Identifying Physician Fraud in Healthcare with Open Data

Brandon Fan[1][0000−0002−3235−0633], Xuan Zhang[2][0000−0003−2563−0597], and Weiguo Fan[3][0000−0003−1272−5538]

[1] Blacksburg High School, Blacksburg VA 24060, USA
[2] Virginia Tech, Blacksburg VA 24060, USA
[3] Tippie College of Business, University of Iowa, Iowa City, IA 52242, USA
[4] Tianjin University, China

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 publicly 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

1 Introduction

Healthcare fraud encompasses multiple fraudulent activities including health insurance fraud, fraudulent billing of insurers for services not provided, and exaggeration of medical services [8]. 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, in percent GDP, has increased from 5% to 18.3% between 1960 and 2017 [5]. With such an intensive demand for healthcare services, healthcare fraud has become a mainstream issue. About 10% of the U.S. healthcare expenditure is produced by fraud, which represents more than 100 billion dollars per year, according to the General Accounting office in the United States [11].The requirement for effective and efficient approaches for fraud identification is necessary considering the serious consequence of healthcare frauds and increasing demand for high quality healthcare. Current methods rely on the manual review of materials by human experts that is extremely labor-intensive and time-consuming, but is still the major approach for healthcare fraud detection in many places [14]. Another problem is that nonpublic and highly domain-specific data is used in current approaches, which greatly hinders generalizability and extensibility in real world applications [4, 10, 14]. Additionally, most preceding methods implement healthcare fraud identification at the claim level [10, 12, 14], but little work has investigated detection of fraudulent physicians utilizing the aggregated comprehensive records (e.g. prescription, payment, patient reviews, etc.). We believe detecting physician fraud could be more effective when we can leverage information cues from different open sources.

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

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 produced the best F1 score of 75.59% when doing prediction with the Prescriber dataset. The F1 score increases to 96.1% if we use physician instances occurring in both social media and Prescriber datasets. 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, and the Conclusion section summarizes the paper, discusses the limitations and future work.

2 Related Work

Due to the significance of detecting healthcare fraud and the problems of manually reviewing materials by human experts, researchers have conducted extensive studies in automatic and effective techniques for detecting healthcare fraud. This existing research focuses on multiple types of frauds, collects data from various sources, and proposes diverse features and models to capture fraudulent cases.

When it comes to fraudulent behaviors, there are three primary groups of people according to Yang and Hwang [14]. The first party consists of service providers, such as physicians, hospitals, ambulance companies, and laboratories. The second party consists of insurance subscribers, including patients and patients employers. The final party consists of 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: the service providers.

Several relevant studies on healthcare fraud prediction have been conducted. Yang and Hwang propose a data-mining framework which utilizes the concept of clinical pathways to develop a healthcare fraud detection model [14]. The proposed approach has 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 invoices for outpatient services. This research was also carried out based on the NHI data [7]. Recently, Thornton et al. built upon the Medicaid environment and developed a Medicaid multidimensional schema that provides a set of multidimensional data models to predict fraudulent activities [12].

The datasets used for fraud identification were collected from insurance carriers [6]. The major government data sources for existing healthcare fraud include: the US Health Care Financing Administration (HCFA) [9], the Bureau of National Health Insurance (NHI) in Taiwan area [1, 5, 13], and the Health Insurance Commission (HIC) in Australia [3, 4].

Although a great deal of effort has been put into developing healthcare fraud detection models, and some progress has been achieved, there are a few limitations. The first and most important

one is that most of these datasets are not publicly available and/or 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 study explores 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.

3 Datasets

3.1 Fraud Label Datasets

Two 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)* [6]. Anyone who hires an individual or entity on the LEIE may be subject to civil monetary penalties (CMP). The physician records present in the LEIE dataset was then combined with the subsequent board action dataset to create a conglomerate fraud label dataset.

Board Action Datasets Medical Boards are established in many states to properly regulate the practice of medicine and surgery. Every year, these 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. As its difficult to collect the board action records of all the 50 states, we chose states with large populations. According to Wikipedia, the top 5 US states with the largest population are CA, TX, FL, NY, and PA. However, its difficult to extract board action records of Texas and Pennsylvania from electronic files, and New York has surprisingly low matches with the payment feature dataset. Therefore, the board action records of California, Florida, and North Carolina were selected for this research.

