Expert Systems With Applications 210 (2022) 118433

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

# Feature engineering to detect fraud using healthcare claims data

Nishamathi Kumaraswamy [a,*,1], Mia K. Markey [b], Jamie C. Barner [a], Karen Rascati [a]

[a] *The University of Texas at Austin College of Pharmacy, 2409 University Avenue Stop A1930, Austin, TX 78712, United States*
[b] *The University of Texas at Austin, Department of Biomedical Engineering, 107 W Dean Keeton Street Stop C0800, Austin, TX 78712, United States*

ARTICLE INFO

*Keywords:*
Medicaid
Fraud detection
Class imbalance
Machine learning
Statistical models

ABSTRACT

Insurance fraud is ranked second in the list of expensive crimes in the United States, with healthcare fraud being the second highest amongst all insurance fraud. Contrary to the popular belief, insurance fraud is not a victimless crime. The cost of crime is passed onto law-abiding citizens in the form of increased premiums or serious harm or danger to beneficiaries. To combat this kind of societal threat, there is an intense need for healthcare fraud detection systems to evolve. Some common roadblocks in implementing digital advancements (as seen in other domains) to healthcare are the complexity, heterogeneity of the data systems, and varied health program models across the United States. In other words, data are not stored in a centralized manner due to the sensitive domain nature, thus making it difficult to implement a robust real-world fraud-detection system. At the same time, in addition to the complexity of the varied systems involved, there is also the need to meet certain standards before a fraud actor can be prosecuted in a litigation setting. Thus, there is a human aspect to the fraud detection process flow in the real-world.

In this article, a novel framework was outlined that converts diverse prescription claims (both fee-for-service and managed care) into a set of input variables/features suitable for implementation of an advanced statistical modeling fraud framework. This article thus aims to contribute to the existing literature by describing a process to transform prescription claims data to secondary features specific to provider fraud detection. The core idea was to focus on three main aspects of fraud (business heuristics on claims, provider-to-prescriber relations, and provider's client populations) to design the input features. A systematic method was proposed to extract features that have the potential to detect billing or behavioral outliers among pharmacy providers using information extracted from a secondary database (outpatient prescriptions). The application of a commonly used dimensionality reduction method, the Principal Component Analysis (PCA), was evaluated. PCA evaluates and reduces the extensive feature subspace to only those that captures the most variance in the data. To evaluate the features extracted from this framework, the application of the engineered features and the principal components to out-of-the-box logistic regression and Random Forest algorithms were considered to identify potential fraud. The engineered features when tested in different experimental settings with a logistic regression model had the highest area under the Receiver Operating Characteristic (ROC) curve of 0.76 and a weighted F score of 0.85 while a random forest model had the highest area under curve of 0.74 and a weighted F score of 0.88.

## 1. Introduction

Healthcare expenses are on the rise across the world, but the United States (US) has a higher growth rate for healthcare expenses relative to its gross domestic product (GDP). As reported by the Centers for Medicare and Medicaid services (CMS), the national health expenditure (NHE) grew 3.9% to $3.5 trillion in 2017, which equates to $10,739 per person. This accounted for 17.9% of the US GDP. US health spending is projected to grow at an average rate of 5.5% per year in the 2018–2027 time period and is expected to reach nearly $6.0 trillion by 2027 (Projections of National Health Expenditures: Forecast Summary, 2018). With such a rapid increase in our nation's healthcare expenses it is imperative to find solutions to contain costs. One such cost containment area is program integrity. A conservative estimated average of $80 billion per year is lost to fraud across all lines of insurance (Coalition against Insurance Fraud, 2022). Fraud occurs in different industries,

* Corresponding author.
*E-mail addresses:* nishamathi.bits@utexas.edu (N. Kumaraswamy), mia.markey@utexas.edu (M.K. Markey), jbarner@austin.utexas.edu (J.C. Barner), krascati@mail.utexas.edu (K. Rascati).

[1] ORCID: 0000-0002-1545-780X.

https://doi.org/10.1016/j.eswa.2022.118433
Received 20 June 2021; Received in revised form 1 August 2022; Accepted 4 August 2022
Available online 8 August 2022

such as finance, telecommunications, credit card, insurance and others. In the United States (US), insurance fraud is ranked second in the list of expensive crimes (Farid Bavi, Insurance Fraud Management Theory and practice, 2016).

The magnitude of healthcare fraud in the United States is however very difficult to estimate, because of the inherent difficulty in determining the extent of fraud within our varied health ecosystem. That is, the nature of fraud in healthcare significantly differs from that of other industries because of the complexity and heterogeneity of the data systems and program models. There is no reported estimate on what percent of healthcare fraud can be attributed to the private sector health insurance or state/federal government healthcare. Considering that a large portion of healthcare expenses come from government administered programs (Hartman, Martin, Espinosa, Catlin, & National Health Expenditure Accounts Team, 2018), it is imperative to mitigate fraudulent activities, especially in government administered programs. The US federal government has reported healthcare fraud judgements and settlements (in addition to other healthcare administrative and imposed sanctions) of 2.6 billion dollars in fiscal year 2019 (The Department of Health and Human Services And The Department of Justice Health Care Fraud and Abuse Control Program Annual Report for Fiscal Year 2019, 2019). The National Health Care anti-fraud Association (NHCAA) conservatively estimates that about 3% of our healthcare spending is lost to fraud (Rosenbaum, Lopez & Stifler, 2009; National Health Care Anti-Fraud Association, 2021).

In this study, available prescription claims variables from a state Medicaid Management Information System (secondary data source) were evaluated to design a unique feature extraction framework. These features have the potential to answer complex analytical problems, one of which is identifying pharmacy provider fraud. The uniqueness of this research lies in the proposed feature engineering framework utilizing real-world prescription claims data and validating the features using a proprietary ground truth database, thus providing confidence in the engineered features. Our main technical contribution is the process of deriving a series of features from real-world transactional claims data attributes using a systematic exploratory analysis. The main technical challenge was to process the raw flattened claims data (due to the size of the data) and convert them to provider-level data. The claims data (2016–2017) contained ~72 million transactions with 30 variables/attributes describing each transaction totaling a reimbursement of ~$7 billion.

## 2. Related work

Data mining and machine learning technologies have been widely used for fraud detection in various domains such as auto-insurance, life-insurance, health insurance, and banking (Brockett, Derrig, Golden, Levine, & Alpert, 2002; Shah, & Asthana, 2013; Fan, Zhang, & Fan, 2019; Muttipati, Viswanadham, Senapathi, & Rao, 2021). In a supervised healthcare fraud detection predictive modeling problem, the primary objective is to learn a function that maps the response variable ('provider fraud' or 'no provider fraud') to features in the provider's billing data. The secondary objective is the ability to explain the model prediction results to a manual reviewer or an end user.

Previous studies have been conducted by several research groups in healthcare fraud detection using publicly available claims (Public Use Files – PUF) data such as Center for Medicare and Medicaid services (CMS) and other databases such as the health insurance commission of Australia, Iran's private health insurance sector, the Health Insurance Review and Assessment Services and the National Health Insurance program in Taiwan (Phua, Lee, Smith, & Gayler, 2010; Shin, Park, Lee, & Jhee, 2012; He, Wang, Graco, & Hawkins, 1997; Bauder, Khoshgoftaar, Richter & Herland, 2016; Joudaki, Rashidian, Minaei-Bidgoli, Mahmoodi, Geraili, Nasiri, & Arab, 2016; Fan, Zhang, & Fan, 2019; Aral, Güvenir, & Akar, 2012; Zhang & Wang, 2018; Castaneda, Morris, & Khoshgoftaar, 2019; Tang, Mendis, Murray, Hu, & Sutinen, 2011). Provider fraud prediction research in the United States commonly uses CMS data that includes Part B, Part D, durable medical equipment, prosthetics, orthotics and supplies, Medicare provider utilization and payment variables. Claims data available to researchers from CMS is provided as aggregated data for each provider, thus the features available for each provider are fixed or limited by the data owners (CMS) who published the data. Scant research explores the usefulness of feature engineering with these publicly available datasets. In addition, there is limited literature on feature engineering using real-world healthcare claims data that are specific indicators for potential provider fraud. The few studies that discuss feature engineering as part of their modeling process are highly domain-specific and in some cases require extensive domain knowledge to perform feature engineering (Aral, Güvenir, Sabuncuoğlu, & Akar, 2012; Tang, Mendis, Murray, Hu, & Sutinen, 2011; Castaneda, Morris, & Khoshgoftaar, 2019; Zhang & Wang, 2018). For example, Aral et al. and Tang et al. use prescription claims from Australia Medicare and only use six features based on availability from their claims database. Castaneda et al. use neural network with several different CMS data sources (not just prescription data) and their research focus was to assess the different activation functions performance on the neural network model. Zhang et al. use a subset of prescription claims from a Chinese hospital whose claim structure is significantly different from that of US prescription claims data. In addition, they only use a fixed set of features like Aral et al. and Tang et al. to assess provider risk.

