AN EVALUATION OF UNSUPERVISED MACHINE LEARNING ALGORITHMS FOR DETECTING FRAUD AND ABUSE IN THE U.S. MEDICARE INSURANCE PROGRAM

by

Raquel C. da Rosa

A Thesis Submitted to the Faculty of

The College of Engineering and Computer Science

in Partial Fulfillment of the Requirements for the Degree of

Master of Science

Florida Atlantic University

Boca Raton, FL

May 2018

AN EVALUATION OF UNSUPERVISED MACHINE LEARNING ALGORITHMS FOR DETECTING FRAUD AND ABUSE IN THE U.S. MEDICARE INSURANCE PROGRAM

by

Raquel C. da Rosa

This thesis was prepared under the direction of the candidate's thesis advisor, Dr. Taghi M. Khoshgoftaar, Department of Computer and Electrical Engineering and Computer Science, and has been approved by the members of her 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 Master of Science.

SUPERVISORY COMMITTEE:

Taghi M. Khoshgoftaar, Ph.D.
Thesis Advisor

Mehrdad Nojoumian, Ph.D.

Dingding Wang, 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, Ph.D.
Interim Dean, Graduate College

April 25, 2018
Date

ACKNOWLEDGEMENTS

I would first like to thank my thesis advisor, Dr. Taghi M. Khoshgoftaar, for the patient guidance, encouragement, and advice he has provided throughout my time as his student. Dr. Khoshgoftaar was thoroughly dedicated to making me a better researcher. Additionally, I would like to thank my thesis committee members, Dr. Mehrdad Nojoumian and Dr. Dingding Wang. I would also like to thank Richard A. Bauder for his help on this research project.

Finally, I must express my very sincere gratitude to my husband Lindino, my daughter Carolina, my parents Diomário and Lédia, and to my sister Isabela, for providing me with constant support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

ABSTRACT

Author:	Raquel C. da Rosa
Title:	An evaluation of Unsupervised Machine Learning Algorithms for Detecting Fraud and Abuse in the U.S. Medicare Insurance Program
Institution:	Florida Atlantic University
Thesis Advisor:	Dr. Taghi M. Khoshgoftaar
Degree:	Master of Science
Year:	2018

The population of people ages 65 and older has increased since the 1960s and current estimates indicate it will double by 2060. Medicare is a federal health insurance program for people 65 or older in the United States. Medicare claims fraud and abuse is an ongoing issue that wastes a large amount of money every year resulting in higher health care costs and taxes for everyone. In this study, an empirical evaluation of several unsupervised machine learning approaches is performed which indicates reasonable fraud detection results. We employ two unsupervised machine learning algorithms, Isolation Forest and Unsupervised Random Forest, which have not been previously used for the detection of fraud and abuse on Medicare data. Additionally, we implement three other machine learning methods previously applied on Medicare data which include: Local Outlier Factor, Autoencoder, and k-Nearest Neighbor. For our dataset, we combine the 2012 to 2015 Medicare provider utilization and payment data and add fraud labels from the List of Excluded Individuals/Entities (LEIE) database. Results show that Local Outlier Factor is the best model to use for Medicare fraud detection.

AN EVALUATION OF UNSUPERVISED MACHINE LEARNING ALGORITHMS FOR DETECTING FRAUD AND ABUSE IN THE U.S. MEDICARE INSURANCE PROGRAM

List of Tables . . . viii

List of Figures . . . ix

1 Introduction . . . 1

1.1 Motivation . . . 2

1.2 Contributions . . . 3

1.3 Organization . . . 3

2 Related work . . . 4

2.1 Isolation Forest . . . 4

2.2 Local Outlier Factor . . . 8

2.3 Unsupervised Random Forest . . . 13

2.4 Autoencoders . . . 15

2.5 k-Nearest Neighbors . . . 18

3 Methodology . . . 21

3.1 Data . . . 21

3.2 Machine Learning . . . 24

3.2.1 Outlier Detection . . . 26

3.2.2 Supervised and Unsupervised Learning . . . 26

3.3 Machine Learning Algorithms . . . 27

3.3.1 Isolation Forest . . . 27

3.3.2 Local Outlier Factor 28
3.3.3 Unsupervised Random Forest 29
3.3.4 Autoencoders 29
3.3.5 k-Nearest Neighbors 30
3.4 Performance Metrics 31
3.4.1 ROC Curve 31
3.4.2 Area Under the Curve, AUC Curve 31

4 Case Studies 33
4.1 Isolation Forest 33
4.2 Local Outlier Factor 35
4.3 Unsupervised Random Forest 37
4.4 Autoencoder 39
4.5 k-Nearest Neighbors 41
4.6 Comparing all Models 43

5 Conclusions and Future Work 45
5.1 Conclusions 45
5.2 Future Work 46

Bibliography 48

LIST OF TABLES

3.1 LEIE Exclusion Rules 23

3.2 Description of Medicare Features 25

3.3 Medicare Datasets 25

4.1 Performance of IF Model 34

4.2 Performance of LOF Models 36

4.3 Performance of URF Model 38

4.4 Performance of AE Models 40

4.5 Performance of KNN Model 42

4.6 Best Performance of all Models 44

LIST OF FIGURES

3.1 Outlier Detection 26

4.1 AUC curve - IF 35

4.2 AUC curve - LOF 37

4.3 AUC curve - URF 39

4.4 AUC curve - AE Models 41

4.5 AUC curve - KNN 43

4.6 AUC curve - All Models 44

CHAPTER 1

INTRODUCTION

Medicare is a federal health insurance program for people in the United States (US) aged 65 or older, younger individuals with specific disability statuses, and people with end-stage renal disease and amyotrophic lateral sclerosis [14]. In 2016, Medicare covered 56.8 million people with 47.8 million aged 65 and older, and 9.0 million disabled [14]. The number of people ages 65 and older has increased steadily in the United States since the 1960s and is projected to more than double from 46 million in 2015 to over 98 million by 2060. Between 2020 and 2030, the number of older persons is projected to increase by almost 18 million as the last of the baby boomers reaches age 65 [63].

Medicare expenditures represented 3.6 percent of Gross Domestic Product (GDP) in 2016 [1]. The funds for Medicare come from Social Security Administration funded by taxpayers [15]. A large amount of these funds are wasted in Medicare fraud every year, eventually increasing health care costs and taxes for everyone [9]. It is estimated by the Federal Bureau of Investigation (FBI), that health care fraud costs tens of billions of dollars per year [8]. For example, in June 2015, 243 people where arrested across the US, charged with submitting fake billing for Medicare that totaled $712 million. The Medicare Fraud Strike Force teams, brings together the efforts of the Office of Inspector General, the Department of Justice, Offices of the United States Attorneys, the Federal Bureau of Investigation, local law enforcement, and other entities to combat and prevent health care fraud, waste, and abuse. These teams have a proven record of success in analyzing data and investigative intelligence to quickly identify fraud and bring prosecutions [10].

Despite the current efforts to combat Medicare fraud, which includes the Medicare Fraud Strike Force teams, fraud is still rampant and thus requires new and innovative ways to continue to reduce financial losses. The Medicare database is extensive, including insurance claims, clinical data, information on providers, health records, and other significant data. To process this data to detect fraudulent activities, such as fake billing and upcoding [25], requires the use of specialized tools and techniques. This study aims to apply machine learning algorithms as an outlier detection tool to identify fraud and abuse in Medicare data.

1.1 MOTIVATION

Outlier detection is a common approach used in machine learning to identify anomalous patterns in data. An outlier often contains useful information about abnormal characteristics of the data. The recognition of such unusual characteristics provides useful application-specific insights. For instance, outlier detection can be useful to law enforcement, particularly in cases where abnormal patterns can only be discovered from the multiple actions of an entity over an extended period. Another example occurs when detecting fraud in trading activities, financial transactions, or insurance claims which usually require the identification of abnormal patterns in the data generated by the activity of the criminal entity [44].

Research related to outlier detection on health care data has explored a variety of machine learning models to detect abnormalities; however, from the currently available literature, we did not identify publications that used Isolation Forest or Unsupervised Random Forest to detect outliers on health care data. Therefore, we test these models together with Local Outlier Factor, Autoencoder, and k-Nearest Neighbor to compare model performance and evaluate its effectiveness in detecting anomalous patterns in Medicare data.

1.2 CONTRIBUTIONS

The contributions of this thesis are:

1. To explore Isolation Forest and Unsupervised Random Forest with Medicare data from the Medicare Part B provided by the Center for Medicare and Medicaid Services, for the first time.

2. To conduct a comparative study of Isolation Forest (IF), Local Outlier Factor (LOF), Unsupervised Random Forest (URF), Autoencoder (AE), and k-Nearest Neighbor (KNN) on the detection of fraud and abuse in the Medicare system.

3. The use of real-world fraud labels from the LEIE database for the evaluation of fraud detection performance of unsupervised methods.

1.3 ORGANIZATION

The remainder of this thesis is organized into the following chapters:

- Chapter 2 provides information on previous studies related to each of the machine learning models used on this project.

- Chapter 3 gives details on the methodology applied including data, machine learning algorithms, and performance metrics.

- Chapter 4 includes the case studies and results.

- Chapter 5 presents the conclusions of our experiments and suggestions for future work.

CHAPTER 2

RELATED WORK

The main focus of this thesis is the evaluation of the five different machine learning algorithms in their ability to detect fraud and abuse using Medicare data, in particular, employing Isolation Forest and Unsupervised Random Forest, which have not been previously evaluated with Medicare outlier detection. Even though this chapter provides information on previous studies related to each of the machine learning models used in this study, we briefly present two other studies that use the same Medicare dataset and apply regression and probability models to detect fraud. In one study by Bauder and Khoshgoftaar [26], the authors use multivariate regression to establish a baseline for expected Medicare payments, per medical specialty. This baseline is then used as the normal case in which to compare the actual payment amounts, with any deviations flagged as outliers. In another study [29], Bauder and Khoshgoftaar incorporate a two-step approach in detecting Medicare fraud by specialty. Their method employs a multivariate regression model with the residuals passed into a Bayesian probability model. This model produces probabilities indicating how likely it is that a particular value is fraudulent for further investigation.

2.1 ISOLATION FOREST

This section presents previous studies performed using IF [56]. IF has been applied within diverse subject areas. When searching for studies using IF related to detecting Medicare fraud, the only work identified, at the time of this research, was Liu et al. [58]. The authors briefly mention the use of IF for outlier detection but do

not present a case study that could prove the efficacy of this method on detecting fraud in Medicare. The authors developed a graph analysis technique to search for fraud, waste, and abuse activities on real-world Medicaid health care datasets as part of their tool known as the Xerox Program Integrity Validator (XPIV). The tool provides two categories of functionalities: automated screening for which an analyst can focus attention on a small list of suspect providers as opposed to a large set, and an interactive drill-down where the analyst starts from a suspicious individual or activity and interacts with the system to navigate through data items and collect evidence to build an investigation case. The two categories have different technical focal points. Automated screening focuses on algorithmic design for detecting diverse forms of outliers and interactive drill-down focuses on database indexing/caching for fast data retrieval and user interface design for intuitive user-system interaction.

Previous studies added different techniques to further improve the performance of IF for outlier detection. Chen et al. [41] proposed an algorithm called Isolation Forest Outlier detection and Subset selection (IOS), which can detect outliers and select representative subsets simultaneously, reducing prediction errors significantly compared with other methods. Ding and Fei [44] proposed an adapted streaming data outlier detection algorithm based on IF, namely iForestASD, which is suited for outlier detection for streaming data. The experiment results performed on four real-world datasets from the UCI repository demonstrate that the proposed algorithm can effectively detect anomalous instances from the streaming data. The authors validated that the proposed method is suitable for streaming data outlier detection but did not compare it with other existing methods. Sun et al. [79] presented an anomalous user behavior detection framework that applies an extended version of the IF algorithm. They propose a simple method for extending the IF algorithm to datasets which include categorical dimensions. The authors showed that the method is relativelly fast and scalable and does not require example anomalies in the training dataset.