Thus, our label dataset is a combination of both the LEIE dataset and the Board Action Dataset. These are then matched with predictor dataset records (discussed in the subsequent section) in order to gather features on physicians with labeled fraudulent activities as well as physicians that are considered unfradulent. The predictor dataset features provide us the features for a vector P_x where we pass through a function $f(x)$ that produces a fraudulent label of $P_y \in 0, 1$. 0 being non-fradulent, and 1 being fraudulent.

3.2 Predictor Datasets

Part D Prescriber Dataset The Part D Prescriber Public Use File (PUF) [7] 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

[6] https://oig.hhs.gov/exclusions/exclusions_list.asp
[7] https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Part-D-Prescriber.html

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 [8], 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 enter detailed information about payments, other transfers of value, or investment interests into CMSs Open Payments system. 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 in place to ensure that ALL payments are reported and database is complete. In addition, 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.

4 Methodology

4.1 Feature Extraction from Open Datasets

Using the open datasets discussed in the previous section, we identify and extract primary features from each dataset to utilize as features for the fraud detection framework. These features are determined based on domain knowledge as well as consultation with insurance companies and are further conglomerated into one comprehensive model for physician fraud detection. Each feature, its associated definition, and dataset is shown in Table 1.

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

Dataset	Feature	Description
Part D Prescriber Dataset	**TOTAL_CLAIM_COUNT**	Number of Medicare Part D Claims, Including Refills
	TOTAL_DAY_SUPPLY	Number of Day's Supply for All Claims.
	TOTAL_DRUG_COST	Aggregate cost paid for all terms
	Average_Day_Supply_Per_Claim	Average day supply per claim of physician.
	TOTAL_CLAIM_COUNT_DEVIATION	Total sum of deviation from average claim count
	TOTAL_DAY_SUPPLY_DEVIATION	Total sum of deviation from day supply
	TOTAL_DRUG_COST_DEVIATION	Total sum of deviation from drug cost
	Average_Day_Supply_Per_Claim_Deviation	Total sum of deviation from average day supply per claim
	Specialty (dummy variable)	Physician's Expertise
	Average_Claim_Count	TOTAL_CLAIM_COUNT / Records per Physician
CMS Open Payment	Unusual Drug Prescription	Presence of unusual drug prescription
	Payment Count	Total payment count over all records
	Payment Amount	Total payment amount over all records
	Primary Type (dummy variable)	Primary type of payment
	Unusual Device Prescriptions	Presence of unusual device prescription
Social Media Dataset	**Average Review Rating**	Average user review rating from 1 to 5
	Rating Count	Total number of user reviews
	Trustworthiness	A rating from 1 to 5 of physician's trustworthiness
	Explains Condition Well	A rating from 1 to 5 of physician's clarity
	Answer Questions	Properly answers questions of patients
	Time Well Spent	A rating from 1 to 5 of physician's appointments
	Scheduling	A rating from 1 to 5 of physician's scheduling habits
	Office Environment	A rating from 1 to 5 of how physician's office environment
	Staff Friendliness	A rating from 1 to 5 of how friendly physician is to staff
	State	State of Physician's Practice

Table 1: Selected Features From The Three Predictor Datasets. Bolded Features Were Used in a Comprehensive Model.

4.2 Physician Fraud Detection Framework

Using the identified features, we then proceed to create a physician fraud detection framework. Our proposed framework for physician fraud detection using open data can be summarized in Figure 1. The detailed steps are explained below.

