Identifying Healthcare Fraud with Open Data

Submission Type: Completed Research Full Papers

Abstract

Health care fraud is a serious problem that impacts every patient and consumer. This fraudulent behavior causes excessive financial losses every year and causes significant patient harm. Healthcare fraud includes health insurance fraud, fraudulent billing of insurers for services not provided, and exaggeration of medical services, etc. To identify healthcare fraud thus becomes an urgent task to avoid the abuse and waste of public funds. Existing methods in this research field usually use classified data from governments, which greatly compromises the generalizability and scope of application. This paper introduces a methodology to use publically available data sources to identify potentially fraudulent behavior among physicians. The research involved data pairing of multiple datasets, selection of useful features, comparisons of classification models, and analysis of useful predictors. Our performance evaluation results clearly demonstrate the efficacy of the proposed method.

Keywords

Healthcare, Fraud Prediction, Machine Learning, Imbalanced Data, Entity Matching.

Introduction

Healthcare frauds include health insurance fraud, fraudulent billing of insurers for services not provided, and exaggeration of medical services (Rudman et al. 2008). Healthcare fraud is a serious problem that impacts every patient and consumer. This fraudulent behavior causes excessive financial losses in the magnitude of billions of dollars of losses every year and significant patient harm. The U.S. national health expenditure as percent of GDP has increased from 5% to 18.3% between 1960 and 2017[1]. With such an intensive demand for the healthcare services, healthcare frauds have become a mainstream issue that impacts every patient and consumer. About 10% of the healthcare expenditure is produced by frauds, which represents more than 100 billion dollars per year, according to the General Accounting office in the United States (Thompson 1992).

The requirement for effective and efficient approaches for fraud identification is urgent, considering the serious consequence of healthcare frauds and increasing demands for high quality healthcare. Nevertheless, manually reviewing materials by human experts, which is extremely labor-intensive and time-consuming, is still the major approach for healthcare fraud detection in many places (Yang and Hwang 2006). Another problem is that nonpublic and highly domain-specific data is used in current approaches, which greatly hindered their generality and extensibility in the real world application (He et al. 1997; Sokol et al. 2001; Yang and Hwang 2006). Also, most existing methods implemented healthcare fraud identification at the claim level (Sokol et al. 2001; Thornton et al. 2013; Yang and Hwang 2006), but few work has investigated how to find the fraudulent physicians using the aggregated comprehensive records (e.g. prescription, payment, patient reviews, etc.). We believe detecting the physicians who commit fraud could be more effective when we can leverage information cues from different open sources.

To fill the research gaps above, we are motivated to develop a methodology that uses open datasets to predict healthcare fraud at the physician level and reduce the workload of human experts. In particular, a list of Excluded Individuals and Entities (LEIE) and board actions were used as labels for fraud cases. Different publically available predictor datasets, such as Part D Prescriber, Open Payment, and Social Media datasets, were consolidated and used for building a predictive model to identify potentially fraudulent behavior among physicians. The research involved data pairing and entity matching of multiple datasets, selection of useful features for modeling, imbalanced data analysis, classification model comparisons, and analysis of useful predictors. Experimental results showed that features from the Part D Prescriber dataset

[1] https://www.statista.com/statistics/184968/us-health-expenditure-as-percent-of-gdp-since-1960/

produced the best F1 score of 75.59% when doing prediction with the Prescriber dataset. Furthermore, the F1 score increases to 96.1% if we combine sources of social media and Prescriber datasets.

Our contributions to literature are 3 folds: 1) To the best of our knowledge, this is among the earliest studies that investigate the effectiveness of features from multiple publicly available datasets for healthcare fraud detection. Because these datasets are all publically available, our approach features a high level of generalizability; 2) Based on data aggregation and pre-processing, we proposed several feature extraction strategies to extract novel features from the original open datasets, which are proved to be effective in predicting healthcare fraud; 3) We conducted extensive evaluations of our fraud detection model after consolidating several publicly available datasets. The performance results clearly demonstrate the efficacy of the proposed method. Our model and results also provide great insights to healthcare regulators for better regulations.