The proposed method was applied to an enterprise dataset of staff accessing the payroll system in a large organization which showed promising results with the system being able to isolate anomalous instances from the baseline user model, using a single feature or combined features. Puggini and McLoone [71] consider the dimensionality and variable correlation problems related to the use of Optical Emission Spectroscopy (OES) data for interpretable outlier detection in semiconductor manufacturing. Dimensionality reduction tailored to outlier detection together with IF for outlier score generation were proposed as an outlier detection methodology. The approach consisted of an outlier diagnosis system based on IF that allows individual contributing variables to be identified. Falk et al. [45] presented novel approaches to detecting outages in a mobile network using non-parametric outlier detection methods. They used performance data from a 4G-LTE network carrier to train two parameter-free models. The first model relies on IF, while the second is histogram based. The trained models represent the data characteristics for normal periods; new data is matched against the trained models to classify the new time period as being normal or abnormal. They show that the proposed methods can gauge the mobile network state with more subtlety than standard success/failure thresholds used in real-world networks. Gamachchi et al. [49] introduced a framework based on graphical and outlier detection approaches for identifying potential malicious insider threats. Their model generates outlier scores based on different input parameters for each user. Considering the nature of insider attacks, a user can be deemed to be suspicious even if a single parameter has been found to be suspicious. They have adopted graph and subgraph properties and statistical methods in generating input parameters for the outlier detection algorithm through multi-domain real-world information. Empirical results reveal this framework to be useful in differentiating the majority of users with typical behavior from the minority of users who show suspicious behavior. They also found that more than 79% of users have common behavioral patterns, whereas the

rest of the group shows suspicious behavior based on different parameters. Users belonging to the minority group can be tagged and directed for further investigation. The IF algorithm is executed for isolating anomalous users within the Anomaly Detection Unit (ADU). Anomaly scores for each user are generated as the output of the ADU. These values are used in identifying and separating possible malicious users from the rest of the workforce. Calheiros et al. [37] proposed a method to code time-series information as extra attributes that enable temporal outlier detection, as well as establish its feasibility to adapt to seasonality and trends in the time-series and to be applied online and in real-time. They investigated the applicability of the IF outlier detection algorithm for detection of abnormal events in resource utilization of large-scale cloud data centers. they demonstrated how time-series information was extracted into extra attributes that enabled temporal outlier detection. Then they investigated the capacity of the method in detecting seasons and trends in the dataset, along with the method's feasibility for online and real-time outlier detection.

Other studies tested the use of IF in different subject areas. He et al. [52] proposed the idea to use outlier detection methods to predict bugs in software code changes. IF was adopted as the outlier detection method used in this study. The prediction method consisted of three steps, including change data extraction, data preprocessing, model construction, and prediction. To validate the effectiveness of the proposed prediction method, eight open source Java projects were chosen for their empirical study. The experimental results showed that IF performed effectively and better than other two traditional classification methods in bug prediction in software code changes. Zhang et al. [88] investigated unsupervised techniques for outlier-based network intrusion detection. They used real-time traffic data from University of Virginia's network. The authors evaluated the performance between LOF and IF by probing the similarities and differences between the result of each approach. Distribution plots indicated there was a greater variation of attributes in outliers

identified by IF than those outliers identified by LOF. With the assumptions that outliers are points that are rare and distinctive, they find that IF performs well in identifying outliers compared to LOF.

2.2 LOCAL OUTLIER FACTOR

This section presents some previous studies performed using LOF [35]. LOF has been widely used, but our data mining and machine learning group at Florida Atlantic University was the first and only to use LOF with the U.S. Medicare data provided by the Centers for Medicare and Medicaid Service [2]. Bauder and Khoshgoftaar [30] proposed a new method for discovering outliers in Medicare payment data using multiple predictors as model inputs. The authors used the 2012-2014 Medicare and Medicaid Service data [2], and compared the new approach with LOF, AE, and KNN. Their multivariate outlier detection approach has two parts: (1) to create a Multivariate Adaptive Regression Splines model to produce studentized residuals; (2) to use the residuals as input into a general univariate outlier detection model, based on full Bayesian inference, using probabilistic programming. Using this approach, they incorporated multiple variables to detect outliers with a model that provided probability distributions with credible intervals. These credible intervals further enhance confidence that the detected outliers are true abnormal values, thus possibly fraudulent activities. The results showed that the successful detection of these potential fraudulent activities can provide effective and meaningful results for further investigation. Bauder and Khoshgoftaar [28] compared several machine learning methods, including LOF and Autoencoders, to detect fraud on Medicare data. The authors performed a comparative study with supervised, unsupervised, and hybrid machine learning approaches using four performance metrics and class imbalance reduction via oversampling and an 80-20 undersampling method. They group the 2015 Medicare data from CMS into provider types, with fraud labels from the List of Excluded

Individuals/Entities database. Results show that the successful detection of fraudulent providers is possible, with the 80-20 sampling method demonstrating the best performance across the learners. Supervised methods performed better than unsupervised or hybrid methods, varying based on the class imbalance sampling technique and provider type. Bauder and Khoshgoftaar [27] proposed a general outlier detection model, based on Bayesian inference, using probabilistic programming. The model provided probability distributions rather than just point values, as with most common outlier detection methods. Credible intervals were also generated to further enhance confidence that the detected outliers should in fact, be considered outliers. Two case studies were presented demonstrating the model's effectiveness in detecting outliers. The first case study used temperature data to provide a clear comparison of several outlier detection techniques, including LOF. The second case study used a Medicare dataset to showcase the proposed outlier detection model. The results showed that the successful detection of outliers, which indicate possible fraudulent activities, can provide effective and meaningful results for further investigation within medical specialties or by using real-world, medical provider fraud investigation cases. Zhang and He [89] proposed a Medicare fraud detection framework based on outlier detection using datasets provided by the medical insurance bureau of a city in Sichuan Province. The method consisted of two parts. The first part was a spatial density-based algorithm, called improved LOF (imLOF), which is more applicable than simple LOF when using medical insurance data. The second part was robust regression to depict the linear dependence between variables. Experiments showed that the new method is effective in detecting anomalies. Shan et al. [75] presented an application of LOF in public health service management. The authors studied the potential of applying the outlier detection method to medical specialist groups to discover inappropriate billings. The results were validated by specialist compliance history and direct domain expert evaluation. The results suggested that LOF is an effective method for

identifying inappropriate billing patterns and is a valuable tool for monitoring medical practitioner billing compliance. Paulauskas and Bagdonas [68] presented a novel approach to detect the network flow outliers. The method relied on aggregated network flow metrics and is based on LOF algorithm, which evaluates each event's uniqueness based on distance from the k-Nearest Neighbors. In their research, 15 different groups of features (a total of 74 features) were suggested to detect anomalous network flows. According to experimental results, the best groups of features were identified with the highest values of precision, recall, and F-measure. Hamlet et al. [51] proposed a novel incremental modification to the Local Outlier Probabilities algorithm to enable it to detect outliers almost instantly in data streams. The presented incremental algorithm's strength was based on denying the insertion of incremental points into the dataset. This method prevented the outlier scores of other points from having to update, saving computational time, while resulting in a small amount of error. The goal of their study was to allow low-resource machines, such as small or older satellites, to perform incremental outlier detection on large static datasets quickly, exchanging accuracy impairment for speed of detection. Ortner et al. [65] presented a novel approach for outlier detection, called local projections, which is based on concepts of LOF and RobPCA [54]. By applying aspects of both methods, the proposed method is robust towards noise variables and can perform outlier detection in multi-group situations. For each observation of a dataset, they identified a local group of dense nearby observations, which were called a core, based on a modification of the k-Nearest Neighbors algorithm. By projecting the dataset onto the space spanned by those observations, two aspects were revealed. First, it was possible to analyze the distance from an observation to the center of the core within the projection space to provide a measure of quality of description of the observation by the projection. Second, they considered the distance of the observation to the projection space to assess the suitability of the core for describing the "outlyingness" [69] of the obser-

vation. These new interpretations lead to a univariate measure of outlyingness based on aggregations over all local projections, which outperforms LOF and RobPCA. Experiments in the context of real-world applications employing datasets of various dimensionality demonstrated the advantages of local projections. Prez-Ra et al. [69] proposed an original method for detecting and localizing anomalous motion patterns in videos from a camera view-based motion representation perspective. As part of their method, they used LOF to detect anomalous motion pattern in any block at any time instant of a video sequence. The threshold value was automatically set in each block by means of statistical arguments. The authors reported comparative experiments on several video datasets demonstrating that their method was highly competitive for the intricate task of detecting different types of anomalous motion in videos. Valentino et al. [81] presented an approach based on Principal Component Analysis and LOF to detect outliers in the loss maps in the Large Hadron Collider (LHC) and provide an automatic check of the collimation hierarchy. Yan et al. [85] proposed a distributed solution for the LOF method. Their study presented a distributed LOF pipeline framework, called DLOF. Each stage of the DLOF computation was conducted in a fully distributed fashion by leveraging their invariant observation for intermediate value management. They also proposed a data assignment strategy which ensures that each machine is self-sufficient in all stages of the DLOF pipeline while minimizing the number of data replicas. Based on the convergence property derived from analyzing this strategy in the context of real-world datasets, they introduced several data-driven optimization strategies. These strategies not only minimize the computation costs within each stage but also eliminated unnecessary communication costs by aggressively pushing the LOF computation into the early stages of the DLOF pipeline. Results, using both real and synthetic datasets confirmed the efficiency and scalability of the proposed approach with Terabyte-level data. Hoang et al. [53] introduced two improved algorithm methods, namely INFLOF (Influenced

Local Outlier Factor) and COF (Connectivity based Outlier Factor), which are based on LOF. INFLOF can solve the problem of edge misjudgment caused by different density cluster's being too close to each other in the dataset, while COF can solve the problem of outliers. Zhang et al. [91] presented a new potential radio frequency interference (RFI) detection and mitigation algorithm that was based on the LOF. Experimental results showed that a satisfactory performance can be obtained by a LOF algorithm even in detecting moderate RFI. Ma et al. [61] introduced a density-based outlier detection method by estimating the LOF on a projected principal component analysis (PCA) domain from real-world spatial-temporal (ST) traffic signals. The goal of their study was to detect traffic data outliers which were errors in data and traffic anomalies in real situations such as accidents, congestions, and low volume. Based on the designed LOF algorithm, a semi-supervised technique was employed to label any embedded outliers. The method reached an average detection success rate of 93.5%. Bhatt et al. [32] studied the impact of k-means and local outlier factor on a given bank dataset. Their research described the comparative study of five distinct methodologies using k-means as the base algorithm, along with the various distances method applied in finding the dissimilarities between the objects. Zhao et al., 2014 [92], proposed the use of LOF in photovoltaic (PV) installations to improve fault detection. Their suggested method demonstrated several advantages over traditional PV monitoring systems, such as simplicity, quick response, easy implementation, and weather information was not necessary. A brief comparison between statistic outlier rules and LOF was presented showing that, with statistic outlier detection rule, the upper and lower bounds provide easy data interpretation but it could cause false alarms. The LOF, contrastingly, was more reliable with no false alarm but it lost physical meaning in the original data and it had a higher computational cost. Feng et al. [47] introduced a two-stage machine learning system to detect outliers. In the first stage, they projected the access logs of cloud file-sharing services onto re-

lationship graphs and used three complementary graph-based unsupervised learning methods, OddBall [18], PageRank [67] and LOF [35], to generate outlier indicators. In the second stage, they generated an ensemble of outlier indicators and introduced the discrete wavelet transform (DWT) method, as well as proposed a procedure to use wavelet coefficients with the Haar wavelet function to identify outliers for insider threat. The proposed system has been deployed in a real business environment and demonstrated to be effective in detecting outliers from several selected case studies. Atzmueller et al. [21] presented a sequential modeling and outlier analytics approach in an industrial application context. They applied LOF for detecting outliers on the numeric sensor data. Due to the nature of the provided sensor data, the concept of a locally sensitive algorithm was found to be useful, because with different set points (for plant operation) range and characteristics of the sensor readings varied greatly in their study. The outlier scores could be calculated for either all available sensors, for certain subgroups, or single sensors, depending on the desired granularity.

2.3 UNSUPERVISED RANDOM FOREST

Random Forest has been widely used for prediction and feature selection; however, the use of Random Forest as an unsupervised model is limited. At the time of this study, we did not find any publication related to using URF to detect fraud in the Medicare system. There are a few publications on the use of URF in other areas. Baron and Poznanski [23] presented an outlier detection algorithm based on URF. They tested the algorithm on more than two million galaxy spectra from the Sloan Digital Sky Survey and examined 400 galaxies with the highest outlier score. They found objects which have extreme emission line ratios and abnormally strong absorption lines and objects with unusual continua, including extremely reddened galaxies. The algorithm can be executed on imaging, time series and other spectroscopic data, operates well with thousands of features, is not sensitive to missing values, and is eas-

ily parallelizable. URF has also been studied in search schemes where Yu et al. [87] developed an URF voting technique for action detection and search. This method has the unique property that it is easy to leverage feedback from the user. The interest point matching is much faster than the existing nearest-neighbor-based method. To handle the computational cost in searching the large video space, they proposed a coarse-to-fine subvolume search scheme, which results in a dramatic speedup over the existing video branch-and-bound method. Cross-dataset experiments demonstrate that the proposed method was not only fast to search higher-resolution videos but also robust to action variations, partial occlusions, and cluttered and dynamic backgrounds. Their interactive action search results show that the detection performance would be improved significantly after only a few rounds of relevance feedback. URF was also used in detecting fraud in telecommunications data as proposed by Saaid et al. [72]. They used URF to detect fraud in telecommunications data which consisted of millions of call records generated each day. The fraud detection was implemented via the construction of user call profiles using the calls detail records (CDR) data. Their work attempted to investigate the reliability of the URF method in building the profiles using its variable importance measure. The study showed that, using different number of input variables for splitting the trees nodes, there is no bias in the selection process. Furthermore, identification of the first five important variables was similar regardless of the number of variables used to split the nodes. Variable importance measures in the RF method were found to be reliable in the rare event or unbalanced datasets, which is one of the characteristics of fraud cases. The URF method in the simulation study indicated that the runtime depends on the number of variables used to split the nodes of each tree.