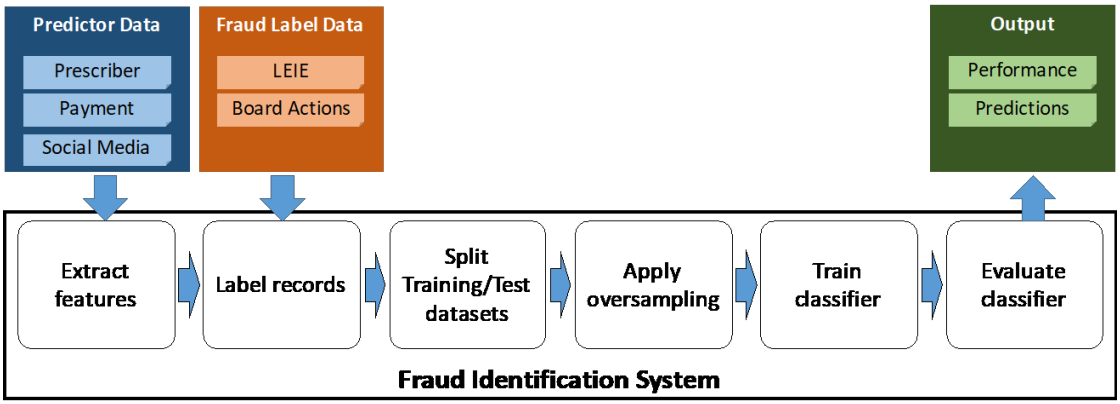

Fig. 1: Fraud Detection Framework using Open Datasets

Step 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). Then, logistic regression was conducted to identify the most relevant features for further analysis. In addition, combinations of features from different datasets were performed to do comprehensive fraud prediction.

Step 2 Data pairing and entity matching were 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). The entire dataset was split into training and test with a ratio of 80:20. The splitting followed a stratified shuffle process, keeping the original class proportions in both training and test datasets.

Step 3 Since the data was extremely imbalanced (e.g. only 0.045% physician records of the LEIE data were fraudulent), SMOTE oversampling [2] was 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).

Step 4 Next, classifiers were trained using different classification algorithms including Logistic Regression, Naive Bayes, Decision Tree, and SVM.

Step 5 Finally, the classification performance was evaluated on the held-out test dataset, which was balanced dataset after oversampling. The Weighted F1 was used as a comprehensive measure of performance.

5 Experiment Results

5.1 Part D Prescriber Dataset

Since one physician may have multiple drug prescription records in this dataset, we need to aggregate the records and create a single record for each physician, which will be used for clas-

sification. In this way, 837,679 physician records were extracted from the Prescriber data for fraudulent behavior prediction.

Because board action data is state-dependent, we were unable to utilize the information to correspond with the prescriber data. In addition, we wished to deliberately test the capability of the prescriber predictor dataset that does not include a state feature, thus reducing a confounding variable. Finally, the LEIE provides the most relevant and reputable source of information that can be easily matched to for 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 Nave Bayes classifier.

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

- Calculate the average value of the 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 physicians value and the specialty-drug average was measured and difference or Deviation was noted.

The performance measure 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.

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

$$Weighted\,F1 = \frac{\sum_{i=1}^{C} W_i \cdot F1_i}{C} \tag{2}$$

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 extracted, 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 Nave Bayes classifier obtained the best weighted F1 of 75.59%. The Deviation features in Table 2 indicates the difference between a physicians 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 physicians value and the Specialty-Drug Average.

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

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 2: Classification Performance Using 800 features of Top 100 Relevant Drugs Extracted for Physicians

- 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 3: Classification Performance Using 10 Features Extracted for Physicians

Most useful features: We examined the co-efficient of the 10 features introduced in Table 3 using a logistic regression analysis. To make this analysis fairer, all the numerical features were normalized before running the logistic regression. Since the Specialty was taken as a dummy variable, it produced 191 features during the classification process. Running correlation analysis, 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. Its 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_Deviation 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. The subsequent classification results are seen in Table 3.

5.2 CMS Open Payment Dataset

The same fraud prediction process used on the Part D Prescriber dataset was applied to the CMS payment datasets. Because CMS is an open, public dataset, we utilized both LEIE and Board Actions to increase the number of matched records in comparison to the previous experiment. 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. In this experiment, we utilized the "State" feature to see if performance can be improved. 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, Payment Amount, and Payment Count*, when taking LEIE and Board Action Records as labels, respectively.

5.3 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. Table 6 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.

Most useful features: Five features (Rating Count, Average Review Rating, Trustworthiness, Explains Condition Well, Answer Questions) were selected based on the p-value in the logistic regression results at significant level of 0.05 for the comprehensive analysis below.

5.4 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 7. The constraint of this method is that it only works on a very small number of instances with complete fields.

