2017 IEEE International Conference on Information Reuse and Integration

Medical Provider Specialty Predictions for the Detection of Anomalous Medicare Insurance Claims

Matthew Herland
Florida Atlantic University
Boca Raton, Florida, USA
mherlan1@fau.edu

Richard A. Bauder
Florida Atlantic University
Boca Raton, Florida, USA
rbauder2014@fau.edu

Taghi M. Khoshgoftaar
Florida Atlantic University
Boca Raton, Florida, USA
khoshgof@fau.edu

***Abstract*—**

Fraud, waste, and abuse in medical insurance contributes to significant increases in costs for providers and patients. One way to reduce costs is through the detection of abnormal medical practices that could indicate possible fraud. In this paper, we expand upon our previous research into medical specialty anomaly detection by validating the efficacy of our model using real-world fraud cases, and then testing three strategies to improve model performance. The three strategies are feature selection (to include adjusting for class imbalance), medical specialty grouping, and the removal of specific, overlapping specialties. We use the publicly available Medicare claims data, released by the Center for Medicare and Medicaid Services, for building and testing our models. In addition to using the 2013 data, we use the 2014 data for model validation and comparisons. We employ the List of Excluded Individuals and Entities (LEIE) database, released by the Office of Inspector General, as well as two other documented fraud cases, for model testing. Multinomial Naïve Bayes is used to build all models. In this work, we confirm our prior model was able to correctly classify 67% of the real-world fraudulent physicians contained in the LEIE database as fraudulent. Furthermore, the three proposed strategies show good results in improving model performance.

Keywords: *Fraud Detection; Anomaly Detection; Machine Learning; Medicare*

I. Introduction

As both healthcare costs and the average life spans continue to increase, more financial tension is applied to public and private institutions which assist in funding doctor visits and treatments [1, 2]. Therefore, the goal of the healthcare system should be offering sufficient and necessary treatment to as many patients as possible, at fair cost to both patients and providers. One such healthcare system is Medicare, which is a government-run program that specifically provides financial support to seniors (and other select groups) [3]. Similar to other insurance-related programs (such as Medicaid [4]), each covered procedure is submitted with a specific procedure code. In general, a physician performs a series of procedures and then submits a claim to Medicare for payment. We do not detail Medicare or the Medicare claims process in this paper, but the interested reader can find additional information in [3, 5].

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

In 2015, Ko et al. [6] estimated that over $125 million could be saved in the field of Urology alone if adequate regulations were put in place. Joudaki et al. [7] and Pawar et al. [8] document that about 10% of all medical claims within the Healthcare system are fraudulent. An article from the Coalition Against Insurance Fraud [9], provides statistics on fraud and abuse found within the U.S. healthcare system which demonstrate the severity of the problem, as well as validate the continued need to combat this ongoing problem. As an example, recently, the largest value single insurance fraud scheme was perpetrated by three individuals, targeting nursing homes, totaling $1 billion in Medicare costs [10]. A few additional highlights on healthcare FWA include the recovery of $17.9 billion to Medicare between 2009 and 2016 [9], the exclusion of 1,662 individuals and entities from Medicare and Medicaid claims and payments, and a nearly five-fold recovery of proceeds (i.e. civil recoveries) over the last three years [11]. In an effort to catch and detain fraudulent physicians, the Office of Inspector General (OIG) created the Medicare Fraud Strike Force [12] who use data analytics to find fraud, waste, and abuse. As of June 30, 2016, this strike force has recovered $1.98 billion dollars. This is an ongoing effort of increasing importance for which we propose ways, using machine learning models and anomaly detection, to help detect possibly fraudulent or wasteful activities within Medicare claims data [13, 14, 15].

In this study, as with our previous work [16], we use data released by the Centers for Medicare and Medicaid Services (CMS) [17, 5]. In response to a new policy declared by the U.S. Department of Health and Human Services [18], CMS has begun releasing datasets in an attempt to assist in

