can be found in these other domains, it is entirely possible there are issues in the domain of Medicare fraud, which can lead CV to select sub-optimal models.

6.1.3 Data

The CMS Medicare datasets used in our experiments are derived from administrative claims data for Medicare beneficiaries enrolled in the Fee-For-Service program. We use three publicly available Medicare datasets maintained by CMS: Part B [27], Part D [29], and DMEPOS [30]. CMS records all claims information after payments are made [39, 40, 41]; therefore, we assume the Part B, Part D and DMEPOS data is already cleansed and correct. Additionally, we create a Combined dataset that contains the features from all three Medicare parts. Our research is focused on predicting fraudulent behavior as it appears in real-world medical practice. Therefore, we utilize the LEIE, which contains physicians that have committed real-world fraud allowing for an accurate assessment of fraud detection performance. All four datasets are supplemented with fraud labels using the LEIE. In practice, a physician can submit claims to multiple Medicare parts and there is no dependable way to determine within which part a malicious actor will commit fraud. Through combining information relating to procedures, drugs and equipment, we can provide additional information for each physician to detect possible fraud. One limitation to combining datasets is that there is no guarantee a physician will submit claims to multiple Medicare parts. Each Medicare dataset contains a number of features, but we are only interested in those specifically related to claims information and a select physician-specific data points that are readily usable by machine learning models. Table 2.10 demonstrates the feature used in this section. The exclusion feature is generated through mapping to the LEIE creating the fraud or non-fraud labels for classifying physicians. For each dataset in this section, we design separate training and test datasets in order to perform Train_Test. The training datasets contain all available years 2015 and prior, and the test datasets are derived from the 2016 data. The CV experiments employ only the training datasets, as they evaluate models built on single datasets and do not employ a separate test dataset. In an effort to provide comprehensive dataset-related details, in Section 2.3.2, we provide discussions on our unique data processing, including the datasets, data preparation and feature engineering, fraud labeling, and the differences between the training and test datasets. Table 2.11 summarizes each training dataset and Table 2.13 summarizes each test dataset used in this section. Both tables list the number of features, number of fraudulent and non-fraudulent instances, and the percentage of fraudulent cases after aggregation, one-hot-encoding, and fraud labeling. The main difference between the training and test datasets are in the provider type labels, which is discussed in Section 2.3.2.

6.1.4 Experimental Design

For running and validating models, we employ Spark on top of a Hadoop [3] YARN cluster which can effectively handle the large datasets presented by Medicare. In this section, we use the Apache Spark [152, 153] Machine Learning Library [102]. From this library, we select one non-tree learner, LR, and two tree-based learners, RF and GTB. A description for each machine learning algorithm is available in Section 3.2. Unless specified otherwise as outlined in Section 3.2, we used default configurations for each learner. We use AUC [16] as our evaluation metric, in this section, to demonstrate the fraud detection capabilities of each model. AUC has been found to be an effective metric for quantifying results for studies employing datasets with class imbalance [77]. A description of AUC is presented in Section 3.3.2.

Class imbalance presents issues for machine learning algorithms, due to vast differences in the number of majority and minority class instances, creating a bias towards the majority class. Table 3.3a in Section 3.1.2 demonstrates the degree of imbalance for each dataset by year. One interesting observation is that the number and percentage of fraudulent instances decrease each year. This decrease could be attributed to a number of reasons including continued efforts to remove fraudulent physicians from practice, physicians being able to avoid fraud detection, law enforcement shifting focus from physician fraud to other kinds of fraud, and technological advances in fraud detection. We have also seen, through our studies using these Medicare datasets, that the non-fraud cases generally increase each year at a larger volume than the fraudulent cases are decreasing, furthering the imbalance. In order to curtail the negative effects of class imbalance, we employ RUS to our datasets to generate new datasets with varying degrees of class representation. When using Big Data, the goal is to find a balance between removing the maximum number of majority instances while incurring minimal information loss. Given this, we selected the following class ratios: 1:99, 10:90, 25:75, 35:65, and 50:50 (minority:majority), including the full datasets (all:all) as a baseline where RUS is not applied. These ratios were chosen because they provide a good distribution of classes, ranging from balanced 50:50 to highly imbalanced 1:99. Note that when applying RUS for Train_Test, only the training datasets are sampled, for model building, with the test sets unaltered for model evaluation.

As mentioned, two different model evaluation methods are compared in this section: CV and Train_Test. For our experiment, we employ stratified k-fold CV where k=5. Both Train_Test and CV are described in detail in Section 3.1.2. In order to validate the CV results, we repeat the CV process 10 times for each learner/dataset pair, where the final AUC score is the average over these repeats. We repeat CV to reduce any bias caused by bad random draws when creating folds. For Train_Test, the instances in the test set are completely new instances, never used in model building, as opposed to CV. Train_Test is necessary for real-world applications answering whether, based on past occurrences, a model accurately predicts new occurrences. This method will determine whether, based on prior information (year $<$ 2016), physicians can be classified as fraud or non-fraud given new information (year = 2016).

6.1.5 Discussion and Results

In this subsection, we discuss our experimental results and provide our recommendations for researchers and practitioners looking to detect Medicare fraud through machine learning. Table 6.1 provides the average AUC scores for the Train_Test and Train_CV methods, without data sampling. The boldfaced values denote the dataset producing the best fraud detection performance per learner, the underlined shows the best performing learner per dataset, and the highlighted value shows the highest overall AUC score. These scores demonstrate that for all but one case, LR is the best learner, with the Combined dataset having the best overall results.

Table 6.1: Average AUC Results by Method

Method	Learner	Part B	Part D	DMEPOS	Combined
Train_Test	Gradient Boosted Trees	0.78636	0.74969	0.78281	**0.83654**
	Logistic Regression	0.82133	0.79566	0.78088	**0.86888**
	Random Forest	0.75725	0.69302	0.77105	**0.80122**
Train_CV	Gradient Boosted Trees	**0.79446**	0.74717	0.72789	0.79399
	Logistic Regression	0.80570	0.78174	0.74120	**0.81252**
	Random Forest	**0.79623**	0.70935	0.70525	0.79345

In Figure 6.1, we present a bar plot comparing differences in average AUC scores. This bar graph represents the average AUC score for Train_Test minus the average AUC score for Train_CV, for each learner. The graph demonstrates that Train_Test outperforms Train_CV in almost all cases, including LR for all datasets, signifying that the models built on previous years and evaluated using a new year provide better results. These results could be due to Train_CV having either a high degree of bias or variance. Another possibility that there is a benefit for Train_Test, with imbalanced data, is that the model is trained using all positive instances (fraudulent) allowing for a lesser chance of overfitting the majority class (non-fraud), whereas CV only builds models with a sub-sample of positive instances in each fold. Therefore, detecting fraud on a new year of Medicare claims data based on previous year's data is preferable to CV.

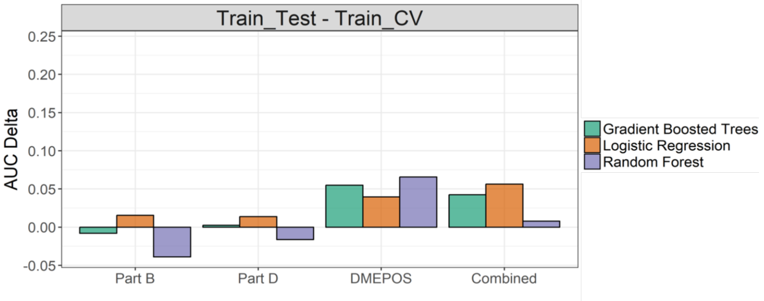


Figure 6.1: Average AUC for Train_Test versus CV

Table 6.2 presents the average AUC scores for Train_Test with the incorporation of RUS, where the boldfaced values indicate the highest scoring learner/ratio pair. Upon applying sampling, we observe that LR consistently delivers the best scores across datasets, aside from DMEPOS where the tree-based algorithms score higher. These results may be due to LR's better handling of class imbalance due to its employment of regularization, which penalizes large coefficients (through Ridge Regression), diminishing the effects of overfitting and increasing model generalization. The tree-based learners also have built-in methods to reduce overfitting, but demonstrate more difficulty in handling class imbalance. The 10:90 ratio scores highest across datasets aside from the Combined, where all:all and 1:99 perform slightly better. The Combined dataset provides the best results for the majority of learner/ratio pairs, with LR and all:all having the highest overall performance. One possible reason for this observation is the Combined dataset allows for a more encompassing view into a physician's behavior over each individual part, which is due to the larger selection of features, and the fact that the Combined is only focusing on providers having claims across all three data sources.

In order to provide additional insight about our AUC results, we performed both ANOVA [55] and Tukey's HSD tests [137] for the Train_Test fraud detection results, delving into the differences between the datasets, learners and class ratios. ANOVA

Table 6.2: Average AUC Scores with RUS: Train_Test

Learner	Ratio	Part B	Part D	DMEPOS	Combined
Gradient Boosted Trees	[1:99]	0.79523	0.76157	0.79202	0.84513
	[10:90]	0.81566	0.78212	**0.79683**	0.84929
	[25:75]	0.81476	0.77174	0.78660	0.83126
	[35:65]	0.81057	0.77054	0.77678	0.82757
	[50:50]	0.80626	0.74277	0.75944	0.81149
	[all:all]	0.78636	0.74969	0.78281	0.83654
Logistic Regression	[1:99]	0.82542	0.79714	0.77819	0.86829
	[10:90]	**0.82901**	**0.79858**	0.77482	0.86157
	[25:75]	0.82675	0.79431	0.76963	0.84583
	[35:65]	0.82768	0.79031	0.76715	0.83743
	[50:50]	0.82109	0.78866	0.75723	0.81778
	[all:all]	0.82133	0.79566	0.78088	**0.86888**
Random Forest	[1:99]	0.77081	0.73303	0.78803	0.82193
	[10:90]	0.78527	0.77433	0.78861	0.82896
	[25:75]	0.78683	0.76349	0.78914	0.81791
	[35:65]	0.78505	0.75418	0.78076	0.81273
	[50:50]	0.78136	0.75602	0.77333	0.81375
	[all:all]	0.75725	0.69302	0.77105	0.80122

is a statistical test determining whether the means of several groups (or factors) are equal. Tukey's HSD test determines factor means that are significantly different from each other. This test compares all possible pairs of means using a method similar to a t-test, where statistically significant differences are grouped by assigning different letter combinations (e.g. group a is significantly better than group b in correlation to the issue). The ANOVA results shown in Table 6.3 determines for the Train_Test experiment that the factors and their interactions are significant, at a 5% significance level. Therefore, each factor significantly contributed to AUC performance, which for this domain translates to successfully detecting fraudulent activities within Medicare.

Figure 6.2 presents the results of the Tukey's HSD test, represented as box plots, for the Train_Test experiment. The Tukey's HSD test exhibits significance levels, or groups, between factors allowing a visual representation as to which factor is better for fraudulent detection using Medicare claims data. The yellow dots indicate the overall average AUC for each factor and the vertical lines represent the range of AUC values. For this Train_Test experiment, the Combined dataset is shown alone in group a with

Table 6.3: ANOVA test results for Train_Test models

Term	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Dataset	3	0.45670	0.15223	1477.21578	3.46E-289
Learner	2	0.10675	0.05337	517.93217	4.23E-135
Ratio	5	0.03912	0.00782	75.91760	1.32E-62
Dataset:Learner	6	0.07258	0.01210	117.38693	4.37E-100
Dataset:Ratio	15	0.03227	0.00215	20.87356	2.91E-46
Learner:Ratio	10	0.03311	0.00331	32.12900	1.28E-50
Dataset:Learner:Ratio	30	0.01348	4.49E-04	4.35869	5.42E-13
Residuals	648	0.06678	1.03E-04	-	-

Part B alone in group b. This would indicate that between the three parts within the Combined dataset that Part B provides the majority of fraud detection, while Part D and DMEPOS contribute less but do improve results compared to Part B. This may indicate that there is a closer correlation between fraud detection and provider type related to details concerning procedures performed (HCPCS) over prescribed drugs or equipment referrals. Expounding on the results shown above, LR is confirmed to be to be significantly better than GBT and RF across each dataset/class ratio pair. In comparing RUS against the non-sampled datasets all:all, RUS performs significantly better except for the 50:50 class ratio. Balanced datasets have been shown to not produce the best results for Big Data in this domain, possibly due to the removal of too many majority class instances (non-fraudulent), resulting in datasets that cannot represent the characteristics of either class as they appear in the real-world [14, 68]. This can lead to an increasingly inaccurate representation of the real-world distribution of physician behavior as the datasets become more balanced. Interestingly, the highest overall AUC score was achieved without sampling, but the 10:90 ratio was superior to all other ratios over each dataset/learner pair, finding a balance between correcting class imbalance and information loss. Due to the high level of imbalance present in these datasets including the 10:90 ratio, there is still a dramatic decrease in the number of non-fraudulent cases, allowing for significantly less computing resources and time to run these models.

Overall, based on these results, the best case scenario for detecting Medicare

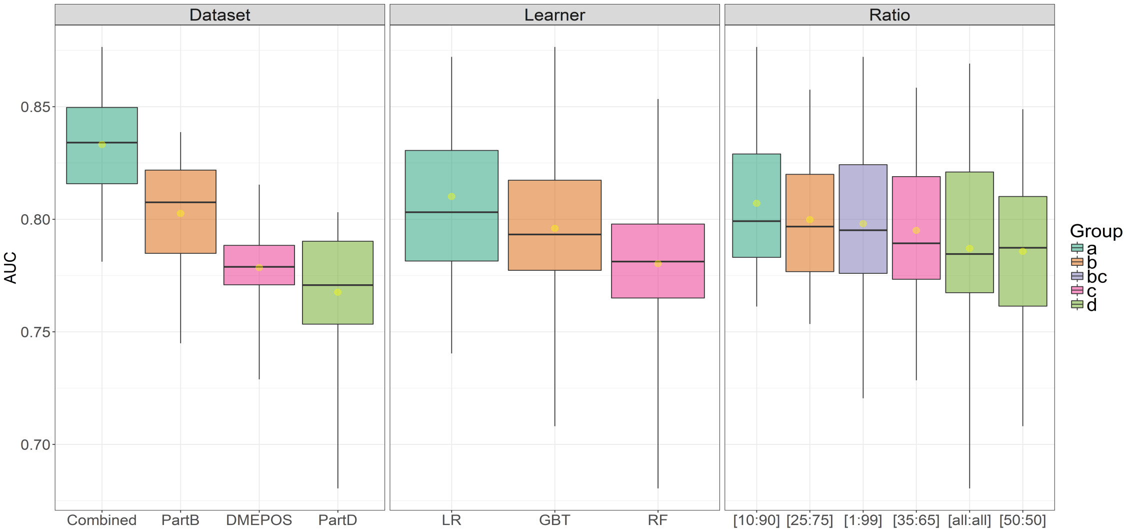


Figure 6.2: Tukey's HSD: Train_Test

fraud with claims data is (Train_Test) over CV. Currently, practitioners could use this Train_Test model to evaluate physicians' actions in 2016. In case a researcher does not have access to datasets with new instances, then CV will offer relatively similar results. We would recommend the Train_Test setup using LR with a ratio no more balanced than 10:90.