2.4 AUTOENCODERS

The use of autoencoders [64] for outlier detection has been applied in several subject areas such as maritime surveillance [70], video monitoring [42], and network intrusion [66]. Our group at Florida Atlantic University was, again, the first and only to apply autoencoders with the U.S. Medicare data provided by the Centers for Medicare and Medicaid Service [2]. As mentioned on previous LOF section, Bauder and Khoshgoftaar [28] compared Autoencoders, LOF and KNN among other machine learning methods, to detect fraud on Medicare data performing a comparative study with supervised, unsupervised, and hybrid machine learning approaches using four performance metrics and class imbalance reduction via oversampling and an 80-20 undersampling method. Supervised methods performed better than unsupervised or hybrid methods. Aygun and Yavuz [22], proposed two deep learning based outlier detection models using autoencoder and denoising autoencoder to detect zero-day attacks. These methods produce higher accuracy over other methods because intrusion detection systems do not perform well when it comes to detecting zero-day attacks. The obtained results show that as a singular model, the proposed outlier detection models outperformed any other singular outlier detection methods and perform almost the same as the newly suggested hybrid outlier detection models. Lu et al. [59] introduced a novel deep structured framework to solve the challenging sequential outlier detection problem. They used autoencoder models to capture the intrinsic difference between outliers and normal instances and integrate the models to recurrent neural networks that allow the learning to make use of the previous context, as well as make the learners more robust to warp along the time axis. Experimental results demonstrated the effectiveness of their model compared with other state-of-the-art approaches. Chen et al. [39] introduced autoencoder ensembles for unsupervised outlier detection. One problem with neural networks is that they are sensitive to noise and often require large datasets to work correctly. Increasing the

data size slows down the model, as a result, there are only a few existing works in the literature on the use of neural networks in outlier detection. Their study showed that neural networks can be a competitive technique versus other existing methods. The authors' basic idea was to produce different types of randomly connected autoencoders with varying densities to obtain significantly better performance. Zhou and Paffenroth [93] demonstrated novel extensions to deep autoencoders which not only maintained a deep autoencoders' ability to discover high quality, non-linear features but was also able to eliminate outliers and noise without access to any clean training data. Their model was inspired by Robust Principal Component Analysis. Zhang et al. [90] proposed an outlier detection method based on autoencoders that perform rumor detection on social networks to include several self-adapting thresholds which facilitated rumor detection. They further discussed how the different number of hidden layers of autoencoder influenced the detection performance. The experimental results showed that the proposed technique improved significantly over the earlier neural network methods for anomaly detection. Schreyer et al. [74] proposed a novel method for detecting outliers in Large-Scale Accounting Data using Deep Autoencoder Networks. They demonstrated that the trained networks reconstruction error regularized by the individual attribute probabilities of a journal entry could be interpreted as a highly adaptive outlier assessment. The empirical study, based on two datasets of real-world journal entries, demonstrated the effectiveness of the approach and outperformed several baseline outlier detection methods. Lv et al. [60] presented a weighted time series fault diagnosis method to learn the deep correlations of faults and reduce the loss of fault information. The model included two key properties: it was able to learn high-level abstract features of faults and the underlying fault patterns, and a mathematical framework of stacked sparse autoencoder-based fault diagnosis. The monitoring performance was compared with multivariate statistical methods and conventional artificial intelligence methods on the Tennessee

Eastman process dataset, which is a well-known chemical industrial benchmark. The experimental results showed performance gains over existing methods, especially for incipient faults that are difficult to detect with traditional technologies. Tran and Hogg [80] proposed a method for video outlier detection using a winner-take-all convolutional autoencoder that gives competitive results in learning for classification task. The method builds on state-of-the-art approaches to outlier detection using a convolutional autoencoder and a one-class SVM to build a model of normality. The key novelties were (1) using the motion-feature encoding extracted from a convolutional autoencoder as input to a one-class SVM rather than exploiting reconstruction error of the convolutional autoencoder, and (2) introducing a spatial winner-take-all step after the final encoding layer during training to introduce a high degree of sparsity. They demonstrated an improvement in performance over the state-of-the-art on UCSD and Avenue datasets. Osada et al. [66] proposed a novel network intrusion detection method for classification with the latent variable, which represented the causes underlying the traffic observed. The proposed model was based on Variational Autoencoder and its strength was a scalability to the amount of training data. They demonstrated that the proposed method was able to dramatically improve the detection accuracy of attack by increasing the amount of unlabeled data. In terms of the false negative rate, it also outperformed the previous work based on semi-supervised learning method, called Laplacian regularized least squares, which has cubic complexity in the number of training data records and consequently is too inefficient to leverage large amounts of unlabeled data. Castellini et al. [38] presented a model based on artificial neural networks to detect fake Twitter profiles. In their method, a denoising autoencoder was implemented as outlier detector, trained with a semi-supervised learning approach. Chong and Tay [42] presented an efficient method for detecting outliers in videos. They proposed a spatiotemporal architecture for outlier detection in videos including crowded scenes. Their architecture included two

main components, one for spatial feature representation, and one for learning the temporal evolution of the spatial features. Experimental results on Avenue, Subway, and UCSD benchmarks confirmed that the detection accuracy of the method was comparable to state-of-the-art methods at a considerable speed of up to 140 fps. Protopapadakis et al. [70] explored the applicability of deep learning techniques for detecting deviations from the norm in behavioral patterns of vessels (outliers) as they are tracked from a High-Frequency Surface-Wave (HFSW) radar or over-the-horizon (OTH) radars, as they are commonly known. The proposed methodology exploited the nonlinear mapping capabilities of deep stacked autoencoders in combination with density-based clustering. A comparative experimental evaluation of the approach showed promising results in terms of the performance of the proposed methodologies.

2.5 K-NEAREST NEIGHBORS

Several researchers have used KNN to detect fraud in Medicare, including our group at Florida Atlantic University. We were the first and only to use KNN with the U.S. Medicare data provided by the Center for Medicare and Medicaid Services [2]. As mentioned on previous sections, Bauder and Khoshgoftaar [28] compared KNN, Autoencoders, and LOF among other machine learning methods, to detect fraud on Medicare data performing a comparative study with supervised, unsupervised, and hybrid machine learning approaches using four performance metrics and class imbalance reduction via oversampling and an 80-20 undersampling method. Supervised methods performed better than unsupervised or hybrid methods. Burr et al. [36] applied traditional and customized multivariate outlier detection methods to detect fraud in Medicare claims. The data used in this study was provided by Floridas' National Claims History (NCH), covering 1995 and the first quarter of 1996. They used two sets of 11 derived features and one set of the 22 combined features. They focused on three issues: (1) outlier masking (e.g. the presence of one outlier can make

it difficult to detect a second outlier), (2) the impact of having an apriori direction to search for fraud, and (3) how to compare the detection methods. Traditional methods include Mahalanobis distances (with and without dimension reduction), KNN, and density estimation methods. Customized methods include ranking methods such as Spearman rank ordering. No two methods agree completely as to which providers are most suspicious, so they focused on ways to compare the different methods. One comparison method used a list of known-fraudulent providers. All comparison methods restricted attention to the most suspicious providers. KNN was also applied to different areas such as cybersecurity and credit card fraud. Weiss et al. [82] described statistical methods for managing health care costs using peer-group models and outlier detection where they applied KNN to predict each target variable. Overall, the prediction of actual outcomes from peer profiles is significantly better than chance, with a reduction of average error by 45.5%. For the 10% of physicians that prescribed the most medications, there were extreme and highly significant differences found between their expected and predicted outcomes. Wu et al. [84] presented a vision-based system to detect intentional attacks on additive manufacturing processes, utilizing machine learning techniques. Particularly, additive manufacturing systems have unique vulnerabilities to malicious attacks, which can result in defective infills but without affecting the exterior. In order to detect such infill defects, the research uses simulated 3D printing process images, as well as actual 3D printing process images to compare accuracies of machine learning algorithms in classifying, clustering, and detecting outliers on different types of infills. Random forest, KNN, and outlier detection were been adopted in the research and shown to be effective in detecting such defects. Akomolafe and Adegboyega [19] presented a methodology for the outlier-based intrusion detection, which uses k-dimensional trees for the description of the system activity and the KNN algorithm during the intrusion detection phase, as well as the evaluations of the results. Malini and Pushpa [62] used KNN to detect

credit card fraud. Compared with power methods and other known outlier detection methods, experimental results indicated that the KNN method was more accurate and efficient. Liu et al [57] proposed an efficient and effective outlier detection algorithm, which was also robust to the parameter k of KNN. Extensive evaluation experiments conducted on twelve public real-world datasets with five popular outlier detection algorithms showed that the performance of the proposed method is competitive and promising. Song et al.[76] proposed a hybrid semi-supervised outlier detection model for high-dimensional data that consisted of two parts: a deep autoencoder (DAE) and an ensemble -nearest neighbor graphs- (-NNG-) based outlier detector. Benefiting from the ability of nonlinear mapping, the DAE was first trained to learn the intrinsic features of a high-dimensional dataset to represent the high-dimensional data in a more compact subspace. Several nonparametric KNN-based outlier detectors are then built from different subsets that were randomly sampled from the whole dataset. The final prediction was made by all the outlier detectors. The performance of the proposed method is evaluated on several real-life datasets, and the results confirm that the proposed hybrid model improves the detection accuracy and reduces the computational complexity.

Outlier detection is important in many domains, but there are a number of challenges involved in finding these anomalies, therefore, this subject has been widely studied. As shown in this chapter, several related studies involve using the machine learning algorithms that are applied in our research: IF, LOF, URF, AE, and KNN. The study of outlier detection on Medicare data is still emerging, with just a few studies comparing outlier detection algorithms on Medicare data. No previous research used IF or URF on Medicare data to detect fraud, also our data mining and machine learning group at Florida Atlantic University is the only ones to use the LOF, AE, and KNN models to detect fraud and abuse on Medicare using the CMS database.

CHAPTER 3
METHODOLOGY

This section describes the datasets used in this study (Medicare and LEIE database), machine learning, the related machine learning algorithms, as well as the performance metrics applied in our analysis.

3.1 DATA

Two datasets are used in this study. The first dataset is the Medicare Provider Utilization and Payment Data: Physician and Other Supplier (2012-2015), also known as Medicare Part B, is provided by the Center for Medicare and Medicaid Services (CMS) [2]. CMS usually releases new data every year, where each new dataset consists of a year of claims that are made available to the public two years after the end of that year. This study combines four years of data, from 2012 to 2015. The 2015 data is the latest provided by CMS at the time of this study and it was released in 2017. The Medicare dataset is not labeled and does not specify which claim is a fraudulent claim. This dataset includes medical claims related to services provided to Medicare beneficiaries. Each claim includes a National Provider Identifier (NPI) [11], which is the provider's or physician's unique identifier code. The dataset has several features such as a provider's specialty or number of procedures/services a provider performed. The Medicare dataset contains values that are registered after claims payments were made, and we assume that this dataset was correctly generated and cleansed by CMS [4]. We filtered the Medicare data for non-prescription data only. The non-prescription data are those codes that are not for specific services listed on

the Medicare Part B Drug Average Sales Price file [13]. Therefore, these are actual provider's services and not drug-specific activities. Furthermore, only providers who are indicated as participants in Medicare program were included.

