2016 IEEE 28th International Conference on Tools with Artificial Intelligence

Predicting Medical Provider Specialties to Detect Anomalous Insurance Claims

Richard A. Bauder, Taghi M. Khoshgoftaar, Aaron Richter, Matthew Herland
Florida Atlantic University
Email: {rbauder2014, khoshgof, arichter, mherlan1} @fau.edu

***Abstract*—**

The healthcare industry is a complex system with many moving parts. One issue in this field is the misuse of medical insurance systems, such as Medicare. In this paper, we build a machine learning model to detect when physicians exhibit anomalous behavior in their medical insurance claims. This new research has the potential to give some insight in determining if, and when, physicians are acting outside the norm of their respective specialty, which could indicate misuse, fraud, or lack of knowledge around billing procedures. We use a publicly available procedure billing dataset, released by the U.S. Medicare system. Due to the large size of the dataset, we sampled the dataset to include all physicians practicing within one state only. The model uses the multinomial Naïve Bayes algorithm and is evaluated by calculating precision, recall, and F-score with 5-fold cross-validation. The model is able to successfully predict several classes of physicians with an F-score over 0.9. These results show that it is possible to effectively use machine learning in a novel way to classify physicians into their respective fields solely using the procedures they bill for. This research provides a model that can identify physicians who are potentially misusing insurance systems for further investigation.

Keywords*—*Fraud Detection, Anomaly Detection, Machine Learning, Healthcare, Medicare

I. Introduction

The human body is a complex system; therefore, it is necessary to have specialized physicians trained to diagnose and treat diseases in different parts of the body. This leads to different types of treatment plans and procedures that doctors perform for patients in various fields. The goal of a healthcare system is to effectively treat as many patients as possible, but there is cost associated with every treatment on various levels. Physicians, drug manufacturers, and medical staff must be paid for their time and expertise, in addition to various medical equipment and facilities. As these costs are frequently much more than an individual patient can afford, insurance plans are used to distribute costs across all patients in the network and pay for the necessary personnel and equipment. As with any insurance system, there is potential for misuse or fraudulent activities. The Federal Bureau of Investigations estimates that fraud accounts for 3-10% of all billings [14]. With $604 billion in costs during 2013 [11], fraudulent activities could account for $18 to $61 billion annually. Clearly, healthcare fraud continues to be a major problem for the government and taxpayers, with a need for more effective detection methods. Given this need, the detection of potential fraudulent behaviors, such as physician referrals or upcoding [4], is the purpose of our research. We attempt to identify such activity within the healthcare system by examining claims for medical procedures performed by healthcare providers.

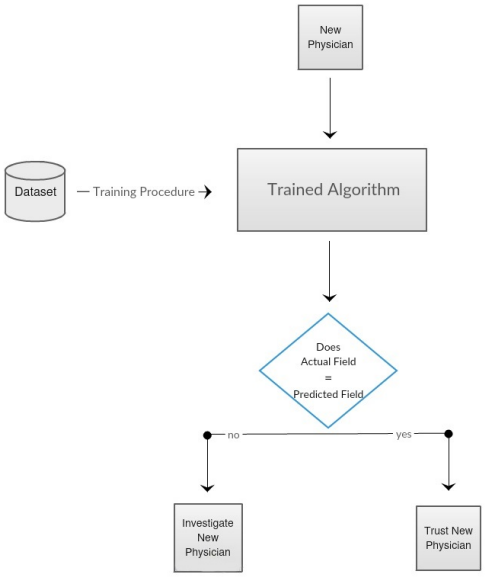
One such insurance provider is Medicare, the government-run organization that provides health insurance for seniors (as well other select groups). The Centers for Medicare and Medicaid Services, [3], in response to a new policy declared by the U.S. Department of Health and Human Services [10], recently started releasing datasets in an attempt to assist in identifying fraud, waste, and abuse within Medicare [7]. One such dataset outlines all procedure claims made by each health provider in the U.S., and the average amount paid for these services, among other data points.