6.1.6 Section Summary

Medicare fraud contributes to significant, unnecessary financial losses, leading to decreased availability and a lower quality of care. Therefore, minimizing Medicare fraud is vitally important in the United States, as Medicare helps facilitate a growing elderly population's often necessary access to quality healthcare. Medicare fraud research, utilizing machine learning methods, has been shown viable in detecting fraudulent activity. Such research carries the benefits of reducing effort and time spent manually investigating possible fraudulent physicians by reducing false positives and false negatives. CMS has contributed to this cause, releasing various Big Data Medicare claims datasets from which we employ three, over all currently available years, and discuss our unique data processing approach for preparing the data and mapping fraud labels.

In order to determine the fraud detection abilities of machine learning models, they need to be evaluated [117]. Two commonly used evaluation methods are Train_Test and CV. The major difference between these two methods is Train_Test builds the model on the entire data and validates the model on a separate test dataset, while CV splits that same data into smaller training and test datasets. In this section, we compared Train_Test and CV evaluation methods, determining which provides better fraudulent prediction. We found that Train_Test outperformed Train_CV across datasets and learners, though the average results were comparable. Based on the results of Train_Test, we show that machine learning models using claims can accurately predict data for a new year based on historical data. The results using LR were the best across each dataset and evaluation method. Upon applying RUS to Train_Test, the 10:90 ratio demonstrated superior performance in lessening the negative effects of class imbalance. We also note that the Combined dataset showed the highest overall scores, presumably due to the encompassing nature of this dataset, containing claims from three parts of Medicare. Therefore, we recommend utilizing the Train_Test evaluation method, and building the model with LR and the Combined dataset at a 10:90 ratio. We also determined that if necessary, CV is a reasonable substitute when a practitioner does not have the appropriate resources to evaluate models using a Train_Test method.

6.2 THE EFFECTS OF CLASS RARITY ON THE EVALUATION OF FRAUD DETECTION MODELS

The healthcare system in the U.S. contains an extremely large number of physicians, who perform numerous services for an even larger number of patients. Every day, there are a massive number of financial transactions generated by physicians administering healthcare services, such as hospital visits, drug prescriptions, and other medical procedures. The vast majority of these financial transactions are conducted without any fraudulent intent, but there are a minority of physicians who maliciously defraud the system for personal gain. In machine learning, when a dataset portrays this discrepancy in class representation (i.e. a low number of actual fraud cases), it is known as class imbalance [62, 127]. The main issue attributed to class imbalance is the difficulty in discriminating useful information between classes due to the over-representation of the majority class (non-fraud) and the limited amount of information available in the minority class (fraud). In real-world medical practice, the number of fraudulent physicians is in the minority, where even fewer are confirmed, and well-documented. To further confound the situation, we found that the number of these known fraudulent physicians are becoming less frequent each year, trending towards class rarity. Class rarity is when the PCC, or number of minority class instances, becomes extremely small. We argue that qualifying for severe class imbalance and rarity does not necessarily rely on the proportion between classes, but on PCC. A common class imbalance percentage is 2% [51], but, for example, a dataset with 10,000,000 instances still has a PCC of 200,000. A machine learning model would most likely be able to effectively determine qualities and patterns in the minority class by using 200,000 instances, and would not be affected by issues normally attributed to class imbalance [13], especially with data sampling. However, these issues would be present if this dataset had an extremely small PCC, such as 100, where a machine learning model would have to contend with 9,999,900 negative instances during training. This latter example indicates the concerns presented with class rarity. These issues are further exacerbated when applied to Big Data, which can dramatically increase the number of majority instances while leaving the minority class representation relatively unchanged [82].

We employ three publicly available Medicare Big Data datasets released by CMS: Part B, Part D, and DMEPOS. The data begins in either 2012 or 2013 and ends in

2016 (which was released June 2018). Note that neither 2017 or 2018 is currently available for these Medicare datasets. These CMS datasets include payment information for claims submitted to Medicare, payments made by Medicare, and other data points related to procedures performed, drugs administered, or supplies issued. Both individually and combined, the Medicare datasets provide an extensive view into a physician's annual claims, across three major parts of Medicare. Furthermore, we utilize the LEIE [108] to generate fraud labels since CMS does not provide any fraud information. For further detail related to Medicare and Medicare fraud, we refer the reader to [15, 36, 26, 38, 32]. Even though there are far fewer fraudulent physicians, they still contribute to large financial losses with the FBI estimating up to 10% of all healthcare expenditure is from fraud [104]. Many citizens depend on healthcare, especially the elderly population, which comprises the majority of Medicare beneficiaries.

A number of studies employed the CMS Medicare datasets to detect fraudulent physician behavior through data mining, machine learning and other analytical methods [132, 144], with a large portion of these studies using only Part B data [8, 9, 21, 49, 66, 86]. Even though the Part B dataset is comprehensive in its own right, employing only one Medicare part limits the comprehensive assessment of fraud detection performance available to a machine learning model compared to multiple parts. There are a few studies that utilized multiple parts of Medicare including [18, 70, 124]. Branting et al. [18] used the Part B (2012 - 2014) and Part D (2013) datasets. They utilize the LEIE for determining fraud labels through their identity-matching algorithm centered around a physician's National Provider Identifier (NPI) [28]. Through this algorithm, they matched over 12,000 fraudulent physicians, but the authors were not as discriminatory in fraud labeling as we are in this section, leading to the possible inclusion of physicians excluded for charges unrelated to fraud. The authors employed sampling, balancing their dataset to a 50:50 class ratio. The authors may have benefited from examining other class ratios, possibly without removing as many non-fraudulent instances. They developed a method for discriminating fraudulent behavior by determining the fraud risk using graph-based features in conjunction with a decision tree learner resulting in good overall fraud detection. In [124], Sadiq et al. employ Part B, Part D, and DMEPOS datasets, but limit their study to Florida only. They venture to find anomalies that possibly indicate fraudulent or other interesting behavior. The authors use an unsupervised method (Patient Rule Induction Method based bump hunting) to try to detect peak anomalies by spotting spaces of higher modes and masses within the dataset. They conclude that their method can accurately characterize the attribute space of CMS datasets. In [70], we conducted an exploratory study to determine which part of Medicare and which learner allows for the better detection of fraudulent behavior. We used all available years from 2015 and before, while fraud labels were generated through LEIE mapping. However, in [70], our experiments were conducted to show the feasibility of Medicare fraud detection, without focusing on the problem of class imbalance, rarity, or any methods to mitigate adverse effects on model performance. Unfortunately, even with these and the other research currently available, current methods are not significantly decreasing monetary losses facing the U.S. healthcare system. Therefore, we continue the efforts to detect Medicare fraud in order to decrease monetary loss due to real-world fraud.

In real-world practice, machine learning models are built on a full training dataset and evaluated on a separate test dataset consisting of new, unseen data points (i.e. hold-out set). We denote this evaluation method as Train_Test, and emulate this process by splitting the CMS datasets into training datasets (all years prior to 2016) and test datasets (the full 2016 year). In essence, the problem can be summarized as: can a model accurately detect new, known fraudulent physicians (from 2016) based on historical patterns of fraud (prior to 2016)? The results from the Train_Test method will provide a clear evaluation of fraud detection performance with Medicare claims data using machine learning. Furthering this sentiment, Rao et al. [119] discuss that: “Any modeling decisions based upon experiments on the training set, even cross validation estimates, are suspect, until independently verified [by a completely new Test dataset].” Unfortunately, in practice, the means to generate both a separate training and test set is limited, such as when the number of positive cases are too few or when only prior data is available. Therefore, we also conduct all of our experiments using CV. CV emulates the Train_Test method by splitting a single dataset into smaller training and test datasets for building and evaluation. This allows practitioners to assess prediction performance without a separate test dataset. For our experiments, we apply CV to our training datasets, and through comparisons with Train_Test results, we determine if CV can be a useful substitute. There are different variants of CV including: k-fold, stratified k-fold, leave-p-out, and holdout [117]. In this section, our experiments employ stratified k-fold CV due to its usefulness and ubiquity in machine learning as well as its focus on creating balanced class distribution across folds.

These Medicare datasets, with the added LEIE fraud labels, have severely imbalanced class distributions. In order to assess the effects of severe class imbalance and rarity, in addition to these original datasets, we generate datasets with growing class imbalance and increasing degrees of rarity by randomly removing positive class instances (i.e. lowering PCC). We also perform data sampling, specifically Random Undersampling (RUS), to determine whether sampling can effectively mitigate the negative effects of severe class imbalance and class rarity. For each original and generated dataset, we created five additional datasets with varying class ratios, ranging from balanced (50:50) to highly imbalanced (1:99). We evaluate our results over three different learners, across all Medicare parts and the Combined dataset, using the AUC and significance testing. This section has several goals [67]. Primarily, we are determining the effects severe class imbalance and rarity have on the Train_Test evaluation method in Medicare fraud detection. We also compare the Train_Test method to CV, across all experimental configurations. Lastly, we examine overall trends across Train_Test and CV to determine the optimal model configuration in terms of data sampling ratio and learner. From the Train_Test results, we determine that for the severely imbalanced Medicare claims, machine learning was able to discriminate between real-world fraudulent and non-fraudulent physician behavior reasonably well, but as the PCC trended toward rarity, the results generally decreased over all experiments. Overall, the results show that Train_Test significantly outperforms CV. Even so, we conclude that when necessary CV can be a viable substitute, but a practitioner should note that estimates may be conservative. Moreover, CV was similarly affected by severe class imbalance and rarity. Data sampling was demonstrated to mitigate the effects of having such a limited number of positive class instances, where the best class ratios were those with slightly larger negative class representation (i.e. less balanced).

The rest of the section is organized as follows. Section 6.2.1 presents related works, focusing on studies that employ datasets with class rarity. In Section 6.2.2, we discuss the Medicare and LEIE datasets, summarize our data processing and fraud label mapping, discuss severe class imbalance and rarity, and outline our implementation of RUS. We explain the Train_Test and CV evaluation methods in the Train_Test and Cross-Validation Section 6.2.3. In Section 6.2.4, we outline the learners, performance metrics, and significance testing. We then present our experimental results in Section 6.2.5. Finally, conclusions are presented in Section 6.2.6.

6.2.1 Related Works

Throughout the previous academic literature, class imbalance has been widely studied, and as in the real-world, there are many cases where there is a large disparity between classes, such as online shopping (making purchase/not making purchase)

[155] and healthcare fraud (fraud/non-fraud). The majority of these studies employ smaller datasets [2, 59, 93, 94, 153, 156]. Experiments using class imbalance with smaller data, could provide a basic understanding of the effects that class imbalance has on Big Data, but will be limited in addressing specific concerns when using Big Data. For instance, we determined throughout our research, when applying sampling in machine learning, a balanced ratio (50:50) is not as beneficial for Big Data as it is for smaller datasets, at least in the Medicare fraud detection domain. Studies that focus on class rarity are far less common [47, 127]. With regards to the research presented in this section, we limit our discussion to studies that employ Big Data for studying class imbalance in relation to rarity.

In [61], Hasanin et al. use four real-world Big Data sources from the sentiment140 text corpus and the UCI Machine Learning Repository in order to assess the impacts of severe class imbalance on Big Data. To supplement the original datasets, the authors generate additional datasets ranging from imbalanced to severely imbalanced with the positive class percentages of 10%, 1%, 0.1%, 0.01%, and 0.001%. They use the RF learner on both Apache Spark and H2O Big Data frameworks to assess classification performance, determining that that 0.1% and 1.0% can provide adequate results. They also experimented with data sampling, specifically RUS, and determined that balanced class ratios provided no benefit over the full datasets. Even though 0.001% is a very large disparity between classes, the authors do not generate any datasets that qualify as rarity. Fernandez et al. [51] provide a literature survey and experimentation focused on Big Data and class imbalance. They employ Hadoop with MapReduce using the Spark Machine Learning Library (MLlib) [102] versions of undersampling (RUS) and oversampling (ROS), and Synthetic Minority Over-sampling Technique (SMOTE). The authors compare RUS, ROS, and SMOTE over two Big Data frameworks using two imbalanced datasets, derived from the ECBDL14 dataset, which consist of 12 million and 600,000 instances, 90 features and class ratios of 98:2

(majority:minority). They compare these methods across two learners, RF and Decision Tree. They determine that as the number of partitions decreases, RUS has better performance, while ROS performs better with a larger number of partitions, while SMOTE performed inadequately across all experiments. They also recommend that newer, more advanced Big Data frameworks, such as Apache Spark, should be used compared to more dated frameworks. The authors do not remove any positive class instances in order to study the effects of rarity nor severe class imbalance. Another work by Rastogi et al. [121], use the ECBDL14 dataset, with 1.7% positive class representation (PCC = 48,637) and a total of 2.8 million instances and 631 features. They split the data 80% for training and 20% for testing. They compare Python SMOTE to their own version of SMOTE based on Locality Sensitive Hashing implemented in Apache Spark, demonstrating their model is superior. The authors do not test their method on a dataset with a rare number of positive class members.

Two studies that employ relatively Big Data with rarity are [46] and [154]. Dong et al. [46] develop a deep learning model for very imbalanced datasets, employing batch-wise incremental minority class rectification along with a scalable hard mining principle. They evaluate their method's performance on a number of datasets, including a clothing attribute benchmark dataset (X-domain) with a PCC of 20 and 204,177 negative class instances (clothing attributes dataset). This dataset contains multiple classes, but they also test a binary dataset with 3,713 positive instances and 159,057 negative. Through their experiments, they found their method was superior, with a minimum 3-5% increase in accuracy, with the additional benefit of being up to seven times faster. In [154], Zhai et al. utilize seven different datasets from various data sources including one artificial dataset with 321,191 negative class instances and 150 positive. They developed an algorithm based on MapReduce and ensemble extreme learning machine (ELM) classifiers, and determine their method superior compared to three different versions of SMOTE. Through Zhai et al's. research, it is hard to infer the effects of rarity of the aforementioned dataset because their datasets are gathered from multiple sources. In order to determine the effects of rarity and the ability of their model, it would have been beneficial if datasets from the same source were tested with varying PCC values.

