



# Data-Centric AI for Healthcare Fraud Detection

Justin M. Johnson<sup>1</sup> · Taghi M. Khoshgoftaar<sup>1</sup>

Received: 24 December 2022 / Accepted: 20 March 2023  
© The Author(s), under exclusive licence to Springer Nature Singapore Pte Ltd 2023

## Abstract

Automated methods for detecting fraudulent healthcare providers have the potential to save billions of dollars in healthcare costs and improve the overall quality of patient care. This study presents a data-centric approach to improve healthcare fraud classification performance and reliability using Medicare claims data. Publicly available data from the Centers for Medicare & Medicaid Services (CMS) are used to construct nine large-scale labeled data sets for supervised learning. First, we leverage CMS data to curate the 2013–2019 Part B, Part D, and Durable Medical Equipment, Prosthetics, Orthotics, and Supplies (DMEPOS) Medicare fraud classification data sets. We provide a review of each data set and data preparation techniques to create Medicare data sets for supervised learning and we propose an improved data labeling process. Next, we enrich the original Medicare fraud data sets with up to 58 new provider summary features. **Finally, we address a common model evaluation pitfall and propose an adjusted cross-validation technique that mitigates target leakage to provide reliable evaluation results.** Each data set is evaluated on the Medicare fraud classification task using extreme gradient boosting and random forest learners, multiple complementary performance metrics, and 95% confidence intervals. Results show that the new enriched data sets consistently outperform the original Medicare data sets that are currently used in related works. Our results encourage the data-centric machine learning workflow and provide a strong foundation for data understanding and preparation techniques for machine learning applications in healthcare fraud.

**Keywords** Healthcare · Big data · Fraud detection · Data preparation · Data labeling · Data quality

## Introduction

There are many factors that influence the costs of healthcare and health insurance, including fraud, waste, and abuse (FWA). The Federal Bureau of Investigation (FBI) estimates that fraud accounts for up to 10% of all billings within the United States (U.S.) Medicare program [1]. Some examples of healthcare fraud include billing for appointments that the patient did not keep, billing for services more complex than those performed, or billing for services not provided [2]. In 2022, for example, the Healthcare Fraud Unit established by the U.S. Department of Justice began targeting a fraud

scenario related to COVID-19 out-of-office testing. Referred to as “bundling office visit fraud”, this increasingly popular fraud scheme entails providers administering COVID-19 tests in a quick drive-through fashion, while billing Medicare for in-office visits and/or complex procedures that were not provided [3]. While these fraudulent practices have clear financial repercussions, there are also abusive practices that place the patient’s well-being at risk. For example, healthcare providers may provide patients with medically unnecessary or inappropriate services that increase the risk of patient harm.

Through the proliferation of electronic health records and advances in data mining and machine learning methods, there is a great opportunity to automate healthcare fraud detection. However, there are several key challenges related to machine learning for healthcare fraud detection. For example, the healthcare fraud problem is an inherently imbalanced problem with a small number of known fraudulent providers and many non-fraudulent providers. This causes machine learning algorithms to become biased towards the majority class, and many times, causes

---

This article is part of the topical collection “Recent Trends on AI for HealthCare” guest edited by Lydia Bouzar-Benlabiod.

---

✉ Justin M. Johnson  
jjohn273@fau.edu  
Taghi M. Khoshgoftaar  
khoshgof@fau.edu

<sup>1</sup> Florida Atlantic University, Boca Raton, FL, USA

the learner to have low classification performance on the fraudulent class. Similarly, healthcare data are characterized by big data challenges [4] and have many high-dimensional categorical variables that require special handling [5]. As a result, most of the related healthcare fraud detection works employ advanced modeling techniques to address these challenges. In general, there is a lack of work that emphasizes the importance of data understanding and data preparation.