The second dataset used in this study is the Office of Inspector Generals (OIG) List of Excluded Individuals/Entities (LEIE) database [12]. This dataset is a list of physicians and other health care entities that are banned from involvement in Medicare for a certain time frame. The LEIE dataset indicates each provider's NPI, which is used to label fraudulent claim. For this study, the datasets available from CMS, 2012 to 2015, were combined, with the exclusions from the LEIE database matched taking into consideration the start and end periods of the exclusions to prevent overlap and possible double counting of fraudulent claims. Excluded providers from the LEIE database [12] were added to the dataset in order to obtain labels indicating fraudulent providers. The LEIE database includes only NPI-level, or provider-level, exclusions, not fraud associated with specific medical procedures performed. The exclusions are categorized by various rule numbers, which indicate severity as well as the length of time of each exclusion. The providers selected were providers excluded for more severe reasons and considered mandatory exclusions by the OIG [6], as listed in Table 3.1. For building and testing our models, we assume that providers on the LEIE are considered fraudulent and those not included are not fraudulent.

The Medicare Part B data contains claims information regarding each provider and performed procedure, as well as other attributes such as the place of service, submitted amounts, and payment amounts. As mentioned, the LEIE data provides exclusion information for a provider but not for any specific procedures performed by that provider. At the time of this publication, there are no known publicly available datasets with fraud labels by provider and by each procedure performed. For this reason, the Medicare data was grouped and aggregated to the NPI-level. The original dataset was reduced to 50%. The reason for that is because the LOF, KNN, and URF,

in particular, could not run the original dataset due to memory limitations.

Table 3.1: LEIE Exclusion Rules

Rule Number	Description
1128(a)(1)	Conviction of program-related crimes.
1128(a)(2)	Conviction relating to patient abuse or neglect.
1128(a)(3)	Felony conviction relating to health care fraud.
1128(b)(4)	License revocation or suspension.
1128(c)(3)(g)(i)	Conviction of 2 mandatory exclusion offenses.
1128(c)(3)(g)(ii)	Conviction on 3 or more mandatory exclusion offenses.

We grouped the data by specialty (provider type), NPI, and gender and aggregated across all procedures and places of services for each year. In order to avoid extra information loss due to aggregation, we generated additional numeric features from the original five, including the mean, sum, median, standard deviation, minimum, and maximum. Table 3.2 shows the original Medicare Part B features from which the aggregated dataset is generated, as well as the categorical and class features. The mapping of the LEIE exclusions, to use as fraud labels, follows the Part B data aggregation. We join the Part B data with the LEIE data by matching NPI only and create the exclusion feature which is initialized to FALSE for non-fraud. We mark any instance where the Medicare year is less than the end of the exclusion period for all matched providers, changing the exclusion value to TRUE for fraud. We also incorporate any waivers or reinstatements which indicate the provider is no longer considered excluded. Note that providers are labeled as fraudulent only for the available 2012 to 2015 Medicare Part B years. For example, if a provider has a five-year exclusion from 2008 to 2013, this overlaps with the available Part B years of 2012 and 2013 which are both labeled as fraud for that particular provider. Furthermore, by using years prior to the end of the exclusion period, we are able to detect fraud

leading up to the exclusion and during the exclusion period. The latter can indicate improper payments made by the excluded provider which could be possible fraud per the federal False Claims Act (FCA) [7].

To be able to build our model with a mixture of numerical and categorical features, we used one-hot encoding, which is a method that uses the categorical values to generate dummy features with binary values which indicate the presence of this variable, assigning a value of one if present, otherwise zero, versus all other dummy features. This translates each of the original categorical values into distinct binary features. Table 3.3 compares the original Medicare data and the NPI-level aggregated data with fraud labels. The final dataset has 3,693,980 rows and 34 columns with 1417 fraud labels, which is 0.038% of all instances. The NPI feature is not used to build or test the models, this is only used to identify physicians.

3.2 MACHINE LEARNING

Machine learning is a sub-field of computer science which is an implementation of artificial intelligence (AI) that provides systems the ability to learn insights from data [83]. Machine Learning provides the scientific basis of data mining, which is the way of solving problems by analyzing data [83]. Machine learning algorithms can make data-driven predictions or decisions by constructing a model from sample inputs. Machine learning is used to obtain meaningful information from raw data that can be used for diverse applications, such as email filtering, detection of network intruders, optical character recognition (OCR), and computer vision [83]. Machine learning techniques can be used to detect data points that are significantly different from the remaining data, known as outliers [34].

Table 3.2: Description of Medicare Features

Feature	Description
npi	Unique provider identification number
provider_type	Medical providers specialty (or practice)
nppes_provider_gender	Gender
line_srvc_cnt	Number of procedures the provider performed
bene_unique_cnt	Number of distinct Medicare beneficiaries receiving the service
bene_day_srvc_cnt	Number of distinct Medicare beneficiary/per day services performed
average_submitted_chrg_amt	Average of the charges that the provider submitted for the service
average_medicare_payment_amt	Average payment made to a provider per claim for the service performed
exclusion	Fraud labels from the LEIE database

Table 3.3: Medicare Datasets

Dataset	Instances	Features
Original	37,147,213	30
NPI-level	3,693,980	34
NPI-level (one-hot-encoded)	3,693,980	125

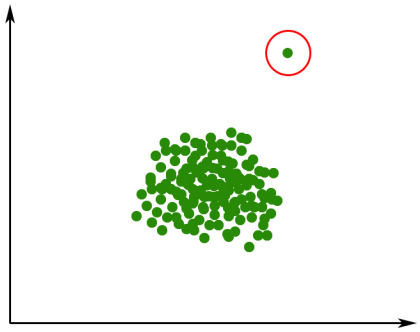

Figure 3.1: Outlier Detection

3.2.1 Outlier Detection

Outliers are also called anomalies, abnormalities, discordants, exceptions, aberrations, contaminants, or deviants. Outlier detection relates to the task of looking for patterns in data that do not conform to expected behavior as shown in Figure 3.1. Outlier detection has been studied within many research areas and application domains such as credit card fraud [20] and intrusion detection on cyber-security [55]. Outlier detection is a significant task as anomalies in data can indicate unusual or harmful activities for different application domains. For instance, an abnormal traffic pattern in a computer network could imply that a hacked computer is providing sensitive data to an unauthorized destination [55]. An anomalous MRI image may show the presence of malignant tumors [77]. Outliers in credit card transaction data could be a sign of credit card fraud or identity theft [20]. An anomalous reading from a spacecraft sensor could imply a fault in some component [48].

3.2.2 Supervised and Unsupervised Learning

Two common techniques used in machine learning are supervised and unsupervised learning, where supervised learning requires the training data to be labeled and unsupervised learning use unlabeled data [83]. As an example of supervised learning, Bauder and Khoshgoftaar [26] used physician specialties as labels and created a model

to predict whether a physician is behaving within the norm of his or her medical specialty. Unsupervised learning can identify hidden patterns and group data to provide useful information [83]. There are a variety of applications where unsupervised learning can be used such as when a user needs to explore the data by splitting it into groups or clusters. In this study, five unsupervised learning models will be used in an effort to detect outliers in Medicare data. These outlier detection methods include the following: Isolation Forest [56], Local Outlier Factor [35], Unsupervised Random Forest [16], Autoencoders [64], and k-Nearest Neighbors [5].

3.3 MACHINE LEARNING ALGORITHMS

The five machine learning models selected for this experiment are: IF [56], LOF [35], URF [16], AE [64], and KNN [5]. These models are briefly described in this section.

3.3.1 Isolation Forest

Most existing model-based approaches to detect outliers construct a profile of normal instances, then identify instances that do not conform to the normal profile as outliers. Liu et al. [56] proposed a fundamentally different model-based method that explicitly isolates outliers instead of profiles normal points. Isolation Forest uses a random forest of decision trees to detect data anomalies. The Isolation Forest isolates observations by randomly selecting a feature, and then randomly selecting a split value between the maximum and the minimum values of the selected feature. Since recursive partitioning can be represented by a tree structure, the number of splittings required to isolate a sample is equivalent to the path length from the root node to the terminating node. This path length averaged over a forest of many random trees is a measure of abnormality. Random partitioning produces noticeable shorter paths for outliers. When a forest of random trees collectively produce shorter path lengths for particular samples, they are highly likely to be outliers [56]. Isolation Forest is an

algorithm with small memory requirements and linear time complexity. It builds a good performing model with a small number of trees using small sub-samples of fixed size, regardless of how large a dataset is. It has constant time and space complexities during training. Liu et al [56] performed an empirical comparison with four state-of-the-art outlier detection algorithms and concluded that IF is superior in terms of runtime, detection accuracy and memory requirements, especially in large datasets, and its ability to deal with high dimensional data with irrelevant attributes. The authors also show that IF is capable of being trained with or without outliers in the training data. There are two input parameters on the IF algorithm, the sub-sampling size and the number of trees t. For this study, we run IF using t=100 which is the recommend number of trees [56].

3.3.2 Local Outlier Factor

The LOF algorithm is a well-known unsupervised outlier detection model which calculates the local density deviation of a given data point with respect to its neighbors [35]. It considers outliers the samples that have a significantly lower density than their neighbors. It is local since the outlier score is determined by how isolated the object is with respect to the surrounding neighborhood. Locality is given by k-Nearest Neighbors, whose distance is used to estimate the local density. By comparing the local density of a sample to the local densities of its neighbors, it is possible to identify samples that have a considerably lower density than their neighbors. These are considered outliers [35]. LOF has been widely used in several different fields such as predicting financial crises [40], wireless sensors networks [73], and network intrusion [68]. For this experiment, we ran LOF model five times, with the parameter k having the values of 10, 20, 40, 80, and 100.

3.3.3 Unsupervised Random Forest

Random Forest (RF) was introduced in 2001 as a regression and classification ensemble learning method that constructs a multiple of decision trees at the training stage and returns a class which is the mean prediction of the individual trees [33]. It has been widely accepted within the data analysis community as a supervised learning data analysis tool [16]. Many are familiar with the use of Random Forest for prediction and feature selection but few are aware of its potential as an unsupervised technique.

URF holds the assumption that if the data holds any structure it should be distinguishable from a randomly generated version of itself [16]. The approach is to consider the original data as class 1 and to create a synthetic second class of the same size and labeled it class 2. For this study, we used the tutorial and case studies available in Afanador et al. [16] where a synthetic dataset is generated randomly based on the original dataset. Then, the two datasets together form a two-class classification problem that can be modeled using classical supervised RF. A proximity matrix is used to analyze the number of instances in which two cases, like samples or observations, are distributed into the same child node of an RF tree. These node distributions are averaged across t trees with the output being an n x n matrix of proximity scores where n is the number of samples. These proximity scores are then used to conduct an unsupervised analysis for detecting significant structure in a dataset [16]. The parameter for URF is the number of trees t. For this experiment, we run URF with t=100.

3.3.4 Autoencoders

An autoencoder neural network is an unsupervised learning algorithm that applies backpropagation to replicate its input in the output. With the sparsity constraint enforced, an autoencoder automatically learns useful features of the unlabeled train-

ing data [64]. Autoencoders are typically used to compress the data into a lower-dimensional representation or feature learning but recently they have also been used for learning generative models of data [50]. A sparse autoencoder is an autoencoder with a sparsity penalty on the code layer. It is usually used to learn features for an additional task such as classification [50]. An autoencoder is composed of an encoder and decoder in order to learn the patterns from the data to create representative features of that data. These features are then used to reconstruct the original data patterns, with their reconstruction error indicating the divergence in the models prediction versus the original input. In this study, we incorporate "bottleneck" training creating a middle-hidden layer that is very small. We use a hidden layer of just 2 nodes for which the autoencoder will have to reduce the dimensionality of the input data. The autoencoder model will then learn the input data patterns. We run the unsupervised autoencoder with two activation functions: Rectifier Linear Unit (RELU) and Hyperbolic Tangent (Tanh). A dropout at a rate of 0.5 was used to create a more generalizable model that is less likely to overfit the training data. Using dropout essentially forces an artificial neural network to learn multiple independent representations of the same data by alternately randomly disabling neurons in the learning phase [78]. We run a total of six autoencoder models, three models, with RELU with dropout, with 50, 100 and 200 nodes, and three models with Tanh with dropout, also with 50, 100 and 200 nodes.

3.3.5 k-Nearest Neighbors

The k-Nearest Neighbors algorithm is a relatively simple and robust method that looks at the k-Nearest Neighbors around some particular value to determine which neighbors are most similar, based on their distance to points in a training dataset [43]. Various metrics can be used to determine the distance such as Euclidean distance, Standardized Euclidean distance, Mahalanobis distance, City block distance,

Minkowski distance, Chebychev distance, Cosine distance, Correlation distance, Hamming distance, Jaccard distance, and Spearman distance [3]. For this study, the distance metric used is the Euclidean distance. Points having large KNN distances are seen as outliers. To determine which of the k instances are most similar to a new input KNN is used in a variety of applications such as economic forecasting, data compression and genetics [5]. For this study, we run the KNN model using k=1 and k=5. In this case for each point, the model first computes the distance to all other points, then find the minimum distance, then store those distances as outlier scores.