978-1-5386-1562-1/17 $31.00 © 2017 IEEE
DOI 10.1109/IRI.2017.29

identifying fraud, waste, and abuse within Medicare [19]. One such dataset outlines every Medicare claim made by healthcare providers throughout the U.S., the average amount paid for these services and several other data points related to the specific procedures. Previously, we built an anomaly detection model to flag outliers using only the physician's procedures performed found in the Medicare claims data. Our model detects possibly fraudulent behavior by predicting a provider's specialty based on the number of procedures performed. If the physician is predicted into a different medical field (e.g. chiropractor classified as a rheumatologist), then the physician in question either has mostly unique patients or is exhibiting possibly fraudulent behaviors. In this paper, we do not consider different types of healthcare fraud, such as upcoding or self-referrals, but the reader is referred to [20] for additional information. For our research, we use the 2013 and 2014 Medicare claims data for testing and validation, and subset the data to include only the Florida-based claims. Additionally, we use data released by the OIG from the List of Excluded Individuals and Entities (LEIE) [21] data repository, which includes all current physicians who have been found unfit to practice thus excluded from practicing in the U.S. for a given period of time. The Office of Inspector General holds authority to exclude individuals and entities from federally funded healthcare programs, such as Medicare, in accordance with Sections 1128 and 1156 of the Social Security Act [21]. We employ the LEIE information to identify fraudulent physicians within the Medicare dataset for model testing and validation. In addition to the LEIE database, we found two documented fraud cases: Michael Burgos, who was indicted on charges of Medicare fraud constituting $13.8 million [22] and Salomon Melgen, an eye doctor who was charged with scamming Medicare out of $105 million [23].

Our contribution includes two related areas. We first expand upon our previous work [16], also referred to as the original model, by testing our model against a dataset of real-world fraudulent physicians. The model performance is determined by how many of physicians are classified into a specialty other than their actual practicing specialty and compared to those known fraudulent physicians found in the LEIE database. We then use our proposed strategies for improving the original model's performance. The first strategy employs feature selection and sampling to account for class imbalance. The second strategy is the removal of a selection of low scoring medical fields with a large number of instances (presumably caused by procedural overlap) in order to help boost model performance. The final strategy is grouping of similar medical specialties that have a relatively large overlapping of procedures, as confirmed through the confusion matrix. Results are shown using the 2013 Medicare data for comparative purposes, with the 2014 data used to validate those results. Note the 2014 data is not used as a test set, per say, but as validation of the 2013 model results to confirm the results are similar (for both 2013 and 2014). We employ the Multinomial Naïve Bayes classifier evaluated using 5-fold cross-validation and three performance metrics. The only attribute used for each model is the number of times each provider performs a particular procedure.

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

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

II. Related Works

The data released by the Centers for Medicare and Medicaid Services, at the point of this publication, is for 2012, 2013 and 2014. Therefore, all research done using this data is still relatively new, with additional future work needed for finding misuse in medical insurance utilization. One such effort by Feldman et al. [19] looked into how a given physician's past schooling determines the way he or she practices, from the 2012 Medicare data. The authors compared medical school charges, procedures, and payments as well as looked to find possible anomalies in the data by presenting a geographical analysis with the national distribution of school procedure payments and charges. The authors attempted to find correlations between educational backgrounds and the practices and procedures physicians perform to help pinpoint those physicians who are misusing or inefficiently using medical insurance systems.

Another study by Ko et al. [6] used the 2012 data for the urology specialty only. The authors analyzed the variability among urologists within the field's service utilization and payment, and determined an estimated savings from a standardized service utilization. They found that the number of patient visits had a strong correlation with reimbursement from Medicare. They also found, in terms of services per visit, there was high utilization variability and a possible 9% savings within the field of urology. This research could culminate in finding rules for better service utilization.

Though CMS was not the only data used, a general coverage paper by Chandola et al. [24] assessed healthcare fraud using data with labels for fraudulent providers, primarily from the Texas Office of Inspector General's exclusion database. The authors employed several techniques including social network analysis, text mining, and temporal analysis in order to translate the problem of healthcare data analysis into some well-known data mining methods. More specifically, the authors discussed the use of typical treatment profiles, i.e. procedures performed, in order to compare among providers and spot possible issues or abuses in procedures to treat particular ailments.

There are additional, related studies utilizing data from sources other than the CMS data. Pawar et al. [8] wrote a survey giving a small overview of fraud detection using publicly available data, specifically through the use of geolocation clustering along with various other clustering algorithms. Joudaki et al. [7] collected claims data, and other

information, for general physicians specifically in the area of drug prescription. They attempted to find indicators of fraud within the healthcare system using Iranian-based data. They employed a total of thirteen indicators in determining abuse and fraud in physician behavior using clustering and discriminate analysis with satisfactory results. In the study [25], Van et al. used Medicaid data for 369 dentists with a total of 650,000 claims and similar to our work employed outlier detection in order to determine fraudulent physicians. In their work they used unsupervised techniques with 14 metrics (which they decided after review of additional sources such as FBI reports) and determined that 17 out of the 369 dentists should be investigated for fraud. These results were then discussed with professionals within the field of dentistry and 12 out of these 17 physicians were determined worthy of further inspection giving them a 71% detection rate.