In a study conducted by Tayal et al. [133], the authors perform experiments with real-world datasets derived from the standard KDD Cup 1999 data, where the largest contained 812,808 instances. The positive class made up 0.098%, with a PCC of around 800, qualifying this dataset as severely imbalanced, but not rarity. This latter point is because 800 instances could still provide a reasonable level of discrimination for a machine learning model. They determined that their RankRC method was able to outperform several SVM methods and was more efficient with processing speed and space required. Maalouf et al. [96] present a truncated Newton method in prior correction Logistic Regression (LR) including an additional regularization term to improve performance. They also employ the KDD Cup 1999 dataset, along with six others. The largest dataset they use has 304,814 instances with the positive representation at 0.34%, translating to a PCC of a little over 1,000. In [20], Chai et al. generate a dataset from a manufacturer and user facility device experience database with the goal of automatically identifying health information technology incidents. The subset consists of 570,272 instances with a PCC of 1,534. They generate two additional subsets, one balanced (50:50) and another with 0.297% class representation. They employ statistical text classification through LR. These studies utilize data that can be described as relatively big, but do not assess the effects of rarity.

Zhang et al. [155] discuss the vast amounts of data created from online shopping websites and the large imbalance between purchases made versus visits made without a purchase. The level of imbalance quickly escalates when considering high spending customers (i.e. over $100). They found that in a week, a retail website had 42 million visits with only 16,000 purchases resulting in a ratio of 1:2,500, while the ratio for high spending customers was 1:10,000. They developed an adaptive sampling scheme that samples from severely imbalanced data. Through this method, the authors ensure that when sampling data, they obtain a satisfactory number of positive class instances by searching through the original data. We would argue that when the effects of imbalance can be solved by searching for more available positive class instances, then the domain in question does not suffer from the traditional effects attributed to class imbalance even if that ratio is (1:10,000) or worse. The real effects of severe class imbalance and rarity are felt when, throughout all available data, the resultant PCC is so minimal, a machine learning algorithm cannot discriminate useful patterns from the positive class. As mentioned, we note that the number of available real-world fraudulent physicians matching with the CMS Medicare datasets are decreasing, moving the Medicare fraud detection domain towards rarity. We assess the effects of class rarity using the Train_Test evaluation method with Medicare Big Data using real-world fraud labels.

6.2.2 Data

We utilize three publicly available Medicare datasets maintained by the CMS: Part B, Part D, and DMEPOS [27, 29, 30]. We provide discussion in Section 2.3.2, to cover all data-related details associated with our unique methods including the datasets, data processing and engineering, fraud labeling and our training and test datasets, outlining the differences, and discuss our processes. This section aims to predict fraudulent behavior as it appears in real-world medical practice. We utilize the LEIE, which currently contains the most comprehensive list of real-world fraudulent physicians throughout the U.S. To the best of our knowledge, there is no publicly available database containing both provider claims activity and fraud labels, and therefore, we use the LEIE to supplement the Medicare datasets, allowing for an accurate assessment of fraud detection performance. From these Medicare datasets, we select the features specifically related to claims information and a select physician-specific data points, as we believe they provide value and are readily usable by machine learning models. Table 2.10 demonstrates the features chosen for this Section. Note the exclusion feature is generated via the LEIE, creating the fraud or non-fraud labels for classifying physicians. There are different categories of exclusions, based on severity of offense. As shown in Table 2.5, we chose only the mandatory non-permissive exclusions. We use these excluded providers as fraud labels in all of our training and test datasets. Note, the LEIE does not provide the program within which a physician perpetrated their offenses (i.e. Medicare), meaning these excluded physicians were not necessarily convicted for committing criminal activities within Medicare, but we assume that a physician who commits such acts would continue their fraudulent behavior when submitting claims to Medicare.

Table 2.11 and 2.13 summarize both the training and test datasets used in this section, detailing the number of features, number of fraudulent and non-fraudulent instances, and the percentage of fraudulent cases after aggregation, one-hot-encoding and fraud labeling. The main difference between the training and test datasets are in the provider type labels within the 2016 CMS datasets, as they were either entered incorrectly, slightly different, or completely changed. We adjusted as many of these provider type labels as possible when processing the 2016 datasets (test datasets) to match them to the training dataset labels. In addition, there were other provider types that were added or removed, which we were unable to match between the training and test datasets. We provide full details in Section 2.3.2 and document these non-matching provider types in Table 2.14.

Having a large difference between the number of majority and minority class instances can create bias towards the majority class when building machine learning models. This is known as class imbalance [140], which presents issues for machine learning algorithms when attempting to discriminate, often complex, patterns between classes, particularly when applied to Big Data. Rarity is an exceptionally severe form of class imbalance. In real-world situations, when severe class imbalance and rarity are present, the minority class is generally the class of interest [127]. A detailed discussion of severe class imbalance and rarity is presented in Section 3.1.2. Table 3.3a demonstrates the level of class imbalance present in each Medicare dataset, split by year. We observe that the number and percentage of fraudulent instances matching between the LEIE and the Medicare datasets decreases every year, across each dataset. We also note from the labeled Medicare datasets that each year, the non-fraudulent instances generally increase at a faster rate than the fraudulent cases are decreasing. These two observances are pushing the imbalance in fraud instances from severe to rarity. Therefore, rarity is an important topic to study in Medicare fraud detection, and in order to study rarity, we generate additional training datasets as shown in Table 3.3b. All non-fraudulent instances are kept, while we remove a number of fraudulent instances, achieving further levels of severe class imbalance and rarity. The PCCs in these new generated datasets range from 1,000 to 100, based on original number of fraudulent instances. These PCCs were further chosen, based on preliminary results, which demonstrate that these adequately represent class rarity in Big Data. In order to get a thorough representation of fraudulent instances, we generate ten different datasets by re-sampling for each dataset/PCC pair. For example, with regard to the 400 PCC for the Part B, we randomly select 400 instances from the original 1,409, with this process repeated ten times. The final result for each dataset/PCC pair is the average score across all ten generated rarity subsets.

RUS is used to mitigate the effects caused by severe class imbalance and rarity. We discuss data sampling in Section 3.1.2. The goal of data sampling is adjusting the datasets to a given ratio of majority and minority representation. As we employ RUS, our goal is to incur minimal information loss while simultaneously removing the maximum number of majority instances (i.e. determine which ratio delivers the best fraud detection). Therefore, we chose the following class ratios: 1:99, 10:90, 25:75, 35:65, and 50:50 (minority:majority), including the full, non-sampled datasets as the baseline (labeled as Full). In applying these ratios, we generate ten datasets for each original and generated training dataset, to reduce bias due to poor random draws. These ratios were chosen because they provide a good distribution, ranging from balanced 50:50 to highly imbalanced 1:99 [10]. Note, for Train_Test, when applying RUS or creating the severe class imbalance and rare subsets, only the training datasets are sampled, as they build the model, while test datasets are kept unaltered for model evaluation.

6.2.3 Train_Test and Cross-Validation

In this section, we employ the Train_Test evaluation method, which uses a training dataset for building the model, and evaluate this model using a separate, distinct test dataset, as demonstrated in Figure 3.3b. We are assessing how rarity effects the Train_Test method's results, which is especially important since known fraudulent instances are decreasing year-over-year. Additionally, we also use CV, specifically stratified k-fold CV with k=5, in order to compare results and determine whether employing CV estimates are similar to results from the Train_Test evaluation method. The process for k-fold CV is demonstrated in Figure 3.3a. For CV, we use training datasets that were not altered to match the test datasets, as CV does not employ the test dataset. Rao et al. [119] recommend validating CV results with a separate test dataset. They also mention that when a model is tuned by a test dataset, this is no longer an accurate simulation of the real-world event. Therefore, by employing the 2016 test datasets with Train_Test, we are evaluating the viability of CV for providing estimates that lead to model selection in Medicare fraud detection. We provide discussion and definitions for Train_Test and CV in Section 3.1.2

6.2.4 Experimental Design

Since our Medicare datasets have such a large volume of data, we required a machine learning network that can handle Big Data. Therefore, we employ Apache Spark [4] on top of a Hadoop [3] YARN cluster for running and validating our models using their implemented MLlib. Apache Spark is a unified analytics engine capable of handling Big Data, offering dramatically quicker data processing over traditional methods or other approaches using MapReduce. The MLlib provided by Apache is a scalable machine learning library built on top of Spark. From the Apache Spark MLlib, we chose LR, and two tree-based models: RF and GTB. We chose these three based on preliminary research where other learners provided relatively worse fraud detection, such as Multilayer Perceptron or Naive Bayes. We used default configurations for each learner, unless noted otherwise. These learners are discussed in Section 3.2.

In order to evaluate the fraud detection performance of each learner, we use the AUC [16, 129]. AUC has demonstrated itself quite capable as a metric for quantifying results for machine learning studies employing datasets with class imbalance [77]. AUC shows performance over all decision thresholds, representing the ROC curve as a single value ranging from 0 to 1. An AUC of 1 denotes a classifier with perfect prediction for both the positive and negative classes, 0.5 represents random guessing, and any score under 0.5 means a learner demonstrated predictions worse than random guessing. We also perform hypothesis testing to demonstrate the statistical significance around our AUC results through ANOVA [55] and Tukey's HSD tests [137]. ANOVA is a statistical test determining whether the means of several groups (or factors) are equal. Tukey's HSD test determines factor means that are significantly different from each other. A complete description of AUC, ANOVA and Tukey's HSD tests are presented in Section 3.3.2.

6.2.5 Discussion and Results

In this subsection, we discuss our experimental results, assessing the impacts of rarity on Medicare fraud detection, as well as provide recommendations for practitioners based on these results. Table 6.4 presents the average AUC scores for Train_Test across PCC, consisting of original class distribution (All) and the selected severe class imbalance and rarity values, split by dataset (sub-tables), learner, and class ratio. The boldfaced values indicate the learner/ratio pair producing the best fraud detection performance per PCC. The effects of class rarity are demonstrated across each dataset, where the boldfaced values decrease as the PCCs decrease, and persist across nearly every learner/ratio pair. We notice that across all datasets, LR frequently presents the best scores, where the only outlier is DMEPOS, with GBT having better results for higher PCCs. Even though DMEPOS demonstrates better results with tree-based learners, we observe that as PCC decreases, LR begins to have better results. We believe that LR's results are due to a more successful strategy for handling class imbalance and rarity through regularization, which penalizes large coefficients (Ridge Regression) minimizing the adversities of noise and overfitting, leading to increased model generalization. Both GBT and RF also employ mechanisms to curtail the effects of noise and overfitting, but appear less robust to class imbalance and especially rarity, for Medicare fraud detection. Among the boldfaced values, we notice the less balanced ratios have the highest scores. Upon closer inspection, the 10:90 ratio most frequently scores higher across PCC/ratio pairs followed closely by 1:99 and Full, especially for the Combined dataset. Note that the Full (non-sampled) results indicate good detection performance, again, showing that a good representation of the majority class is beneficial. We believe the diminishing results when approaching a balanced configuration are due to the removal of too many negative class instances, deterring the learner's ability to discriminate the details of the non-fraudulent class. The Combined dataset has higher scores compared to the individual datasets, as indicated by the boldfaced values. Possible contributing factors are that the Combined dataset contains a larger selection of attributes, which facilitates a broader view of physician behavior over the individual parts, and that it only concentrates on providers who submitted claims to all three parts of Medicare.

Additionally, we perform this same experiment for Train_CV, and provide the results in Table 6.5. Figure 6.3 presents multiple bar graphs, comparing the differences in average AUC scores between Train_Test and Train_CV across each learner, dataset, PCC, and ratio configuration, where each bar represents the average AUC score for Train_Test minus the average AUC score for Train_CV. These bar graphs demonstrate that Train_Test outperforms Train_CV in almost all cases, the only notable contradiction being for the Part B dataset when employing the tree-based learners, in particular RF. We note that Train_Test had superior results for every configuration when employing LR. This signifies that the models built using previous years (training dataset) and evaluated on a separate, new year (test dataset) provide better fraud detection over applying CV on the training dataset alone. As mentioned above, CV is susceptible to bias and variance, which could contribute to the moderate results compared to Train_Test. We observe that as PCC decreases, becoming more rare, the delta between Train_Test and Train_CV generally increases. We surmise that Train_Test handles imbalanced data and class rarity better due to the models being trained with all available positive class (fraudulent) instances. Thus, with Train_Test, there is a decreased chance of overfitting, compared to CV, where models are built using a sub-sample of instances in each fold, bringing the already small PCC even lower for each training dataset.