In this study, analysis and processing of real-world healthcare claims transaction data was used to engineer features that identify potential provider fraud. In addition, the engineered features were assessed for their performance in predicting fraud using a real-world healthcare dataset. Therefore, this section is limited to the small body of previous work (Kumaraswamy, Markey, Ekin, Barner & Rascati, 2022) conducted to identify potential fraudulent behavior using claims data in the United States. At the time of writing this article, only seven studies (He, Wang, Graco, & Hawkins, 1997; Shin, Park, Lee, & Jhee, 2012; Joudaki, Rashidian, Minaei-Bidgoli, Mahmoodi, Geraili, Nasiri, & Arab, 2016; Zafari & Ekin, 2019; Capelleveen, Poel, Mueller, Thornton & Hillegerberg, 2016; Thornton, Capelleveen, Poel, Hillegerberg, 2014; Fan, Zhang, & Fan, 2019) fell under that category, with only three studies (Shin, Park, Lee, & Jhee, 2012; Fan, Zhang, & Fan, 2019; Zafari & Ekin, 2019) providing details of how the features were generated from the raw claims data. He et al. generated their model's feature space from subject matter experts, while Shin et al. obtained their main feature space from claims data; however, details on converting raw claims to these indicators (as they stated) were not discussed. Joudaki et al. developed their features based on expert interviews and logical inference (fraud schemes). Capelleveen et al. and Thornton et al. both explored data mining techniques such as univariate, multi-variate and clustering techniques to identify outliers using a fixed set of features derived from dental subject matter experts. Zafari et al. applied a topic modeling concept using a limited subset of three features (provider, prescriber, and prescribed drug) from all prescription claims variables, as their goal was to handle the hierarchical nature of prescriptions, rather than feature engineering for the best features indicative of provider fraud. Fan et al. performed their feature engineering in addition to the usually available CMS data features using other data sources such as social media and CMS open payment data. Although Fan and colleagues explained the feature engineering process to some extent, they were more focused on combining features from three different data sources and were interested in the assessment of a combination of features on model performance.

The types of features available in these public claims data were: billing amount, number of services received, and patient demographics, average amount billed/paid by Medicare for each physician/procedure, place of service on the claim with the provider, type of provider, count of beneficiaries, total claim count, total 30 day fill count, total day supply, and total drug cost (Peng, Kou, Sabatka, Chen, Khazanchi, & Shi, 2006; Bauder, Khoshgoftaar, Richter, & Herland, 2016). These are pre-processed features available in the public claims data and hence previous researchers were limited to feature engineering within the pre-processed features. However, in this manuscript, we engineer features directly from the prescription claim transactions and transform each element/attribute from claims into several pre-processed features. This is the first study to publish the data transformation and feature engineering process from real-world healthcare claims data using commonly available prescription claim elements in a Medicaid management information system.

## 3. Methods

The study methods consisted of a series of sequential steps that included pre-processing, feature engineering and feature extraction to prepare data for modeling. In the predictive modeling context, features are elements/attributes available in the structured data (e.g., age of patients, prescriptions utilized by members). Feature engineering is the process of creating new features based on the available structured raw data elements and the model prediction goals. This process typically requires domain expertise to identify aspects of the data/features that are most relevant to the objective of an analysis. Feature engineering is commonly driven quantitatively, where features are iteratively created and refined based on predictive modeling output (e.g., splitting age into coarser age groups after examining age as a whole).

The current set of engineered features was designed with the goal of delineating pharmacy provider fraud from the non-fraud cases. The pharmacy domain actors who can be involved in a fraud scheme are shown in Fig. 1. Here, actors were defined as the different participants who are involved in any pharmacy prescription claim transaction and hence have the potential to be involved in fraud. Three types of candidate feature categories were examined based on interactions between these actors (in case of outpatient prescription claims) namely:

(a) Provider-client population – A provider-client population is a strong indicator dictating the business of a pharmacy in a quantitative fashion depending on their clientele's overall health.

(b) Provider-prescriber relation – A provider-prescriber feature determines any relation that might exist between these 2 actors. For example, if most of a pharmacy provider's business is from a prescriber and the prescriber's business is also mostly with the pharmacy provider, they might be involved in a kick-back scheme that would benefit both provider and pharmacy, which may or may not affect a Medicaid client.

(c) Claim-focused features – Claims-focused features are commonly used as a summary measure for a pharmacy and are also used as indicators in univariate outlier fraud detection in the literature (e.g., unique number of beneficiaries and prescribers per pharmacy provider).

In this exploratory study, comprehensive, in-depth data processing details were provided and features were assessed and validated using two representative learners/models (logistic regression and random forest). The data experiments performed were as follows:

1. Using aggregated features engineered from prescription claims
2. Using aggregated features engineered from prescription claims + Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance
3. Using the first 'n' principal components of a principal component analysis
4. Using the first 'n' principal components of a principal component analysis + Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance

The two representative learners/models were trained with these combinations and the performance metrics of the model were compared with those available in the literature, leaving other possible learners/algorithms/models such as decision trees, neural network, gradient boosting for future work. These two learners/models were chosen such that the models' objective function can account for both a linear and a non-linear separation of classes (fraud vs non-fraud).

The focus of this article was on the analytical methods to design

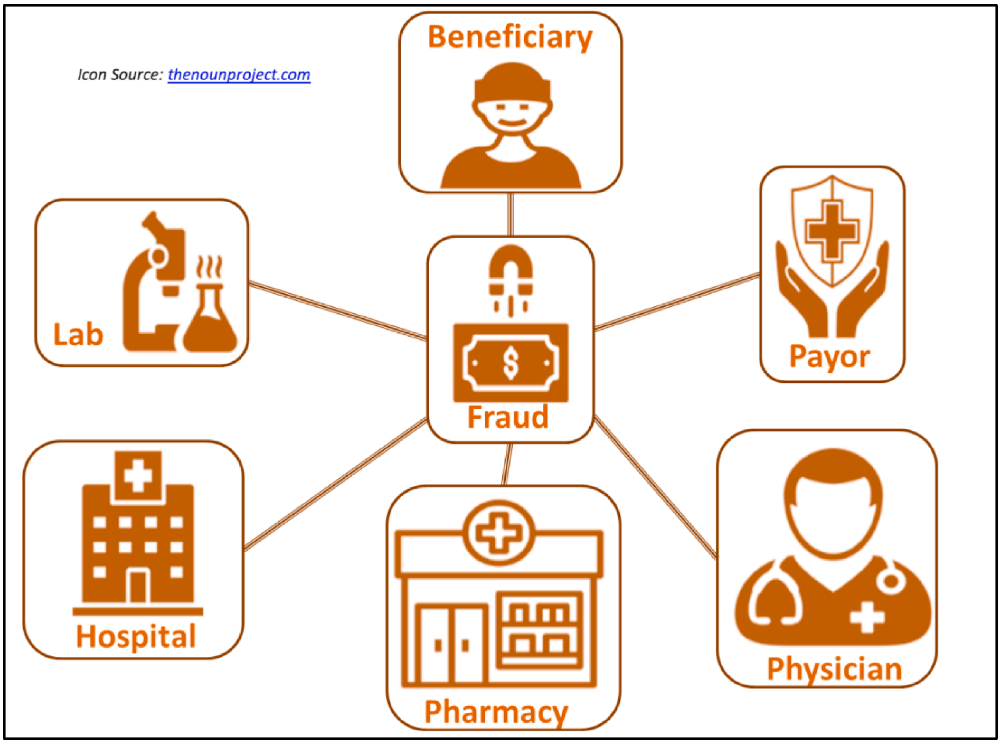


**Fig. 1.** Potential Actors in Pharmacy fraud.

engineered features for fraud prediction. Fig. 2 depicts a high-level overview of the different components in the proposed feature engineering model framework, followed by validation of features' performance using a logistic regression and random forest model. However, the performance of machine learning or statistical algorithms is usually evaluated using metrics such as predictive accuracy or area under curve (AUC). In case of fraud detection algorithms, it is important to note that an imbalanced data set (i.e., an unequal distribution of samples among the classes) is seen. The class having the majority of data (i.e., non-fraud) is called the majority class and the other, fraud, is considered the minority class. Machine learning algorithms when trained on imbalanced data, will have greater classification error on the minority class (fraud). Therefore, the logistic regression model was trained using a sampling technique called the synthetic minority oversampling technique or SMOTE (Chawla, Bowyer, & Hall, 2002; Fernández, Garcia, Galar, Prati, Krawczyk, & Herrera, 2018) to correct for the imbalance inherent in the data set. The Random Forest model was trained using a balanced overall training data and a balanced sub-sample training data.