Malicious or wasteful use of any medical financial system makes healthcare inefficient, potentially leaving patients without the treatment they need. Ko et al. [12] estimate a 9% savings just in the field of Urology (over $125 million) by regulating the use of Medicare. In order to both set and control regulations, there needs to be a process to determine when regulations are broken. Anomaly detection, as part of a machine learning process, can determine when certain physicians do not practice in a similar manner as his or her peers [15]. While this model would not be able to guarantee that a physician is practicing maliciously, it can help insurance systems flag outliers that would require further investigation. To achieve this, we explore the use of a machine learning model to predict the specialty of a physician, based solely on the procedures that he or she performs, as depicted in Figure 1. If the predicted specialty matches the provider's actual specialty, then the assumption is the provider is practicing within the norm of his or her field. If not, the provider could have very unique patients or could be sending wasteful or malicious claims. For the latter case, these providers exhibit aberrant, and possibly fraudulent, behaviors, warranting additional scrutiny into their practicing habits. To the best of our knowledge, we are the first to propose such a method.

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

2375-0197/16 $31.00 © 2016 IEEE
DOI 10.1109/ICTAI.2016.120

Fig. 1. Machine Learning Workflow



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

The rest of the paper is organized as follows. Section II discusses works related to the current research in this domain. In Section III, the experimental methods used in this paper are detailed to include 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 that the Centers for Medicare and Medicaid Services (CMS) has released, at the point of this publication, is only for 2012 and 2013 and were released in 2014 and 2015, respectively. Therefore, all research done using this data is in the preliminary stages with additional future work needed for finding misuse in medical insurance utilization. One such research effort, which uses the 2012 data, looked into how a given physician's past schooling determines the way he or she practices [7]. They compare medical school charges, procedures, and payments as well as look to find possible anomalies in the data by presenting a geographical analysis with the national distribution of school procedure payments and charges. By following this line of research, the authors attempt to find correlations between educational backgrounds and the practices and procedures physicians perform to help pinpoint those physicians who are misusing or inefficiently using medical insurance systems.

Another study that used the 2012 CMS data specifically looked at one field: Urology [12]. The authors analyze variability among Urologists within the field's service utilization and payment and determine an estimated savings from a standardized service utilization. They found that the number of patient visits had a strong correlation with reimbursement from Medicare. They also found, in terms of services per visit, there was a high utilization variability and a possible 9% savings within the field of Urology. This research can lead to finding rules for service utilization.

Though not only using CMS data, a general coverage paper by Chandola et al. [5] assesses healthcare fraud using data with labels for fraudulent providers, primarily from the Texas Office of Inspector General's exclusion database. The authors employ several techniques including social network analysis, text mining, and temporal analysis in order to translate the problem of healthcare data analysis into some well-known data mining methods. More specifically, Chandola et al. [5] discuss the use of typical treatment profiles, i.e. procedures performed, in order to compare among providers and spot possible misuses or abuses in procedures to treat particular ailments.

III. Methodology

This section discusses the dataset, learner and metrics used in this experiment. We chose one learner, multinomial Naïve Bayes, for research on our edited CMS dataset and used F-score as our main performance metric. Figure 1 explains the workflow of our proposed model, where the CMS procedure data is leveraged to create a model which takes the physician in question and determines whether they fit into the norm of their respective field. If they are determined to fit into the norm of their respective field, then they are considered trustworthy; if not, there may be reason to investigate the physician for possible misuse. Multi-class classification of each provider specialty is produced, using the counts for procedures performed as the model inputs.

For this paper, we decided to use the 2013 data and to the best of our knowledge, we are the first to try determining whether or not it is possible to predict a physician's field of expertise based on the procedures they perform. Our research could lead to assisting other researchers, and eventually the government, finding discrepancies in the everyday dealings of physicians who are abusing the system, committing insurance fraud or wasting funds through the detection and flagging of outliers by finding physicians that do not fit into the norm of their respective field. This study can help determine which doctors should be investigated and assist insurance systems (such as Medicare) with setting up rules and regulations for physicians to run more cost effective practices free of abuse and mistreatment.

A. CMS Data

The Physician and Other Supplier Data CY 2013 dataset, found at CMS.gov [3], outlines how many times each provider,

TABLE I. COMPARISON OF DATASET STATISTICS

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

or physician, billed a specific procedure. The dataset represents 5,983 distinct types of procedures done by 909,606 physicians throughout the United States. Due to the large size of the dataset, we decided to only use data from office clinics in Florida (as opposed to larger facilities, such as hospitals and academic institutions). This resulted in a subset of the dataset composed of 82 classes of physicians, 40,940 instances (individual physicians) and 2,789 distinct procedures. Due to Florida's unique demographic, the use of this subset is not necessarily representative of the entire US population. Even so, Florida is a good candidate for testing our method for several reasons to include having the second highest number of Medicare beneficiaries and being second in total Medicare spending [8]. Table I illustrates the size of the original CMS dataset in comparison to our subset, in terms of classes, instances, and procedures, in order to demonstrate the proportion of the original used to validate this work. Each physician is denoted by his or her National Provider Identifier (NPI) and each procedure is labeled by its Healthcare Common Procedure Coding System (HCPCS) code [1]. The dataset contains a number of other features, such as the average amount billed and paid by Medicare for each physician/procedure.