Additionally, we performed hypothesis testing to demonstrate the significance of our results. We used a one-factor ANOVA test for Evaluation Method (Train_Test and Train_CV), and assess significance over learners, datasets, ratios and PCC as shown in Table 6.6. Evaluation Method was significant at a 95% confidence interval,

Table 6.4: AUC Results for Train_Test

Learner	Ratio	200	400	1000	All
GBT	[Full]	0.77604	0.78207	0.79080	0.78636
	[1:99]	0.78453	0.79440	0.80158	0.79523
	[10:90]	0.78575	0.79954	0.81324	0.81566
	[25:75]	0.75562	0.78570	0.80824	0.81476
	[35:65]	0.74548	0.78412	0.80523	0.81057
	[50:50]	0.72767	0.76659	0.79002	0.80626
LR	[Full]	0.80406	0.81686	0.82063	0.82133
	[1:99]	**0.80803**	0.82062	0.82441	0.82542
	[10:90]	0.80501	0.81933	**0.82601**	**0.82901**
	[25:75]	0.79251	**0.82101**	0.82347	0.82675
	[35:65]	0.77845	0.81197	0.82049	0.82768
	[50:50]	0.76491	0.80030	0.81863	0.82109
RF	[Full]	0.70754	0.73765	0.75130	0.75725
	[1:99]	0.73952	0.75813	0.76869	0.77081
	[10:90]	0.76287	0.77653	0.78856	0.78527
	[25:75]	0.75999	0.77324	0.78079	0.78683
	[35:65]	0.75139	0.76720	0.77863	0.78505
	[50:50]	0.74320	0.76689	0.77646	0.78136

(a) Part B

Learner	Ratio	100	200	400	All
GBT	[Full]	0.69437	0.70885	0.74099	0.74969
	[1:99]	0.69356	0.71879	0.74821	0.76157
	[10:90]	0.68142	0.70006	0.75261	0.78212
	[25:75]	0.65184	0.69628	0.73418	0.77174
	[35:65]	0.64126	0.67560	0.71313	0.77054
	[50:50]	0.62193	0.65445	0.70517	0.74277
LR	[Full]	**0.74339**	0.76832	0.78592	0.79566
	[1:99]	0.73766	**0.76943**	**0.78623**	0.79714
	[10:90]	0.72541	0.76436	0.78491	**0.79858**
	[25:75]	0.71602	0.75240	0.77726	0.79431
	[35:65]	0.69825	0.74023	0.77341	0.79031
	[50:50]	0.68920	0.72922	0.75804	0.78866
RF	[Full]	0.60202	0.62445	0.64317	0.69302
	[1:99]	0.66243	0.68387	0.69370	0.73303
	[10:90]	0.70282	0.72050	0.73803	0.77433
	[25:75]	0.69181	0.71398	0.74121	0.76349
	[35:65]	0.68118	0.70447	0.73372	0.75418
	[50:50]	0.66406	0.68312	0.71714	0.75602

(b) Part D

Learner	Ratio	100	200	400	All
GBT	[Full]	0.72688	0.75808	0.78221	0.78281
	[1:99]	0.73083	0.75805	**0.78426**	0.79202
	[10:90]	0.71749	0.74800	0.77617	**0.79683**
	[25:75]	0.67911	0.73027	0.76314	0.78660
	[35:65]	0.66527	0.69776	0.75773	0.77678
	[50:50]	0.65155	0.66424	0.74161	0.75944
LR	[Full]	0.75220	**0.76024**	0.77622	0.78088
	[1:99]	**0.74545**	0.75646	0.77403	0.77819
	[10:90]	0.73002	0.75079	0.76687	0.77482
	[25:75]	0.70978	0.73345	0.75993	0.76963
	[35:65]	0.68578	0.72187	0.75741	0.76715
	[50:50]	0.67933	0.70508	0.74394	0.75723
RF	[Full]	0.65576	0.71649	0.75220	0.77105
	[1:99]	0.67756	0.72877	0.76250	0.78803
	[10:90]	0.70214	0.73717	0.77153	0.78861
	[25:75]	0.71244	0.74737	0.76301	0.78914
	[35:65]	0.70956	0.72159	0.76735	0.78076
	[50:50]	0.69299	0.73423	0.75302	0.77333

(c) DMEPOS

Learner	Ratio	100	200	All
GBT	[Full]	0.76056	0.78431	0.83654
	[1:99]	0.75698	0.79823	0.84513
	[10:90]	0.74038	0.79609	0.84929
	[25:75]	0.73296	0.78145	0.83126
	[35:65]	0.69390	0.77744	0.82757
	[50:50]	0.70015	0.75962	0.81149
LR	[Full]	**0.81514**	**0.85430**	**0.86888**
	[1:99]	0.80496	0.84899	0.86829
	[10:90]	0.76965	0.82737	0.86157
	[25:75]	0.73072	0.80287	0.84583
	[35:65]	0.71753	0.77810	0.83743
	[50:50]	0.69273	0.74712	0.81778
RF	[Full]	0.62150	0.72501	0.80122
	[1:99]	0.71805	0.78308	0.82193
	[10:90]	0.74251	0.78836	0.82896
	[25:75]	0.74442	0.77432	0.81791
	[35:65]	0.73639	0.77206	0.81273
	[50:50]	0.72878	0.76666	0.81375

(d) Combined

Table 6.5: AUC Results for Train_CV

Learner	Ratio	200	400	1000	All
GBT	[Full]	0.75982	0.78328	0.79120	0.79569
	[1:99]	0.76740	0.79520	0.80378	0.80373
	[10:90]	0.76032	0.79377	0.81847	0.82064
	[25:75]	0.74964	0.78624	0.81464	0.81948
	[35:65]	0.73271	0.77326	0.80600	0.81434
	[50:50]	0.71530	0.75244	0.79563	0.80499
LR	[Full]	0.77162	0.78921	0.80019	0.80516
	[1:99]	**0.78282**	0.79295	0.81119	0.81238
	[10:90]	0.77752	0.79680	0.81465	0.81881
	[25:75]	0.76797	**0.79746**	0.81507	0.81686
	[35:65]	0.75771	0.79061	0.81336	0.81806
	[50:50]	0.73414	0.78012	0.80964	0.81415
RF	[Full]	0.71510	0.73806	0.78110	0.79604
	[1:99]	0.74846	0.77197	0.80579	0.81586
	[10:90]	0.76661	0.79117	**0.81933**	**0.83012**
	[25:75]	0.76031	0.79187	0.81641	0.82703
	[35:65]	0.75699	0.78061	0.81299	0.82156
	[50:50]	0.74994	0.77298	0.80448	0.81496

(a) Part B

Learner	Ratio	100	200	400	All
GBT	[Full]	0.68101	0.71044	0.73932	0.74851
	[1:99]	0.68871	0.70412	0.74731	0.75727
	[10:90]	0.66033	0.69299	0.74381	0.76756
	[25:75]	0.65692	0.67700	0.73008	0.76538
	[35:65]	0.63219	0.66694	0.71228	0.75996
	[50:50]	0.62040	0.65461	0.70773	0.74506
LR	[Full]	0.72516	**0.75436**	0.77369	0.78164
	[1:99]	**0.71200**	0.75396	**0.77575**	0.78486
	[10:90]	0.71031	0.75129	0.77481	**0.78657**
	[25:75]	0.70331	0.73115	0.77009	0.78540
	[35:65]	0.67880	0.72835	0.76340	0.78216
	[50:50]	0.67158	0.70696	0.74834	0.77557
RF	[Full]	0.62721	0.63364	0.66818	0.70888
	[1:99]	0.67627	0.68215	0.70816	0.73706
	[10:90]	0.67777	0.69735	0.73538	0.75857
	[25:75]	0.67634	0.69916	0.72832	0.75838
	[35:65]	0.65126	0.68992	0.72510	0.74904
	[50:50]	0.64951	0.68343	0.70771	0.74088

(b) Part D

Learner	Ratio	200	400	1000	All
GBT	[Full]	0.67203	0.68827	0.72125	0.73129
	[1:99]	0.66654	0.68611	0.72516	0.73591
	[10:90]	0.65411	0.68073	0.72241	0.73777
	[25:75]	0.64571	0.66342	0.71327	0.73389
	[35:65]	0.61468	0.64740	0.70118	0.72090
	[50:50]	0.60699	0.63259	0.68728	0.70598
LR	[Full]	0.68783	**0.70311**	0.73615	0.74063
	[1:99]	**0.68960**	0.69565	**0.73853**	**0.74085**
	[10:90]	0.67423	0.69604	0.73498	0.74421
	[25:75]	0.66667	0.68769	0.72912	0.73715
	[35:65]	0.65088	0.68259	0.72463	0.73488
	[50:50]	0.64590	0.66432	0.71445	0.72225
RF	[Full]	0.61229	0.64745	0.69381	0.70756
	[1:99]	0.64896	0.66998	0.70598	0.72245
	[10:90]	0.65829	0.67671	0.72066	0.73767
	[25:75]	0.65636	0.67337	0.71790	0.72889
	[35:65]	0.64239	0.67054	0.71756	0.72390
	[50:50]	0.63938	0.66152	0.70306	0.72379

(c) DMEPOS

Learner	Ratio	100	200	All
GBT	[Full]	0.73906	0.76623	0.79047
	[1:99]	0.73626	0.78562	0.80373
	[10:90]	0.72482	0.76730	0.81675
	[25:75]	0.68806	0.75833	0.80405
	[35:65]	0.68275	0.74855	0.79127
	[50:50]	0.65960	0.72675	0.77587
LR	[Full]	0.74260	0.80043	0.81554
	[1:99]	0.73814	**0.80060**	0.82011
	[10:90]	0.72508	0.78653	0.81868
	[25:75]	0.69117	0.77479	0.81553
	[35:65]	0.67940	0.76854	0.80998
	[50:50]	0.67567	0.74588	0.79415
RF	[Full]	0.64769	0.71098	0.79383
	[1:99]	0.71813	0.76663	0.81515
	[10:90]	0.73110	0.79011	**0.82793**
	[25:75]	**0.74162**	0.77822	0.81503
	[35:65]	0.72834	0.76699	0.80619
	[50:50]	0.71446	0.76228	0.79546

(d) Combined

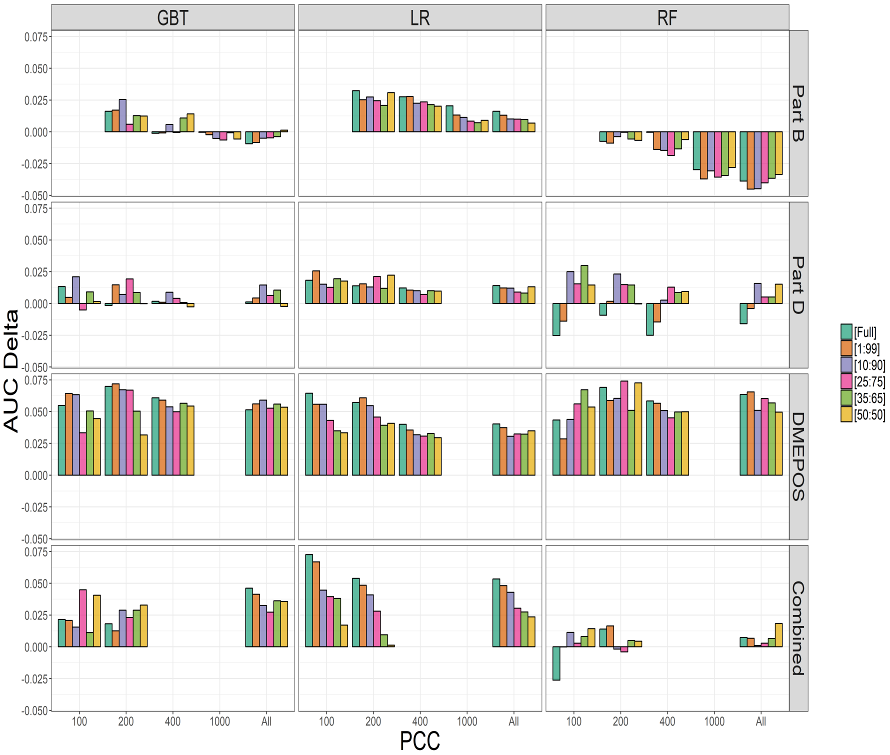


Figure 6.3: Average AUC: Comparing Train_Test - Train_CV

and therefore, we further conducted a Tukey's HSD test, presented in Table 6.7, to determine the significance between fraud detection results garnered from Train_Test and Train_CV. The Tukey's HSD test placed Train_Test in group 'a' and Train_CV in group 'b' signifying that evaluating a model on a segregated test set provides significantly better results over building and evaluating a model through CV.

Term	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Evaluation Method	1	0.45439	0.45439	103.9411081	2.59E-24
Residuals	12598	55.07319	0.00437	-	-

Table 6.6: One Factor: Train_Test and Train_CV

Table 6.7: Tukey's HSD test results for evaluation methods

Factor	Level	AUC	std	r	Min	Max	Group
Evaluation Method	Train_Test	0.74604	0.05212	1,980	0.56483	0.87266	a
Evaluation Method	Train_CV	0.72954	0.06841	10,620	0.40944	0.88532	b

Even though the Tukey's HSD test determined the evaluation methods are one group apart, we can argue that CV provides comparable results to Train_Test, albeit conservative. Therefore, we present the following results for both Train_Test and Train_CV in order to provide a thorough claim as to which learner and ratio yield the best results for class imbalance and rarity, as well as provide further insight into comparing these evaluation methods. We perform a 4-factor ANOVA test for both Train_Test and Train_CV, in Tables 6.8a and 6.8b, and evaluate the differences between datasets, learners, class ratios and PCCs. All factors and their interactions are shown as significant, at a 5% significance level. We perform further Tukey's HSD tests for each PCC, and assess any significant differences for learners (across ratios and datasets) and ratios (across learners and datasets), shown in Figures 6.4a and 6.4b, respectively. We concentrate only on the results for the best scoring group (group 'a'). For these graphs, there can be multiple combinations within a group designation, such as in Figure 6.4a, for the Part D dataset with 200 and 400 PCC, Train_Test and Train_CV with a class ratio of 10:90, are each in group 'a', respectively. The results in Figure 6.4a, show that LR contains the vast majority of group 'a' members, while Train_Test and Train_CV both have almost an identical distribution with the same number of group 'a' members. Therefore, regardless of model configuration across all PCCs, LR is able to provide the highest levels of discernment between fraudulent and non-fraudulent behavior patterns within each of our Medicare datasets. Figure 6.4b shows that the less balanced ratios contain the majority of group 'a' membership, with 10:90 having more representation than all other ratios combined. The more balanced ratios have significantly less group 'a' representation, where 50:50 has zero members. As seen with the learner results, Train_Test and Train_CV have similar distributions. However, Train_Test handles the more balanced datasets better, which is potentially due to the fact that Train_Test employs the entire training datasets whereas Train_CV splits the data, minimizing the already small fraud and non-fraud instances. Overall, from these results, we observe that for PCC, although the rarity experiments have similar group 'a' representation compared to the original class distribution, the overall AUC scores diminish as the level of rarity is increased. The main difference between evaluation methods from the learner and ratio Tukey's test is that the Train_Test generally has higher average AUC scores over comparable configurations. The complete Tukey's HSD results for all configurations for both learners and ratios are listed in Tables 6.9 and 6.10.