Figure 1 illustrates the Cross Industry Standard Process for Data Mining (CRISP-DM) [6]. Most related works focus heavily on the modeling stage of the process [7–9]. In practice, however, 80% of the machine learning workflow consists of data understanding and data preparation [10]. Furthermore, most would argue that the representation and quality of the input data is the most important factor in the success of any machine learning application [11]. This shift of focus from the modeling stage to the data understanding and preparation stages aligns with recent data-centric artificial intelligence (DCAI) initiatives [12]. DCAI encourages us to spend more time iterating on data, and less time iterating on models. In addition to traditional data preprocessing tasks, DCAI covers topics such as fairness and bias in labeled data sets [13], experiment reproducibility [14], and improving overall data quality [15]. While these data improvements are labor intensive, they have been shown to significantly improve classification performance [16–18].

This study presents a data-centric approach to improving healthcare fraud classification rates within the U.S. Medicare program. The U.S. Medicare program provides affordable health insurance to individuals 65 years and older, and other

select individuals with permanent disabilities [19]. In 2020 alone, there were more than 62 million Medicare beneficiaries and expenditures exceeded \$926 billion [20]. The massive scale and complexity of the program make it an easy target for fraud [21]. In an effort to reduce fraud, the Centers for Medicare & Medicaid Services (CMS) makes Medicare data sets publicly available for analysis. We provide a comprehensive understanding of publicly available Medicare data sources and the preprocessing steps required to create labeled data sets for supervised learning. Most importantly, we present new techniques for data labeling, introduce three new feature-rich data sets, and provide insights on fair model evaluation. Topics of interest for this special issue include data transparency, data profiling, data cleaning, and reproducible data preparation for machine learning and healthcare fraud detection.

We first provide a comprehensive review of the publicly available Medicare data sources that we use for fraud classification. These data sets include six Medicare provider and claims data sets from the CMS [22], and real-world fraud labels from the List of Excluded Individuals and Entities (LEIE) [23]. Three different Medicare programs are used in this study: (1) Physician and Other Practitioners (Part B), (2) Part D Prescriber (Part D), and (3) Durable Medical Equipment, Prosthetics, Orthotics, and Supplies (DMEPOS). The Part B, Part D, and DMEPOS data sets include the latest CMS claims data available at the time of experimentation, including claims for years 2013–2019. We leverage the CMS methodology documentation, data dictionaries, and exploratory data analysis to provide a review of each data set and to identify the necessary data preparation steps.

Next, we describe the data preparation techniques that we have used to prepare the CMS Medicare data sets for supervised learning. These include imputing missing values, transforming features, normalizing columns, encoding categorical variables, feature selection, mitigating redundancy, and data labeling. In addition to these common preprocessing techniques, data aggregation is used to compress the size and dimensionality of the data sets significantly, while simultaneously introducing new features that better capture provider billing patterns. Next, we introduce three new Medicare Summary by Provider fraud classification data sets, i.e., one for each Medicare program, that have not been used previously in related works. Finally, we leverage the new Summary by Provider data to enrich the existing Medicare data sets with up to 58 new features that include provider-level, claims-level, and beneficiary-level statistics. These are referred to as the aggregated, summary by provider, and aggregated-enriched data sets throughout this study.

The quality of the nine resulting Medicare data sets is evaluated on the fraud classification task using the extreme gradient boosting (XGB) [24] and random forest (RF)

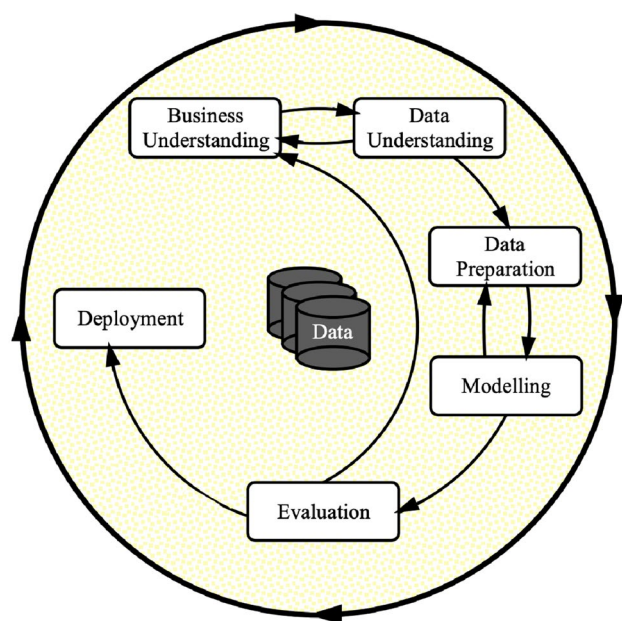


Fig. 1 Phases of the CRISP-DM process for data modeling [6]