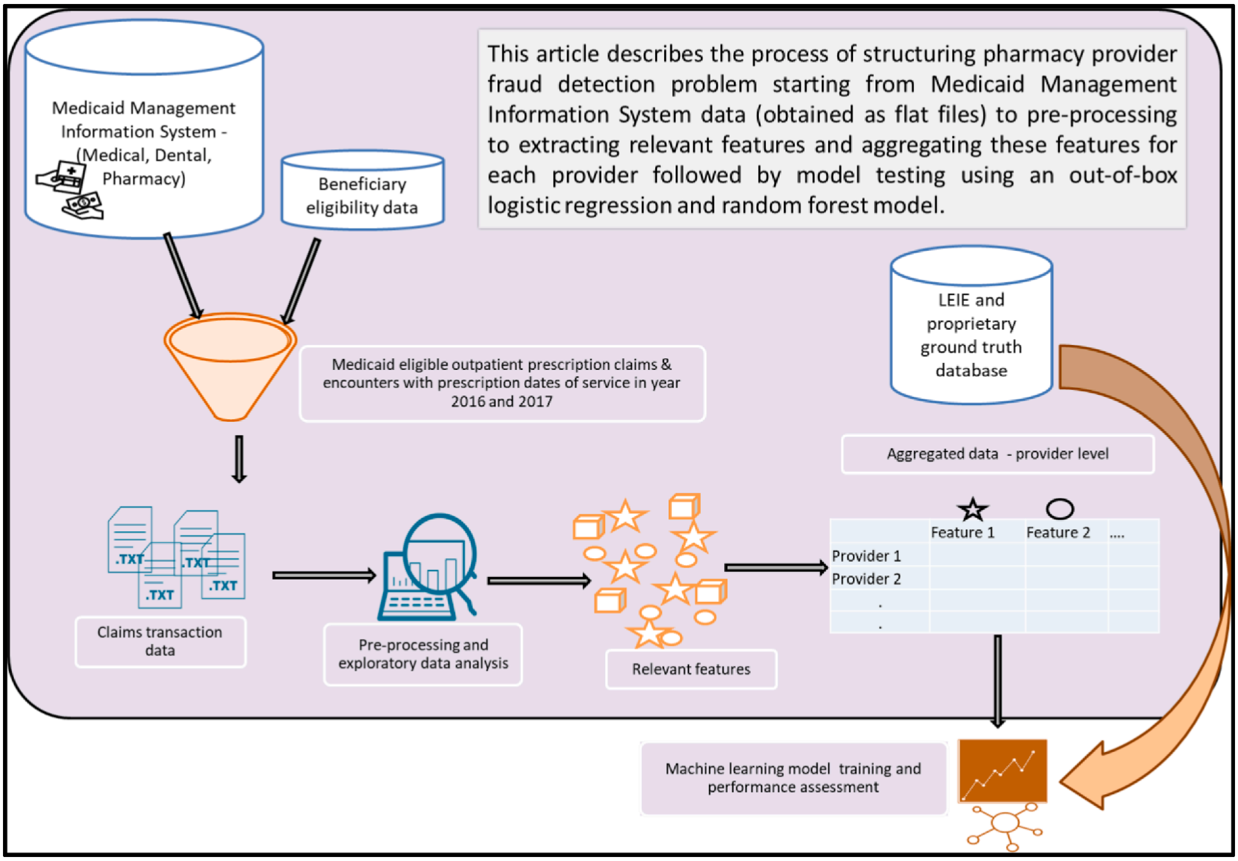
### 3.1. Data

All appropriate approvals as per the Health and Human Services (HHS) circular C-055 form were obtained prior to data release to the requestor (first author). The data request was made via a C-055 form, the request form that an employee of Texas Health and Human Services can complete to request de-identified limited data for purposes of research. De-identified limited Texas Medicaid outpatient prescription claims and encounters data for two calendar years (2016 and 2017) were released to the requestor. The datasets used for fraud labels in this study were obtained from the 'list of excluded individuals and entities' (LEIE) published by office of inspector general (OIG) and from the 'case management' database owned by Texas HHS OIG. LEIE is a complete database listing all exclusions such as providers who submitted false or fraudulent claims to any federal health care program or providers who had convictions relating to patient abuse or neglect and others. The datasets used in this study (claims transaction data and case management data) are not open source; they are proprietary data owned by Texas Health and Human Services. However, any member of the public can request such a dataset from Texas Health and Human Services by submitting a research proposal and an open records request.

Due to the large volume involved, prescription claims data were released to the requestor as a set of eight '.txt' files with each file containing de-identified prescription claims data from each quarter of the requested years, i.e., 2016 and 2017. The 17 transaction elements that were used from prescription claims are shown in Table 1. The structured data elements in the table were further processed to engineer features for modeling purposes.

Researchers have found that feature engineering is an important step in fraud detection algorithms (Fan, Zhang, & Fan, 2019). Many authors (Shin, Park, Lee, & Jhee, 2012; Fan, Zhang, & Fan, 2019; Zafari & Ekin, 2019) approached this process by using features that are readily available and aggregated in a holistic manner (for a provider) such as from CMS Part D Medicare claims data. However, these features fail to capture relationships between the three main actors: the pharmacy provider, pharmacy prescriber and the pharmacy beneficiary. It is hypothesized that pharmacy provider fraud detection could be made effective by leveraging information from the transactional claims that occur between the three main actors.

A systematic methodology was developed to analyze and engineer features for pharmacy provider fraud detection from transactions data, using the LEIE data and the proprietary case management data for our labels (fraud and non-fraud cases).



This article describes the process of structuring pharmacy provider fraud detection problem starting from Medicaid Management Information System data (obtained as flat files) to pre-processing to extracting relevant features and aggregating these features for each provider followed by model testing using an out-of-box logistic regression and random forest model.

Medicaid Management Information System - (Medical, Dental, Pharmacy)
Beneficiary eligibility data
Medicaid eligible outpatient prescription claims & encounters with prescription dates of service in year 2016 and 2017
LEIE and proprietary ground truth database
Aggregated data - provider level
.TXT
.TXT
.TXT
.TXT
Claims transaction data
Pre-processing and exploratory data analysis
Relevant features
Feature 1
Feature 2
....
Provider 1
Provider 2
Machine learning model training and performance assessment

Fig. 2. A high level overview of feature engineering process.

**Table 1**
Seventeen data attributes/elements from outpatient pharmacy prescription data and their description.

| Data element | Data element description |
|---|---|
| Prescription fill date | This variable identifies the first date on which services were rendered |
| Prescription date | This variable identifies the date, the prescription was written by the prescribing provider |
| Days supply | This variable identifies the number of days' supply dispensed for the prescription and is a value submitted by the pharmacy provider. |
| American Hospital Formulary Service code | This variable provides a classification that allows grouping of drugs with similar pharmacologic, therapeutic and chemical characteristics. |
| Dispense fee | This variable identifies the dollar amount on the claim detail that was paid as dispensing fee. The amount paid to the pharmacy for dispensing (packaging, filling, counseling) the prescription. |
| Paid amount | This variable identifies the dollar amount on the claim detail that was paid for the prescription |
| Total paid amount | This variable identifies the total dollar amount on the claim detail that was paid for the prescription (Paid amount + Dispensing fee) |
| National Drug Code source | This variable identifies if the drug on the claim is a branded generic or generic or single source or an innovator drug. |
| DEA code | This variable identifies the Drug Enforcement Agency (DEA) Controlled Substance Abuse (CSA) Act scheduling code. This is a 1-digit value specifying the DEA Controlled Substance Abuse Act Scheduling. Examples for each possible value are below:<br>0 - No control<br>1 – Lysergic acid diethylamide (LSD), heroin, marijuana (research only)<br>2 - Morphine, meperidine, amphetamines, etc.<br>3 - Aspirin/codeine etc., (less abused)<br>4 - Valium (potential abuse)<br>5 – Cough preparation with less than 200 mg of codeine. Usually controlled sale by pharmacy |
| Compound code | This variable indicates whether or not a prescription is compounded or non-compounded type prescription. |
| DAW code | This variable indicates whether a substitution is allowed by the prescriber. A claim can have Dispense As Written (DAW) code value between 0 and 9 |
| Thera class code | This variable indicates the Texas Therapeutic Class associated with an NDC. |
| Client category | Client's age is categorized in 33 bins. A client category value of p[0,2) meaning all PCNs aged from 0 years to under 2 years are included in this bin. |
| MCO code | Managed Care Organization (MCO) program code such as STAR, STAR + PLUS etc. |
| Beneficiary ID | This variable is a de-identified beneficiary identification to whom the drug prescription was prescribed |
| Pharmacy ID | This variable is a de-identified pharmacy identification who dispensed the drug on the claim |
| Prescriber ID | This variable is a de-identified prescriber identification who wrote the prescription for the beneficiary |

### 3.2. *Descriptive statistics and feature engineering:*

To identify features for each pharmacy provider, descriptive statistics of each data attribute/element were used to derive features that describe the provider billing, based on the three categories (see section 3). Our current study details the aggregation strategy used in our feature engineering framework for fraud detection. Each data attribute/element from the raw adjudicated claims data is discussed, and the aggregate method was applied to each data attribute/element that forms the initial feature space matrix. For example, Table 2 provides a hypothetical case scenario on how features used together can identify an out-of-norm provider. In this case-scenario, for purposes of clarity and simplicity, the prescriptions billed by the 5 pharmacy providers are assumed to be the same. Here P1 and P4 are providers who seem to have out-of-norm billing based on their average paid dollar amount and the standard deviation of this amount in comparison to their peers. The idea is that given a unique set of prescription distributions among the clients for each provider, P1 and P4 stand out due to their high average and high standard deviation when considering the billing of these 5 providers in a holistic fashion.

**Table 2**
Hypothetical case of a suspicious provider. P1 and P4 are considered suspicious based on the four features.

| Provider ID | Average paid | Unique Beneficiaries | STD | Per Member Per Month exposure |
|---|---|---|---|---|
| P1 | 100 | 10 | 10 | 120 |
| P2 | 10 | 10 | 1 | 120 |
| P3 | 5 | 5 | 0.05 | 60 |
| P4 | 10,000 | 10 | 10 | 120 |
| P5 | 20 | 20 | 1 | 240 |

A summary of descriptive statistics of prescription claims data (~35 million transactions; training data) for year 2016 can be found in Table 3. The dataset provided had 30 variables that together described each transaction in year 2016, from all Texas Medicaid clients and pharmacies. The initial 30 variables also included descriptions of coded variables (e.g., therapeutic class code is a variable and description is another variable). From the original 30 variables, 17 variables contained meaningful information on each transaction and were aggregated to create 17 transaction attributes. Different measures of the 17 variables, such as mean paid amount, median paid amount or the third quartile of paid amount for a pharmacy provider, were aggregated based on the providers' client configuration and business (billing) configuration. This corresponds to a total of 176 features with collinearity and these features together describe the overall billing configuration of a pharmacy provider. Categorical variables with a few (less than 5) categorical values were aggregated, but other categorical variables such as the type of program billed in the claim or encounter were further aggregated based on the overall distribution of claims and encounters for different value of program types such as STAR, STAR + PLUS, and others. The distribution of claims (counts by a variable) and the dollars corresponding to the claims are shown for the different variables available in our data in Figs. 3–6, and Tables 4 and 5.