In our previous work [16], we investigated whether or not it is possible to predict a physician's field of expertise based on only the procedures they perform. The idea was that if we could perform this prediction accurately (determined by F-Score), then we could find potentially anomalous behavior in a particular physician's procedures (compared to the norm of other physicians in their field). If they are anomalous then they are more likely to be fraudulent, wasteful or abusive. Even though the initial results were satisfactory, considering the large number of fields present in the dataset, the model could be improved to better detect possibly fraudulent behaviors. This study was a proof-of-concept endeavor and did not confirm whether or not the model can actually predict a physician's fraudulent behavior. As such, we intend to improve upon this prior study by evaluating the performance of our original model, using real-word known fraud cases, and assess the improvements made through our three proposed strategies.

III. Methodology

This section details the datasets, learner, and performance metrics used. Additionally, we explain feature selection and sampling as well as the three improvement strategies. For our research, we chose Multinomial Naïve Bayes to build each model using the Medicare claims datasets (2013 and 2014), with Precision, Recall and F-Score as the performance metrics.

A. Data

The Physician and Other Supplier Data CY 2013 and 2014 dataset [5] outlines how many times each provider billed a specific procedure. Due to the large size of the dataset, we decided to only use data from office clinics in Florida (as opposed to larger facilities, such as hospitals and academic institutions). Table I summarizes the 2013 and 2014 Medicare data, nationally and for Florida only. Each physician is denoted by his or her National Provider Identifier (NPI) [26] and each procedure is labeled by its Healthcare Common Procedure Coding System (HCPCS) code [27]. For this study, we are only interested in the provider's specialty, procedures, and number of procedures performed. Even so, the Medicare dataset contains a number of other features, such as the average amount billed and paid for each physician and procedure.

Because we are only interested in the procedures physicians perform, we transformed each physician entry into a vector

TABLE I: Dataset Summary

2013 Statistic	Full Dataset	Florida Only
Number of Physicians	909,606	40,040
Number of Procedures	5,983	2,789
Provider Types (Specialties)	90	82

2014 Statistics	Full Dataset	Florida Only
Number of Physicians	938,147	41,896
Number of Procedures	5,973	2,563
Provider Types (Specialties)	90	82

where the key value, or class label, for each instance is the physician's specialty and the features are all available procedures (identified via unique HCPCS codes). The value for each feature is the number of times a given provider billed Medicare for that given procedure. This results in a sparse vector, since most physicians only use a small number of codes necessary for their own practice. Table II shows a small example of the sparse vector where each line is a physician, PROVIDER_TYPE is the class attribute and every other attribute (codes 99222 through 64482 in this example) are the procedures. For every instance, there is a value for the number of times the given physician performed that procedure.

TABLE II: Sample of Dataset used for this Study

Specialty	99222	99223	88304	...	64482
Internal Medicine	142	96	0	...	0
Pathology	0	0	209	...	0
Anesthesiology	0	0	0	...	16

Additionally, for model testing and validation, we incorporate the List of Excluded Individuals and Entities data, which includes physicians who have been found to be in violation of one or more items within Sections 1128 and 1156 of the Social Security Act. After reviewing the violations under these sections, we decided to only keep physicians that violated the codes described in Table III. Only physicians who violated these codes were used in our study. We matched physicians from our 2013 procedure code dataset with the LEIE data. The LEIE is constantly changing as physicians are added or removed; therefore, we accessed the LEIE database twice during our research, for 2016 and 2017, and found 16 physicians in the Florida Medicare data. Unfortunately, the LEIE does not contain NPI numbers for all physicians, and first and last names are not reliable, so we were left with only those physicians with NPI numbers listed to identify matches between datasets. We supplemented these 16 physicians with two other documented fraud cases [23, 22], found in the 2013 data, giving us 18 fraudulent physicians. Note that the LEIE dataset is only used to identify fraudulent doctors corresponding to the information in the Medicare data. Only the Medicare dataset is used for model training and testing.

B. Original Model Testing and Validation

In order to create and assess improvements to our original model, we need to validate, through the use of real-

TABLE III: LEIE 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 healthcare fraud.
1128(b)(4)	License revocation or suspension.
1128(b)(7)	Fraud, kickbacks, other prohibited activities.

world fraudulent physicians, the hypothesis that by using only procedural data, a prediction model can successfully detect possibly fraudulent or wasteful behaviors. For this, we use the 2013 Medicare claims data with labels consisting of known fraudulent physicians from the LEIE database and the two documented cases. The procedure used for testing our original model is exclusion testing, where we removed the fraudulent physicians from the training dataset and created a test dataset composed of the 18 known fraudulent physicians. The model was built using Multinomial Naïve Bayes, in Weka [28], from the modified training data (after the 18 were removed) and evaluated on the test dataset (including only the 18 removed instances). The test evaluation was done by reviewing the resulting confusion matrix, where the number of instances predicted into a class other than their actual field are denoted as possibly fraudulent.