learners [25]. Despite not having any personally identifiable information (PII) in our modeling feature set, preliminary results showed that large models are capable of de-anonymizing and memorizing providers from rich feature sets. This led to overly optimistic performance results that would not generalize to real-world applications. To provide a fair evaluation that is representative of real-world fraud detection systems, we introduce the k-fold-by-npi cross-validation technique which ensures that providers in the training folds do not exist in the test folds. Notably, the k-fold-by-npi technique proposed in this study can be generalized to other problem domains that may suffer from target leakage. Performance is reported over six runs of cross-validation using the area under the receiver-operating characteristic curve (AUC), true-positive rate (TPR), true-negative rate (TNR), and geometric mean (G-Mean). This combination of threshold-agnostic and complementary performance metrics provides a holistic view of overall performance. Across all metrics and Medicare programs, we find that the newly introduced aggregated-enriched data sets significantly outperform the original Aggregated data sets. Finally, we report feature contribution scores to illustrate the importance of the newly introduced features. Through this data-centric exercise, we have significantly increased the performance of the CMS Medicare fraud classification task with public data sources, reproducible data preparation, and a fair evaluation strategy.

The remainder of this paper is outlined as follows. The next section related works on data preprocessing for machine learning and healthcare fraud classification. Following section thoroughly reviews the CMS data sets used in this study and provides a detailed explanation of all data preparation steps. The next section outlines the experiment design and evaluation strategy used in this study. Following section highlights the importance of our proposed evaluation strategy and presents the fraud classification results obtained using the aggregated and aggregated-enriched data sets. Finally, the last section concludes with a summary of our findings and recommendations for future works.

## Related Works

Data preparation techniques are thoroughly covered in academic textbooks for general purpose data mining and machine learning [25, 26]. Topics covered include data ingestion, encoding numeric and categorical variables, feature standardization and normalization, handling missing values, addressing class imbalance, and handling data with incorrect values. Related works survey these topics and summarize opportunities and challenges related to data preprocessing [11, 27, 28]. Domain-specific preprocessing techniques are also of importance, e.g., computer vision (CV),

natural language processing (NLP), and other deep learning applications [29–31]. For example, NLP applications require a separate pipeline of feature transformations to convert free text to numeric values for machine learning, i.e., tokenization, stop-word removal, stemming, case folding, and token embedding [32, 33]. In CV, image transformation pipelines include resizing, cropping, and padding images to achieve a fixed size, as well as per-channel pixel normalization and image augmentation techniques [34]. Therefore, the set of data preparations techniques depend heavily on the problem domain, the data understanding stage, and the model selection stage. In this study, we focus on traditional preprocessing techniques for structured tabular data. The remainder of this section discusses the preprocessing steps employed by related works in the area of healthcare fraud detection.

Many of the related works provide extensive details on modeling methodologies. Ko et al. [9] model Medicare payments for Urologists as a function of the total number of patient visits using a linear regression model and the 2012 CMS Medicare Part B data. Actual payments are compared to estimated payments to identify areas of over-utilization and potential savings. In a previous study [35], we evaluate deep neural networks and various techniques for addressing class imbalance using the Part B data set. Data-level techniques for addressing class imbalance during the preprocessing stage are covered thoroughly. Bauder and Khoshgoftaar [36] compare supervised and unsupervised methods for detecting fraud within 2015 CMS Part B data and find that supervised learners perform significantly better. The authors filter the data to exclude prescription drug claims and apply manual feature selection to select a subset of features that characterize provider claims. Branting et al. [8] extract features from graph structures using 2012–2014 Part B and 2013 Part D claims data from the CMS and fraud labels from the LEIE. Behavioral similarity and geospatial co-location features are extracted from the graph and modeled with decision tree learners to classify fraud. Chandola et al. [7] explore three different approaches for detecting fraud within Medicaid and Medicare claims data: using provider-diagnosis matrices to model hidden topics, modeling provider relations with graph networks, and temporal analysis for identifying fraudulent providers. While these related works advance the state of healthcare fraud detection through machine learning, and touch on several specific preprocessing steps, we believe that they are lacking sufficient details to reliably reproduce their results.