The rest of the paper is organized as follows. The "Related Work" section reviews related work in healthcare fraud detection and highlight the research gap. The "Approach" section describes our proposed research framework. The "Datasets" section introduces the open datasets we have investigated. The "Experiment Results" section demonstrates the experimental details and related discussions. The "Conclusion" section summarizes the paper and discusses the limitations of this study and directions for future work.

Related Work

Recognizing the significance of detecting healthcare fraud and problems of manually reviewing materials by human experts, researchers had conducted extensive studies in automatic and effective techniques for detecting healthcare fraud. These existing researches targeted on different kinds of fraud, collected data from various sources, and proposed diverse features and models to capture fraudulent cases.

Three groups of people may commit fraudulent behaviors, according to Yang and Hwang (2006). The first party is service providers, such as physicians, hospitals, ambulance companies, and laboratories; another group is insurance subscribers, including patients and patients' employers; the third group is insurance carriers, who receive regular premiums from their subscribers and pay health care costs on behalf of their subscribers, such as government departments on healthcare and private insurance companies. This research focuses on the first group of people, service providers.

Several relevant studies on healthcare fraud prediction were reported in the literature. Yang and Hwang propose a data-mining framework which utilizes the concept of clinical pathways to develop a healthcare fraud detection model (Yang and Hwang 2006). The proposed approaches have been evaluated objectively by a real-world data set gathered from the National Health Insurance (NHI) program in Taiwan. Liou et al. utilize data mining techniques to detect fraudulent or abusive reporting by healthcare providers using their invoices for outpatient services, which was also carried out based on the NHI data (Liou et al. 2008). Recently, Thornton et al. builds upon the Medicaid environment and to develop a Medicaid multidimensional schema and provide a set of multidimensional data models to predict fraudulent activities (Thornton et al. 2013).

The datasets used for fraud identification were mainly collected from insurance carriers (Li et al. 2008). The US Health Care Financing Administration (HCFA) (Shapiro 2002), the Bureau of National Health Insurance (NHI) in Taiwan (Chan and Lan 2001; Hwang et al. 2004; Wei et al. 2000), and the Health Insurance Commission (HIC) in Australia (He et al. 1997, 2000) are the major governmental data sources for existing healthcare fraud research studies.

Although a great deal of efforts have been put to develop healthcare fraud detection models, and some progresses have been achieved, there are a few limitations. The most important one is that most of these datasets are not publicly available and/or they are highly domain-specific and require extensive background knowledge to conduct feature engineering. Models developed using these proprietary data sets have limited generalizability and are hard to replicate in reality. Almost no research studies explore the usefulness of publicly available datasets, how to extract useful features from these open data sets, and lastly how to combine multiple datasets to improve performance.

Methodology

We follow a typical data science approach to tackle the healthcare fraud detection problem at physician level. In this research, the fraud detection is solved as a binary classification problem: a physician will be classified as Fraud or Non-Fraud based on the features extracted for this physician. The overall research framework is shown in Figure 1.

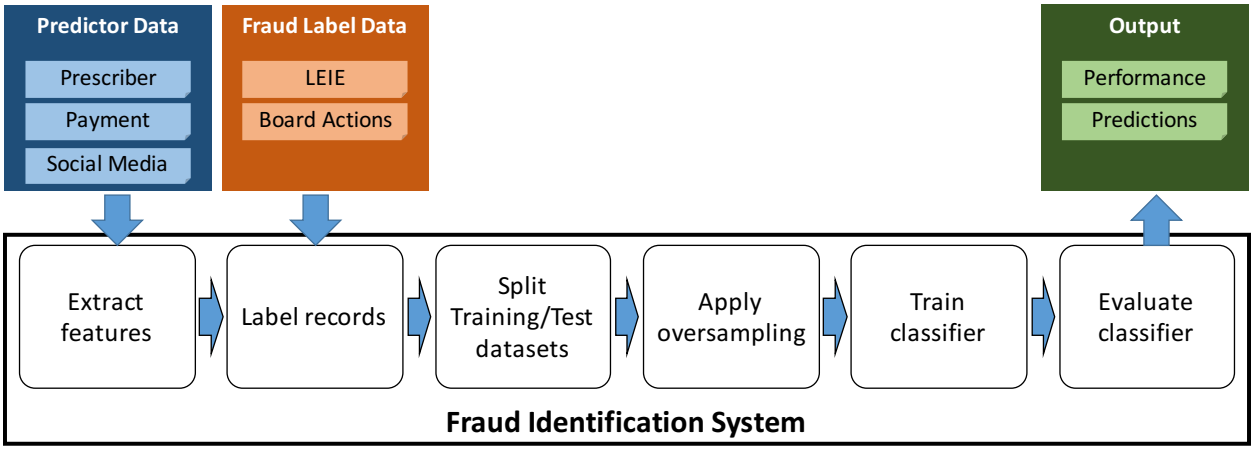

Figure 1 Fraud Detection Framework using Open Datasets

The detailed steps are explained below.