For this study, we are only interested in the procedures and various physician attributes. Therefore, we transformed each physician entry into a vector where the key value, or class label, for each instance is the physician's field and the features are all the procedures done in each field (even if the procedure is only done once by one physician in a given field). The value for each feature is the number of times a given provider billed Medicare for the given procedure. This results in a sparse vector, since most physicians only use a small number of codes necessary for their own practice. The rest of the features are then zero for that physician. Table II shows a small sample of the the original CMS data and for comparison, Table III shows a sample of the dataset after it was converted to the sparse vector for this study (NPIs are masked).

B. Learners

For our multi-class classification experiment, we use the multinomial Naïve Bayes classifier. 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 [16]. We used the WEKA [9] machine learning toolkit to perform the experiment, with 5-fold cross-validation for model evaluation.

C. Performance Metrics

In the calculation of the following metrics, we took the average over each fold of the 5-fold cross-validation for each metric. We use the one-vs-all approach for calculating the error rates, which considers the class in question as the positive class then considers the rest of the classes as being in the same negative class. True positive (tp in the subsequent equations) is when a class is correctly identified in the positive class, whereas true negative (tn) identifies a negative class correctly. We determined, through preliminary experimentation, that the following selected metrics are the most suitable, because the dataset contains numerous classes where a metric such as accuracy would not be useful due to the data sparseness and multi-class imbalances. Techniques for handling class imbalance are not currently used but are considered for future work.

Recall measures the ability of a classifier to determine the rate of positively marked instances that are in fact positive; therefore, for this dataset, recall is showing the proportion that a given physician is labeled correctly as its field and not as any of the other 81 fields.

$$Recall = \frac{tp}{tp + fn} \tag{1}$$

Precision tells how well a classifier has predicted a class by finding the ratio of actually positive instances from the pool of instances that it has marked as part of the positive class; therefore precision here is showing the proportion that a given physician is marked correctly against the amount of physicians, from any of the other 81 fields, marked also as the class in question.

$$Precision = \frac{tp}{tp + fp} \tag{2}$$

F-score, or F1 score, is the harmonic mean of both precision and recall which generates a number between 0 to 1. This is the metric that the results are organized by in the following section in order to use one concise metric.