Table 6.8: ANOVA Tests

Term	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Dataset	3	1.96516	0.65505	1600.558663	0
Learner	2	0.66409	0.33205	811.3215437	1.29E-270
PCC	4	2.33669	0.58417	1427.367028	0
Ratio	5	0.30092	0.06018	147.0525251	1.17E-136
Dataset:Learner	6	0.22145	0.03691	90.18221558	2.05E-102
Dataset:PCC	7	0.05788	0.00827	20.2044874	1.61E-26
Learner:PCC	8	0.03291	0.00411	10.0523189	7.04E-14
Dataset:Ratio	15	0.08724	0.00582	14.21108373	2.17E-35
Learner:Ratio	10	0.46537	0.04654	113.7087576	2.69E-194
PCC:Ratio	20	0.05443	0.00272	6.649302133	3.98E-18
Dataset:Learner:PCC	14	0.02401	0.00171	4.189763574	2.52E-07
Dataset:Learner:Ratio	30	0.11577	0.00386	9.428942232	2.99E-40
Dataset:PCC:Ratio	35	0.01306	0.00037	0.911606914	0.61790442
Learner:PCC:Ratio	40	0.10682	0.00267	6.52539348	3.70E-32
Dataset:Learner:PCC:Ratio	70	0.04795	0.00068	1.673573178	4.56E-04
Residuals	2430	0.99452	0.00041	-	-

(a) Four Factor: Train_Test

Term	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Dataset	7	26.98719	3.85531	2641.872038	0
Learner	2	1.34356	0.67178	460.3389767	5.33E-194
PCC	3	6.30292	2.10097	1439.702651	0
Ratio	5	1.25714	0.25143	172.2919289	3.65E-178
Dataset:Learner	14	0.79698	0.05693	39.00963177	2.86E-105
Dataset:PCC	4	0.34344	0.08586	58.83679308	2.58E-49
Learner:PCC	6	0.07181	0.01197	8.201310311	7.04E-09
Dataset:Ratio	35	0.32743	0.00936	6.410645904	3.43E-29
Learner:Ratio	10	1.36440	0.13644	93.49620948	1.03E-187
PCC:Ratio	15	0.13786	0.00919	6.298014086	1.64E-13
Dataset:Learner:PCC	8	0.04741	0.00593	4.061371638	7.71E-05
Dataset:Learner:Ratio	70	0.42329	0.00605	4.143770726	3.52E-28
Dataset:PCC:Ratio	20	0.01735	0.00087	0.594299086	0.919852682
Learner:PCC:Ratio	30	0.08452	0.00282	1.9304991	0.001663456
Dataset:Learner:PCC:Ratio	40	0.07660	0.00192	1.312296909	0.089594623
Residuals	13230	19.30669	0.00146	-	-

(b) Four Factor: Train_CV

In summary, we assessed the effects that class imbalance and rarity have on Medicare claims data, and compared how various machine learning techniques handle detecting fraudulent behavior when being subjected to these effects. Machine learning was able to improve results, with RUS, for the majority of the class imbalance and rarity experiments presented in this work. However, we found that the rarer fraudulent instances become, the less machine learning can effectively discern fraudulent behavior from non-fraudulent behavior. Therefore, if the PCC qualifies a Medicare claims dataset as rare, we recommend that a practitioner gather additional, qual-

Table 6.9: Tukey's HSD: Learner

Part B				Part D				DMEPOS				Combined			
PCC	Learner	AUC	Group	PCC	Learner	AUC	Group	PCC	Learner	AUC	Group	PCC	Learner	AUC	Group
all	LR	0.82521	a	all	LR	0.79411	a	all	GBT	0.78241	a	all	LR	0.84996	a
1000	LR	0.82227	a	400	LR	0.77763	a	all	RF	0.78182	a	200	LR	0.80979	a
400	LR	0.81501	a	200	LR	0.75399	a	400	GBT	0.76752	a	100	LR	0.75512	a
200	LR	0.79216	a	100	LR	0.71832	a	400	LR	0.76307	a	all	GBT	0.83355	b
all	GBT	0.80481	b	all	GBT	0.76307	b	400	RF	0.76160	a	200	GBT	0.78286	b
1000	GBT	0.80152	b	400	GBT	0.73238	b	200	LR	0.73798	a	100	GBT	0.73082	b
400	GBT	0.78540	b	200	GBT	0.69234	b	100	LR	0.71709	a	100	RF	0.71527	b
200	GBT	0.76252	b	200	RF	0.68840	b	200	RF	0.73094	ab	all	RF	0.81608	c
all	RF	0.77776	c	100	RF	0.66739	b	all	LR	0.77132	b	200	RF	0.76825	c
1000	RF	0.77407	c	100	GBT	0.66406	b	200	GBT	0.72607	b	-	-	-	-
400	RF	0.76327	c	all	RF	0.74568	c	100	GBT	0.69519	b	-	-	-	-
200	RF	0.74408	c	400	RF	0.71116	c	100	RF	0.69174	b	-	-	-	-

(a) Train_Test

Part B				Part D				DMEPOS				Combined			
PCC	Learner	AUC	Group	PCC	Learner	AUC	Group	PCC	Learner	AUC	Group	PCC	Learner	AUC	Group
all	RF	0.81760	a	all	LR	0.78270	a	all	LR	0.73666	a	all	LR	0.81233	a
1000	LR	0.81068	a	400	LR	0.76768	a	400	LR	0.72964	a	all	RF	0.80893	a
400	LR	0.79119	a	200	LR	0.73768	a	200	LR	0.68823	a	200	LR	0.77946	a
200	LR	0.76530	a	100	LR	0.70019	a	100	LR	0.66918	a	100	RF	0.71356	a
all	LR	0.81424	b	all	GBT	0.75729	b	all	GBT	0.72763	b	100	LR	0.70868	a
1000	RF	0.80668	b	400	GBT	0.73009	b	all	RF	0.72404	b	100	GBT	0.70509	a
1000	GBT	0.80495	b	200	GBT	0.68435	b	400	GBT	0.71176	b	all	GBT	0.79702	b
400	GBT	0.78070	b	200	RF	0.68094	b	400	RF	0.70983	b	200	RF	0.76253	b
200	RF	0.74957	b	100	RF	0.65973	b	200	RF	0.66659	b	200	GBT	0.75880	b
200	GBT	0.74753	b	100	GBT	0.65659	b	200	GBT	0.66642	b	-	-	-	-
all	GBT	0.80981	c	all	RF	0.74213	c	100	GBT	0.64334	b	-	-	-	-
400	RF	0.77444	c	400	RF	0.71214	c	100	RF	0.64295	b	-	-	-	-

(b) Train_CV

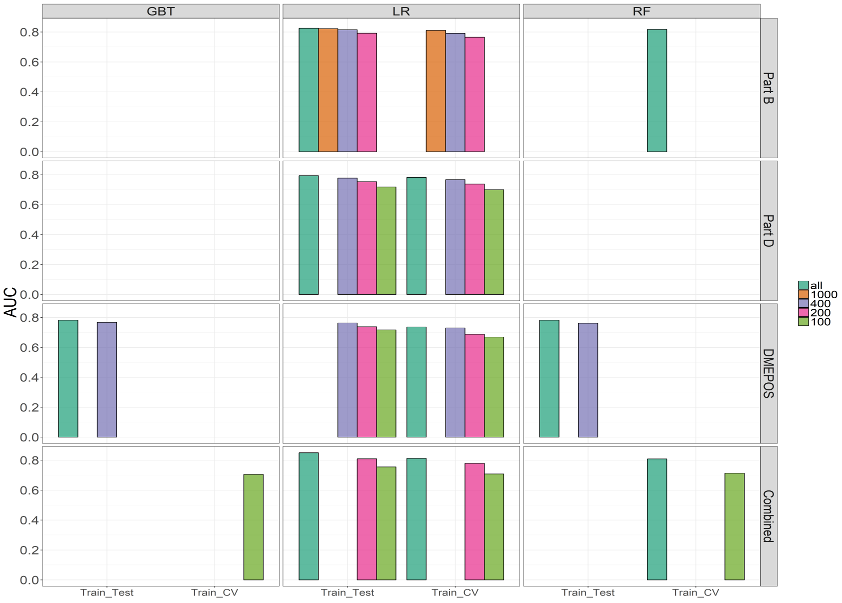
Table 6.10: Tukey's HSD: Ratio

Part B				Part D				DMEPOS				Combined			
PCC	Ratio	AUC	Group	PCC	Ratio	AUC	Group	PCC	Ratio	AUC	Group	PCC	Ratio	AUC	Group
all	[10:90]	0.80998	a	all	[10:90]	0.78501	a	all	[10:90]	0.78675	a	all	[10:90]	0.84661	a
all	[25:75]	0.80945	a	400	[10:90]	0.75851	a	400	[1:99]	0.77359	a	200	[1:99]	0.81010	a
1000	[10:90]	0.80927	a	200	[10:90]	0.72831	a	400	[10:90]	0.77152	a	100	[1:99]	0.76000	a
all	[35:65]	0.80777	a	200	[1:99]	0.72403	a	400	[all:all]	0.77021	a	100	[10:90]	0.75085	a
400	[10:90]	0.79846	a	100	[10:90]	0.70322	a	400	[25:75]	0.76203	a	all	[1:99]	0.84512	ab
200	[10:90]	0.78454	a	400	[25:75]	0.75088	ab	400	[35:65]	0.76083	a	200	[10:90]	0.80394	ab
1000	[25:75]	0.80416	ab	200	[25:75]	0.72089	ab	200	[1:99]	0.74776	a	100	[25:75]	0.73603	ab
400	[25:75]	0.79332	ab	100	[1:99]	0.69788	ab	200	[10:90]	0.74532	a	100	[all:all]	0.73240	abc
400	[1:99]	0.79105	ab	100	[25:75]	0.68656	abc	200	[all:all]	0.74494	a	all	[all:all]	0.83554	bc
200	[1:99]	0.77736	ab	all	[25:75]	0.77651	b	200	[25:75]	0.73703	a	200	[all:all]	0.78787	bc
all	[50:50]	0.80290	b	all	[35:65]	0.77168	b	100	[1:99]	0.71795	a	100	[35:65]	0.71594	bc
1000	[35:65]	0.80145	bc	400	[1:99]	0.74272	b	100	[10:90]	0.71655	a	all	[25:75]	0.83167	c
400	[35:65]	0.78776	bc	400	[35:65]	0.74009	b	100	[all:all]	0.71162	a	all	[35:65]	0.82591	c
200	[25:75]	0.76937	bc	200	[35:65]	0.70677	bc	all	[1:99]	0.78608	ab	200	[25:75]	0.78622	c
all	[1:99]	0.79715	c	100	[all:all]	0.67992	bc	100	[25:75]	0.70045	ab	200	[35:65]	0.77587	c
200	[all:all]	0.76255	c	all	[1:99]	0.76392	c	all	[25:75]	0.78179	abc	100	[50:50]	0.70722	c
200	[35:65]	0.75844	c	all	[50:50]	0.76248	c	400	[50:50]	0.74619	b	all	[50:50]	0.81434	d
1000	[1:99]	0.79823	cd	400	[50:50]	0.72678	c	200	[35:65]	0.71374	b	200	[50:50]	0.75780	d
400	[all:all]	0.77886	cd	400	[all:all]	0.72336	c	200	[50:50]	0.70118	b	-	-	-	-
1000	[50:50]	0.79504	d	200	[all:all]	0.70054	cd	all	[all:all]	0.77825	bc	-	-	-	-
all	[all:all]	0.78832	d	100	[35:65]	0.67357	cd	100	[35:65]	0.68687	bc	-	-	-	-
400	[50:50]	0.77793	d	all	[all:all]	0.74612	d	all	[35:65]	0.77490	c	-	-	-	-
200	[50:50]	0.74526	d	200	[50:50]	0.68893	d	100	[50:50]	0.67462	c	-	-	-	-
1000	[all:all]	0.78758	e	100	[50:50]	0.65839	d	all	[50:50]	0.76333	d	-	-	-	-

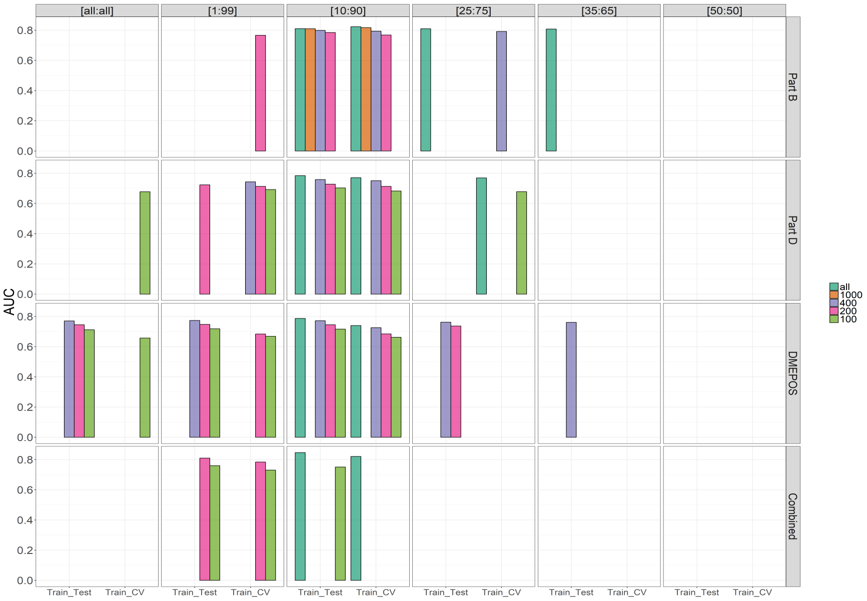
(a) Train_Test

Part B				Part D				DMEPOS				Combined			
PCC	Ratio	AUC	Group	PCC	Ratio	AUC	Group	PCC	Ratio	AUC	Group	PCC	Ratio	AUC	Group
all	[10:90]	0.82319	a	all	[10:90]	0.77090	a	all	[10:90]	0.73988	a	all	[10:90]	0.82112	a
1000	[10:90]	0.81748	a	all	[25:75]	0.76972	a	400	[10:90]	0.72602	a	200	[1:99]	0.78428	a
400	[10:90]	0.79391	a	400	[10:90]	0.75133	a	200	[10:90]	0.68450	a	100	[1:99]	0.73084	a
400	[25:75]	0.79186	a	400	[1:99]	0.74374	a	200	[1:99]	0.68391	a	200	[10:90]	0.78131	ab
200	[10:90]	0.76815	a	200	[10:90]	0.71388	a	100	[1:99]	0.66837	a	100	[10:90]	0.72700	ab
200	[1:99]	0.76623	a	200	[1:99]	0.71341	a	100	[10:90]	0.66221	a	all	[1:99]	0.81300	b
all	[25:75]	0.82113	ab	100	[1:99]	0.69233	a	100	[all:all]	0.65738	a	all	[25:75]	0.81154	b
1000	[25:75]	0.81538	ab	100	[10:90]	0.68280	a	400	[1:99]	0.72322	ab	200	[25:75]	0.77045	bc
400	[1:99]	0.78671	ab	100	[25:75]	0.67886	a	400	[25:75]	0.72010	ab	100	[all:all]	0.70978	bc
200	[25:75]	0.75931	ab	100	[all:all]	0.67779	a	400	[all:all]	0.71707	ab	100	[25:75]	0.70695	bc
all	[35:65]	0.81799	b	400	[25:75]	0.74283	ab	200	[all:all]	0.67961	ab	all	[35:65]	0.80248	c
400	[35:65]	0.78149	b	200	[25:75]	0.70243	ab	200	[25:75]	0.67483	ab	all	[all:all]	0.79995	c
200	[35:65]	0.74914	b	all	[35:65]	0.76372	b	100	[25:75]	0.65625	ab	200	[35:65]	0.76136	c
200	[all:all]	0.74885	b	all	[1:99]	0.75973	b	all	[25:75]	0.73331	b	200	[all:all]	0.75921	c
1000	[35:65]	0.81078	bc	200	[all:all]	0.69948	b	all	[1:99]	0.73307	b	100	[35:65]	0.69683	cd
all	[50:50]	0.81137	c	100	[35:65]	0.65408	b	400	[35:65]	0.71446	b	all	[50:50]	0.78849	d
all	[1:99]	0.81066	c	100	[50:50]	0.64716	b	200	[35:65]	0.66684	b	200	[50:50]	0.74497	d
400	[all:all]	0.77018	c	400	[35:65]	0.73359	bc	100	[35:65]	0.63598	bc	100	[50:50]	0.68324	d
400	[50:50]	0.76851	c	200	[35:65]	0.69507	bc	all	[35:65]	0.72656	c	-	-	-	-
200	[50:50]	0.73313	c	all	[50:50]	0.75384	c	all	[all:all]	0.72649	c	-	-	-	-
1000	[1:99]	0.80692	cd	200	[50:50]	0.68167	c	400	[50:50]	0.70160	c	-	-	-	-
1000	[50:50]	0.80325	d	400	[all:all]	0.72706	cd	200	[50:50]	0.65281	c	-	-	-	-
all	[all:all]	0.79896	d	all	[all:all]	0.74634	d	100	[50:50]	0.63076	c	-	-	-	-
1000	[all:all]	0.79083	e	400	[50:50]	0.72126	d	all	[50:50]	0.71734	d	-	-	-	-