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

6 Limitation and Future Work

Identifying healthcare fraud is the primary task of this research, thus we have concentrated on acquiring better fraud prediction accuracy, including the attempts on various features and algorithms. As future work, well investigate potential interesting findings, to find features which are significant indicators of frauds.

Because the data was imbalanced and sparse, it was challenging to make accurate prediction on the complete. 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 as well.

7 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, while the best Weighted F1 of 75.59% is obtained using data from single source. Our main findings include the following:

a) In contrast to the annual CMS open payment datasets, the Part-D Prescriber dataset has more records, more physicians, and more matched excluded physicians. According to these

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 6: Classification Performance Using 11 Features from Social Media

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

Table 7: Classification Performance Using 8 Features from Social Media, Open-payment Datasets and Prescriber Datasets

facts, the Part-D Prescriber dataset is more reliable and provides more useful information in term of fraud prediction.

b) Taking LEIE fraud labels as the dependent variable, the important signals in the Part-D Prescriber dataset that indicate fraud include Physician Specialty, TOTAL_CLAIM_COUNT, TOTAL_CLAIM_COUNT_DEVIATION, Unusual Drug Prescription, Average_Day_Supply_Per_Claim, Unusual Drug Prescription, and Average_Day_Supply_Per_Claim.

c) The important signals in the payment dataset indicating fraud include Physician Specialty such as Hepatology, and other features such as Unusual Drug Prescription and Payment Amount.

d) The combination of the Part-D prescriber dataset, open-paymnet dataset, and the social media dataset gives the best performance.

8 Acknowledgements

This publication was made possible by the support of Dr. Robin Russell, and Dr. Nottingham Quinton from the Pamplin College of Business, Virginia Tech for their help during this research.

References

1. Chan, C., Lan, C.: A data mining technique combining fuzzy sets theory and bayesian classifieran application of auditing the health insurance fee. In: Proceedings of the international conference on artificial intelligence. vol. 402408 (2001)
2. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)
3. He, H., Hawkins, S., Graco, W., Yao, X.: Application of genetic algorithm and k-nearest neighbour method in real world medical fraud detection problem. JACIII **4**(2), 130–137 (2000)
4. He, H., Wang, J., Graco, W., Hawkins, S.: Application of neural networks to detection of medical fraud. Expert systems with applications **13**(4), 329–336 (1997)
5. Hwang, S.Y., Wei, C.P., Yang, W.S.: Discovery of temporal patterns from process instances. Computers in industry **53**(3), 345–364 (2004)
6. Li, J., Huang, K.Y., Jin, J., Shi, J.: A survey on statistical methods for health care fraud detection. Health care management science **11**(3), 275–287 (2008)
7. Liou, F.M., Tang, Y.C., Chen, J.Y.: Detecting hospital fraud and claim abuse through diabetic outpatient services. Health care management science **11**(4), 353–358 (2008)
8. Rudman, W.J., Eberhardt, J.S., Pierce, W., Hart-Hester, S.: Healthcare fraud and abuse. Perspectives in Health Information Management/AHIMA, American Health Information Management Association **6**(Fall) (2009)
9. Shapiro, A.F.: The merging of neural networks, fuzzy logic, and genetic algorithms. Insurance: Mathematics and Economics **31**(1), 115–131 (2002)
10. Sokol, L., Garcia, B., Rodriguez, J., West, M., Johnson, K.: Using data mining to find fraud in hcfa health care claims. Topics in health information management **22**(1), 1–13 (2001)
11. Thompson, L.: 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) (1992)
12. Thornton, D., Mueller, R.M., Schoutsen, P., Van Hillegersberg, J.: Predicting healthcare fraud in medicaid: a multidimensional data model and analysis techniques for fraud detection. Procedia technology **9**, 1252–1264 (2013)
13. Wei, C., Hwang, S., Yang, W.S.: Mining frequent temporal patterns in process databases. In: Proceedings of international workshop on information technologies and systems, Australia. vol. 175180 (2000)
14. Yang, W.S., Hwang, S.Y.: A process-mining framework for the detection of healthcare fraud and abuse. Expert Systems with Applications **31**(1), 56–68 (2006)