$$F_1 = 2 \times \frac{Recall \times Precision}{Recall + Precision} \tag{3}$$

Fig. 2. Histogram of F-Scores

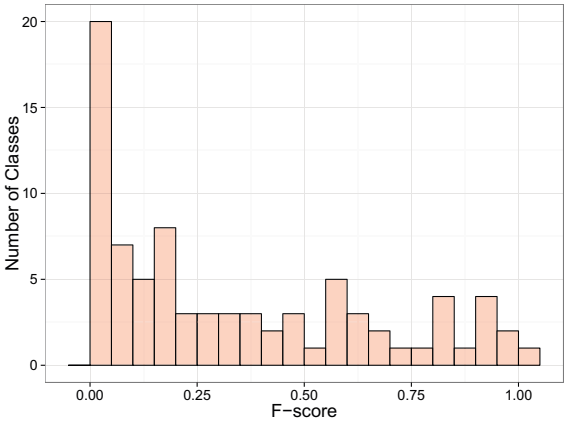


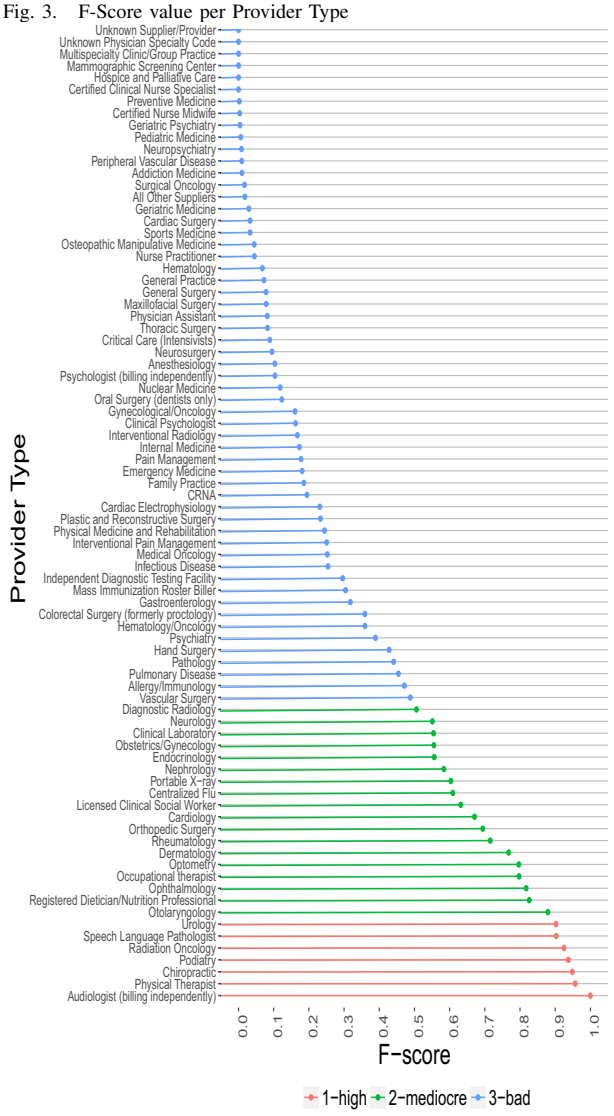
TABLE II. SAMPLE OF THE FULL DATASET

NPI	PROVIDER_TYPE	HCPCS_CODE	LINE_SRVC_CNT
123456789	Internal Medicine	99222	142
123456789	Internal Medicine	99223	96
111111111	Pathology	88304	209
111111111	Pathology	88305	5786
222222222	Anesthesiology	62311	56
222222222	Anesthesiology	64483	16

TABLE III. SAMPLE OF DATASET USED FOR THIS STUDY

NPI	PROVIDER_TYPE	99222	99223	88304	88305	...	62311	64483
123456789	Internal Medicine	142	96	0	0	...	0	0
111111111	Pathology	0	0	209	5786	...	0	0
222222222	Anesthesiology	0	0	0	0	...	56	16

Fig. 3. F-Score value per Provider Type



IV. RESULTS AND DISCUSSION

Figure 2 presents the distribution of F-score values throughout the dataset indicating visible groups that classes of physicians fall into, informing the categories with which our results are organized and discussed. Figure 3 shows the model F-score values per provider type, indicating the chosen groupings by F-score. In order to find physicians that are misusing insurance, we need to determine the conditions that achieve the best possible prediction for each field of expertise. Along with the performance metrics outlined above, two other explanatory values are shown for each provider type, which are the number of instances and number of codes.

The results are organized in three groups, as they have similar needs, based on Figure 2 and Figure 3: the group that scored very high in terms of F-score (above 0.90), the group that scored mediocre but needs work (between 0.5 and 0.90) and the group that had bad results (between 0.0 and 0.50). There are 7 classes that scored very high, 18 that scored mediocre and 57 that scored unfavorable. In an attempt to not clutter the results with 82 classes, the tables below will only display a sample of the classes within each breakdown, with the exception of the high scoring partition.

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

TABLE IV. RESULTS FOR .90+ F-SCORE

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

Most of the classes within this partition have an adequate number of instances (more than 100) but there is one standout, Speech Language Pathologist, that has a very small number of instances (19). Speech Language Pathologists use 6 unique procedures codes (92506, 92507, 92526, 92610,

TABLE V. COUNT OF 6 UNIQUE PROCEDURES BY PROVIDER TYPE

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

92611, 92612), but these same procedures overlap with 9 other specialties. As seen in Table V, Speech Language Pathologists performed about 50% of the procedures, across the 6 unique procedure codes, with the other 9 providers accounting for the rest, indicating overlap in the procedures performed.