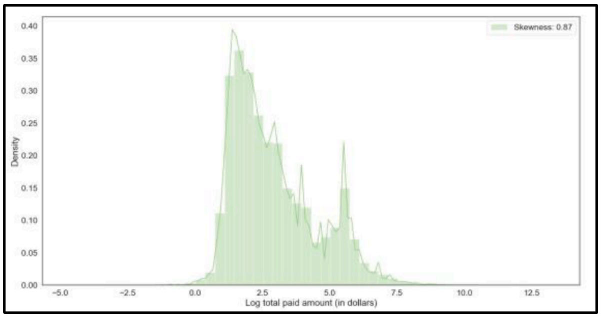
Each data attribute, or combination of attributes, from the raw transactions were analyzed to understand the distribution of each attribute and then aggregated to create features that defined the billing configuration of a provider. The conversion of each available raw variable from the transactional claims to a feature is explained and discussed in detail in subsequent sections.

De-identified Client ID: This variable represents the de-identified Medicaid beneficiary who obtained a prescription from the pharmacy provider. The unique count of Medicaid clients who received service from a pharmacy provider forms the proxy-population of the pharmacy. The number of unique beneficiaries for each provider is calculated from 2016 transactions data.

De-identified prescribing provider ID: The claim transactions have the prescribing provider detail for each transaction. A count of unique prescribers for each pharmacy provider details the variance in the prescriptions received by a pharmacy provider.

Compounded or non-compounded claim: This variable identifies a compounded prescription claim (prepared from ingredients in the pharmacy) from a non-compounded prescription claim (Table 5). As compounded drugs involve a higher dispensing fee, features based on paid amount, dispensing fee amount and the total paid amount for such transactions were aggregated by provider. The sum of total paid amount, mean of total paid amount, median of total paid amount, standard deviation in the total paid amount, and third quartile of the total paid amount for each provider were calculated. This feature based on total paid amount gives a distribution of each pharmacy's Medicaid reimbursements for compounded and non-compounded prescriptions in a

**Table 3**
Descriptive statistics of year 2016 Medicaid prescription (Rx) claims.

| | | |
|---|---|---|
| No. of claim details | 35,573,103 (excluding compounded claim details) | |
| Log (total paid amount) distribution: Claim density vs log total paid amount |  | |
| | Rx paid amount | Rx Dispensing Fee |
| Sum or Total paid ($$) | $3,376,177,233 | $70,941,464 |
| Minimum paid dollar ($$) | $0.00 | $0.00 |
| Paid dollar P25 ($$)/Quartile1 | $4.18 | $1.00 |
| Paid dollar P50 ($$)/Median | $10.22 | $1.00 |
| Paid dollar P75 ($$)/Quartile3 | $46.00 | $1.75 |
| Maximum paid dollar ($$) | $552,200.00 | $200.00 |
| Mean paid dollar ($$) | $94.91 | $1.99 |
| Standard deviation of paid dollar | $828.96 | $4.63 |

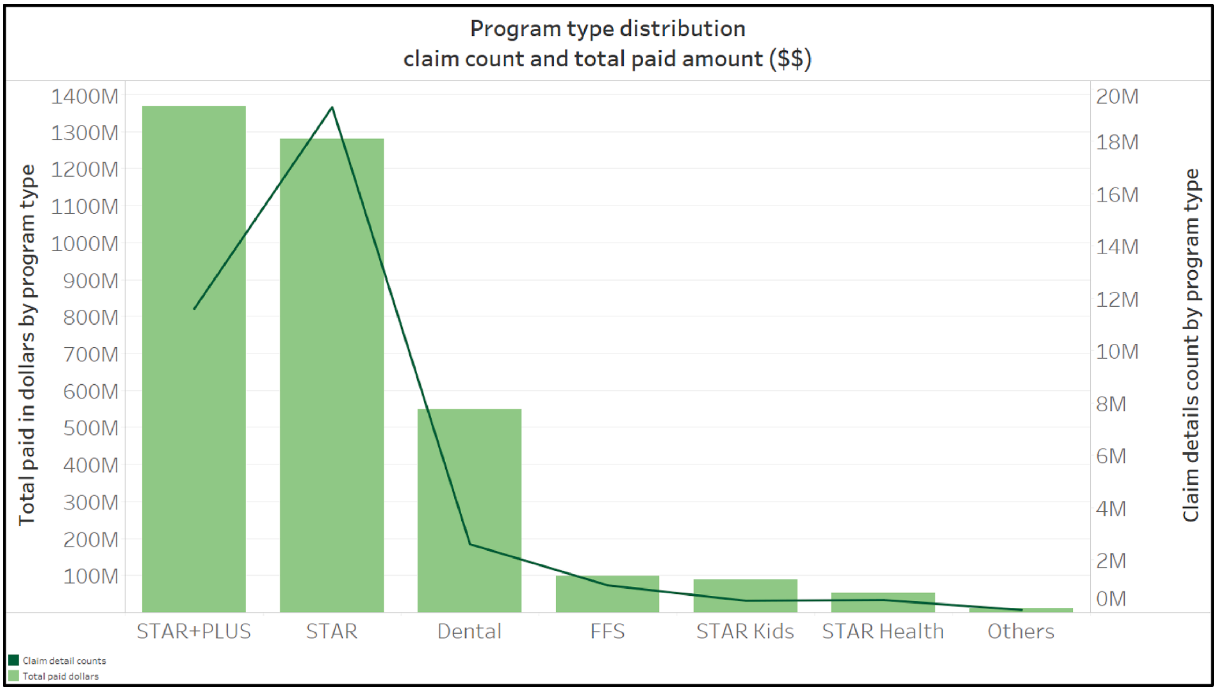


**Fig. 3.** Program type – distribution for claims and dollars reimbursed to Medicaid clients in 2016.

year's time period. This was grouped by whether a claim was a compounded prescription or non-compounded prescription for each provider and then aggregated. This led to a total of 32 features for each provider.

Program type: The distribution of claims by different program types for 2016 is shown in Fig. 3. The different program types relate to the type of clients associated with a pharmacy provider. Accounting for the program types in feature creation will help evaluate the distribution of client population in the different programs for each provider. The top 5 expensive programs (STAR + PLUS, STAR, DENTAL, FFS, and STAR Kids) within Medicaid were considered. Aggregation of the number of claims, sum of the total paid amount, average of the total paid amount, standard deviation of the total paid amount, median of the total paid amount, third quartile of the total paid amount was conducted. This led to a total of 30 features for each provider.

Age category/bin: The distribution of claims by different client age categories for 2016 is shown in Fig. 4. Although the number of pharmacy prescription transactions among younger clients (<10) were high, the dollars associated with such transactions are lower in comparison to clients aged 10–19 and those 20 and above. Therefore, we further collapsed age categories to three final categories (0 to under 10, 10 to under 20, 20 and above) of interest for each provider. For each age bin category we calculated the count of claims, sum of total paid amount, average of total paid amount, median of total paid amount, standard deviation of total paid amount, and third quartile of paid amount, leading to a total of 18 features for each provider.

Dispense as written (DAW) code: The distribution of claims by all DAW codes for 2016 is shown in Fig. 5. When prescribers indicated that