Ekin et al. [37] provide relatively comprehensive preprocessing steps for their Medicare fraud classification study that uses a subset of the 2014 CMS Part B data. They use feature engineering to create two new predictors: provider aggressiveness and the mean provider aggressiveness for each provider type–procedure pair. The provider aggressiveness is an interaction term that is taken as the ratio of the average

submitted charge and the average payment amount. Missing values are imputed using the multivariate imputation by chained equations technique [38], and class imbalance is addressed using the synthetic minority over-sampling technique (SMOTE) [39]. The authors also enrich the Part B data set with locality features that capture patterns related to geographical variations in payments. Capelleveen et al. [21] explore fraud detection through a variety of outlier detection methods using 11 months of Medicaid dental claims. Records containing null values, zero-dollar payments, and future service dates are removed from the data set. Additional preprocessing steps include validating the number of rows and the schemas against supporting metadata, and removing duplicate records.

Our study is most closely related to the Medicare fraud classification work from Herland et al. [40]. The authors construct labeled fraud data sets from the 2012–2015 CMS Part B, Part D, and DMEPOS data sets using real-world fraud labels from the LEIE. Compared to related works, the authors provide one of the most comprehensive data understanding sections. Column normalization is used to match columns across each year of data and remove any columns that are not common across all years. Rows missing the provider NPI, procedure code, or prescription drug name are filtered from the data sets. Provider-level features that are not associated with fraud detection are removed. Data aggregation is used as a feature engineering method to increase the total number of numeric claims-level features by a factor of six. We build on this work from Herland et al. by expanding their methodologies to the latest 2019 CMS Medicare data that are available. We provide a detailed section on data understanding, review all data preprocessing steps required to curate the Part B, Part D, and DMEPOS data sets for classification, and propose an improved method for assigning fraud labels. Most importantly, we introduce three new enriched data sets with 47–58 new features, and show how this data enrichment process enables significantly better classification results.

## Medicare Data Preparation

Fraud classification data sets are created from publicly available CMS data sources. This section describes these data sources and the various preprocessing techniques employed to enable supervised learning. Topics covered include normalizing data inconsistencies, imputing missing values, fraud labeling, and data aggregation. We begin with the Summary by Provider and Service data sets that are used in related works. Next, we introduce a new Summary by Provider data source. Finally, we leverage this new data source to enrich the existing data sets from related works.

## Medicare Summary Data by Provider and Service

Each CMS Summary by Provider and Service source summarizes the utilization and payments for procedures, services, products, and prescription drugs that have been provided to Medicare fee-for-service beneficiaries by health-care professionals. A new summary is released by the CMS each year, and public use files are currently available for 2013–2019. In this section, we introduce the Part B [41], Part D [42], and DMEPOS [43] Summary by Provider and Service data sets and discuss the data preprocessing steps used to prepare them for classification.

The Summary by Provider and Service data sets include provider-level and claims-level statistics for the Part B, Part D, and DMEPOS Medicare programs. The CMS website provides both a user interface and an application programming interface to explore these data sets and perform basic search and filter queries. For our purposes, the Medicare data sets were downloaded in a comma-delimited format for further processing. For data understanding, the CMS provides methodology documentation outlining their data collection and processing procedures, along with data dictionaries that define all available attributes.

The 2013–2019 Medicare Part B Summary by Provider and Service set contains approximately 67 million records and 29 columns. Provider-level attributes include the provider's national provider identifier (NPI), provider specialty type, gender, first and last name, credentials, and geographic details. The CMS aggregates the claims each year on: (1) national provider identifier (NPI), (2) Healthcare Common Procedure Coding System (HCPCS) [44] code, and (3) the place of service. As such, the remaining claims-level attributes summarize the provider's billing activity relative to a specific HCPCS code and place of service. The place of service attribute distinguishes between services that are provided within an office (O) or a facility (F). The HCPCS code attribute includes 7738 procedure codes that identify specific procedures performed by a provider. For example: G9964 identifies a child wellness visit, V5008 identifies a hearing screening, and M1003 identifies a Tuberculosis test. Examples of claims-level attributes include the number of services performed, the number of beneficiaries seen, and the average amount charged to Medicare. A preview of this data is listed in Table 1 using a subset of columns.