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

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

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

Table VI shows a sample of the classes (14/18) with mediocre results, as previously defined. The classes that fall into this category are of the most interest for future research, because they have the best chance of being improved through various data mining techniques and the general number of instances are high. Even with the simple multinomial Naïve Bayes learner and the inclusion of the other classes' procedures in the dataset, the model still produced a result better than 0.50. Only Portable X-ray and Occupational Therapist have a small number of instances; therefore, the results from these fields could benefit more complex learners and/or additional data. There are a few fields that have a large difference in their Recall and Precision (leading to a low F-score), but a high number of instances. These fields could use some extra attention such as Cardiology, Licensed Social Worker, Centralized Flu, and Diagnostic Radiology. These can benefit from further research using techniques that will increase classifier performance, such as feature selection or ensemble techniques, or future methods to address procedure overlap.

TABLE VI. SAMPLE OF RESULTS .50 - .90 F-SCORE

Provider Type	# of Instances	# of Codes	Recall	Precision	**F-score**
Otolaryng- ology	475	2453	0.84	0.92	**0.88**
Optha- mology	1138	227	0.75	0.90	**0.82**
Occupational Therapist	211	37	0.96	0.69	**0.80**
Optometry	1364	74	0.81	0.78	**0.80**
Dermatology	864	276	0.73	0.81	**0.77**
Rheuma- tology	267	315	0.65	0.80	**0.72**
Cardiology	1528	525	0.53	0.90	**0.67**
Licensed Clinical Social Worker	560	8	0.98	0.47	**0.63**
Centralized Flu	713	12	0.90	0.50	**0.61**
Portable X-ray	27	79	0.81	0.50	**0.60**
Obstetrics/ Gynecology	1283	285	0.45	0.73	**0.56**
Clinical Laboratory	174	773	0.42	0.86	**0.55**
Neurology	748	339	0.42	0.81	**0.55**
Diagnostic Radiology	950	499	0.36	0.86	**0.51**

Table VII shows a sample of the fields (19/57) that did not have good results under these basic conditions but could still possibly benefit from testing other learners as well as applying various other methodologies in order to improve the results [6]. The results with F-scores between 0.25 and 0.50 contain only two classes (Colorectal Surgery and Medical Oncology) with a small number of instances; thus, the results for that range are fairly dependable for each class. As with the previous group, there are a few classes that have a high precision or recall but low F-score values for classes, such as Allergy/Immunology, Hand Surgery, Hematology/Oncology and Mass Immunization Roster Biller. For these fields, classifiers have a greater chance of being improved over their counterparts with evenly low precision and recall.

In general, the classes that received an F-score below 0.25 have a low number of instances making it indiscernible whether their low results were due to the lack of representation in the dataset or difficulty with distinguishing each class. There are four standouts in this groups that have a quite large number of instances: Family Practice, Internal Medicine, Physician Assistant and Nurse Practitioner which, on their own, make up a large percent of the dataset. Since these four classes have so many instances and each have over 500 performed procedures but low F-score results, it would lead one to believe that there are many overlapping procedure codes between other classes. Meaning that the physicians in Internal Medicine, for example, perform a lot of the same procedures done by the other 81 fields. Several other fields have an adequate amount of instances within this group, such as Clinical Psychologist and General Surgery, but still exhibit overlapping procedures

amongst providers. None of the classes that scored an F-score of 0.0 are shown here as all of them have very few instances.

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

TABLE VII. SAMPLE OF < .50 AVERAGE F-SCORE

Provider Type	# of Instances	# of Codes	Recall	Precision	**F-score**
Vascular Surgery	155	180	0.56	0.44	**0.49**
Allergy/ Immunology	188	169	0.84	0.33	**0.47**
Hand Surgery	84	127	0.77	0.30	**0.43**
Mass Immunization Roster Biller	1915	15	0.26	0.85	**0.30**
Medical Oncology	96	202	0.48	0.17	**0.25**
Interventio- nal Pain Management	209	318	0.19	0.37	**0.25**
CRNA	34	21	0.53	0.18	**0.19**
Family Practice	3806	790	0.12	0.45	**0.19**
Emergency Medicine	270	416	0.19	0.18	**0.18**
Pain Management	103	225	0.30	0.13	**0.18**
Internal Medicine	4216	961	0.10	0.61	**0.17**
Clinical Psychologist	733	20	0.09	0.84	**0.16**
Physician Assistant	1479	555	0.05	0.29	**0.08**
General Surgery	845	386	0.04	0.31	**0.08**
Nurse Practitioner	2462	589	0.02	0.36	**0.04**
Sports Medicine	34	57	0.27	0.02	**0.03**
Geriatric Medicine	89	86	0.32	0.02	**0.03**
Peripheral Vascular Disease	60	60	0.20	0.00	**0.01**