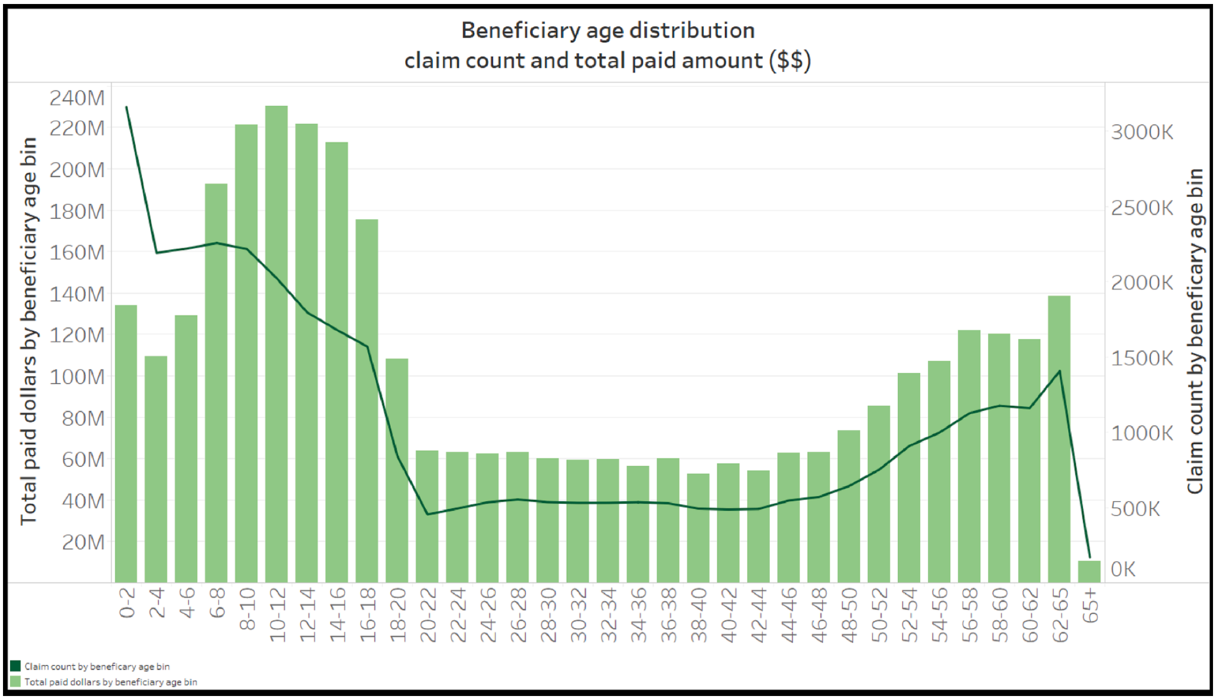


**Fig. 4.** Client age – distribution for claims and dollars reimbursed to Medicaid clients in 2016.

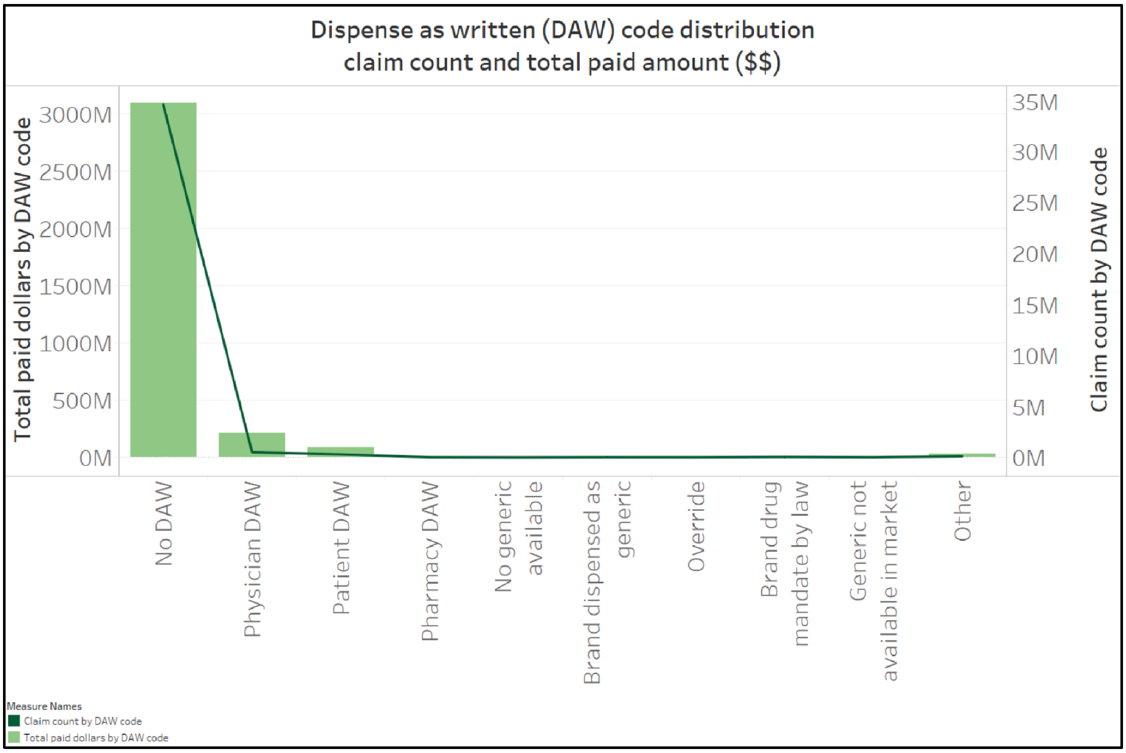


**Fig. 5.** DAW code – distribution for claims and dollars reimbursed to Medicaid clients in 2016.

the prescription had to be dispensed as written (DAW), this prohibited the pharmacy provider from dispensing a generic version of the branded product. For each provider the count of physician DAW claims, sum of the total paid amount, average of the total paid amount, standard deviation of the total paid amount, median of the total paid amount, third quartile of the total paid amount of physician DAW claims was aggregated, leading to a total of 6 features for each provider.

Drug enforcement administration (DEA) code: The distribution of claims by all DEA codes for 2016 is shown in Fig. 6. DEA code 2 (morphine and amphetamines etc.) claims were focused on and were aggregated to a provider level. For each provider, the count, sum of the total paid amount, average of the total paid amount, standard deviation of the total paid amount, median of the total paid amount, third quartile of the total paid amount of morphine claims was aggregated, leading to a

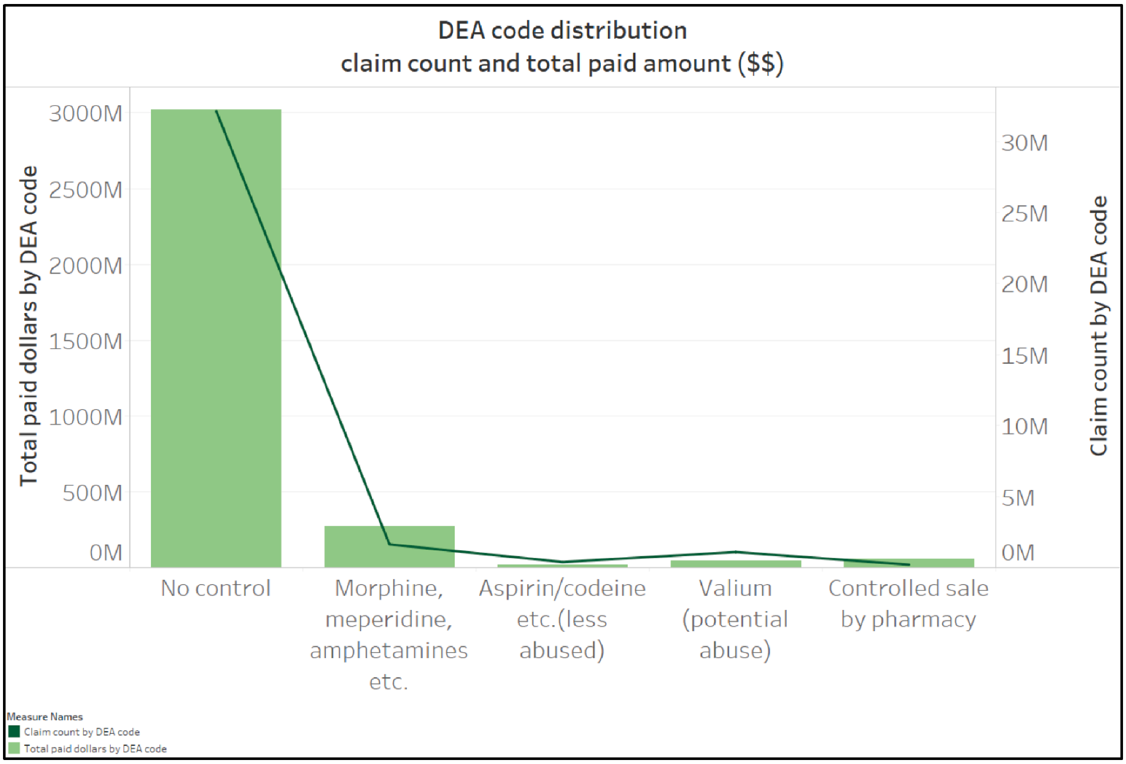


**Fig. 6.** DEA code – distribution for claims and dollars reimbursed to Medicaid clients in 2016.

**Table 4**
Descriptive statistics of 2016 claims by National Drug Code (NDC) source value.

| NDC source | Claim count (%) | Total paid $$ (%) |
|---|---|---|
| Generic | 26,467,984 (74 %) | 657,414,839 (19 %) |
| Innovator | 4,343,931 (12 %) | 1,290,216,765 (37 %) |
| Single Source | 2,892,028 (8 %) | 1,441,090,935 (42 %) |
| Branded Generic | 1,757,069 (5 %) | 26,145,042 (1 %) |
| Compounded prescriptions | 112,091 (1 %) | 32,251,117 (1 %) |

**Table 5**
Descriptive statistics of 2016 transactions by compounded drug status.

| Compounded | Claim count (%) | Total paid $$ (%) |
|---|---|---|
| **No** | 35,461,012 (99.7 %) | 3,414,867,580 (99.1 %) |
| **Yes** | 112,091 (0.3 %) | 32,251,117 (0.9 %) |

total of 6 features for each provider.

Per member per month exposure (PMPM exposure): This is a derived feature that quantifies the measure of client population variance in a pharmacy. For each provider, PMPM exposure is calculated based on the client's eligibility period in a year. For each beneficiary, their eligibility for the month is calculated based on whether they had at least one outpatient prescription claim for that month. If a beneficiary had at least 1 prescription claim for all 12 months in a year, then the beneficiary is considered as having 12 months of exposure. When a provider has 10 such beneficiaries the PMPM exposure for a pharmacy provider is calculated as below:

$$PMPM\ exposure\ for\ a\ provider = \frac{Sum\ of\ total\ paid\ dollars}{Total\ exposure\ of\ all\ beneficiaries}$$

Pharmacy-prescriber relation (PPR) – This feature is designed to establish or quantify the relationship between a pharmacy and prescriber. This variable thus assesses the number of prescribers who have a high relationship between a pharmacy and a prescriber based on their prescription patterns. If a pharmacy has more than n% of their prescriptions coming from a prescriber and if the same prescriber has more than k% of their prescriptions going to the same pharmacy, together it is noted that there is a relationship between the pharmacy and prescriber. It is important to note that any kind of ties/relationship between the pharmacy provider and prescriber does not warranty an improper relationship (for example, if there is only one pharmacy in a small town, or a pharmacy is located in a medical complex); however, PPR along with other features can increase the confidence in fraud identification.

Days of supply – The days of supply is a claim indicator that ensures a client is receiving the correct amount/dosage of medication. The distribution of days of supply for non-compounded claims is shown in Fig. 7. For each provider, count of the number of clients who received 30-day supply prescriptions, count of all 30-day supply prescription claims, along with the total paid dollars, mean and median paid dollars were aggregated leading to a total of 5 features for each provider.