C. Improvement Strategies

In this subsection, we discuss the three improvement strategies to include feature selection and sampling, grouping of similar physician types, and removing specific classes. From preliminary analysis, we found that specialties with comparable characteristics are similarly improved by a particular strategy, thus chose a small, representative number of specialties to focus our research. It is important to note that the Medicare 2013 dataset is identical to that used in our previous work. Only the exclusion testing involved the removal of non-fraudulent physicians.

1) Feature Selection and Sampling: The 2013 Florida data subset was altered using the R language [29] into two classes, or specialties, rather than the original 82 specialties. The specialty that we wish to focus on for classification was kept as is in a single class (the positive class), while each instance of the remaining 81 fields were grouped into a single class (the negative class or *other class*), creating a one-versus-all scenario. For example, we can choose Podiatry as the positive class and then group all other classes in the *other class* (negative class) for binary classification. We use the Weka platform to apply the Gain Ratio feature selection technique [30] to build each Multivariate Naïve Bayes model. Gain Ratio returns the top n features (procedure codes in our case) to keep, while removing the rest. Through initial experimentation, we chose a range of procedures for n from all the procedures (2,789) to 500 procedures, in increments of 500. The performance results will indicate the optimal number of procedures to keep, as well as any model improvements due to feature selection.

In addition to feature selection, we also experiment with under- and over-sampling techniques to improve model performance. Sampling could be beneficial for feature selection since the field of interest (positive class) is relatively small compared to the other class (negative class). For instance, even for the largest field of interest in our research, Internal Medicine only has 4,243 instances versus 35,797 in the other class. Thus, incorporating sampling techniques could help improve the overall prediction results. Under-sampling is the preprocessing technique of removing instances from the majority class in order to balance the two classes. Over-sampling is another method for balancing classes, but rather than removing instances from the majority class, over-sampling adds instances to the minority class. For under-sampling, we use the preprocessing algorithm SpreadSubsample [31] and for over-sampling, we use the Synthetic Minority Over-sampling Technique (SMOTE) [32]. Both sampling techniques are done in Weka, where we assume the default configurations unless specifically stated otherwise.

2) Removing Classes: For this approach, we remove classes based on unique procedures that have both a high number of instances and poor classification performance. For this experiment, four classes were chosen for removal from the dataset: Family Practice, nurse practitioner, Internal Medicine, and physician assistant. Together, these four specialties make up a large portion of the original 2013 dataset with 7,015 out of 40,040, or 17.5% of all instances. The choice of these four classes was determined by reviewing the confusion matrix and confirming that all removed classes do indeed cause a relatively large number of misclassifications. Specialties with high misclassification rates would be considered generic, meaning that they most likely perform HCPCS codes that a number of other fields also perform (i.e. overlapping procedures). We repeated the procedure to validate the model (with removed classes) based on the known fraudulent physicians to compare performance to the original model.

3) Grouping Specialties: This improvement strategy groups a specialty with other similar specialties. Similar specialties, or classes, are ones that share a significant number of HCPCS codes, thus overlapping procedures. Table IV shows the total number of procedures performed and the overlap between ophthalmologist and optometrist, across five randomly chosen codes. The process for determining similar classes was done using the confusion matrix and selecting the specialties which were confused for other specialties, and those with no confusion in classification. Note that classes such as Internal Medicine, which is confused for many other classes, were not considered due to the large grouping created (including too many other classes) and thus not useful for this experiment. A very large group that consists of many classes would defeat the purpose of grouping, as we want to find smaller groups of specialties where every class within this group is actually similar to each other in practice. For instance, Internal Medicine could share a number of procedures with both Cardiology and Optometry, but Cardiology and Optometry are not similar nor is Internal Medicine. Additional anecdotal confirmation of these similar class groupings was found based on whether these groups were reasonable or not. For example, Ophthalmology and Optometry provide medical procedures focused on the eyes, thus is a reasonable group. For this study, classes were grouped manually on a class-by-class basis.