3.4 PERFORMANCE METRICS

The models used in this study are evaluated using the Area Under the ROC (Receiver Operating Characteristics) curve (AUC).

3.4.1 ROC Curve

A ROC curve, is commonly used to visualize the performance of a binary classifier. The ROC curve is generated by plotting the true positive rate (TPR), also called sensitivity, against the false positive rate (FPR), which is 1- specificity, at different threshold settings. The higher the area under the curve is, the better is the classifier's ability to distinguish between positive and negative class [31].

3.4.2 Area Under the Curve, AUC Curve

The area under the ROC curve (AUC) is a concise measure of ROC curve performance. AUC can sort models by overall performance overcoming the difficulties encountered when comparing the differences between ROC curves, especially in cases were the curves intersect, therefore, the AUC is favored in models assessment [24]. One of its properties is that the AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly

chosen negative instance [46]. A higher AUC value means better model performance. To estimate the optimal threshold on the ROC curve, Youden's Index is employed [86]. The optimal cut-off is the threshold that maximizes the distance to the identity (diagonal) line, represented by max(sensitivities + specificities). These thresholds are not included as there is no direct basis for comparison. The best threshold for each method will be wildly different, so including this value would be confusing and not directly comparable.

CHAPTER 4
CASE STUDIES

In this chapter, we discuss the results of our case studies to test the performance of different machine learning algorithms, with varied configurations, to detect Medicare fraud. The methodology and datasets discussed in Chapter 3 are applied to all experiments described herein. IF, LOF, URF, AE, and KNN models are first presented, followed by a comparison of all models in an effort to determine which is the best performer.

We compare the best sensitivity and specificity at the best threshold, where sensitivity indicates the number of positive class correctly identified as positive, also known as the true positive rate (TPR), and specificity, known as false positive rate (FPR), indicates the correct number of non-fraudulent physicians out of those who are truly not fraudulent.

4.1 ISOLATION FOREST

In this case study, we seek to analyze the impact of using IF to detect fraud on Medicare data. For the IF model, we used 100 trees. Table 4.1 shows the results of this experiment and Figure 4.1 displays the outlier detection ROC curve for the IF model.

Contrary to what was shown on Liu et al. [56], IF did not outperform LOF in our study. IF isolates an observation by splitting at each node, where shorter path lengths indicate anomalies [56]. When the forest produces shorter paths for certain samples, these are most likely to be anomalous since the data points are usually located in

sparse regions. The distance from the leaf to the root is used as the outlier score [56]. In our case it seems that there are too many splits occurring to isolate most fraud observations, suggesting that the data points are not in such sparse regions.

It was observed in this experiment that out of the fraudulent physicians, 71% were correctly classified as fraudulent by the IF model, but only 44% of the non-fraudulent physicians were correctly identified as non-fraudulent. Therefore, too many points are seen as possible outliers by this model. The resulting AUC value of 0.55430 is low across all decision thresholds. Table 4.1 summarizes IF performance showing learner sensitivity and specificity at the best decision threshold.

Table 4.1: Performance of IF Model

Model	Sensitivity	Specificity	AUC
IF100	0.71206	0.43613	0.55430

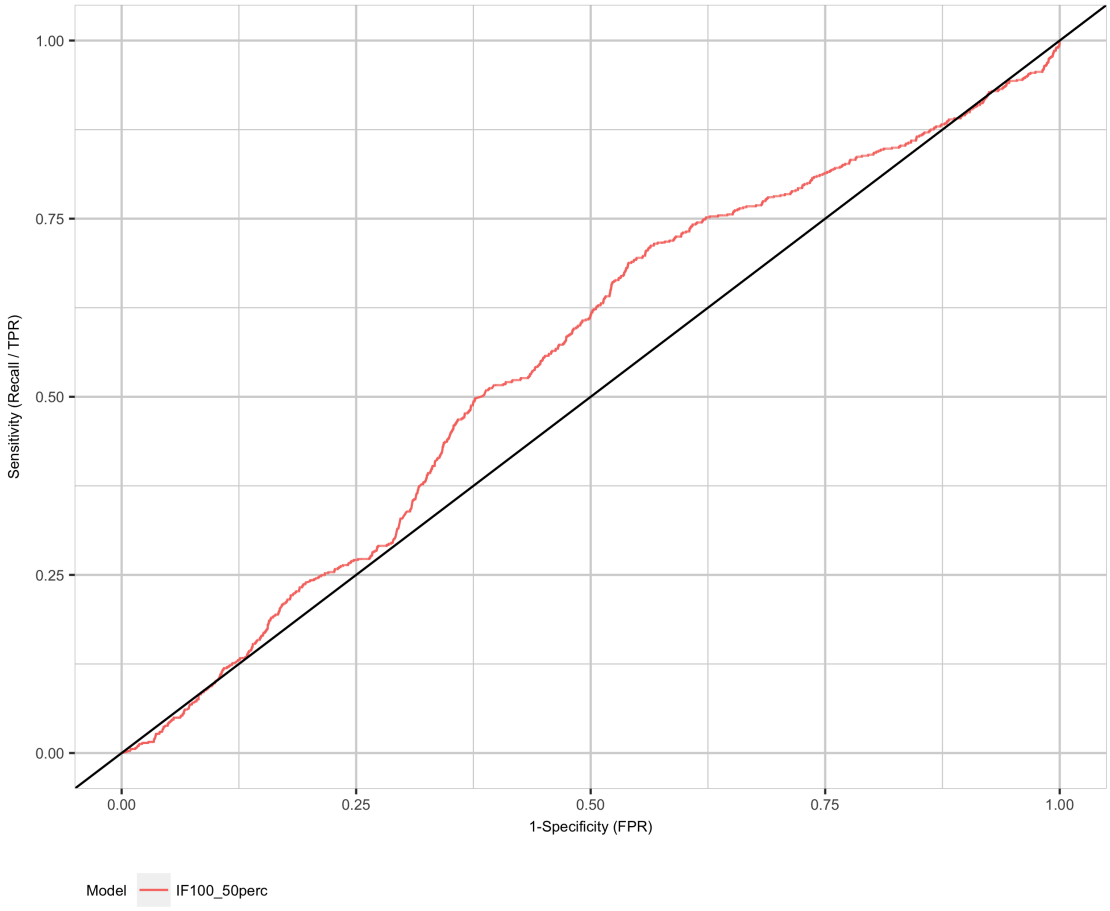

Figure 4.1: AUC curve - IF

4.2 LOCAL OUTLIER FACTOR

The performance of LOF was tested in this study to detect fraud on Medicare data. We ran the LOF model five times, with parameter k having the values of 10, 20, 40, 80, and 100. Table 4.2 shows the results of these experiments and Figure 4.2 presents the outlier detection ROC curve for the LOF model.

LOF was the best performer among all models tested. LOF is a local density-based detection method which relies on the local density of its neighbors to determine normal or outlying points [35]. Locality is given by k-Nearest Neighbors, whose distance is used to estimate the local density. By comparing the local density of a sample to

the local densities of its neighbors, it is possible to identify samples that have a considerably lower density than their neighbors which are considered outliers [35]. The local density measurement property of the LOF is possibly the reason why this model was the most successful. One known issue of LOF is that it can sometimes be ineffective when regions of different density are not clearly separated [17] which may have influenced the results of this model.

Our study shows that LOF had better performance than IF in every configuration tested, in terms of AUC value. The best AUC of 0.62985 was obtained using k=40. Table 4.2 shows the best sensitivity, specificity, and AUC results for each of the LOF models. The LOF configuration with k=20 was able to correctly classify as fraudulent 66% of the fraudulent physicians. The model with k=10 classified 77% of the non-fraudulent physicians correctly.

Table 4.2: Performance of LOF Models

Model	Sensitivity	Specificity	AUC
LOF10	0.41560	0.77495	0.62012
LOF20	0.65957	0.54430	0.62840
LOF40	**0.53617**	**0.67676**	**0.62985**
LOF80	0.59433	0.61066	0.62872
LOF100	0.53617	0.65729	0.62543

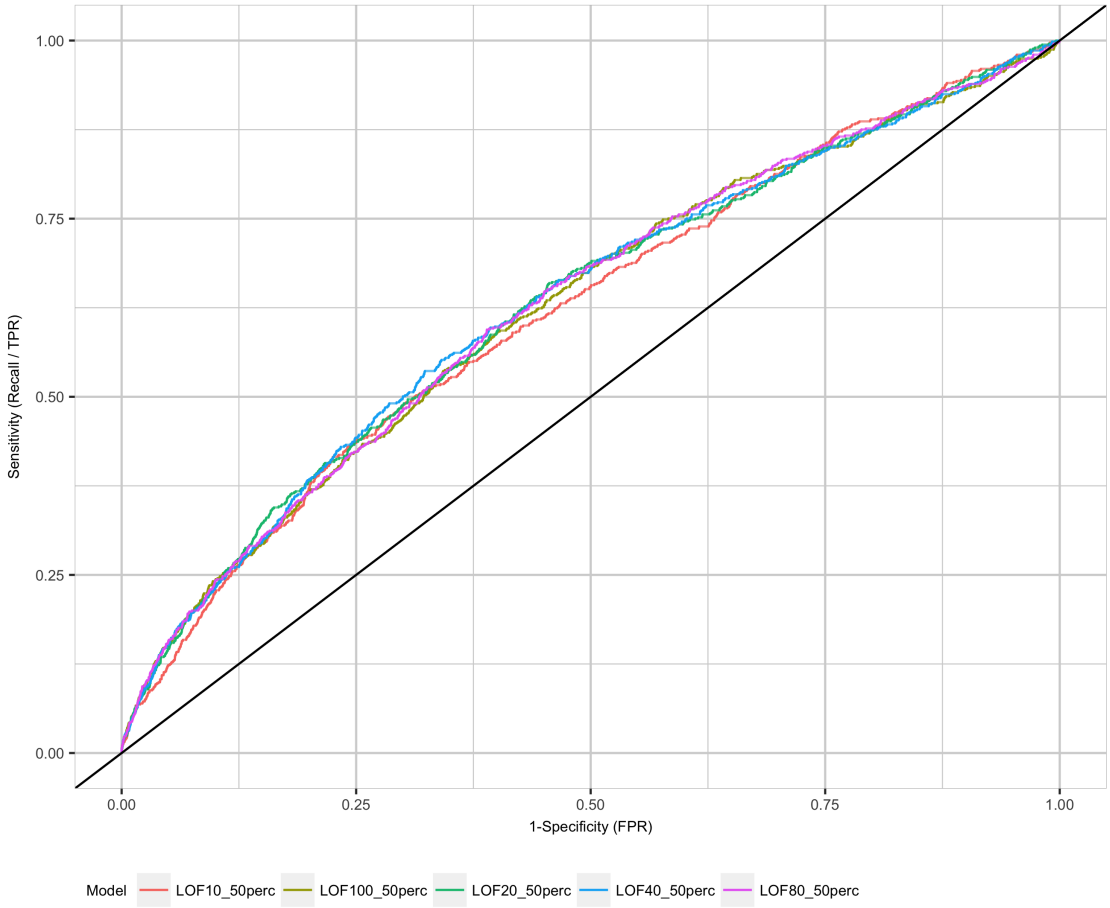

Figure 4.2: AUC curve - LOF

4.3 UNSUPERVISED RANDOM FOREST

The URF model was also tested in this experiment with the goal to observe its impact on the detection of fraud in Medicare data. For this experiment, we run URF with 100 trees. Table 4.3 shows the result for this experiment and Figure 4.3 displays the outlier detection ROC curve for the URF model.

URF holds the assumption that if the data holds any structure it should be distinguishable from a randomly generated version of itself so it produces a random version of the original dataset by splitting the data into two classes creating a proximity matrix used to assess outliers. It generates a second class of instances by sampling from

each of the original, univariate features, retaining a similar distribution to the original data. This two-class classification problem can then be modeled using classical supervised Random Forest [16].

The use of Random Forest as an unsupervised model is very limited. Our experiment shows URF as a good performer, ranking third in our experiment, after LOF and KNN1. URF was able to correctly classify as fraudulent 53% of the fraudulent physicians and it classified 65% of the non-fraudulent physicians correctly. The AUC value and the best sensitivity and specificity are provided in Table 4.3.