National drug code (NDC) source - The NDC source is a claim indicator that identifies a prescription transaction as generic drug, branded generic drug, innovator drug or a single source drug. The distribution of NDC source for non-compounded claims is shown in Table 4. Provider's claims were aggregated to calculate the number of clients, count of such claims, along with the total paid dollars, mean and median paid dollars for such prescriptions within each source drug category (branded generic, generic, innovator and single source).

Therapeutic class drugs – There were approximately 500 therapeutic classes of drugs dispensed in the year 2016. To focus on specific class/category only non-compounded claims were considered. The percent of dollars utilized by a therapeutic class and the percent of claims for the same therapeutic class were calculated. The top 5 of therapeutic class drug categories that had the highest difference between the percent of claims and percent of dollars were identified. For example, if 5 percent of the claims are from a therapeutic class composition of TC1, but TC1 occupies 10 percent of the total paid dollars amongst all therapeutic classes, this therapeutic class category would be flagged. For each class the count of claims, sum of total paid dollars, median of total paid

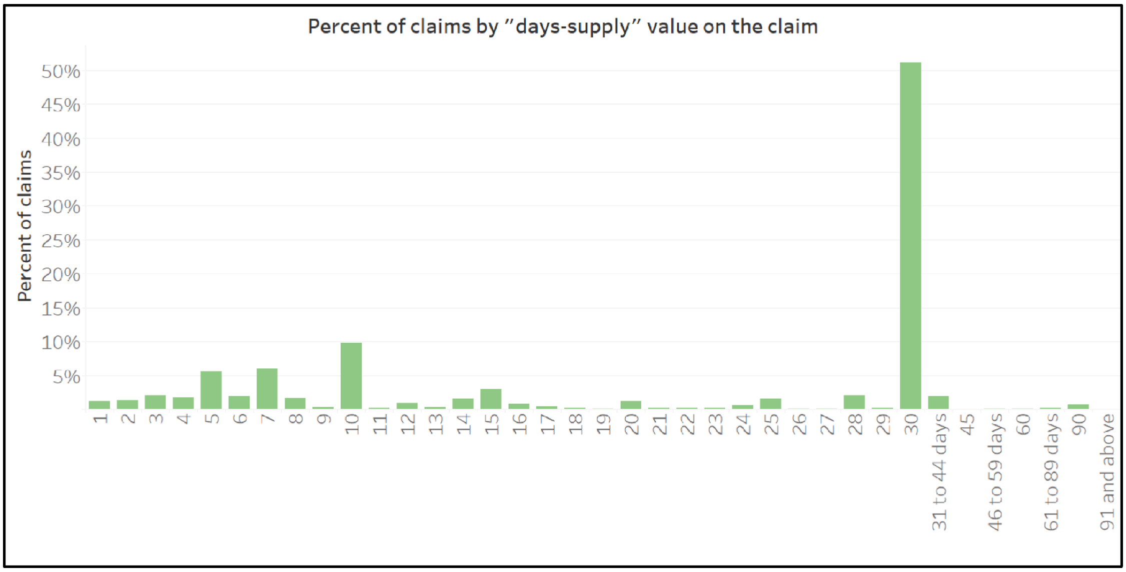


**Fig. 7.** Days' supply – % distribution for claims in 2016.

dollars, mean of total paid dollars, standard deviation of total paid dollars and 3rd quartile of total paid dollars by provider were aggregated leading to a total of 30 features for each provider. The top 5 therapeutic class categories selected as features were:

1. H7X – Antipsychotics, Atypicals, D2 partial agonist/5HT
2. C4G – Insulins
3. H7T – Antipsychotics, atypical, dopamine, serotonin, antagonist
4. M0E – Antihemophilic factors
5. P1A – Growth hormones

American Hospital Formulary Service (AHFS) class drugs – There were approximately 3,500 AHFS classes of drugs dispensed in the year 2016. Only non-compounded claims were considered. The percent of dollars utilized by an AHFS drug class and the percent of claims for the same AHFS drug class, and the top 5 with the largest difference in percentage were identified, similar to the therapeutic class calculations explained above. The top five non-overlapping drug classes between therapeutic class drugs and AHFS drugs were considered. The count of claims, sum of total paid dollars, median of total paid dollars, mean of total paid dollars, standard deviation of total paid dollars and 3rd quartile of total paid dollars by provider was aggregated to a total of 30 features for each provider. The top 5 AHFS class categories selected as features were:

1. 28:20.04.30 – Amphetamines
2. 92:00.00.44 – Miscellaneous therapeutic agents e.g., insulin agents
3. 56:40.00.73 – Miscellaneous GI drugs
4. 86:04.08.00 – Muscle relaxants
5. 52:08.00.00 – anti-inflammatory agents

### 3.3. *Model experiments using engineered features:*

This section describes the modeling experiments performed with the engineered features. The individual variables when aggregated to a provider level led to a feature space with 176 features, some of which might be redundant. However, both features (mean and median of a variable) add value to a provider's billing in comparison to their peers. The data were de-identified and allowed consideration of all of these variables in developing the final set of features for provider fraud detection. As expected, three of the engineered variables – average paid cost (dollars) per provider, median paid cost (dollars) per provider, and number of Medicaid clients per provider – were highly correlated, so a transformation was performed using principal component analysis (PCA) in order to capture the deviance in data represented by such correlated features. PCA is a dimension reduction technique whose input is represented by the original set of 176 features. After PCA, the reduced sets of features that are linear combination of original features were obtained that were used as inputs to the classifier (logistic regression).

Two commonly used classification algorithms (Logistic regression and Random Forests models) were tested on the engineered feature based data matrix. Logistic regression is one of the widely used model approaches for binary outcomes in healthcare (Kleinbaum & Klein, 2010). Here a linear relationship between the logarithm of odds of the outcome (fraud versus non-fraud) and the predictors (PCA components or the full feature set) is assumed. Random Forest is another commonly used ensemble model robust to noise and to variable interactions/redundancy. Random forest model implements the random-split selection and thus averages the variance from the individual component decision trees (Breiman, 2001; Aggarwal, 2015; Wu, Ye, Zhang, Ng, and Ho, 2014). To simulate similar experiments as with logistic regression models, Random Forest model was also tested with all the 176 features to understand its discriminatory power of fraudulent vs non-fraudulent providers. On these experiments, the focus was on the performance of the algorithm on the same dataset with both a balanced training sample and a balanced training sub-sample. The results from these trials are shown in Table 6.

Considering the focus is on feature engineering and identifying relevant features for any model, the engineering framework was tested by performing 4 experiments (see section 3) that address two main issues. The two main issues addressed here were: 1) class imbalance and 2) feature redundancy. PCA teases out multicollinearity and dependency that existed in the main feature subspace. A min–max scalar normalization was performed for each feature before performing a PCA. PCA thus reduced the feature space from 176 to 15 principal components that captured 85% of the variance in the data. The principal components were then used as inputs in a logistic regression model (principal component regression model) and the model was tested for performance using F1 scores (weighted and class specific) and AUC metrics. For the Random Forest learner, a scaled feature set (176 features) was used to train the model and was tested on a separate stand-alone test dataset to obtain performance metric scores.

Provider-level aggregated data from the year 2016 was used for training the models and provider-level aggregated data from the year

**Table 6**
Learner performance highlights from the study experiments utilizing engineered features/components.

| Learner performance on 2016 claims (Training data) | Area Under ROC Curve | F1-score (Fraud class) | F1-score (Weighted class average) |
|---|---|---|---|
| Logistic regression with 176 features (No re-sampling) | 0.79 | 0.07 | 0.95 |
| Logistic regression with 15 principal components as input features (No re-sampling) | 0.77 | 0.04 | 0.94 |
| Logistic regression with 176 features and SMOTE re-sampling technique | 0.83 | 0.76 | 0.76 |
| Logistic regression with 15 principal components as input features and SMOTE re-sampling technique | 0.77 | 0.17 | 0.82 |
| Standard Random Forest (No re-sampling) | 0.72 | 0.07 | 0.95 |
| Random Forest (balanced) | 0.80 | 0.19 | 0.86 |
| Random Forest (balanced sub-sample) | 0.79 | 0.20 | 0.86 |
| **Learner performance on 2017 claims (Testing data)** | **Area Under ROC Curve** | **F1-score (Fraud class)** | **F1-score (Weighted class average)** |
| Logistic regression with 176 features (No re-sampling) | 0.75 | 0.01 | 0.94 |
| Logistic regression with 15 principal components as input features (No re-sampling) | 0.74 | N/A | 0.94 |
| Logistic regression with 176 features and SMOTE re-sampling technique | 0.76 | 0.18 | 0.85 |
| Logistic regression with 15 principal components as input features and SMOTE re-sampling technique | 0.76 | 0.16 | 0.84 |
| Standard Random Forest (No re-sampling) | 0.66 | 0.02 | 0.94 |
| Random Forest (balanced) | 0.73 | 0.16 | 0.84 |
| Random Forest (balanced sub-sample) | 0.74 | 0.18 | 0.88 |

**Table 7**
Confusion matrices from the study experiments utilizing engineered features/components on testing data (2017 prescription claims).