(b) Train_CV



(a) Learner



(b) Ratio

Figure 6.4: Tukey's HSD test results for group 'a'

ity data until there is a sufficient PCC. The Train_Test and Train_CV method, in general, had similar results, but the latter's results were somewhat conservative in comparison. We recommend practitioners employ the Train_Test evaluation method with LR, after applying RUS with a 10:90 class distribution.

6.2.6 Section Summary

Two significant challenges facing healthcare fraud detection are the large amounts of data generated (Big Data) and the significant imbalance in fraudulent versus non-fraudulent behavior (class imbalance). The combination of these two issues leads to datasets that contain an extremely large volume of negative class instances (non-fraudulent) and very small numbers of positive class instance (fraudulent). In the case of fraud detection, the data is severely imbalanced. This section employs three Big Data datasets released by CMS, specifically Part B, Part D, and DMEPOS (individually and combined), as well as the LEIE from the OIG in order to map our real-world fraud labels. We notice that the fraudulent physicians from the LEIE had less matching physicians between the Medicare datasets, each year, since CMS started releasing these datasets. Because of this, we experimented with further severe class imbalance leading into class rarity. We do this by generating additional datasets and randomly remove fraudulent instances in order to determine the effects of increasing rarity on real-world fraud detection performance (Train_Test). In order to minimize the effects of severe class imbalance and rarity, we also employ data sampling, with various class ratios. In applying RUS, we created a new dataset for each original, severe class imbalanced and rare dataset. Detecting fraudulent behavior is the first step towards eliminating, or at least minimizing, fraud in healthcare, which would allow programs such as Medicare the ability to provide medical funding to a larger number of beneficiaries in the United States.

Throughout this section, we employ three learners and assess model performance using AUC and significance testing. When using the Train_Test evaluation method for severely imbalanced and rare datasets, we recommend building the model with LR and applying RUS with a 10:90 ratio. We noticed that as ratios approached balance (i.e. 50:50), performance decreased, and as such, determine that larger non-fraudulent representation is beneficial, with 10:90 being optimal. In practice though, a separate test dataset to evaluate a machine learning model may not be available due to a shortage of positive cases or lack of new data, and thus requires the use of other methods. To address this, we re-ran all experiments with CV, using the training datasets. CV emulates the Train_Test method, providing model generalization and error estimates on a single dataset by sub-setting the dataset into smaller training and test datasets, allowing all instances to both build and evaluate performance. We found that Train_Test results were significantly better than CV, but we determine that CV can be used a reliable substitute, when necessary, but a practitioner should keep in mind that results will be conservative. CV also showed similar patterns to Train_Test in terms of observed effects due to severe class imbalance and rarity, as well as the improvement garnered upon applying RUS. Overall, we noticed that prediction performance decreased as the number of fraudulent instances trended towards rarity, and therefore, we recommend that when PCC becomes too small (rare), then a practitioner should search for more quality data in order to appropriately allow for proper discrimination between fraudulent and non-fraudulent instances when applying machine learning.

6.3 CHAPTER SUMMARY

Throughout this chapter we studied the effects of severe class imbalance and rarity on Medicare fraud detection, employing both the Train_Test and CV evaluation methods. We used three Big Data Medicare datasets both individually and combined. After matching the Medicare datasets to the LEIE, we found that the datasets qualify as severely class imbalanced. Thus, in order to evaluate the effects of rarity, we generate additional datasets and randomly remove a varying number of fraudulent instances. Additionally, we applied RUS in an effort to mitigate the negative effects of severe class imbalance and rarity. Overall, we determined that the Train_Test method was superior to CV, but CV could be used as a reliable substitute if a test set is not available or there are not enough positive class representation. Between both experiments covered in this chapter, the 90:10 ratio performed significantly better than the other tested ratios and LR performed significantly better than other machine learning models used. For rarity, we recommend that when the PCC becomes too small, practitioners should find more quality data before devising a model that will be implemented in a real-world situation.

CHAPTER 7
CONCLUSION AND FUTURE WORK

Fraudulent activity within healthcare contributes to significant monetary losses, resulting in higher premiums [112], and ultimately reducing the number of patients receiving much needed care. Medicare represents a considerable portion of overall healthcare spending, and in 2017, comprised 3.7% of the total U.S. gross domestic product. It is predicted to increase to roughly 6% by 2042 [135]. Considering that Medicare primarily funds the elderly population, a demographic which is rapidly increasing, coupled with the fact that only 14.5% continue to work as of 2014 [106], it is imperative that Medicare minimize unnecessary expenditures. Decreasing fraud within Medicare would lessen the financial burden on beneficiaries, potentially help lower premiums and increase the number of elderly patients receiving healthcare, ultimately improving the quality of life for many throughout the population. Due to technological advances and an increasing volume of healthcare data, there are more opportunities for fraud detection than ever before. The combination of machine learning and frameworks that can handle Big Data, enable the leveraging of vast amounts of healthcare data, which can provide considerable improvement over manual fraud detection efforts in terms of processing speed, accuracy of results and ability to pinpoint unique patterns within the data. In an effort to encourage transparency and minimize fraud in Medicare, CMS released a number of Big Data datasets that include information for all claims submitted to Medicare by physicians throughout the U.S. and its commonwealths. These datasets do not include fraud labels; therefore, in order to determine the viability of our models to detect real-world fraud, we incorporated the LEIE to identify physicians who have been convicted of real-world fraudulent behavior.

After cross-referencing the Medicare data with the LEIE, we found that there is a larger number of non-fraudulent physicians than fraudulent. In machine learning, this is known as class imbalance, which causes learners to have difficulty discriminating patterns between classes, with a bias towards the overwhelming majority class. Additionally, we have determined that every year since the CMS has begun releasing their Medicare claims datasets, the number and percentage of fraudulent physicians matching with the LEIE has decreased steadily every year. This decrease indicates that the Medicare fraud detection problem can become one of class rarity. Rarity is an extreme form of class imbalance, occurring when the PCC is so minimal that machine learning algorithms may not have enough positive class membership to find any unique patterns, causing inaccurate predictions. Big Data can exacerbate the problem of class imbalance and rarity due to the overwhelming number of majority class instances. In this dissertation, we focus on evaluating machine learning models with the goal of determining the best practices for detecting fraudulent behavior in Medicare by employing imbalanced Big Data datasets.

7.1 CONCLUSIONS

Throughout this research, we evaluate the performance of both anomaly and traditional fraud detection, through the application of various machine learning techniques employing big Medicare datasets from CMS, in conjunction with the LEIE for obtaining real-world fraudulent physicians. We discuss our unique data processing and feature engineering approaches for both fraud detection approaches: anomaly detection and traditional fraud detection. For anomaly detection, we experiment with a number of improvement techniques including: feature selection (and sampling), grouping by specialty, removal of select specialties (i.e. general specialties), and a traditional method we designate as class isolation. In our traditional fraud detection experiments, we tested the viability of multiple Medicare datasets, both individually and combined. Additionally, we evaluated the prevalently used CV method compared to Train_Test (hold-out), determining whether CV estimates are reliable. Through evaluating Train_Test, we determine how viable traditional fraud detection can be for predicting real-world Medicare fraud. Rarity is also studied by generating additional datasets and randomly removing fraudulent physicians, demonstrating the effects on model performance. We also perform data sampling to mitigate the negative effects of severe class imbalance and rarity.

In Chapter 4, we employ an approach to predict a physician's expected specialty based on the type and number of procedures performed, via the Part B dataset, across three different research efforts. The experiments conducted determine whether physicians acting outside the norm of their respective specialty can indicate fraudulent behavior. In Section 4.1, we performed a preliminary study with a small subset of the full Part B dataset including physicians practicing only within Florida for the 2013 calendar year, and further split by office only. The Multinomial Naïve Bayes algorithm was used to build the model and was evaluated by calculating precision, Recall, and F-score with 5-fold cross-validation. The model was able to successfully predict several classes of physicians with an F-score over 0.9. From these results we demonstrate that it is possible to effectively use machine learning in a novel way through classifying physicians into their respective fields solely using the procedures they bill for. This research provides a model that can identify physicians who are potentially misusing insurance systems for further investigation.

The experiments covered in Section 4.2 expand upon our research into medical specialty anomaly detection by validating the efficacy of our model using real-world fraud cases, and then testing three strategies to improve model performance. The three strategies are feature selection and sampling, class grouping (specialty grouping), and class removal (removal of selected specialties). In addition to using the 2013 data, we use the 2014 data for model validation and comparisons. We employed the LEIE database, released by the OIG, as well as two other documented fraud cases, for testing the fraud detection of our model. Multinomial Naïve Bayes is used to build all models. We were able to demonstrate our model was able to correctly classify 67% of the real-world fraudulent physicians contained in the LEIE database as fraudulent. Furthermore, the three proposed strategies show good results in improving model performance, evaluated by F-score. Due to the small size of the dataset, we were unable to evaluate the fraud detection abilities of the improvement strategies as there were only 18 total fraudulent physicians.

Section 4.3 further builds onto this line of research as we incorporate the Part B dataset across the entire U.S. for the 2012 through 2015 calendar years, allowing for the identification of 1,312 fraudulent physicians. We generated a new baseline model, comparing LR and MNB, in order to test and assess several improvement strategies to increase the accuracy of Medicare fraud detection. Additionally, in determining the baseline, we also compared SOF and COF methods of combining HCPCS codes. Our results indicate that our proposed improvement strategies (specialty grouping, class removal, and class isolation), applied to different medical specialties, have mixed results over the selected LR/SOF baseline model's fraud detection performance. Class isolation is actually a traditional fraud detection method, and although a direct comparison with our anomaly experiments was not made, we found considerably better results with class isolation. The highest percentage obtained through anomaly detection was 26.1% (class grouping) and class isolation was able to obtain at best a Type I error rate of 0.193 and a Type II error rate of 0.126 with Internal Medicine. Thus, we determined that our anomaly detection approach is inadequate when compared to traditional fraud detection, and thus, we shifted our experimentation to traditional fraud detection.

In Chapter 5, we determined the viability of traditional fraud detection with three Big Data Medicare datasets, both individually and with our unique Combined dataset. The datasets used differ from anomaly detection where they do not consider HCPCS counts, but instead used aggregated payment information as discussed in Section 2.3.2. Our exploratory analysis involved building and assessing three learners on each dataset. Based on the Area Under the Receiver Operating Characteristic (ROC) Curve performance metric, our results show that the Combined dataset with the LR learner yielded the best overall score at 0.816, closely followed by the Part B dataset with LR at 0.805. Overall, the Combined and Part B datasets produced the best fraud detection performance with no statistical difference between these datasets, over all the learners. Therefore, based on our results and the assumption that there is no way to know within which part of Medicare a physician will commit fraud, we determined using the Combined dataset is best for detecting fraudulent behavior, given that a physician has submitted payments through any or all of the three Medicare parts used.

In Chapter 6, we further our experimentation with traditional fraud detection and compare Train_Test and CV evaluation methods on datasets with severe class imbalance and class rarity. We create a training and test dataset for all three Medicare parts, both separately and combined, and assess fraud detection performance, evaluating model performance before deployment into real-world situations. We determine which evaluation method performs better between CV and having separate, distinct training and test datasets (Train_Test). Before a machine learning model can be distributed for real-world use, a performance evaluation is necessary to determine the best configuration (e.g. learner, class sampling ratio) and whether the associated error rates are low, indicating good detection rates. With this chapter, we demonstrate the effects of class imbalance using a Train_Test evaluation method via a hold-out set, by evaluating the machine learning results. The Train_Test evaluation method is the procedure applied in a real-world setting, where a model is trained on a full training dataset and evaluated on a separate test dataset. We repeat the same experiments using CV, and determine it is a viable substitute for Medicare fraud detection. Additionally, we apply data sampling, specifically RUS with varying class ratios, in an effort to mitigate the negative effects of class imbalance.

In Section 6.1, our results show that Train_Test has better performance over cross-validation. We determine that, if necessary, cross-validation is a reliable substitute in showing how well a model detects possible Medicare fraud, though practitioners should keep in mind that CV results are conservative. After applying Random Undersampling to Train_Test datasets we found that LR and the 10:90 ratio had significantly the highest scores.