Table 4.3: Performance of URF Model

Model	Sensitivity	Specificity	AUC
URF100	0.52766	0.64577	0.60389

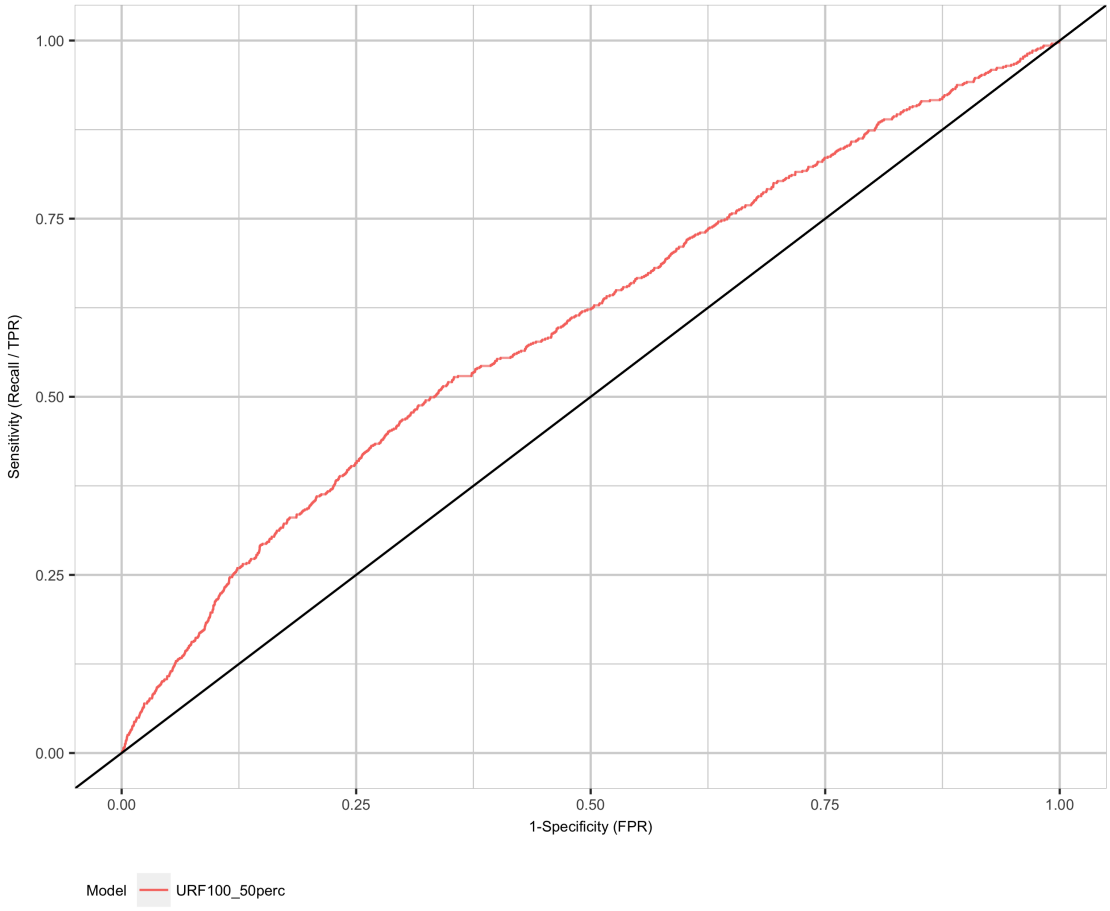

Figure 4.3: AUC curve - URF

4.4 AUTOENCODER

The performance of AE was evaluated in this study regarding its ability to detect fraud on Medicare data. We applied six configurations of the AE model, three using Rectifier with dropout, containing 50, 100, and 200 nodes, and three with Hyperbolic Tangent (Tanh) with dropout, also containing 50, 100, and 200 nodes. Table 4.4 shows the results of the experiments and Figure 4.4 presents the outlier detection ROC curve for the AE models.

In this study, we incorporated bottleneck training creating a very small middle-hidden layer composed of just 2 nodes for which the Autoencoder had to reduce the

dimensionality of the input data. The autoencoder model then learned the input data patterns [64]. Part of the reason this model had a poor performance result, is that the AE reduces the initial feature set to two abstract features which are supposed to represent the fraud/non-fraud cases but the abstraction of the features does not represent well the original features.

Our study shows that the AE's ability to replicate its input in the output does not work well on detecting outliers. In general, AE misses actual outliers, it was able to correctly classify 46% of fraudulent physicians. It worked slightly better at classifying 65% of the non-fraudulent physicians correctly. Results are shown in Table 4.4. The best models were AE with Tanh having the best AUC value at the sensitivity and specificity provided in Table 4.4 with AUC value of 0.55507.

Table 4.4: Performance of AE Models

Model	Sensitivity	Specificity	AUC
AE50_Rect	0.43972	0.66647	0.54184
AE100_Rect	0.46241	0.65348	0.55250
AE200_Rect	0.46383	0.63042	0.53859
AE50_Tanh	**0.46383**	**0.65030**	**0.55507**
AE100_Tanh	**0.46383**	**0.65030**	**0.55507**
AE200_Tanh	**0.46383**	**0.65030**	**0.55507**

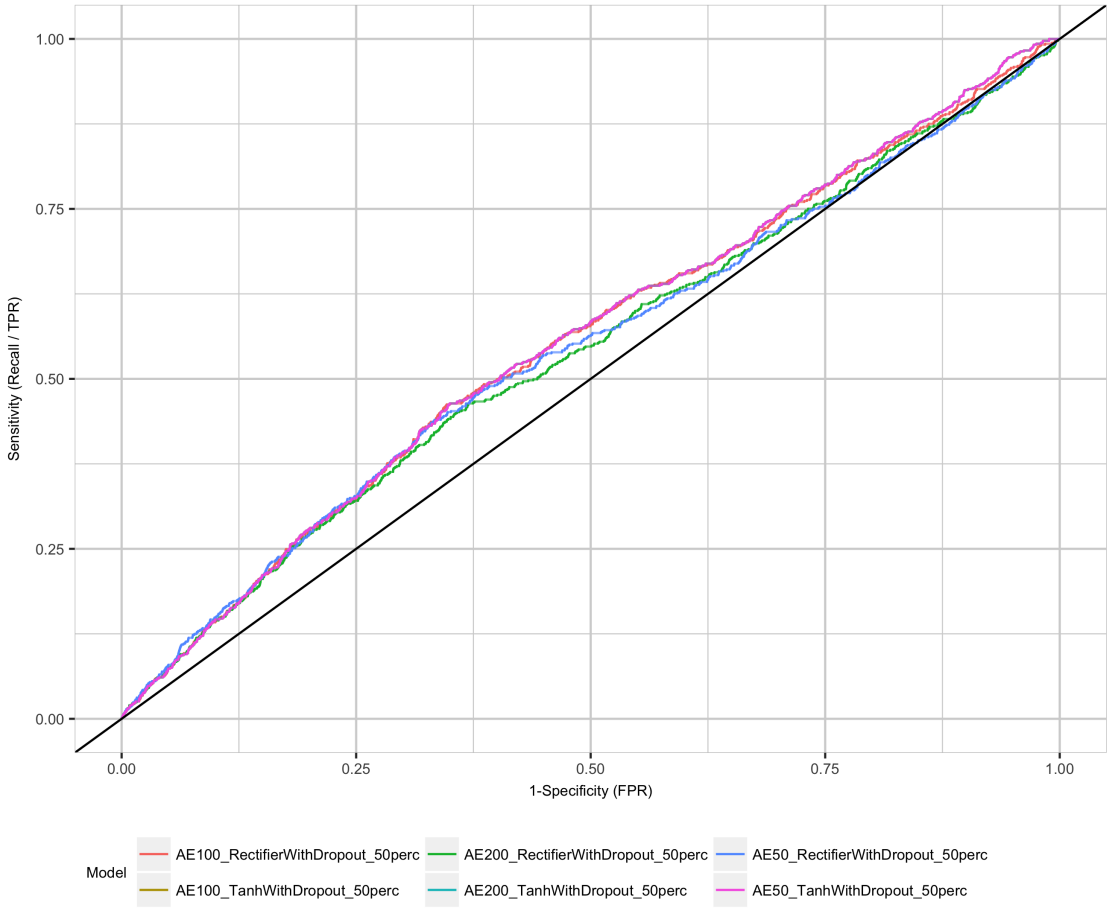

Figure 4.4: AUC curve - AE Models

4.5 K-NEAREST NEIGHBORS

KNN was the last model tested to detect fraud on Medicare data. For this study, we run the KNN model using k=1 and k=5. Table 4.5 shows the results of these experiments and Figure 4.5 displays the outlier detection ROC curve for the KNN model.

The KNN model, similar to LOF, uses neighbors to evaluate outliers. The KNN analyzes the k-Nearest Neighbors around some particular value to decide which neighbors are most similar based on their distance to points in a training set [43].

Our study shows that KNN with k=1 (KNN1) is the second best performing model

after LOF, both had similar AUC value. The KNN1 model was able to correctly classify 50% of the fraudulent physicians and classified 68% of the non-fraudulent physicians correctly. The AUC, sensitivity, and specificity are shown in Table 4.5.

Contrary to KNN1, KNN with k=5 (KNN5) is the worst model tested in terms of AUC value. The KNN5 model was able to correctly classify as fraudulent 60% of the fraudulent physicians, which was a better result than KNN1, although it classified only 44% of the non-fraudulent physicians correctly. Introducing more neighbors may have contributed to the poor performance of the model. The AUC value with sensitivity and specificity provided in Table 4.5 was 0.51862.

Table 4.5: Performance of KNN Model

Model	Sensitivity	Specificity	AUC
KNN1	**0.49787**	**0.67957**	**0.61338**
KNN5	0.59660	0.43812	0.51862

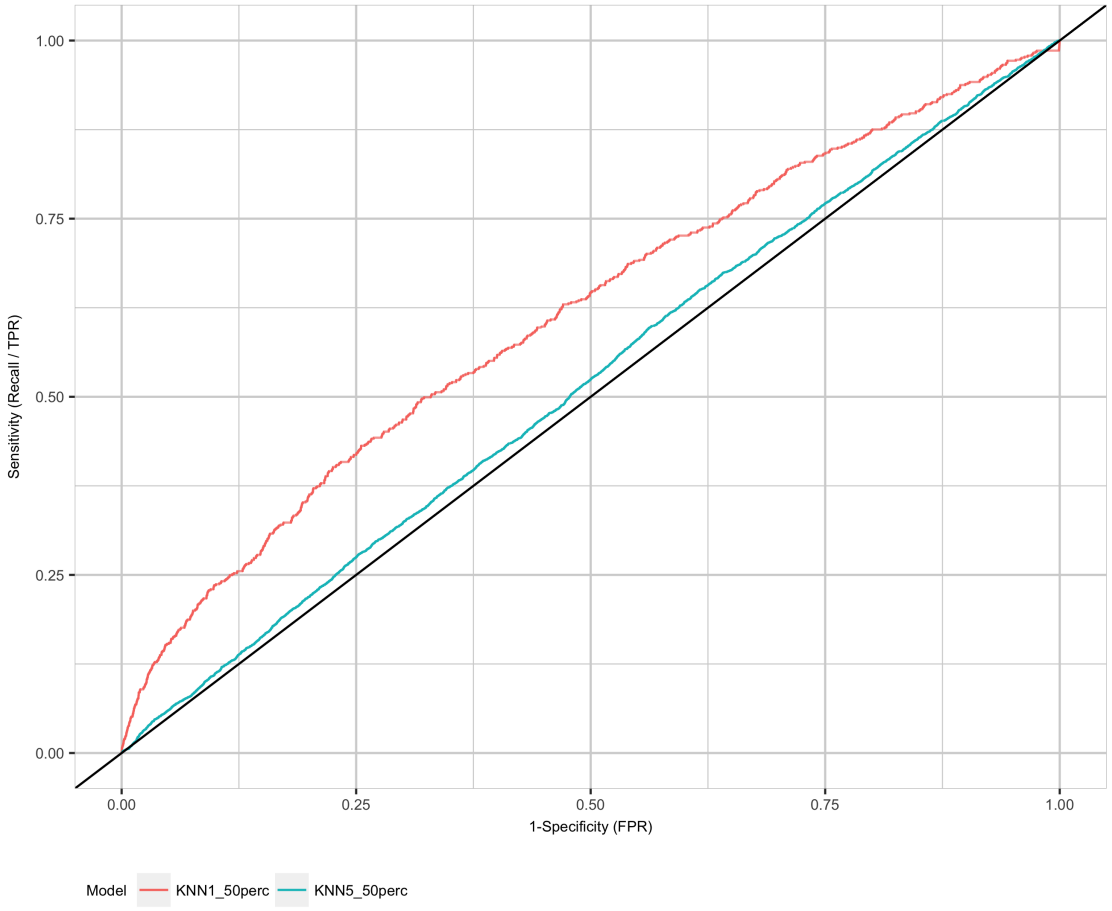

Figure 4.5: AUC curve - KNN

4.6 COMPARING ALL MODELS

Our study shows that all configurations of LOF outperform the remaining outlier detection models. This could be due to its ability to effectively find local outliers. Figure 4.6 shows the ROC curve for the best performers of all models and Table 4.6 presents the sensitivity, specificity and AUC results for these models ordered by the best performer to worst.