1) First, various features were extracted from multiple predictor datasets. Some features were obtained through special calculation (e.g. deviation features) or data aggregation (e.g. average or sum). Logistic regression was conducted to select the most relevant features for further analysis. In addition, combinations of features across different datasets were performed to implement comprehensive fraud prediction.
2) Then, data pairing and entity matching was performed to match the fraud labels extracted from LEIE (2015-2016) and board actions datasets with data records in the predictor datasets (e.g. Part D Prescriber and open payment datasets).
3) Subsequently the entire dataset was split into training and test with a ratio of 80:20. The splitting followed a stratified shuffle process, which will keep the original class proportions in both training and test datasets.
4) Since the data was extremely imbalanced (e.g. only 0.045% physician records of the LEIE data were fraudulent), SMOTE oversampling (Chawla et al. 2002) is applied to both datasets before training a classifier to prevent the classifier from predicting all physicians in the test set as the major class (Non-Fraud).
5) Next, classifiers were trained using different classification algorithms, such as Logistic Regression, Naive Bayes, Decision Tree, and SVM.
6) Finally, the classification performance was evaluated on the held-out test dataset, which is a balanced dataset after oversampling. The Weighted F1 was used as a comprehensive measure of the performance.

Datasets

Fraud Label Datasets

Two kinds of datasets were used as fraud labels for fraud prediction in this research design: the LEIE dataset and the Board Action datasets.

LEIE Dataset

The Office of Inspector General (OIG) of the U.S. has the authority to exclude individuals and entities from federally funded health care programs pursuant to sections 1128 and 1156 of the Social Security Act and maintains *a list of all currently excluded individuals and entities called the List of Excluded Individuals and Entities (LEIE)*[2]. Anyone who hires an individual or entity on the LEIE may be subject to civil monetary penalties (CMP).

Board Action Datasets

Medical Board sanctions are used as proxy for potential fraudulent billing. In addition, due to the variance in state regulations, different boards across states may be more or less lenient in their sanctions. Medical Boards are established in many states to properly regulate the practice of medicine and surgery. Every year, those boards take administrative actions to address possible cases of professional misconduct, license term violations, improper prescriptions, etc., and make this information available to the public. The basic principle to choose states are their population. According to Wikipedia, the top 5 US states with the largest population are CA, TX, FL, NY, and PA. However, it's difficult to extract board action records of Texas and Pennsylvania automatically, and New York has surprisingly low matches with the payment dataset. Therefore, the board action records of California, Florida, and North Carolina were selected for this research.

Predictor Datasets

Part D Prescriber Dataset

The Part D Prescriber Public Use File [3](PUF) provides information on prescription drugs prescribed by individual physicians and other health care providers and paid for under the Medicare Part D Prescription Drug Program. The Part D Prescriber PUF is based on information from the Chronic Conditions Data Warehouse of the Centers for Medicare & Medicaid Services (CMS), which contains Prescription Drug Event records submitted by Medicare Advantage Prescription Drug (MAPD) plans and by stand-alone Prescription Drug Plans (PDP). The dataset identifies providers by their National Provider Identifier (NPI) and the specific prescriptions that were dispensed at their direction, listed by brand name (if applicable) and generic name. For each prescriber and drug, the dataset includes the total number of prescriptions that were dispensed and the total drug cost. The total drug cost includes the ingredient cost of the medication, dispensing fees, sales tax, and any applicable administration fees and is based on the amount paid by the Part D plan, Medicare beneficiary, government subsidies, and any other third-party payers.