| | | | Performance on testing data | |
|---|---|---|---|---|
| Logistic regression with 176 features (No re-sampling) | **Predicted Fraud** | | **Yes** | **No** |
| | **Actual** | **Yes** | 1 | 193 |
| | **Fraud** | **No** | 3 | 4664 |
| Logistic regression with 15 principal components as input features (No re-sampling) | **Predicted Fraud** | | **Yes** | **No** |
| | **Actual** | **Yes** | 0 | 194 |
| | **Fraud** | **No** | 0 | 4667 |
| Logistic regression with 176 features and SMOTE re-sampling technique | **Predicted Fraud** | | **Yes** | **No** |
| | **Actual** | **Yes** | 112 | 82 |
| | **Fraud** | **No** | 944 | 3723 |
| Logistic regression with 15 principal components as input features and SMOTE re-sampling technique | **Predicted Fraud** | | **Yes** | **No** |
| | **Actual** | **Yes** | 108 | 86 |
| | **Fraud** | **No** | 1024 | 3643 |
| Standard Random Forest (No re-sampling) | **Predicted Fraud** | | **Yes** | **No** |
| | **Actual** | **Yes** | 2 | 192 |
| | **Fraud** | **No** | 2 | 4665 |
| Random Forest (balanced) | **Predicted Fraud** | | **Yes** | **No** |
| | **Actual** | **Yes** | 105 | 89 |
| | **Fraud** | **No** | 1043 | 3624 |
| Random Forest (balanced sub-sample) | **Predicted Fraud** | | **Yes** | **No** |
| | **Actual** | **Yes** | 91 | 103 |
| | **Fraud** | **No** | 724 | 3943 |

2017 (for first six months) was used as testing data. The claims dataset was split into training and testing datasets based on the prescription dates in the transactional data. The training dataset was used to perform the feature engineering and the same data transformations were applied to the testing data. However, the SMOTE technique was only used in model training phase. Python programming language was used to load, pre-process the transactional data and for training/testing machine learning or statistical models on the processed data.

## 4. Results

This section discusses the results of our study using testing data (2017 prescription claims data for 6 months), PCA feature reduction, and results from the trained learner performance for pharmacy provider fraud detection. The practices of individual pharmacies are unique and based on the distribution of clients and the billing distributions. Four model experiments were performed using logistic regression, and three additional model experiments were performed using Random Forest models. Both learners were trained with 2016 claims data and tested with separate testing data (six month claims from year 2017). The learner performance results from these experiments on both training and testing data are given in Table 6. The confusion matrix on the testing data for all models is summarized in Table 7 for purposes of completeness.

Of the 176 features (30, 976 pairwise comparisons), there were 770 paired features that were found to be correlated with a degree of>0.8. Based on PCA results, a set of linear combinations of 15 principal components were found to explain 85 % of the variance seen in the claims data. The variance scree plot and the correlation plot (only > 0.8) are shown in Fig. 8 and Fig. 9.

Three different performance measures/metrics (AUC, the F1 metric for class of interest, and the weighted F1 metric) were considered in this study for each experiment with a logistic regression and random forest learner. The logistic regression model (with scaled 176 features + resampling technique) and the random forest model (with balanced sub-sample) performed the best and had an F1 score equal to 0.18 for our class of interest on the testing data and a weighted F1-score of 0.85 (logistic regression) and 0.88 (random forest). The F1 score here is obtained on a stand-alone testing dataset, and hence has no direct comparison from studies in the related work section. A weighted F1-score of 0.69 and accuracy of 0.54 was obtained by a logistic regression model using CMS data (Sadiq & Shyu, 2019) and another previous study (Fan, Zhang, & Fan, 2019) reported a weighted F1-score as 0.60 using a CMS open-payment dataset and a weighted F1-score as 0.92 using features from 3 datasets (social media, CMS's open-payment datasets and prescriber datasets) combined for specific states in the United States. Our weighted F1 scores from the two learners are well within reported ranges in comparison to previous literature (weighted F1 = 0.60 to 0.92) and performs comparably considering the single data source we utilize in this study.

The accuracy and AUC values reported in Castaneda et al. and Herland et al. (2018) respectively provides some means of comparison, to the success of feature engineering on model performance. Herland et al. (2018) and Castaneda et al. use a limited subset of attributes (six excluding provider id and fraud label) from the open-source aggregated healthcare claims data. In addition, both referenced articles do not use a stand-alone testing dataset to evaluate the model performance from the claims dataset. This could lead to data leakage as both evaluation and training of the models are performed on the same datasets. Even so, a simple comparison of Herland et al. (2018)'s AUC results for Logistic Regression (AUC = 0.78) and Random Forest (AUC = 0.71) to our work suggests that our model based features perform similarly (Logistic Regression training data AUC = 0.83 and stand-alone testing data AUC = 0.76, Random Forest stand-alone testing data AUC = 0.74). Castaneda

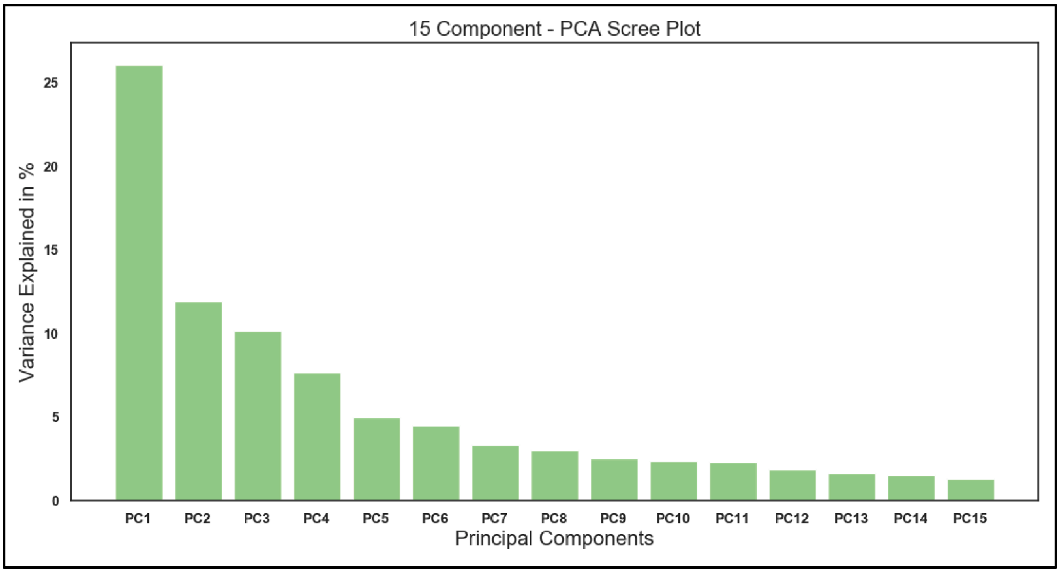


**Fig. 8.** Principal component analysis scree plot.

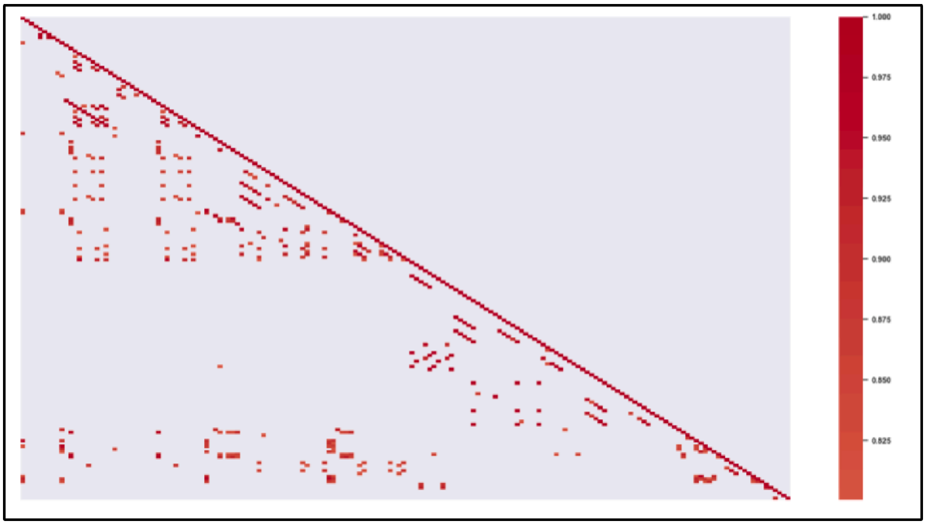


**Fig. 9.** Correlation matrix with >0.8 correlation on the design (input) matrix.

et al. used accuracy as their metric for the Part-D Medicare prescriptions data population with a highest reported accuracy of 0.70. Our work suggests that our model-based features perform comparably (Logistic Regression training data accuracy = 0.96 and testing data accuracy = 0.74).

## 5. Conclusions and discussion

Healthcare is multi-faceted with varying domains from inpatient to outpatient to pharmacy and many more, so domain specific features based on adjudication-rich variables will yield a higher success with an out-of-box machine learning or statistical algorithm such as logistic regression or others. Based on claims adjudication in the pharmacy prescriptions domain, there are many raw transaction elements available that describe the different prescription patterns and transaction patterns of a pharmacy provider. This study provided a way to aggregate these claims variables to features that define the different billing practices of pharmacies accounting for their client distribution and business rules as mentioned in section 2. There are many variables available in a Medicaid management information system and can be aggregated at the provider level, but results differ depending on the end goal of a learner or model. In this study, one-year prescription claims data were used to engineer the features that were then used to predict pharmacy provider fraud. This feature subset can be applied directly to other state or federal agency or private payers and can also be expanded to include features based on availability of additional variables. Based on the limited set of experiments performed as part of this study, we can conclude that logistic regression and Random Forest model perform reasonably well with the engineered features in comparison to those available in literature. However, we do not claim that the logistic regression model or random forest model is the best model that could be achieved, but rather that the results with the 2 models demonstrate the utility of the feature extraction approach introduced in this study. Further experiments with other interpretable learners would be needed to optimize the choice of model for analyzing prescription claims data encoded using the feature extraction approach introduced in this study.