Table 4.6: Best Performance of all Models

Model	Sensitivity	Specificity	AUC
LOF40	**0.53617**	**0.67676**	**0.62985**
KNN1	0.49787	0.67957	0.61338
URF100	0.52766	0.64577	0.60389
AE50_Tanh	0.46383	0.65030	0.55507
IF100	0.71206	0.43613	0.55430

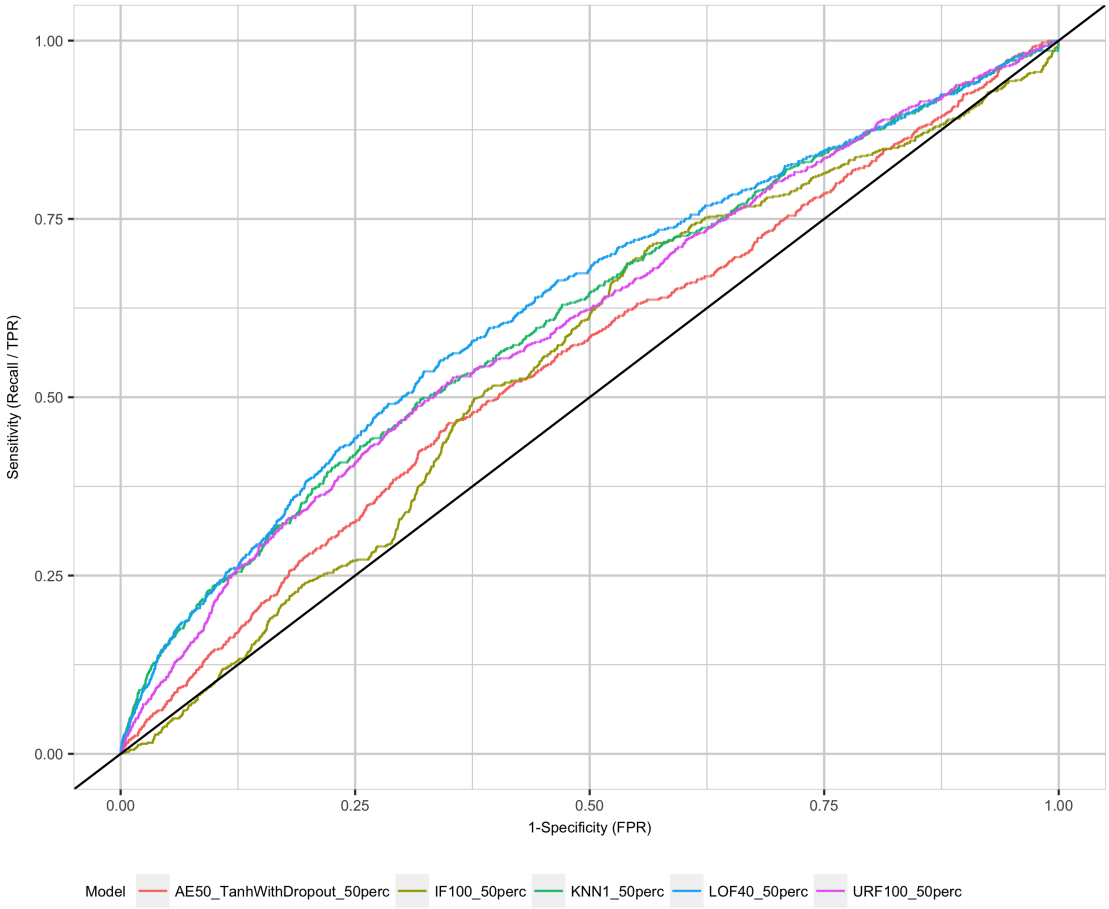

Figure 4.6: AUC curve - All Models

CHAPTER 5
CONCLUSIONS AND FUTURE WORK

Health care is a tempting target for thieves due to the large amount of money involved and the overall complexity of the system. Fraud in the Medicare program is an ongoing problem that researchers and authorities are constantly trying to minimize. The waste of Medicare money with fraud and abuse results in higher Medicare costs for the elderly population that will continue to increase, doubling by 2060. Research using Medicare data with known fraudulent physicians and machine learning approaches is still in its early stages. In our study, we provide a robust empirical analysis and evaluate five unsupervised machine learning algorithms for detecting Medicare fraud and abuse, using the combine 2012 to 2015 Medicare Part B data. We evaluated IF and URF, which were not previously tested on Medicare data, and compared them with LOF, URF, AE, and KNN. We incorporated the LEIE database for fraudulent physician labels and used AUC to assess model fraud detection performance.

5.1 CONCLUSIONS

Our data mining and machine learning group at Florida Atlantic University are the pioneers on using IF, LOF, URF, AE, and KNN with CMS data to detect fraud and abuse in Medicare. In this study, we tested IF and LOF with the CMS data for the first time. We also compared results with URF, AE, and KNN.

The IF model, which is one of the models not previously tested on Medicare data, was our first tested model and was able to find 71% of the fraudulent physicians, but did not perform well on classifying the non-fraudulent physicians. The IF model was

the fifth best classifier in terms of AUC. The second model analyzed was the LOF model, which was the best performer among all models tested. The best configuration was LOF with k=40. Local Outlier Factor works well finding local outliers which could be one reason for the higher performance. The third model tested was URF, which is another model not previously tested on Medicare data. URF was the third best model in our study, it performed better than IF on correctly classifying non-fraudulent physicians and its performance was similar to LOF. The fourth model was AE, which was also the fourth best performer. AE with Tanh performed better than AE with Rectifier, for which the latter ended up placing sixth in our experiments. The last model tested was KNN, with both k=1 and k=5 neighbors. KNN1 was the second-best performer, with results very close to LOF and URF. KNN5 was the worst performer across all models.

Overall, LOF outperformed the other unsupervised machine learning approaches with a 0.62985 AUC. The previously untested methods, IF and URF, both performed poorly in detecting Medicare fraud. LOF, KNN1, and URF had very similar AUC values. The remaining models performed significantly worse. When using unsupervised models to detect fraud on Medicare data, we recommend using LOF as a first choice.

5.2 FUTURE WORK

The empirical analysis presented in this thesis provides a basis for future experimentation in the domain of Medicare fraud detection or related application domains, such as Medicaid. The main limitation in evaluating fraud detection performance was due to the low number of available fraud labels in the LEIE, and the size of the dataset which was reduced to 50%. Potential future work, building on the research presented in this thesis, are presented below:

- In this thesis, we investigated the following unsupervised models: IF, LOF,

URF, AE, and KNN. It is also important to investigate other unsupervised models and compare performances using the same dataset. Different configurations could also be applied to these models.

- Unsupervised machine learning models were tested in this study. It is recommended to apply similar experiments using supervised models in an effort to compare the performance of supervised and unsupervised models.

- Our studies were conducted on datasets from the CMS 2012-2015 Medicare Part B data combined with the LEIE for fraud labels. As more data becomes available from these and other sources, such as Medicare Part D, new experiments should be performed. Additionally, the full dataset can also be used in future work, since we only used 50% of the data in this study.

BIBLIOGRAPHY

[1] 2017 annual report of the boards of trustees of the federal hospital insurance and federal supplementary medical insurance trust funds [online]. available: https://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/reportstrustfunds/downloads/tr2017.pdf. [accessed 06-Nov-2017].

[2] Centers for medicare and medicaid services: Research, statistics, data, and systems [online]. available: https://www.cms.gov/research- statistics- data- and-systems/ research- statistics- data- and- systems.html. [accessed 06-Nov-2017].

[3] Classification using nearest neighbors.[online]. available:https://www.mathworks.com/help/stats/classification-using-nearest-neighbors.html. [accessed 06-Feb-2018].

[4] Cms office of enterprise data and analytics. (2017) medicare fee-for-service provider utilization & payment data physician and other supplier [online]. available: 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. [accessed 06-Nov-2017].

[5] A complete guide to k-nearest-neighbors with applications in python and r. [online]. available: https://kevinzakka.github.io/2016/07/13/k-nearest-neighbor/. [accessed 31-Jan-2018].

[6] Exclusions program. [online]. available: https://oig.hhs.gov/exclusions/. [accessed 06-Feb-2018].

[7] Federal false claims act. [online]. available: https://www.gpo.gov/fdsys/granule/uscode-2011-title31/uscode-2011-title31-subtitleiii-chap37-subchapiii-sec3729. [accessed 06-Feb-2018].

[8] Health care fraud. [online]. available: https://www.fbi.gov/investigate/white-collar-crime/health-care-fraud. [accessed 06-Feb-2018].

[9] Help fight medicare fraud [online]. available: https://www.medicare.gov/forms-help-and-resources/report-fraud-and-abuse/fraud-and-abuse.html. [accessed 06-Nov-2017].

[10] Medicare fraud strike force. [online]. available: https://oig.hhs.gov/fraud/strike-force/. [accessed 06-Feb-2018].

[11] National provider identifier standard (npi) [online]. available: https://www.cms.gov/regulations-and-guidance/administrative-simplification/nationalprovidentstand/. [accessed 30-Jan-2018].

[12] Office of inspector general leie downloadable databases [online]. available: https://oig.hhs.gov/exclusions/index.asp. [accessed 06-Nov-2017].

[13] Profile of older americans: 2015, 2015. [online]. available: https://www.acl.gov/sites/default/files/aging[accessed 28-Mar-2018].

[14] Us medicare program [online]. available: https://www.medicare.gov/. [accessed 06-Nov-2017].

[15] What is medicare? [online]. available: http://time.com/money/collection-post/2791232/what-is-medicare/. [accessed 06-Nov-2017].

[16] Nelson Lee Afanador, Agnieszka Smolinska, Thanh N Tran, and Lionel Blanchet. Unsupervised random forest: a tutorial with case studies. *Journal of Chemometrics*, 30(5):232–241, 2016.

[17] Charu C. Aggarwal. *Outlier analysis*. Springer, 2017.

[18] Leman Akoglu, Mary McGlohon, and Christos Faloutsos. Oddball: Spotting anomalies in weighted graphs. *Advances in Knowledge Discovery and Data Mining*, pages 410–421, 2010.

[19] Oladeji Patrick Akomolafe and Adeleke Ifeoluwa Adegboyega. An improved knn classifier for anomaly intrusion detection system using cluster optimization. *An Improved KNN Classifier for Anomaly Intrusion Detection System Using Cluster Optimization*, 8(2):3438, 2017.

[20] Emin Aleskerov, Bernd Freisleben, and Bharat Rao. Cardwatch: A neural network based database mining system for credit card fraud detection. In *Computational Intelligence for Financial Engineering (CIFEr), 1997., Proceedings of the IEEE/IAFE 1997*, pages 220–226. IEEE, 1997.

[21] Martin Atzmueller, David Arnu, and Andreas Schmidt. Anomaly detection and structural analysis in industrial production environments. In *Data Science–Analytics and Applications*, pages 91–95. Springer, 2017.

[22] R Can Aygun and A Gokhan Yavuz. Network anomaly detection with stochastically improved autoencoder based models. In *Cyber Security and Cloud Computing (CSCloud), 2017 IEEE 4th International Conference on*, pages 193–198. IEEE, 2017.

[23] Dalya Baron and Dovi Poznanski. The weirdest sdss galaxies: results from an outlier detection algorithm. *Monthly Notices of the Royal Astronomical Society*, 465(4):4530–4555, 2016.

[24] Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM Sigkdd Explorations Newsletter*, 6(1):20–29, 2004.

[25] Richard Bauder, Taghi M. Khoshgoftaar, and Naeem Seliya. A survey on the state of healthcare upcoding fraud analysis and detection. *Health Services and Outcomes Research Methodology*, 17(1):31–55, Mar 2017.

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

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

[28] Richard A Bauder and Taghi M Khoshgoftaar. Medicare fraud detection using machine learning methods. In *Machine Learning and Applications (ICMLA), 2017 16th IEEE International Conference on*, pages 858–865. IEEE, 2017.

[29] Richard A Bauder and Taghi M Khoshgoftaar. Multivariate outlier detection in medicare claims payments applying probabilistic programming methods. *Health Services and Outcomes Research Methodology*, 17(3-4):256–289, 2017.