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

V. CONCLUSION

The misuse of medical insurance, whether malicious or not, can lead to many undesirable outcomes such as patients not getting the funding they need or physicians not getting reimbursed for their time. This is unacceptable for the field of healthcare and there needs to be misuse and fraud-detection rules that are actionable. The purpose of this paper is to effectively use machine learning to determine whether or not using procedure data could accurately predict a physicians' field of practice. This research explores the possibility of creating a machine learning model for assessing fraudulent provider behaviors based on their medical procedure history. The process, as seen in Figure 1, provides validation that a provider is performing normally within their specialty, or indicates possibly aberrant medical practices when the model classifies that provider into another specialty for which they do not practice. The results show that there is certainly promise for this research, as results were good for several specialties, even with using a relatively simple learner and a dataset with a large number of classes.

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

The classes with lower than 0.90 F-score need additional research work. One of the noted reasons for lower F-scores, across several provider types, is the overlap in the procedures performed. Even with overlapping procedures, as discussed in Section IV, the model does not always produce low F-scores, but this overlap could make it difficult to detect fraudulent behaviors across providers with very similar procedure profiles. Additionally, classes with a low number of instances, with the majority having low F-scores, make determining their results unverifiable. In order to alleviate this situation, tests need to be performed on a dataset with more instances of these classes, which can be done using a larger subset or the full dataset released by CMS. Furthermore, for classes in the second partition (0.50-0.90), which generally have a high number of instances per provider, more sophisticated machine learning techniques, such as feature selection or clustering [13], could be used to improve classification results.

Throughout our paper, we have discussed possible avenues for future work. With regards to our fraud detection method, given the novelty of our research, there is still work yet to be done before all physicians can be accurately classified into their respective fields in a reliable manner. This research, along with other research done using publicly-available Medicare datasets [2], can lead to a more cost effective healthcare system by detecting and flagging potentially fraudulent behaviors exhibited by healthcare providers. The hope for this line of research is to develop a generalizable model that finds physicians that work outside the norm of their fields through the successful application of anomaly detection methods. One noted potential threat to this line of research is the fact that

many different types of physicians can perform the same procedures. Future models will account for this procedure overlap, as this does not represent an anomaly or fraud, merely a physician practicing everyday medicine. As no standard dataset for insurance fraud exists, there needs to be a set of rules and regulations that establish a baseline behavior for physicians in terms of insurance utilization for each specialty. We believe that by continuing the research started herein, we can help determine these baselines for providers through the detection of anomalous medical procedures. Therefore, the next step is to incorporate the discussed future work and find ways of improving upon the results shown by our model to better detect fraudulent provider behaviors.

ACKNOWLEDGMENT

Acknowledgement: The authors gratefully acknowledge partial support by the National Science Foundation, under grant number CNS-1427536. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Centers for Medicare and Medicaid Services: HCPCS General Information. [Online]. Available: https://www.cms.gov/Medicare/Coding/MedHCPCSGenInfo/index.html?redirect=/medhcpcsgeninfo/

[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

[3] Physician and Other Supplier Data CY 2013 - Centers for Medicare and 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

[4] 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 Methodolog*, pp. 1–25, 2016. [Online]. Available: http://dx.doi.org/10.1007/s10742-016-0154-8

[5] 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*, ser. KDD '13. New York, NY, USA: ACM, 2013, pp. 1312–1320. [Online]. Available: http://doi.acm.org/10.1145/2487575.2488205

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

[7] 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.

[8] T. H. J. K. F. Foundation, "State Health Facts - Medicare," 2015. [Online]. Available: http://kff.org/state-category/medicare/

[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009. [Online]. Available: http://doi.acm.org/10.1145/1656274.1656278

[10] HHS. Hhs.gov. [Online]. Available: http://www.hhs.gov/

[11] K. M. King, "Medicare fraud: Progress made, but more action needed to address Medicare fraud, waste, and abuse," 2014. [Online]. Available: http://www.gao.gov/products/GAO-14-560T

[12] 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.

[13] Q. Liu and M. Vasarhelyi, "Healthcare fraud detection: A survey and a clustering model incorporating geo-location information," in *29th world continuous auditing and reporting symposium*, 2013.

[14] L. Morris, "Combating Fraud In Health Care: An Essential Component Of Any Cost Containment Strategy," 2009. [Online]. Available: http://content.healthaffairs.org/content/28/5/1351.full

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

[16] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.