The 2013–2019 Part D Summary by Provider and prescription drug describes a provider's prescription drug activity within the Medicare program over a given year and contains specific details about the drugs being prescribed. The Part D data set has approximately 172 million records and 22 columns, making it more than two times larger than the Part B data set. Many of the Part D provider-level attributes are the same as those in the Part B data set, e.g., NPI, gender, and provider type, but it does not include a place of service



**Table 1** Sample of Part B data set

NPI	Provider type	HCPSC code	Number of services	Avg. submitted charge	.
1003000142	Anesthesiology	20611	15	137.20	.
1003000142	Anesthesiology	62311	88	145.00	.
1003000142	Anesthesiology	99205	11	305.00	.
1003000142	Anesthesiology	99213	65	109.00	.
1003000142	Anesthesiology	77003	95	48.00	.

column. Since the place of service is not applicable, the CMS has aggregated these data each year by: (1) provider NPI, and (2) prescription drug name. In other words, each record describes a provider's prescription drug activity for a given year and drug name. The drug brand name attribute is a categorical attribute with 3907 unique values, and it defines the brand name of the drug being prescribed to patients by a provider. Examples of claims-level attributes include the number of beneficiaries receiving the drug, the total number of prescriptions written, and the associated costs. A sample of the Part D data is provided in Table 2 with a subset of columns.

The 2013–2019 DMEPOS Summary by Provider and Service data include aggregated data for referring providers ordering DMEPOS products and services for Medicare beneficiaries. With just over 10 million records and 34 columns, it is the smallest of the three Medicare programs. Provider-level attributes are relatively consistent with the other two data sets and include the provider's specialty type,

gender, etc. The products and services ordered by providers are encoded using the HCPSC system, similar to the Part B services. The CMS aggregates the data each year by: (1) provider NPI, (2) HCPSC procedure code, and (3) rental indicator. The rental indicator is a Boolean flag that identifies whether the DMEPOS product or service submitted on the claim is a rental or not. Therefore, the remaining claims-level attributes summarize the supplier's billing activity with respect to each HCPSC code and rental indicator. Examples of claims-level attributes include the number of suppliers rendering products for the referring provider, the total number of claims submitted by the supplier, the total number of services rendered by the supplier, and the associated costs. A sample of the DMEPOS data is provided in Table 3 with a subset of columns.

## Data Preparation

Data cleaning and preprocessing steps are applied to the Part B, Part D, and DMEPOS summary by provider and service data sets. These steps include merging years of data, normalizing columns, imputing missing values, transforming values, reconciling inconsistencies, feature selection, and removing duplicate entries.

The public use files are exported by the CMS portal in a comma-delimited format with one file per year. Therefore, the first step is to combine all years of data for each of the Part B, Part D, and DMEPOS data sets. When merging files, we maintain a year column that we will require during the fraud labeling process in “[Fraud labeling](#)”. During the merging process, we find that some column names do not match for the DMEPOS data set. For example, the DMEPOS uses the column names HCPSC\_CD and Rfrg\_Crdntls in several years, but changes the name to HCPSC\_Cd and Rfrg\_Prldr\_Crdntls in other years. These are corrected by

**Table 2** Sample of Part D data set

NPI	Provider type	Drug name	Bene count	Claim count	.
1003000126	Internal medicine	Levofloxacin	26	26	.
1003000126	Internal medicine	Lisinopril	17	19	.
1003000126	Internal medicine	Metoprolol tartrate	28	30	.
1003000126	Internal medicine	Prednisone	14	14	.

**Table 3** Sample of DMEPOS data set

NPI	Provider type	HCPSC code	# of claims	Avg. submitted charge	.
1003000126	Internal medicine	E0431	23	52.82	.
1003000126	Internal medicine	E1390	27	344.29	.
1003000407	Family practice	E0570	23	32.53	.
1003000407	Family practice	G0333	11	57.00	.
1003000407	Family practice	J7613	12	0.31	.

case folding column names and making manual corrections where necessary.