TABLE IV: Overlapping HCPCS Code Example

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

D. Performance Metrics

The calculations for model performance are from the confusion matrices using the Multivariate Naïve Bayes model with a single run of 5-fold cross-validation. As previously mentioned, we use a one-vs-all approach for calculating the error rates, which considers the class in question as the positive class (true positive) and the rest of the classes as being in the same negative class (true negative). We leverage Recall, Precision and F-Score to assess model performance, with F-Score being the primary metric for comparison.

Recall measures the ability of a classifier to determine the rate of positively marked instances that are in fact positive; therefore, for this dataset, recall is the fraction of physicians labeled correctly and not as any of the other 81 specialties. Precision indicates how well a classifier has predicted a class by finding the ratio of actually positive instances from the pool of instances that it has marked as part of the positive class; therefore, precision shows the fraction of physicians marked correctly against the number of physicians, from any of the other 81 fields, also marked as the class in question.

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

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

E. Learner

Multinomial Naïve Bayes (MNB) classifier is used to build each model for our experiments. This learner classifies new instances by finding the posterior probabilities of class membership based on each feature value, which is learned from a set of labeled training instances. The approximation is done using Bayes' rule by assuming conditional independence. Conditional independence is the idea that each feature in the dataset is independent from one another which is rarely true in practice, however, the model is very effective and is used extensively in the field of data mining and machine learning [33]. Specifically, we used the implementation in the Weka machine learning toolkit, with 5-fold cross-validation for model evaluation.

F. Feature Selection Technique

Gain Ratio is a slightly altered version of the information gain feature selection technique, born out of a weakness of the latter, where it tends to have bias toward attributes with many values [30]. In the Medicare claims data, each attribute can potentially be any value given that each attribute being a procedure, and a physician can perform a procedure any number of times. Therefore, using the Gain Ratio feature selection technique, that considers removing such bias, is beneficial to our research.

IV. Results and Discussion

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

The original model is validated against a real-world exclusion dataset to confirm the successful detection of fraudulent behaviors. Due to the limited number of physicians in the LEIE corresponding to the Florida-only Medicare data, testing the strategies would not be reliable; therefore, with these strategies, we focus on improving F-Score results over the original model results found in [16]. To assess performance changes, we chose to focus on a few select fields from which we found that different strategies improve model performance differently based on the specialty. As mentioned, the three improvement strategies are 1) feature selection and sampling, 2) removing classes, and 3) grouping similar classes.

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

A. Original Model Testing and Validation

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

TABLE V: Classification of Fraudulent Physicians

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

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

B. Strategies

In this section, we present the results of our three strategies, assess improvements made, and discuss any caveats and limitations regarding the application of these strategies per specialty.

TABLE VI: Chosen Fields for 2013 CMS Data

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

Additionally, we provide comparative testing values with the 2014 Medicare data to validate and confirm the results and improvements seen using the 2013 Medicare data.

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

TABLE VII: Dermatology Feature Selection Results

# of Procedures	Specialty	F-Score	Avg F-Score
Full	Dermatology	0.39	0.68
	Other Class	0.97	
2500	Dermatology	0.39	0.68
	Other Class	0.97	
2000	Dermatology	0.40	0.68
	Other Class	0.96	
1500	Dermatology	0.42	0.69
	Other Class	0.96	
1000	Dermatology	0.42	0.70
	Other Class	0.97	
500	Dermatology	0.43	0.70
	Other Class	0.97	

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

Unfortunately, feature selection alone does not produce

TABLE VIII: Feature Selection Results for the Remaining Specialties

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

sufficiently improved performance results. Opthalmology and Internal Medicine show increased performance over the original model, as seen in Table VI, but the others have lower average F-Scores. The lack of improvement is most likely due to the large difference between the size of the classes, which, conversely, is why Internal Medicine performs better as it has a high number of instances relative to the entire dataset versus the other tested specialties. Since class imbalance is seen to be an issue, balancing the class sizes through sampling could be used to further increase model performance.

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

TABLE IX: Results of Random Under- and Over-sampling

Specialty	Under-sampling		Over-sampling	
	Reduction Setting	Average F-Score	Addition Setting	Average F-Score
Ophthalmology	20	0.86	1800	0.97
Dermatology	20	0.79	2200	0.96
Cardiology	10	0.73	1200	0.91
Otolaryngology	30	0.65	4000	0.90
Internal Medicine	4	0.50	450	0.77

2) Removing Classes: In Table X, we present the results based on the removal of specific classes. In this table, we show the original model's 2013 F-Scores, update scores after class removal, the 2013 change in F-Scores, and the changes in scores using the 2014 data for confirmation and validation of model improvements. To reiterate, the 2014 data was not used as a test set but rather to replicate the procedure we used on the 2013 data, creating a new model, in order to show similar results between years. We do not show the original and post class removal F-Scores for 2014 due to space, but the scores are very similar the 2013 data results. Additionally, Fig 1 in the Appendix shows the results for the remaining 78 specialties.