The feature space explored here was 176-dimensional. These features were then funneled through a feature reduction method based on PCA to determine fewer components that accounted for most of the variance in the data. A Pearson pairwise correlation matrix was computed to assess the degree of relations or dependency between all features.

To our knowledge, feature engineering using prescriptions claims data to detect pharmacy provider fraud has not been studied. Some close-enough comparisons of the performance metrics of the current models and those from literature are provided in the last paragraph of section 4. These comparisons show the effect of the proposed feature engineering approach leading to a comparable model performance than those available in the literature on a stand-alone testing dataset. Our proposed framework analyzed the raw prescriptions data to identify a subset of features that can help delineate between fraudulent and non-fraudulent providers. A detailed account of how to convert the transactional prescription claims data to derive features that can be used in any machine learning or statistical model was provided. A certain level of understanding into the Medicaid reimbursement system is necessary to build any substantial hypothesis for the feature variables in being indicators for fraud. Therefore, the data of individual prescription claims identifying potential provider fraud can be tracked. Employing an "off-the-shelf" machine learning or statistical algorithm using aggregated data is relatively straightforward; however, the application to real world practice involves using raw claims data to extract features before a machine learning model can be trained using these features.

Identifying potential fraudulent providers (true positive cases) and saving cost/time incurred due to false positive cases (not fraud) with a trade-off on false negatives is the final goal of this research. False positive fraud cases incur different costs and time than that of false negative cases. False negative cases can be very costly and is domain dependent in healthcare. In the case of false positives and true positives, the costs are estimated equal in that only the initial administrative costs of opening an investigation is usually considered. When a false negative occurs, meaning that fraud is not detected, the losses are equal to an estimated value from the provider billings and can vary provider to provider. There is thus a natural trade-off between false positive and false negative provider cases in the real-world. The first step towards this goal of achieving a balance in trade-off is presented in this manuscript, which is feature engineering. Future investigations of other features available in transactional claims data that might be significant indicators of fraud will be conducted. Model optimizations are also needed to make the predictions more robust and adaptable.

Following are the main contributions from this study to healthcare fraud literature:

1. A set of features were engineered following a logical inference of interactions between potential fraudulent actors. Logical inference here refers to an instance where the prescriber and pharmacy provider having a kick-back scheme might trigger an abnormality in billing criteria from a data perspective. Some features engineered also overlap with features from literature studies and thus have been extensively used also in the open-source CMS datasets.
2. An analytical framework to convert the raw prescription transactions to features or indicators to identify fraud was provided.
3. Features were tested using two commonly used learners (logistic regression and random forest models). These features were validated based on model performance metric comparisons (on a stand-alone hold-out testing data) with literature model performance metrics (as available).

## CRediT authorship contribution statement

**Nishamathi Kumaraswamy:** Conceptualization, Methodology, Formal analysis, Writing – original draft. **Mia K. Markey:** Supervision, Formal analysis, Writing – review & editing. **Jamie C. Barner:** Supervision, Formal analysis, Writing – review & editing. **Karen Rascati:** Supervision, Formal analysis, Writing – review & editing.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

The authors do not have permission to share data.

## Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

## References

Aggarwal, C. C. (2015). *Data mining: the textbook* (Vol. 1). New York: Springer.

Aral, K. D., Güvenir, H. A., Sabuncuoğlu, İ., & Akar, A. R. (2012). A prescription fraud detection model. *Computer Methods and Programs in Biomedicine, 106*(1), 37–46.

Bauder, R. A., Khoshgoftaar, T. M., Richter, A., & Herland, M. (2016). Predicting medical provider specialties to detect anomalous insurance claims. In 2016 IEEE 28th international conference on tools with artificial intelligence (ICTAI) (pp. 784-790). IEEE.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5–32.

Brockett, Derrig, R. A., Golden, L. L., Levine, A., & Alpert, M. (2002). Fraud classification using principal component analysis of RIDITs. *The Journal of Risk and Insurance, 69*(3), 341–371. https://doi.org/10.1111/1539-6975.00027

Capelleveen, G., Poel, M., Mueller, R. M., Thornton, D., & van Hillegersberg, J. (2016). Outlier detection in healthcare fraud: A case study in the Medicaid dental domain. *International Journal of Accounting Information Systems, 21*, 18–31.

Castaneda, G., Morris, P., & Khoshgoftaar, T. M. (2019, April). Maxout neural network for big data medical fraud detection. In 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService) (pp. 357-362). IEEE.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of artificial intelligence research, 16*, 321–357.

Coalition against Insurance Fraud. (2021). Retrieved from https://insurancefraud.org/fraud-stats/ Accessed on April 24, 2022.

Fan, B., Zhang, X., & Fan, W. (2019). Identifying Physician Fraud in Healthcare with Open Data. In International Conference on Smart Health (pp. 222-235). Springer, Cham.

Farid Bavi, Insurance Fraud Management Theory and practice, Academia.edu. https://www.academia.edu/25672392/Insurance_Fraud_Management_Theory_and_practice Accessed on April 18, 2021.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from imbalanced data sets* (Vol. 11). Berlin: Springer.

Hartman, M., Martin, A. B., Espinosa, N., Catlin, A., & National Health Expenditure Accounts Team. (2018). National health care spending in 2016: Spending and enrollment growth slow after initial coverage expansions. *Health Affairs, 37*(1), 150–160.

He, H., Wang, J., Graco, W., & Hawkins, S. (1997). Application of neural networks to detection of medical fraud. *Expert Systems with Applications, 13*(4), 329–336.

Herland, M., Khoshgoftaar, T. M., & Bauder, R. A. (2018). Big data fraud detection using multiple medicare data sources. *Journal of Big Data, 5*(1), 1–21.

Joudaki, H., Rashidian, A., Minaei-Bidgoli, B., Mahmoodi, M., Geraili, B., Nasiri, M., & Arab, M. (2016). Improving fraud and abuse detection in general physician claims: A data mining study. *International Journal of Health Policy and Management, 5*(3), 165.

Kleinbaum, D. G., & Klein, M. (2010). Introduction to logistic regression. In D. G. Kleinbaum, & M. Klein (Eds.), *Logistic regression: a self-learning text* (pp. 1–39). New York: Springer.

Kumaraswamy, N., Markey, M. K., Ekin, T., Barner, J. C., & Rascati, K. (2022). Healthcare fraud data mining methods: A look back and look ahead. *Perspectives in Health Information Management, 19*(1).

Muttipati, A. S., Viswanadham, S., Senapathi, R., & Rao, K. B. (2021). Recognizing credit card fraud using machine learning methods. *Turkish Journal of Computer and Mathematics Education, 12*(12), 3271–3278.

National Health Care Anti-Fraud Association, "The Challenge of Health Care Fraud." 2021. https://www.nhcaa.org/tools-insights/about-health-care-fraud/the-challenge-of-health-care-fraud/.

Peng, Y., Kou, G., Sabatka, A., Chen, Z., Khazanchi, D., & Shi, Y. (2006). Application of clustering methods to health insurance fraud detection. In 2006 International Conference on Service Systems and Service Management (Vol. 1, pp. 116-120). IEEE.

Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. arXiv preprint arXiv:1009.6119.

Projections of National Health Expenditures: Forecast Summary, Centers for Medicare and Medicaid Services (2018). Baltimore (MD): CMS from: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/Downloads/ForecastSummary.pdf Accessed on April 18, 2021.

Rosenbaum, S., Lopez, N., & Stifler, S. (2009). Health insurance fraud: An overview. Washington, D.C.: Department of Health Policy, School of Public Health and Health Services, The George Washington University.

Sadiq, S., & Shyu, M. L. (2019). Cascaded propensity matched fraud miner: Detecting anomalies in medicare big data. *Journal of Innovative Technology, 1*(1), 51–61.

Shah, D., & Asthana, A. K. (2013). Life insurance fraud-risk management and fraud prevention. *International Journal of Marketing, Financial Services & Management Research, 2*(5).

Shin, H., Park, H., Lee, J., & Jhee, W. C. (2012). A scoring model to detect abusive billing patterns in health insurance claims. *Expert Systems with Applications, 39*(8), 7441–7450.

Tang, M., Mendis, B. S. U., Murray, D. W., Hu, Y., & Sutinen, A. (2011, December). Unsupervised fraud detection in Medicare Australia. In Proceedings of the Ninth Australasian Data Mining Conference-Volume 121 (pp. 103-110).

"The Department of Health and Human Services And The Department of Justice Health Care Fraud and Abuse Control Program Annual Report for Fiscal Year 2019" from https://oig.hhs.gov/publications/docs/hcfac/FY2019-hcfac.pdf Accessed on April 18, 2021.

Thornton, D., van Capelleveen, G., Poel, M., van Hillegersberg, J., & Mueller, R. M. (2014, April). Outlier-based Health Insurance Fraud Detection for US Medicaid Data. In ICEIS (2) (pp. 684-694).

Wu, Q., Ye, Y., Zhang, H., Ng, M. K., & Ho, S. S. (2014). ForesTexter: An efficient random forest algorithm for imbalanced text categorization. *Knowledge-Based Systems, 67*, 105–116.

Zafari, B., & Ekin, T. (2019). Topic modelling for medical prescription fraud and abuse detection. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 68*(3), 751–769.

Zhang, H., & Wang, L. (2018). Prescription fraud detection through statistic modeling. In *Proceedings of 2018 International Conference on Mathematics and Artificial Intelligence* (pp. 85–89).