The advantage of these data is the fact physicians are mandated to report their Part D prescription activities to the CMS, since they have to submit a claim in order to be paid. Therefore, the Prescriber dataset is less biased in contrast to the CMS payment dataset, whose payment records are submitted voluntarily.

CMS Open Payment Dataset

Open Payments[4], which is managed by the CMS, is a national disclosure program created by the Affordable Care Act (ACA). The program promotes transparency and accountability by helping consumers understand the financial relationships between pharmaceutical and medical device industries, and physicians and teaching hospitals. These financial relationships may include consulting fees, research grants, travel reimbursements, and payments made from the industry to medical practitioners. It is important to note that financial ties between the health care industry and health care providers do not necessarily indicate an improper relationship. Applicable manufacturers and applicable GPOs must enter detailed information about payments, other transfers of value, or investment interests into CMS's Open Payments system.

[2] https://oig.hhs.gov/exclusions/exclusions_list.asp

[3] https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Part-D-Prescriber.html

[4] https://www.cms.gov/openpayments/

Among the three types of payments (i.e. General Payments, Research Payments, and Physician Ownership or Investment Interest Information), we used the "General Payments" in this research, which saves the most common payment records.

One concern about this dataset is that the data is self-reported. While there is a great care taken to ensure that the reported payments are correct, there are no checks and balances in place to ensure that ALL payments are reported and database is complete. And Table 4 shows that the fraud prediction accuracy using payment features is lower than using prescription features.

Social Media Dataset

The Healthgrades.com website contains rich information about physicians, hospitals and health care providers. It has amassed information on over 3 million U.S. health care providers, with more than 9 million ratings and reviews over 18-year period of time. Healthgrades has built the first comprehensive physician rating and comparison database.

We developed automated crawlers to download the ratings and reviews for all doctors in California, Florida, and North Carolina. The key fields include overall rating, number of ratings, detailed ratings ("Trustworthiness", "Explains condition well", "Answer questions", "Time well spent", "Scheduling", "Office environment", and "Staff friendliness"), text reviews and corresponding ratings, etc.

Experiment Results

Part D Prescriber Dataset

Since one physician may have multiple drug prescription records in this dataset, we need to aggregate them and create a single record for each physician, which will be used for classification. In this way, 837,679 physician records were extracted from the Prescriber data for fraudulent behavior prediction. Among this large number of physicians, only 383 (0.045%) matched the LEIE fraud records.

We tried two methods for data aggregation and feature creation:

- Take the factors (e.g. TAL_CLAIM_COUNT, TOTAL_DAY_SUPPLY, etc.) related to a drug as features of a physician. If there are M types of drugs and N factors for each drug, a physician will have N * M features. This feature was created to identify which drug prescription is most highly correlated with fraud. A potential problem of this method is, the features of a physician might be very sparse, as one physician only have prescription records on a small number of drugs.
- For the K Prescription records of each physician, mean values were taken on key factors (e.g. TOTAL_CLAIM_COUNT, TOTAL_DAY_SUPPLY, etc.). Next, these mean values were added as features of a physician.

For the first data aggregation method, 8 types of features were tried for each drug and the corresponding fraud prediction performance are shown in Table 1. As 873 drugs are related with the prescription records of 383 fraud physicians, which will produce too many features, we used the Chi-Square feature selection algorithm to pick out the top 100 relevant drugs. Together, they will form 8 * 100 = 800 features. The best Weighted F1 was produced by the Naïve Bayes classifier.

Among these 8 types of features, half of them are "Deviation" features, which are calculated as below.

- The average = "TOTAL_CLAIM_COUNT", "TOTAL_DAY_SUPPLY", "TOTAL_DRUG_COST", and "Average_Day_Supply_Per_Claim" of each specialty-drug pair
- For each of the above features, the difference between each physician's value and the specialty-drug average was measured and difference or "Deviation" was noted.