After removing the four classes, as described in Section III-C2, we notice that there are large improvement in the low scoring specialties, as well in Otolaryngology and Dermatology, with Ophthalmology being the only class to have a decrease in F-Score. From these results, the classes with relatively large improvements most likely have procedures that are easily confused with those in the removed classes, i.e. a large number of overlapping procedures. The classes with little to no improvements would indicate more specialized services with minimal overlapping procedures, with regards to the removed classes. Furthermore, we tested the class removal strategy after removing the same four specialties.

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

TABLE X: Improvements from Removing Classes

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

3) Grouping Specialties: The results for the grouping of specialties are outlined in Tables XI and XII. These results

summarize model performance (recall, precision, and F-Score) for the individual specialties and the grouping of the specialties which have similar practices. The group score is the weighted average of the individual performance results. Classes were grouped manually on a class-by-class basis, where, for example, Otolaryngology shows the results for Otolaryngology and its similar classes. As in Section IV-B2, we used the 2014 data in order to corroborate and validate the 2013 model results. The original models and the grouped models were built using Multinomial Naïve Bayes, to fairly compare the results and show any improvements due to our specialty grouping strategy.

F-Scores improved for each grouping, with Cardiology and Ophthalmology having the most significant improvement. These improvements appear to be heavily dependent on the specialty for which the grouping is applied; therefore, this strategy will be effective for some classes but less effective for others. By grouping Ophthalmology with Optometry, the F-Score increased over the individual scores resulting in a very high F-Score of 0.96. The Cardiology grouping shows two of the individual specialties with F-Scores below 0.5 and Cardiology only with a score of 0.82. Even with two of the group members having low individual F-Scores, the Cardiology group had a good F-Score of 0.9. Dermatology indicated negligible improvement with grouping, while the Otolaryngology group had a slight decrease in performance. The improvements shown by grouping specialties may not actually increase the effectiveness of our model to detect frauduelent and wasteful behavior. For instance, the fraudulent Ophthalmology case in our test dataset was confused due to the inclusion of the Optometry specialty. This particular grouping would decrease the 67% fraud detection rate of our model. Thus, additional experimentation is needed to confirm whether this strategy, not only improves F-Scores, but also improves real-world detection. In particular, we would need more real-world fraudulent Ophthalmology cases to determine if they are commonly confused for Optometry, thus demonstrating that some groupings can mask possible fraud activities.

TABLE XI: Class Grouping Results

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

TABLE XII: Improvements from Grouping Classes

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

V. Conclusion And Future Work

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

TABLE XIII: Summary of F-Score Results

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

TABLE XIV: Best Strategies for Classes

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

In Table XIII, we summarize the F-Score results from the original model and all proposed improvement strategies. With feature selection only, we found that using 1,000 to 1,500 procedures gave the best F-Scores across the classes, but none of the classes demonstrated improvement over the original model. In addition to feature selection, we used under- and over-samplig to improve model performance. The results showed that under-sampling did not positively change the F-Score, whereas over-sampling improved results across the board. Our second strategy, class removal, showed large improvements for a few fields (e.g. Dermatology) implying this method could

improve model performance relative to certain specialties. As mentioned, there is a concern with removing too many specialties reducing potential fraud detection capabilities. Grouping of similar specialties is our third strategy which showed some significant improvements over individual specialty results, such as the combination of Ophthalmology and Optometry, but less noticeable changes to other groupings. Overall, our results indicate that different strategies can improve model performance depending on the selected specialties; therefore, the choice of strategy is, in large part, determined by the specialty. With that, we provide specific characteristics or reasons why a specialty would improve given one of the three proposed strategies in Table XIV. As the LEIE dataset contains very few NPIs, with only 18 were available for Florida, future work will incluce adding other states so that more comparisons of real-world fraudulent cases can be done using the LEIE database. Additionally, using different machine learning methods and performance metrics will be pursued.

ACKNOWLEDGMENT

We acknowledge partial support by the NSF (CNS-1427536). Opinions, findings, conclusions, or recommendations in this material are the authors' and do not reflect the views of the NSF.