We also found that the number of documented, real-world fraudulent physicians available has decreased every year since 2012, widening the disparity of class representation for Medicare fraud detection. Therefore, these Medicare datasets are moving from severe class imbalance to class rarity. Thus, we conducted the experiments shown in Section 6.2, which evaluate how class rarity will affect machine learning results for Medicare fraud detection. To emulate class rarity, we generate additional datasets by removing fraud (positive class) instances, lowering the already severely small PCC. Again, RUS was able to improve results, but we found that, as expected, fraud detection performance decreased as the fraudulent instances became rarer as the machine learning algorithms had more difficulty discriminating between classes. In cases where PCC qualify as class rarity, we determine that more quality data should be gathered before models are applied for use in real-world situations.

Overall, traditional fraud detection performed better in our experiments, but that does not mean that other implementations of anomaly detection could not provide better fraudulent detection. Both approaches show the potential impact that machine learning can have on detecting fraud within healthcare. They could allow for vast improvements in processing time and data coverage over manual inspection. These advancements have the potential to increase what should be the bottom line goal of the healthcare system, which is facilitating necessary and quality care for the largest number of patients as possible.

7.2 FUTURE WORK

Potential future work for healthcare fraud detection through machine learning can include:

- Incorporating additional Medicare datasets, such as Part A, either individually or in combination.
- Finding reliable methods beyond mapping by NPI between the LEIE and the Medicare datasets that adequately determine physician's identities in order to increase the number of quality fraudulent physicians available for data analytics.
- Finding additional sources in order to obtain more quality real-world fraudulent physicians.
- Exploring different ways of processing the Medicare datasets, including designing a dataset that contains both claims counts and payment information and evaluating the improvement of them individually.
- For traditional fraud detection, experimentation with separating out physicians by individual specialty, providing potential improvement in discrimination between fraudulent and non-fraudulent physicians.
- Further experimentation with varying the costs of error rates for fraudulent and non-fraudulent instances, which could improve misclassification rates.

BIBLIOGRAPHY

[1] Administration for Community Living. Profile of older Americans: 2015. `https://acl.gov/sites/default/files/Aging%20and%20Disability%20in%20America/2015-Profile.pdf`, 2019.

[2] H. Alhammady and K. Ramamohanarao. Using emerging patterns and decision trees in rare-class classification. In *Fourth IEEE International Conference on Data Mining (ICDM)*, pages 315–318. IEEE, 2004.

[3] Apache. Apache Hadoop. `http://hadoop.apache.org/`, 2018.

[4] Apache. Apache Spark. `http://spark.apache.org/`, 2018.

[5] R. A. Bauder and T. M. Khoshgoftaar. A novel method for fraudulent medicare claims detection from expected payment deviations (application paper). In *2016 IEEE 17th International Conference on Information Reuse and Integration (IRI)*, pages 11–19. IEEE, 2016.

[6] R. A. Bauder and T. M. Khoshgoftaar. A probabilistic programming approach for outlier detection in healthcare claims. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 347–354. IEEE, 2016.

[7] R. A. Bauder and T. M. Khoshgoftaar. Multivariate outlier detection in medicare claims payments applying probabilistic programming methods. *Journal of Health Services and Outcomes Research Methodology*, pages 1–34, Jun 2017.

[8] R. A. Bauder and T. M. Khoshgoftaar. The detection of medicare fraud using machine learning methods with excluded provider labels. In *The Thirty-First International Flairs Conference*, pages 404–409, 2018.

[9] R. A. Bauder and T. M. Khoshgoftaar. The effects of varying class distribution on learner behavior for medicare fraud detection with imbalanced big data. *Health Information Science and Systems*, 6(1):9, 2018.

[10] R. A. Bauder and T. M. Khoshgoftaar. Medicare fraud detection using random forest with class imbalanced big data. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, pages 80–87. IEEE, 2018.

[11] R. A. Bauder and T. M. Khoshgoftaar. A survey of medicare data processing and integration for fraud detection. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, pages 9–14. IEEE, 2018.

[12] R. A. Bauder, T. M. Khoshgoftaar, and T. Hasanin. Data sampling approaches with severely imbalanced big data for medicare fraud detection. In *2018 30th IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 137–142. IEEE, 2018.

[13] R. A. Bauder, T. M. Khoshgoftaar, and T. Hasanin. An empirical study on class rarity in big data. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 785–790. IEEE, 2018.

[14] R. A. Bauder, T. M. Khoshgoftaar, A. Richter, and M. Herland. Predicting medical provider specialties to detect anomalous insurance claims. In *2016 28th IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 784–790. IEEE, 2016.

[15] R. A. Bauder, T. M. Khoshgoftaar, and N. Seliya. A survey on the state of healthcare upcoding fraud analysis and detection. *Journal of Health Services and Outcomes Research Methodology*, 17(1):31–55, 2017.

[16] M. Bekkar, H. K. Djemaa, and T. A. Alitouche. Evaluation measures for models assessment over imbalanced datasets. *J Inf Eng Appl*, 3(10), 2013.

[17] Y. Bengio and Y. Grandvalet. No unbiased estimator of the variance of k-fold cross-validation. *The Journal of Machine Learning Research*, 5(Sep):1089–1105, 2004.

[18] L. K. Branting, F. Reeder, J. Gold, and T. Champney. Graph analytics for healthcare fraud risk estimation. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 845–851. IEEE Press, 2016.

[19] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[20] K. E. Chai, S. Anthony, E. Coiera, and F. Magrabi. Using statistical text classification to identify health information technology incidents. *Journal of the American Medical Informatics Association*, 20(5):980–985, 2013.

[21] V. Chandola, S. R. Sukumar, and J. C. Schryver. Knowledge discovery from massive healthcare claims data. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1312–1320. ACM, 2013.

[22] N. V. Chawla. Data mining for imbalanced datasets: An overview. In *Data Mining and Knowledge Discovery Handbook*, pages 875–886. Springer, 2009.

[23] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[24] Chris Fawcett and Holger H. Hoos. Ablation Analysis. http://www.cs.ubc.ca/labs/beta/Projects/Ablation/, 2019.

[25] CMS. Medicare Claim Submission Guidelines Fact Sheet. http://www.nacns.org/wp-content/uploads/2016/11/CMS_ReimbursementClaim.pdf, 2012.

[26] CMS. Research, Statistics, Data, and Systems. https://www.cms.gov/research-statistics-data-and-systems/research-statistics-data-and-systems.html, 2015.

[27] CMS. Medicare Provider Utilization and Payment Data: Physician and Other Supplier. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Physician-and-Other-Supplier.html, 2016.

[28] CMS. National Provider Identifier Standard (NPI). https://www.cms.gov/Regulations-and-Guidance/Administrative-Simplification/NationalProvIdentStand/, 2016.

[29] CMS. Medicare Provider Utilization and Payment Data: Part D Prescriber. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Part-D-Prescriber.html, 2017.

[30] CMS. Medicare Provider Utilization and Payment Data: Referring Durable Medical Equipment, Prosthetics, Orthotics and Supplies. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/DME.html, 2017.

[31] CMS. Other Entities Frequently Asked Questions. https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/sharedsavingsprogram/Downloads/other-entities-faqs.pdf, 2017.

[32] CMS. The offical U.S. Governement site for Medicare. https://www.medicare.gov/, 2017.

[33] CMS. Center for Medicare and Medicaid Services. https://www.cms.gov/, 2018.

[34] CMS. HCPCS - General Information. https://www.cms.gov/Medicare/Coding/MedHCPCSGenInfo/index.html, 2018.

[35] CMS. Historical. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NationalHealthAccountsHistorical.html, 2018.

[36] CMS. Medicare Fraud Abuse: Prevention, Detection, and Reporting. https://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNProducts/downloads/Fraud_and_Abuse.pdf, 2018.

[37] CMS. National Health Expenditure Projections 2017-2026. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/Downloads/ForecastSummary.pdf, 2018.

[38] CMS. What's Medicare. https://www.medicare.gov/what-medicare-covers/your-medicare-coverage-choices/whats-medicare, 2018.

[39] CMS, Office of Enterprise Data and Analytics. Medicare Fee-For Service Provider Utilization Payment Data Part D Prescriber Public Use File: A Methodological Overview. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Downloads/Prescriber_Methods.pdf, 2018.

[40] CMS, Office of Enterprise Data and Analytics. Medicare Fee-For-Service Provider Utilization Payment Data Physician and Other Supplier. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Downloads/Medicare-Physician-and-Other-Supplier-PUF-Methodology.pdf, 2018.

[41] CMS, Office of Enterprise Data and Analytics. Medicare Fee-For-Service Provider Utilization Payment Data Referring Durable Medical Equipment, Prosthetics, Orthotics and Supplies Public Use File: A Methodological Overview. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Downloads/DME_Methodology.pdf, 2018.

[42] Coalition Against Insurance Fraud. By the numbers: fraud statistics. http://www.insurancefraud.org/statistics.htm, 2017.

[43] K. Dembczynski, A. Jachnik, W. Kotlowski, W. Waegeman, and E. Hüllermeier. Optimizing the f-measure in multi-label classification: Plug-in rule approach versus structured loss minimization. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1130–1138, 2013.

[44] Department of Justice. Florida Doctor Indicted for Role in $13.8 Million Medicare Fraud Scheme. https://www.justice.gov/opa/pr/florida-doctor-indicted-role-138-million-medicare-fraud-scheme, 2017.

[45] Department of Justice U.S. Attorney's Office. Federal Jury Convicts Tinley Park Physician in Medicare Fraud Scheme. https://www.justice.gov/usao-ndil/pr/federal-jury-convicts-tinley-park-physician-medicare-fraud-scheme, 2016.

[46] Q. Dong, S. Gong, and X. Zhu. Imbalanced deep learning by minority class incremental rectification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[47] S. S. Dongre and L. G. Malik. Rare class problem in data mining. *International Journal of Advanced Research in Computer Science*, 8(7):1102–1105, 2017.

[48] M. A. H. Eibe Frank and I. H. Witten. *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.* Morgan Kaufmann, 2016.

[49] K. Feldman and N. V. Chawla. Does medical school training relate to practice? evidence from big data. *Big Data*, 3(2):103–113, 2015.

[50] M. Feldstein. Balancing the goals of health care provision and financing. *Health Affairs*, 25(6):1603–1611, 2006.

[51] A. Fernández, S. del Río, N. V. Chawla, and F. Herrera. An insight into imbalanced big data classification: outcomes and challenges. *Complex & Intelligent Systems*, 3(2):105–120, 2017.

[52] Forbes. Healthcare – 5, 10, 20 years in the past and future. `https://www.forbes.com/sites/singularity/2012/07/02/healthcare-5-10-20-years-in-the-past-and-future/#4d2c89b4310b`, 2017.

[53] G. Forman and M. Scholz. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. *ACM SIGKDD Explorations Newsletter*, 12(1):49–57, 2010.

[54] Fox News. Authorities: $1B Medicare fraud nursing home scam, 3 charged. `http://www.foxnews.com/us/2016/07/22/authorities-1b-medicare-fraud-nursing-home-scam-3-charged.html`, 2017.

[55] A. Gelman. Analysis of variance–why it is more important than ever. *The Annals of Statistics*, 33(1):1–53, 2005.

[56] Google. Life Expectancy. `https://www.google.com/publicdata/explore?ds=d5bncppjof8f9_&met_y=sp_dyn_le00_in&idim=country:USA:GBR:JPN&hl=en&dl=en`, 2017.

[57] GPO. 31 U.S.C. 3729 - False claims. `https://www.gpo.gov/fdsys/granule/USCODE-2011-title31/USCODE-2011-title31-subtitleIII-chap37-subchapIII-sec3729`, 2018.

[58] M. Grimaldi, P. Cunningham, and A. Kokaram. An evaluation of alternative feature selection strategies and ensemble techniques for classifying music. In *Workshop on Multimedia Discovery and Mining*. Citeseer, 2003.

[59] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73:220–239, 2017.

[60] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[61] T. Hasanin and T. M. Khoshgoftaar. The effects of random undersampling with simulated class imbalance for big data. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, pages 70–79. IEEE, 2018.

[62] H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge & Data Engineering*, 21.

[63] Henry J. Kaiser Family Foundation. State Health Facts - Medicare. `http://kff.org/state-category/medicare/`, 2015.

[64] Henry J. Kaiser Family Foundation. Medicare Advantage. `https://www.kff.org/medicare/fact-sheet/medicareadvantage/`, 2017.

[65] Henry J. Kaiser Family Foundation. The Facts on Medicare Spending and Financing. `https://www.kff.org/medicare/issue-brief/the-facts-on-medicare-spending-and-financing/`, 2018.

[66] M. A. Herland, R. A. Bauder, and T. M. Khoshgoftaar. Approaches for identifying us medicare fraud in provider claims data. *Journal of Health Care Management Science*, pages 1–18.

[67] M. A. Herland, R. A. Bauder, and T. M. Khoshgoftaar. The effects of class rarity on the evaluation of supervised healthcare fraud detection models. *Journal of Big Data*, 6(1):21.

[68] M. A. Herland, R. A. Bauder, and T. M. Khoshgoftaar. Medical provider specialty predictions for the detection of anomalous medicare insurance claims. In *2017 IEEE International Conference on Information Reuse and Integration (IRI)*, pages 579–588. IEEE, 2017.

[69] M. A. Herland, R. A. Bauder, and T. M. Khoshgoftaar. The evaluation of medicare fraud predictive models. *Journal of Information Systems Frontiers*, 2019 (Under Review).

[70] M. A. Herland, T. M. Khoshgoftaar, and R. A. Bauder. Big data fraud detection using multiple medicare data sources. *Journal of Big Data*, 5(1):29, 9 2018.

[71] M. A. Herland, T. M. Khoshgoftaar, and R. Wald. Survey of clinical data mining applications on big data in health informatics. In *2013 12th International Conference on Machine Learning and Applications (ICMLA)*, volume 2, pages 465–472. IEEE, 2013.

[72] M. A. Herland, T. M. Khoshgoftaar, and R. Wald. A review of data mining using big data in health informatics. *Journal of Big Data*, 1(1):2, 2014.

[73] HHS. U.S. Department of Health & Human Services. http://www.hhs.gov/, 2015.

[74] HHS. HHS REPORT: Average Health Insurance Premiums Doubled Since 2013. https://www.hhs.gov/about/news/2017/05/23/hhs-report-average-health-insurance-premiums-doubled-2013.html, 2017.

[75] HHS. Health, United States, 2016. https://www.cdc.gov/nchs/data/hus/hus16.pdf, 2018.

[76] N. Japkowicz. Concept-learning in the presence of between-class and within-class imbalances. In *Conference of the Canadian Society for Computational Studies of Intelligence*, pages 67–77. Springer, 2001.