The Weighted F1 was calculated as below. Here C is the number of classes, while W_i is the number of true instances of class i.

$$\mathrm{F1} = 2\frac{precision \cdot recall}{precision + recall} \tag{1}$$

$$\text{Weighted F1} = \frac{\sum_{i=1}^{C} W_i \cdot F1_i}{C} \quad (2)$$

Features(800)	Classifier	Weighted F1
8 Categories of features: TOTAL_CLAIM_COUNT TOTAL_DAY_SUPPLY TOTAL_DRUG_COST Average_Day_Supply_Per_Claim TOTAL_CLAIM_COUNT_DEVIATION TOTAL_DAY_SUPPLY_DEVIATION TOTAL_DRUG_COST_DEVIATION Average_Day_Supply_Per_Claim_Deviation	Logistic Regression	59.04%
	Naïve Bayes	67.69%
	SVM	50.33%

Table 1 Classification Performance Using 800 features of Top 100 Relevant Drugs Extracted for Physicians

For the second data aggregation method, we calculated the mean value of key factors (e.g. TOTAL_CLAIM_COUNT, TOTAL_DAY_SUPPLY, etc.) related to each physician. This research proves the Specialty (as a dummy variable) is an important feature for each physician. Ten features for each physician were used, which produced the best fraud prediction performance on the Part D Prescriber data. The performance of fraud prediction with these features are shown in Table 2. Again, the Naïve Bayes classifier obtained the best weighted F1 of 75.59%.

The "Deviation" features in Table 2 indicates the difference between a physician's value and the "Specialty Average", which is a slightly different from the "Deviation" features in the previous table. Those deviations mean the difference between a physician's value and the "Specialty-Drug Average".

The calculation of the "Unusual Drug Prescription" feature was accomplished as follows.

- Find the "Unusual Drug Prescription" patterns, by identifying the top 5% rare specialty-drug prescription events.
- For each physician, count how many prescription records match those "Unusual drug prescription" patterns.

Features (10)	Classifier	Weighted F1
TOTAL_CLAIM_COUNT TOTAL_DAY_SUPPLY TOTAL_DRUG_COST Average_Day_Supply_Per_Claim TOTAL_CLAIM_COUNT_DEVIATION TOTAL_DAY_SUPPLY_DEVIATION TOTAL_DRUG_COST_DEVIATION Average_Day_Supply_Per_Claim_Deviation Specialty (dummy variable) Unusual Drug Prescription	Logistic Regression	69.08%
	Naïve Bayes	75.59%
	SVM	44.77%

Table 2 Classification Performance Using 10 Features Extracted for Physicians

Most useful features: We examined the co-efficient of the 10 features introduced in Table 2 using the logistic regression analysis. To make this analysis fairer, all the numerical features are normalized before running the logistic regression. Since the "Specialty" was taken as a dummy variable, it produced 191 features during the classification process. In Table 3, the top 10 most relevant specialties were shown with the most relevant other features. High coefficients in the 10 specialties indicate physicians in these specialties are more likely to commit fraud. For example, if a physician practices specialties such as "Personal Emergency Response Attendant", "Osteopathic Manipulative Medicine", and "Neurological Surgery", he or she has a higher fraud probability. It's not surprising to see "Legal Medicine" here. An unexpected case is "Family Medicine". Physicians in this specialty has a positive association with fraud risks. Besides those specialties, other features such as "Total Claim Count", "Total Claim Count Deviation", "Unusual Drug Prescription", "Average Day Supply per Claim", "Unusual Drug Prescription", and "Average Day Supply per Claim" also have high coefficients. For instance, physicians with high "Total Claim Count" has a higher fraud probability. In addition, a physician may have a high fraud risk if he or she made a high "Average_Day_Supply_Per_Claim" or an "Unusual Drug Prescription".

Physician Specialties	Coefficient	Other Features	Coefficient
Personal Emergency Response Attendant	10.7194	TOTAL_CLAIM_COUNT	49.4308
Legal Medicine	6.0965	TOTAL_CLAIM_COUNT _DEVIATION	48.9248
Osteopathic Manipulative Medicine	5.9022	Average_Day_Supply_Per_Claim	1.8668
Neurological Surgery	5.6206	UnusualDrugPrescription	1.5793
Family Medicine	5.4934	Average_Day_Supply_Per_Claim_Deviation	0.2407
Psychiatry	5.0648		
Specialist	4.8325		
Family Practice	4.7811		
Internal Medicine	4.7772		
Psychiatry & Neurology	4.7003		

Table 3 The Most Relevant Features in the Part D Prescriber Dataset

CMS Open Payment Dataset

The same fraud prediction process used on the Part D Prescriber dataset was applied to the CMS payment datasets. Both LEIE and Board Action (records of CA, NC, and FL) were tried as fraud labels in this research.