[30] Richard A Bauder and Taghi M Khoshgoftaar. Multivariate outlier detection in medicare claims payments applying probabilistic programming methods. *Health Services and Outcomes Research Methodology*, 17(3-4):256–289, 2017.

[31] Mohamed Bekkar, Hassiba Kheliouane Djemaa, and Taklit Akrouf Alitouche. Evaluation measures for models assessment over imbalanced data sets. *Iournal Of Information Engineering and Applications*, 3(10), 2013.

[32] Vishal Bhatt, Mradul Dhakar, and Brijesh Kumar Chaurasia. Filtered clustering based on local outlier factor in data mining. *International Journal of Database Theory and Application*, 9(5):275–282, 2016.

[33] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[34] Ronald Bremer. Outliers in statistical data, 1995.

[35] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *ACM sigmod record*, volume 29, pages 93–104. ACM, 2000.

[36] T Burr, C Hale, and M Kantor. Fraud detection in medicare claims: A multivariate outlier detection approach. Technical report, Los Alamos National Lab., NM (United States), 1997.

[37] Rodrigo N Calheiros, Kotagiri Ramamohanarao, Rajkumar Buyya, Christopher Leckie, and Steve Versteeg. On the effectiveness of isolation-based anomaly detection in cloud data centers. *Concurrency and Computation: Practice and Experience*, 2017.

[38] Jacopo Castellini, Valentina Poggioni, and Giulia Sorbi. Fake twitter followers detection by denoising autoencoder. In *Proceedings of the International Conference on Web Intelligence*, pages 195–202. ACM, 2017.

[39] Jinghui Chen, Saket Sathe, Charu Aggarwal, and Deepak Turaga. Outlier detection with autoencoder ensembles. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pages 90–98. SIAM, 2017.

[40] Mei-Chih Chen, Ren-Jay Wang, and An-Pin Chen. An empirical study for the detection of corporate financial anomaly using outlier mining techniques. In *Convergence Information Technology, 2007. International Conference on*, pages 612–617. IEEE, 2007.

[41] Wo-Ruo Chen, Yong-Huan Yun, Ming Wen, Hong-Mei Lu, Zhi-Min Zhang, and Yi-Zeng Liang. Representative subset selection and outlier detection via isolation forest. *Analytical Methods*, 8(39):7225–7231, 2016.

[42] Yong Shean Chong and Yong Haur Tay. Abnormal event detection in videos using spatiotemporal autoencoder. In *International Symposium on Neural Networks*, pages 189–196. Springer, 2017.

[43] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

[44] Zhiguo Ding and Minrui Fei. An anomaly detection approach based on isolation forest algorithm for streaming data using sliding window. *IFAC Proceedings Volumes*, 46(20):12–17, 2013.

[45] Eric Falk, Ramino Camino, Radu State, and Vijay K Gurbani. On non-parametric models for detecting outages in the mobile network. In *Integrated Network and Service Management (IM), 2017 IFIP/IEEE Symposium on*, pages 1139–1142. IEEE, 2017.

[46] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[47] Wangyan Feng, Wenfeng Yan, Shuning Wu, and Ningwei Liu. Wavelet transform and unsupervised machine learning to detect insider threat on cloud file-sharing. In *Intelligence and Security Informatics (ISI), 2017 IEEE International Conference on*, pages 155–157. IEEE, 2017.

[48] Ryohei Fujimaki, Takehisa Yairi, and Kazuo Machida. An approach to spacecraft anomaly detection problem using kernel feature space. In *Proceedings of the*

eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, pages 401–410. ACM, 2005.

[49] Anagi Gamachchi, Li Sun, and Serdar Boztas. Graph based framework for malicious insider threat detection. 2017.

[50] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. The MIT Press, 2017.

[51] Connor Hamlet, Jeremy Straub, Matthew Russell, and Scott Kerlin. An incremental and approximate local outlier probability algorithm for intrusion detection and its evaluation. *Journal of Cyber Security Technology*, 1(2):75–87, 2017.

[52] Yueyang He, Xiaoyan Zhu, Guangtao Wang, Heli Sun, and Yong Wang. Predicting bugs in software code changes using isolation forest. In *Software Quality, Reliability and Security (QRS), 2017 IEEE International Conference on*, pages 296–305. IEEE, 2017.

[53] Nuong Hoang, Khon Loi Nguyen, and Bach Huynh Dunnigan. Improvement of outliers detection algorithm based on density. *Journal of Applied Science and Engineering Innovation*, 4(3):72–75, 2017.

[54] Mia Hubert, Peter J Rousseeuw, and Karlien Vanden Branden. Robpca: a new approach to robust principal component analysis. *Technometrics*, 47(1):64–79, 2005.

[55] Vipin Kumar. Parallel and distributed computing for cybersecurity. *IEEE Distributed Systems Online*, 6(10), 2005.

[56] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 413–422. IEEE, 2008.

[57] Huawen Liu, Xuelong Li, Jiuyong Li, and Shichao Zhang. Efficient outlier detection for high-dimensional data. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2017.

[58] Juan Liu, Eric Bier, Aaron Wilson, John Alexis Guerra-Gomez, Tomonori Honda, Kumar Sricharan, Leilani Gilpin, and Daniel Davies. Graph analysis for detecting fraud, waste, and abuse in healthcare data. *AI Magazine*, 37(2):33, Apr 2016.

[59] Weining Lu, Yu Cheng, Cao Xiao, Shiyu Chang, Shuai Huang, Bin Liang, and Thomas Huang. Unsupervised sequential outlier detection with deep architectures. *IEEE Transactions on Image Processing*, 2017.

[60] Feiya Lv, Chenglin Wen, Meiqin Liu, and Zhejing Bao. Weighted time series fault diagnosis based on a stacked sparse autoencoder. *Journal of Chemometrics*, 31(9), 2017.

[61] Mathew X Ma, Henry YT Ngan, and Wei Liu. Density-based outlier detection by local outlier factor on largescale traffic data. *Electronic Imaging*, 2016(14):1–4, 2016.

[62] N Malini and M Pushpa. Analysis on credit card fraud identification techniques based on knn and outlier detection. In *Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), 2017 Third International Conference on*, pages 255–258. IEEE, 2017.

[63] Linda A. Jacobsen Mark Mather and Kelvin M. Pollard. Aging in the united states [online]. available: http://www.prb.org/pdf16/aging-us-population-bulletin.pdf. [accessed 06-Nov-2017].

[64] Andrew Ng. "sparse autoencoder". *CS294A Lecture notes*, 72(2011):1–19, 2011.

[65] Thomas Ortner, Peter Filzmoser, Maia Zaharieva, Sarka Brodinova, and Christian Breiteneder. Local projections for high-dimensional outlier detection. *arXiv preprint arXiv:1708.01550*, 2017.

[66] Genki Osada, Kazumasa Omote, and Takashi Nishide. Network intrusion detection based on semi-supervised variational auto-encoder. In *European Symposium on Research in Computer Security*, pages 344–361. Springer, 2017.

[67] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[68] Nerijus Paulauskas and uolas Faustas Bagdonas. Local outlier factor use for the network flow anomaly detection. *Security and Communication Networks*, 8(18):4203–4212.

[69] Juan-Manuel Pérez-Rúa, Antoine Basset, and Patrick Bouthemy. Detection and localization of anomalous motion in video sequences from local histograms of labeled affine flows. *Frontiers in ICT*, 4:10, 2017.

[70] Eftychios Protopapadakis, Athanasios Voulodimos, Anastasios Doulamis, Nikolaos Doulamis, Dimitrios Dres, and Matthaios Bimpas. Stacked autoencoders for outlier detection in over-the-horizon radar signals. *Computational Intelligence and Neuroscience*, 2017, 2017.

[71] Luca Puggini and Seán McLoone. An enhanced variable selection and isolation forest based methodology for anomaly detection with oes data. *Engineering Applications of Artificial Intelligence*, 67:126–135, 2018.

[72] Fatimah Almah Saaid, Robert King, Darfiana Nur, et al. Development of users call profiles using unsupervised random forest. In *Third Annual ASEARC Conference*, 2009.

[73] Mahsa Salehi, Christopher Leckie, James C Bezdek, and Tharshan Vaithianathan. Local outlier detection for data streams in sensor networks: Revisiting the utility problem invited paper. In *Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), 2015 IEEE Tenth International Conference on*, pages 1–6. IEEE, 2015.

[74] Marco Schreyer, Timur Sattarov, Damian Borth, Andreas Dengel, and Bernd Reimer. Detection of anomalies in large scale accounting data using deep autoencoder networks. *arXiv preprint arXiv:1709.05254*, 2017.

[75] Yin Shan, D Wayne Murray, and Alison Sutinen. Discovering inappropriate billings with local density based outlier detection method. In *Proceedings of the Eighth Australasian Data Mining Conference-Volume 101*, pages 93–98. Australian Computer Society, Inc., 2009.

[76] Hongchao Song, Zhuqing Jiang, Aidong Men, and Bo Yang. A hybrid semi-supervised anomaly detection model for high-dimensional data. *Computational intelligence and neuroscience*, 2017, 2017.

[77] Clay Spence, Lucas Parra, and Paul Sajda. Detection, synthesis and compression in mammographic image analysis with a hierarchical image probability model. In *Mathematical Methods in Biomedical Image Analysis, 2001. MMBIA 2001. IEEE Workshop on*, pages 3–10. IEEE, 2001.

[78] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, 15(1):1929–1958, 2014.

[79] Li Sun, Steven Versteeg, Serdar Boztas, and Asha Rao. Detecting anomalous user behavior using an extended isolation forest algorithm: an enterprise case study. *arXiv preprint arXiv:1609.06676*, 2016.

[80] Hanh TM Tran and DC Hogg. Anomaly detection using a convolutional winner-take-all autoencoder. In *Proceedings of the British Machine Vision Conference 2017*. Leeds, 2017.

[81] Gianluca Valentino, Roderik Bruce, Sonja Jaster-Merz, Stefano Redaelli, Roberto Rossi, and Panagiotis Theodoropoulos. Anomaly detection for beam loss maps in the large hadron collider. In *8th Int. Particle Accelerator Conf.(IPAC'17), Copenhagen, Denmark, 14â 19 May, 2017*, pages 92–95. JACOW, Geneva, Switzerland, 2017.

[82] Sholom M Weiss, Casimir A Kulikowski, Robert S Galen, Peder A Olsen, and Ramesh Natarajan. Managing healthcare costs by peer-group modeling. *Applied Intelligence*, 43(4):752–759, 2015.

[83] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[84] Mingtao Wu, Heguang Zhou, Longwang Lucas Lin, Bruno Silva, Zhengyi Song, Jackie Cheung, and Young Moon. Detecting attacks in cybermanufacturing systems: additive manufacturing example. In *MATEC Web of Conferences*, volume 108, page 06005. EDP Sciences, 2017.

[85] Yizhou Yan, Lei Cao, Caitlin Kulhman, and Elke Rundensteiner. Distributed local outlier detection in big data. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1225–1234. ACM, 2017.

[86] William J Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.

[87] Gang Yu, Junsong Yuan, and Zicheng Liu. Unsupervised random forest indexing for fast action search. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 865–872. IEEE, 2011.

[88] Julina Zhang, Kerry Jones, Tianye Song, Hyojung Kang, and Donald E Brown. Comparing unsupervised learning approaches to detect network intrusion using netflow data. In *Systems and Information Engineering Design Symposium (SIEDS), 2017*, pages 122–127. IEEE, 2017.

[89] Weijia Zhang and Xiaofeng He. An anomaly detection method for medicare fraud detection. In *Big Knowledge (ICBK), 2017 IEEE International Conference on*, pages 309–314. IEEE, 2017.

[90] Yan Zhang, Weiling Chen, Chai Kiat Yeo, Chiew Tong Lau, and Bu Sung Lee. Detecting rumors on online social networks using multi-layer autoencoder. In *Technology & Engineering Management Conference (TEMSCON), 2017 IEEE*, pages 437–441. IEEE, 2017.

[91] Zhongjun Zhang, Huimin Lan, and Tianjie Zhao. Detection and mitigation of radiometers radio-frequency interference by using the local outlier factor. *Remote Sensing Letters*, 8(4):311–319, 2017.

[92] Ye Zhao, Florent Balboni, Thierry Arnaud, Jerry Mosesian, Roy Ball, and Brad Lehman. Fault experiments in a commercial-scale pv laboratory and fault detection using local outlier factor. In *Photovoltaic Specialist Conference (PVSC), 2014 IEEE 40th*, pages 3398–3403. IEEE, 2014.

[93] Chong Zhou and Randy C Paffenroth. Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 665–674. ACM, 2017.