[77] L. A. Jeni, J. F. Cohn, and F. De La Torre. Facing imbalanced data–recommendations for the use of performance metrics. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 245–251. IEEE, 2013.

[78] S. John Walker. Big data: A revolution that will transform how we live, work, and think. *International Journal of Advertising*, 33(1):181–183, 2014.

[79] H. Joudaki, A. Rashidian, B. Minaei-Bidgoli, M. Mahmoodi, B. Geraili, M. Nasiri, and M. Arab. Improving fraud and abuse detection in general physician claims: a data mining study. *International Journal of Health Policy and Management*, 5(3):165, 2016.

[80] Justin Domke. Overfitting, model selection, cross validation, bias-variance. https://people.cs.umass.edu/~domke/courses/sml2011/02overfitting.pdf, 2018.

[81] A. Kankanhalli, J. Hahn, S. Tan, and G. Gao. Big data and analytics in healthcare: Introduction to the special section. *Journal of Information Systems Frontiers*, 18(2):233–235, Apr 2016.

[82] A. Katal, M. Wazid, and R. Goudar. Big data: issues, challenges, tools and good practices. In *2013 Sixth International Conference on Contemporary Computing (IC3)*, pages 404–409. IEEE, 2013.

[83] T. M. Khoshgoftaar, M. Golawala, and J. Van Hulse. An empirical study of learning from imbalanced data using random forest. In *2007 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, volume 2, pages 310–317. IEEE, 2007.

[84] T. M. Khoshgoftaar, C. Seiffert, J. Van Hulse, A. Napolitano, and A. Folleco. Learning with limited minority class data. In *2013 6th International Conference on Machine Learning and Applications (ICMLA)*, pages 348–353. IEEE, 2007.

[85] T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano. Comparing boosting and bagging techniques with noisy and imbalanced data. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 41(3):552–568, 2011.

[86] N. Khurjekar, C.-A. Chou, and M. T. Khasawneh. Detection of fraudulent claims using hierarchical cluster analysis. In *IIE Annual Conference. Proceedings*, page 2388. Institute of Industrial and Systems Engineers (IISE), 2015.

[87] J. S. Ko, H. Chalfin, B. J. Trock, Z. Feng, E. Humphreys, S.-W. Park, H. B. Carter, K. D. Frick, and M. Han. Variability in medicare utilization and payment among urologists. *Urology*, 85(5):1045–1051, 2015.

[88] J. Kodovskỳ. On dangers of cross-validation in steganalysis. Technical report, Citeseer, 2011.

[89] S. le Cessie and J. van Houwelingen. Ridge estimators in logistic regression. *Applied Statistics*, 41(1):191–201, 1992.

[90] J. L. Leevy, T. M. Khoshgoftaar, R. A. Bauder, and N. Seliya. A survey on addressing high-class imbalance in big data. *Journal of Big Data*, 5(1):42, 2018.

[91] LEIE. Office of inspector general leie downloadable databases, 2016.

[92] J. Li, K.-Y. Huang, J. Jin, and J. Shi. A survey on statistical methods for health care fraud detection. *Journal of Health Care Management Science*, 11(3):275–287, 2008.

[93] J. Li, L.-S. Liu, S. Fong, R. K. Wong, S. Mohammed, J. Fiaidhi, Y. Sung, and K. K. Wong. Adaptive swarm balancing algorithms for rare-event prediction in imbalanced healthcare data. *PloS One*, 12(7):e0180830, 2017.

[94] S.-C. Lin, C. Wang, Z.-Y. Wu, and Y.-F. Chung. Detect rare events via mice algorithm with optimal threshold. In *2013 Seventh International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS)*, pages 70–75. IEEE, 2013.

[95] Q. Liu and M. Vasarhelyi. Healthcare fraud detection: A survey and a clustering model incorporating geo-location information. In *29th World Continuous Auditing and Reporting Symposium*, 2013.

[96] M. Maalouf, D. Homouz, and T. B. Trafalis. Logistic regression in large rare events and imbalanced data: A performance comparison of prior correction and weighting methods. *Computational Intelligence*, 34(1):161–174, 2018.

[97] R. Malouf. A comparison of algorithms for maximum entropy parameter estimation. In *Proceedings of the 6th Conference on Natural Language Learning-Volume 20*, pages 1–7. Association for Computational Linguistics, 2002.

[98] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers. Big data: The next frontier for innovation, competition, and productivity. 2011.

[99] Martin Schmitz. When cross validation fails. `https://towardsdatascience.com/when-cross-validation-fails-9bd5a57f07b5`, 2018.

[100] A. McAfee, E. Brynjolfsson, T. H. Davenport, D. Patil, and D. Barton. Big data: the management revolution. *Harvard Business Review*, 90(10):60–68, 2012.

[101] MedicalBillingAndCoding.org. 4.03: Common problems in medical coding. `https://www.medicalbillingandcoding.org/common-problems-coding/`, 2019.

[102] X. Meng, J. Bradley, B. Yavuz, E. Sparks, S. Venkataraman, D. Liu, J. Freeman, D. Tsai, M. Amde, S. Owen, et al. Mllib: Machine learning in apache spark. *The Journal of Machine Learning Research*, 17(1):1235–1241, 2016.

[103] Missouri Department of Social Services. Provider Sanctions. `https://mmac.mo.gov/providers/provider-sanctions/`, 2018.

[104] L. Morris. Combating fraud in health care: an essential component of any cost containment strategy. *Health Affairs*, 28(5):1351–1356, 2009.

[105] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic. Deep learning applications and challenges in big data analytics. *Journal of Big Data*, 2(1):1, 2015.

[106] OECD Data. Elderly population. `https://data.oecd.org/pop/elderly-population.htm#indicator-chart`, 2019.

[107] F. J. Ohlhorst. *Big Data Analytics: Turning Big Data Into Big Money*. John Wiley & Sons, 2012.

[108] OIG. Exclusions Program. `https://oig.hhs.gov/exclusions/index.asp`, 2016.

[109] OIG. Medicare Fraud Strike Force. `https://www.oig.hhs.gov/fraud/strike-force/`, 2017.

[110] OIG. Office of Inspector General Exclusion Authorities US Department of Health and Human Services. `https://oig.hhs.gov/`, 2018.

[111] OIG. Criminal and Civil Enforcement. `https://oig.hhs.gov/fraud/enforcement/criminal/`, 2019.

[112] Pacific Rim. Health insurance fraud and its impact on the health care system. `https://www.pacificprime.com/blog/health-care-system-fraud-impacts.html`, 2019.

[113] Palm Beach Post. North Palm Beach eye doctor Melgen jailed on Medicare fraud charges. http://www.palmbeachpost.com/news/crime--law/north-palm-beach-eye-doctor-melgen-jailed-medicare-fraud-charges/czyT7d9jhcVrZfbQpyUUcP/, 2017.

[114] V. Pande and W. Maas. Physician medicare fraud: characteristics and consequences. *International Journal of Pharmaceutical and Healthcare Marketing*, 7(1):8–33, 2013.

[115] M. P. Pawar. Review on data mining techniques for fraud detection in health insurance. *International Journal on Emerging Trends in Technology (IJETT)*, 3(2), 2016.

[116] J. Platt. Fast training of support vector machines using sequential minimal optimization. In B. Schoelkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1998.

[117] Prashant Gupta. Cross-validation in machine learning. https://towardsdatascience.com/cross-validation-in-machine-learning-72924a69872f, 2018.

[118] Python Software Foundation. Python. https://www.python.org/, 2017.

[119] R. B. Rao, G. Fung, and R. Rosales. On the dangers of cross-validation. an experimental evaluation. In *Proceedings of the 2008 SIAM International Conference on Data Mining*, pages 588–596. SIAM, 2008.

[120] A. Rashidian, H. Joudaki, and T. Vian. No evidence of the effect of the interventions to combat health care fraud and abuse: a systematic review of literature. *PloS One*, 7(8):e41988, 2012.

[121] A. K. Rastogi, N. Narang, and Z. A. Siddiqui. Imbalanced big data classification: a distributed implementation of smote. In *Proceedings of the Workshop Program of the 19th International Conference on Distributed Computing and Networking*, page 14. ACM, 2018.

[122] C. K. Reddy and C. C. Aggarwal. *Healthcare data analytics*, volume 36. CRC Press, 2015.

[123] J. Roski, G. W. Bo-Linn, and T. A. Andrews. Creating value in health care through big data: opportunities and policy implications. *Health Affairs*, 33(7):1115–1122, 2014.

[124] S. Sadiq, Y. Tao, Y. Yan, and M.-L. Shyu. Mining anomalies in medicare big data using patient rule induction method. In *2017 IEEE Third International Conference on Multimedia Big Data (BigMM)*, pages 185–192. IEEE, 2017.

[125] Santa Clara, Oct 6, 2013, in Conjuction with the IEEE International Conference on BigData. Big data in bioinformatics and health care informatics. http://www.ittc.ku.edu/~jhuan/BBH/, 2017.

[126] B. Sawyer and C. Cox. How does health spending in the U.S. compare to other countries? https://www.healthsystemtracker.org/chart-collection/health-spending-u-s-compare-countries/#item-relative-size-wealth-u-s-spends-disproportionate-amount-health, 2018.

[127] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano. Mining data with rare events: a case study. In *2007 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, volume 2, pages 132–139. IEEE, 2007.

[128] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano. Rusboost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 40(1):185–197, 2010.

[129] N. Seliya, T. M. Khoshgoftaar, and J. Van Hulse. A study on the relationships of classifier performance metrics. In *2009 21st IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 59–66. IEEE, 2009.

[130] S. Senthilkumar, B. K. Rai, A. A. Meshram, A. Gunasekaran, and S. Chandrakumarmangalam. Big data in healthcare management: A review of literature. *American Journal of Theoretical and Applied Business*, 4(2):57–69, 2018.

[131] J. G. Shanahan and L. Dai. Large scale distributed data science using apache spark. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2323–2324. ACM, 2015.

[132] A. Sheshasaayee and S. S. Thomas. A purview of the impact of supervised learning methodologies on health insurance fraud detection. In *Information Systems Design and Intelligent Applications*, pages 978–984. Springer, 2018.

[133] A. Tayal, T. F. Coleman, and Y. Li. Rankrc: Large-scale nonlinear rare class ranking. *IEEE transactions on knowledge and data engineering*, 27(12):3347–3359, 2015.

[134] Texas Office of Inspector General. Texas Exclusion List. https://oig.hhsc.state.tx.us/oigportal/EXCLUSIONS/tabid/81/ctl/DOW/mid/407/Default.aspx, urldate = 2018-01-04, year=2018.

[135] The Motley Fool. Medicare Could Cost You More Than You Think in 2019. https://www.fool.com/investing/2018/10/14/medicare-could-cost-you-more-than-you-think-in-201.aspx, 2019.

[136] The R Foundation. What is R? https://www.r-project.org/about.html, 2017.

[137] J. W. Tukey. Comparing individual means in the analysis of variance. *Biometrics*, 5(2):99–114, 1949.

[138] M. P. A. C. (US). *Report to the Congress, Medicare Payment Policy*. Medicare Payment Advisory Commission, 2007.

[139] G. Van Capelleveen, M. Poel, R. M. Mueller, D. Thornton, and J. van Hillegersberg. Outlier detection in healthcare fraud: A case study in the medicaid dental domain. *International Journal of Accounting Information Systems*, 21:18–31, 2016.

[140] J. Van Hulse, T. M. Khoshgoftaar, and A. Napolitano. Experimental perspectives on learning from imbalanced data. In *Proceedings of the 24th International Conference on Machine Learning*, pages 935–942. ACM, 2007.

[141] J. Van Hulse, T. M. Khoshgoftaar, A. Napolitano, and R. Wald. Feature selection with high-dimensional imbalanced data. In *2009 IEEE International Conference on Data Mining Workshops*, pages 507–514. IEEE, 2009.

[142] G. Varoquaux. Cross-validation failure: small sample sizes lead to large error bars. *Neuroimage*, 2017.

[143] Verify Comply. Exclusion Lists. `https://verifycomply.com/exclusion-lists.asp`, 2018.

[144] S. S. Waghade and A. M. Karandikar. A comprehensive study of healthcare fraud detection based on machine learning. *International Journal of Applied Engineering Research*, 13(6):4175–4178, 2018.

[145] G. M. Weiss. Mining with rarity: a unifying framework. *ACM SIGKDD Explorations Newsletter*, 6(1):7–19, 2004.

[146] K. Weiss, T. M. Khoshgoftaar, and D. Wang. A survey of transfer learning. *Journal of Big Data*, 3(1):9, 2016.

[147] Weka. CostSensitiveClassifier. `https://weka.wikispaces.com/CostSensitiveClassifier`, 2017.

[148] Weka.sourceforge. Class SpreadSubsample. `http://weka.sourceforge.net/doc.stable/weka/filters/supervised/instance/SpreadSubsample.html`, 2017.

[149] J. A. Westerhuis, H. C. Hoefsloot, S. Smit, D. J. Vis, A. K. Smilde, E. J. Van Velzen, J. P. Van Duijnhoven, and F. A. Van Dorsten. Assessment of plsda cross validation. *Metabolomics*, 4(1):81–89, 2008.

[150] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2016.

[151] WSIL. Healthcare big data analytics market 2018 ready to reach upto us$ 9.5 billion with cagr of 11.5% | by component, deployment type, application, software type and by end user- forecast to 2023, 2017.

[152] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M. J. Franklin, S. Shenker, and I. Stoica. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In *Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation*, NSDI'12, pages 2–2, Berkeley, CA, USA, 2012. USENIX Association.

[153] M. Zaharia, R. S. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave, X. Meng, J. Rosen, S. Venkataraman, and M. J. Franklin. Apache spark: a unified engine for big data processing. *Communications of the ACM*, 59(11):56–65, 2016.

[154] J. Zhai, S. Zhang, and C. Wang. The classification of imbalanced large data sets based on mapreduce and ensemble of elm classifiers. *International Journal of Machine Learning and Cybernetics*, 8(3):1009–1017, 2017.

[155] W. Zhang, S. Kobeissi, S. Tomko, and C. Challis. Adaptive sampling scheme for learning in severely imbalanced large scale data. In *Asian Conference on Machine Learning*, pages 240–247, 2017.

[156] X. Zhang, Y. Li, R. Kotagiri, L. Wu, Z. Tari, and M. Cheriet. Krnn: k rare-class nearest neighbour classification. *Pattern Recognition*, 62:33–44, 2017.