1) Take LEIE records as fraud labels
   In this experiment, 233 matches were made with physicians in the LEIE data using "First Name", "Last Name" and "State". If stricter matching conditions were applied, such as "First Name", "Last Name", "State", and "City", 28 matched for physicians were attained. Table 4 shows that using more matched cases produces much better prediction performance (Weighted F1 increases from 53.70% to 71.42%) for Naive Bayes classifier.
2) Take Board Action records as fraud labels
   In contrast to the LEIE dataset, the Board Action records identified a larger number of matched physicians. As shown in Table 5, 55, 235, and 153 disciplined providers were found in the 2016 board action records of NC, CA, and FL, respectively. The payment datasets of 2013-2015 were used as independent variables. "First Name", "Last Name", "State", and "City" were used as matching

condition. Just like the LEIE case, the prediction performance increases along with the number of disciplined providers. California had the best prediction performance.

3) Most useful features
Among the features extracted from the Open Payment datasets, the most relevant features are "Unusual Drug Prescription" and "Payment Amount", and "Payment Count" (shown in Table 6), when taking LEIE and Board Action Records as labels, respectively.

Features	Matching Condition	Fraud cases	Classifier	Weighted F1
Specialty (dummy variable) Payment Count Payment Amount Unusual drug prescriptions Unusual device prescriptions	FN+LN+State	233	Logistic Regression	59.01%
			Naïve Bayes	71.42%
	FN+LN+State+ City	28	Logistic Regression	60.13%
			Naïve Bayes	53.70%

Table 4 Classification Performance Using 5 Features from Payment Data and LEIE Labels

Features	Fraud Dataset	Fraud cases	Classifier	Weighted F1
Primary Type (dummy variable) Specialty (dummy variable) Payment Count Payment Amount Unusual drug prescriptions Unusual device prescriptions	NC Board Actions	55	Logistic Regression	57.29%
			Naïve Bayes	54.96%
	CA Board Actions	235	Logistic Regression	70.31%
			Naïve Bayes	65.81%
	FL Board Actions	153	Logistic Regression	56.55%
			Naïve Bayes	56.41%

Table 5 Classification Performance Using 6 Features from Payment Data and Board Action Labels

Labels	Features	Co-Efficient
LEIE Records	Unusual Drug Prescription	7.6601
	Payment Amount	5.0813
Board Action Records	Payment Count	8.7040

Table 6 The Most Relevant Features in the Open Payment Datasets

Social Media Dataset

Following the similar procedures introduced earlier, we combined all the cases from LEIE and board actions as healthcare frauds. After conducting data matching based on first name, last name, city and state, only 555 (1.86%) cases out of 29,843 are found fraudulent.

The table below shows that the classification performance was not satisfactory. The best performance was obtained by using decision tree classifier, which give F1 score of 0.646, indicating the review data can be used to predict healthcare frauds, but this single dataset is not enough for accurate predicting.

Features (11)	Classifier	Weighted F1
Average review rating Rating count Average rating Trustworthiness Explains condition well Answer questions Time well spent Scheduling Office environment Staff friendliness State	Decision tree	64.6%
	Logistic regression	46.6%

Table 7 Classification Performance Using 11 Features from Social Media

Most useful features: From the results in Table 8, five features ("Rating count", "Average rating", "Trustworthiness", "Explains condition well", "Answer question") were selected based on the *p*-value in the logistic regression results at significant level of 0.05 for the comprehensive analysis below.

Feature	Feature coefficients	Feature	Feature coefficients
Average review rating	0.0057	Time well spent	0.0726
Rating count	0.1997	Scheduling	-0.1538
Average rating	-0.3706	Office environment	-0.1613
Trustworthiness	-0.1053	Staff friendliness	-0.0651
Explains condition well	0.4984	State	-0.1276
Answer questions	-0.2913		

Table 8 The Most Relevant Features in the Social Media Dataset

Comprehensive Datasets

Lastly, all three predictor datasets are merged. Only 265 (1.43%) cases out of 22,770 with complete fields in all three datasets are found fraudulent. Oversampling is applied to both datasets before training classifier to prevent the classifier from predicting all physicians in the test set as the major class (Non-Fraud). The classification performance with high weighted F1 using features from the merged dataset is shown in Table 9. The constraint of this method is, it only works on a very small number of instances with complete fields.