REFERENCES

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

[2] Google. Life expectancy. [Online]. Available: https://www.google.com/publicdata/explore?ds=d5bncppjof8f9_&met_y=sp_dyn_le00_in&idim=country:USA:GBR:JPN&hl=en&dl=en

[3] U.S. Government, U.S. Centers for Medicare & Medicaid Services. The offical u.s. governement site for medicare. [Online]. Available: https://www.medicare.gov/

[4] A federal government managed website by the Centers for Medicare & Medicaid Services. Medicaid.gov. [Online]. Available: https://www.medicaid.gov

[5] CMS. Research, statistics, data, and systems. [Online]. Available: https://www.cms.gov/research-statistics-data-and-systems/research-statistics-data-and-systems.html

[6] J. S. Ko, H. Chalfin, B. J. Trock, Z. Feng, E. Humphreys, S.-W. Park, H. B. Carter, K. D. Frick, and M. Han, "Variability in medicare utilization and payment among urologists," *Urology*, vol. 85, no. 5, pp. 1045–1051, 2015.

[7] H. Joudaki, A. Rashidian, B. Minaei-Bidgoli, M. Mahmoodi, B. Geraili, M. Nasiri, and M. Arab, "Improving fraud and abuse detection in general physician claims: a data mining study," *International journal of health policy and management*, vol. 5, no. 3, p. 165, 2016.

[8] M. P. Pawar, "Review on data mining techniques for fraud detection in health insurance," *IJETT*, vol. 3, no. 2, 2016.

[9] Coalition Against Insurance Fraud. By the numbers: fraud statistics. [Online]. Available: http://www.insurancefraud.org/statistics.htm

[10] Fox News. Authorities: $1b medicare fraud nursing home scam, 3 charged. [Online]. Available: http://www.foxnews.com/us/2016/07/22/authorities-1b-medicare-fraud-nursing-home-scam-3-charged.html

[11] Wikipedia. Civil recovery. [Online]. Available: https://en.wikipedia.org/wiki/Civil_recovery

[12] Medicare Fraud Strike Force. Office of inspector general. [Online]. Available: https://www.oig.hhs.gov/fraud/strike-force/

[13] R. A. Bauder and T. 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*. IEEE, 2016, pp. 11–19.

[14] R. A. Bauder and T. M. Khoshgoftaar, "Multivariate anomaly detection in medicare using model residuals and probabilistic programming," in *FLAIRS Conference*, 2017, pp. 417–422.

[15] R. A. Bauder and T. M. Khoshgoftaar, "Multivariate outlier detection in medicare claims payments applying probabilistic programming methods," *Health Services and Outcomes Research Methodology*, pp. 1–34, Jun 2017. [Online]. Available: http:https://doi.org/10.1007/s10742-017-0172-1

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

[17] CMS. Physician and other supplier data cy 2013- centers for medicare & medicaid services. [Online]. Available: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Physician-and-Other-Supplier2013.html

[18] HHS. U.s. department of health & human services. [Online]. Available: http://www.hhs.gov/

[19] K. Feldman and N. V. Chawla, "Does medical school training relate to practice? evidence from big data," *Big Data*, vol. 3, no. 2, pp. 103–113, 2015.

[20] R. Bauder, T. M. Khoshgoftaar, and N. Seliya, "A survey on the state of healthcare upcoding fraud analysis and detection," *Health Services and Outcomes Research Methodology*, vol. 17, no. 1, pp. 31–55, 2017.

[21] LEIE. Office of inspector general leie downloadable databases. [Online]. Available: https://oig.hhs.gov/exclusions/authorities.asp

[22] Department of Justice. Florida doctor indicted for role in $13.8 million medicare fraud scheme. [Online]. Available: https://www.justice.gov/opa/pr/florida-doctor-indicted-role-138-million-medicare-fraud-scheme

[23] Palm Beach Post. North palm beach eye doctor melgen jailed on medicare fraud charges. [Online]. Available: http://www.palmbeachpost.com/news/crime--law/north-palm-beach-eye-doctor-melgen-jailed-medicare-fraud-charges/czyT7d9jhcVrZfbQpyUUcP/

[24] V. Chandola, S. R. Sukumar, and J. C. Schryver, "Knowledge discovery from massive healthcare claims data," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, pp. 1312–1320.

[25] G. van Capelleveen, M. Poel, R. M. Mueller, D. Thornton, and J. van Hillegersberg, "Outlier detection in healthcare fraud: A case study in the medicaid dental domain," *International journal of accounting information systems*, vol. 21, pp. 18–31, 2016.