Features (10)	Classifier	Weighted F1
Average review rating Rating count Average rating Trustworthiness Explains condition well Payment count	Decision tree	96.1%
	Logistic regression	91.5%

Total payment amount Average claim amount Claim count State		

Table 9 Classification Performance Using 8 Features from Social Media, Open-payment Datasets and Prescriber Datasets

Conclusion

This paper introduces a methodology to use publically available data sources (e.g. Prescriber, Payment, and Social media) to identify potentially fraudulent behavior among physicians. Fraud and other misconduct records in LEIE and Board action datasets are used as fraud cases. The research involved data pairing and entity matching of multiple datasets, selection of useful features, comparisons of classification models, and analysis of useful predictors. Our performance evaluation results clearly demonstrate the efficacy of the proposed method. The best weighted F1 score of 96.5% is achieved using the merged datasets.

However, to make accurate prediction on the complete dataset is still very challenging, due to the extreme data imbalance and sparsity issues. More data collection is needed from other states to make the predictive model more robust and general across states for fraud examination. We will leave this for future research.

REFERENCES

Chan, C., and Lan, C. 2001. "A Data Mining Technique Combining Fuzzy Sets Theory and Bayesian Classifier—an Application of Auditing the Health Insurance Fee," *Proceedings of the International Conference on Artificial*.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. 2002. "SMOTE: Synthetic Minority over-Sampling Technique," *Journal of Artificial Intelligence Research* (16), pp. 321–357.

He, H., Hawkins, S., Graco, W., and Yao, X. 2000. "Application of Genetic Algorithm and K-Nearest Neighbour Method in Real World Medical Fraud Detection Problem.," *JACIII*.

He, H., Wang, J., Graco, W., and Hawkins, S. 1997. "Application of Neural Networks to Detection of Medical Fraud," *Pergamon Expert Systems With Applications* (13:4), pp. 329–336.

Hwang, S. Y., Wei, C. P., and Yang, W. S. 2004. "Discovery of Temporal Patterns from Process Instances," *Computers in Industry* (53:3), pp. 345–364.

Li, J., Huang, K.-Y., Jin, J., and Shi, J. 2008. "A Survey on Statistical Methods for Health Care Fraud Detection," *Health Care Management Science* (11:3), pp. 275–287.

Liou, F. M., Tang, Y. C., and Chen, J. Y. 2008. "Detecting Hospital Fraud and Claim Abuse through Diabetic Outpatient Services," *Health Care Management Science* (11:4), pp. 353–358..

Rudman, W. J., Iii, J. S. E., Pierce, W., and Hart-hester, S. 2008. "Healthcare Fraud and Abuse," *Perspectives in Health Information Management*, pp. 1–24.

Shapiro, A. 2002. "The Merging of Neural Networks, Fuzzy Logic, and Genetic Algorithms," *Insurance: Mathematics and Economics* (31), pp. 115–131..

Sokol, L., Garcia, B., Rodriguez, J., West, M., and Johnson, K. 2001. "Using Data Mining to Find Fraud in HCFA Health Care Claims.," *Topics in Health Information Management* (22:1), pp. 1–13.

Thompson, L. H. 1992. "Health Insurance, Vulnerable Payers Lose Billions to Fraud and Abuse," *Report to Chairman, Subcommittee on Human Resources and Intergovernmental Operations. United States General Accounting Office, Washington, DC (May)*.

Thornton, D., Mueller, R. M., Schoutsen, P., and van Hillegersberg, J. 2013. "Predicting Healthcare Fraud in Medicaid: A Multidimensional Data Model and Analysis Techniques for Fraud Detection," *Procedia Technology* (9), Elsevier B.V., pp. 1252–1264.

Wei, C., Hwang, S., and Yang, W. 2000. "Mining Frequent Temporal Patterns in Process Databases," *Proceedings of International*.

Yang, W. S., and Hwang, S. Y. 2006. "A Process-Mining Framework for the Detection of Healthcare Fraud and Abuse," *Expert Systems with Applications* (31:1), pp. 56–58..