[26] CMS. National provider identifier standard (npi). [Online]. Available: https://www.cms.gov/Regulations-and-Guidance/Administrative-Simplification/NationalProvIdentStand/

[27] CMS. HCPCS - General Information. [Online]. Available: https://www.cms.gov/Medicare/Coding/MedHCPCSGenInfo/index.html?redirect=/medhcpcsgeninfo/

[28] I. H. W. Eibe Frank, Mark A. Hall, "Data mining: Practical machine learning tools and techniques," http://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf, 2016.

[29] The R Foundation. What is r? [Online]. Available: https://www.r-project.org/about.html

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

[31] Weka.sourceforge. Class spreadsubsample. [Online]. Available: http://weka.sourceforge.net/doc.stable/weka/filters/supervised/instance/SpreadSubsample.html

[32] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[33] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

APPENDIX

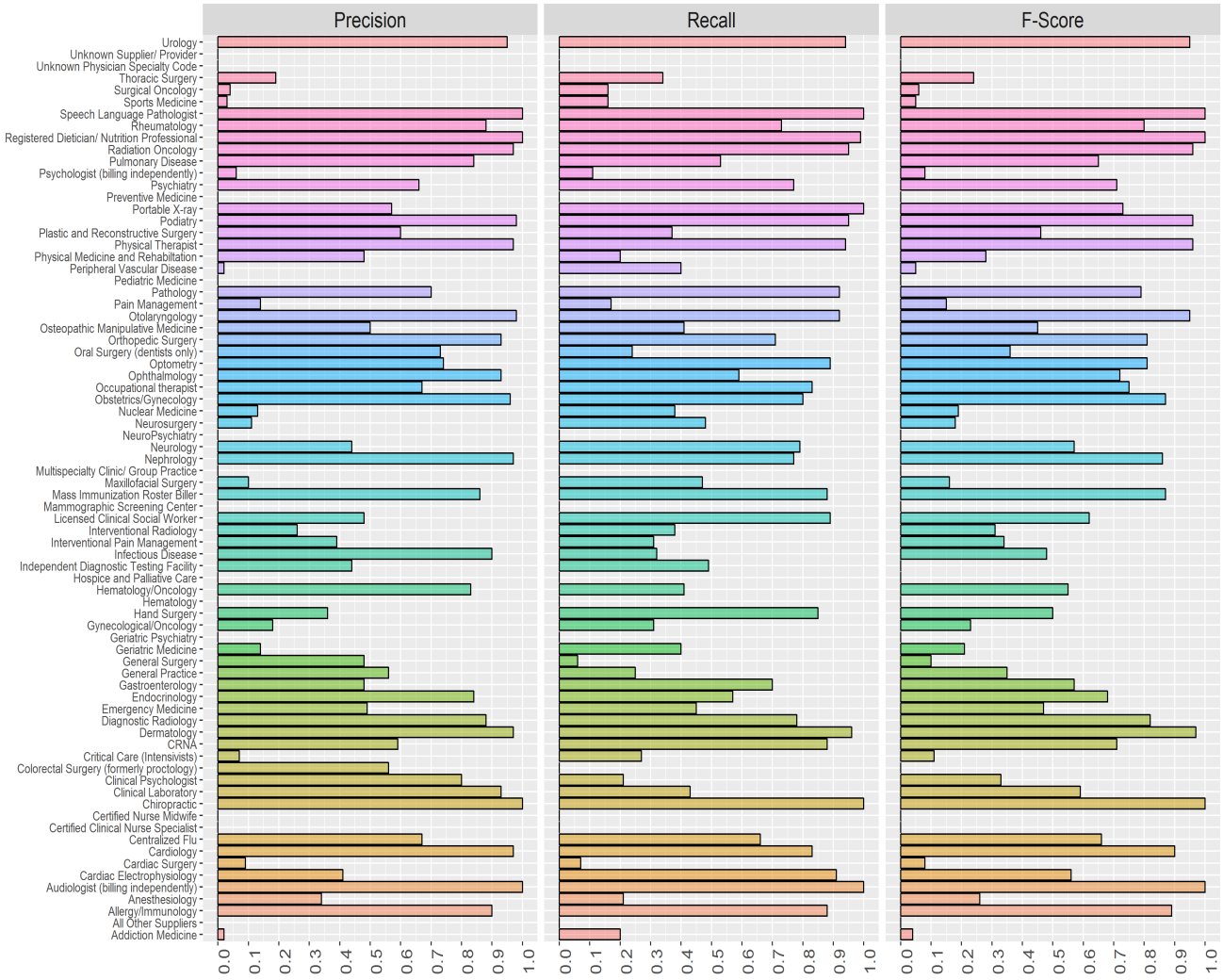

Fig. 1: Class Removal for All Physician Types (2013)