Missing values are identified and imputed for each data set using the data dictionary files from the CMS as a guide for determining missing values. Missing provider genders are imputed with a third gender, *U* for unknown, across all three data sets. The Part D and DMEPOS data sets contain missing values for the Tot\_Benes and Tot\_Suplr\_Benes columns, respectively. The CMS methodology states that these fields are left blank for all records with a value less than 10. Therefore, we impute these missing entries with the median value of 5. The remainder of the columns with missing values are ignored, e.g., provider name and geographic details, because we treat them as PII and do not include them in modeling.

Most of the values within the Medicare data sets are consistently typed, match the expected schema, and require little intervention. For the Part D data set, however, several dollar-amount columns are a mix of numeric values and strings representing dollar amounts. For example, the total drug costs column included string values like “\$1,500.20”. These are cleaned of non-numeric characters and cast to floating point numbers.

The provider type attribute is a categorical variable that describes the provider or supplier’s medical specialty, e.g., Internal Medicine, Family Medicine, Cardiology, etc. Through EDA, we observe that these values are inconsistent across multiple years of data and in some cases contain typos. For example, some years of data include the provider type of “Allergy/ Immunology”, while other years use “Allergy/Immunology”. An example of a less obvious collision is the specialty of “Certified Registered Nurse Anesthetist” that is abbreviated as “CRNA” in other years. We manually correct 26 of these inconsistencies to reduce the overall cardinality of the provider type variable, which in turn enables us to reduce the dimensionality of the final encoded data set. The Part B provider type cardinality is reduced from 127 to 102, the Part D provider type cardinality is reduced from 269 to 249, and the DMEPOS provider type cardinality is reduced from 218 to 204. We leave automated techniques for merging similar provider types as an opportunity for future work.

Provider-level attributes that can be used to identify a specific provider are excluded from the modeling stage. These include the provider name, address, geographic details, and any other PII data. These attributes describe providers, not fraudulent behavior, and including them would only increase model complexity. The resulting set of features selected for modeling are outlined in Table 4. We retain the provider NPI and year attributes, so that we can correctly map fraudulent class labels to the data set, but these are removed for training and inference.

**Table 4** Common medicare by provider and service features

Feature	Type	Part B	Part D	DMEPOS
NPI	Categorical	✓	✓	✓
Year	Categorical	✓	✓	✓
Gender	Categorical	✓		✓
Provider_type	Categorical	✓	✓	✓
Place_of_service	Categorical	✓		
Line_srvc_cnt	Numeric	✓		
Bene_unique_cnt	Numeric	✓	✓	✓
Bene_day_srvc_cnt	Numeric	✓		
Average_submitted_chrg_amt	Numeric	✓		✓
Average_medicare_payment_amt	Numeric	✓		✓
Total_claim_count	Numeric		✓	
Total_30_day_fill_count	Numeric		✓	
Total_day_supply	Numeric		✓	
Total_drug_cost	Numeric		✓	
Number_of_suppliers	Numeric			✓
Number_of_supplier_claims	Numeric			✓
Number_of_supplier_services	Numeric			✓

For the modeling stage, the Part B and Part D data sets each have 5 numeric predictors and the DMEPOS data set has 6 numeric predictors. The Part B data set has 3 categorical variables, the DMEPOS data set has 2 categorical variables, and the Part D data set has only one categorical variable. The provider type is the only common categorical variable that all three data sets share. Categorical variables are one-hot encoded during the modeling stage. We do not perform any normalization or standardization for the numeric attributes, as the XGB and RF learner are not sensitive to scaling [45]. In related works that employ learners which benefit from normalization, e.g., gradient-descent methods, we have had success using min-max scaling to bound numeric variables to a range between 0 and 1 [35].

## Feature Engineering with Data Aggregation

The three Medicare fraud data sets are large-scale data sets with 12–172 million rows each. This magnitude of data increases compute and storage requirements and compounds the class imbalance problem [4]. We address this by creating compressed representations of each data set with data aggregation. The primary motivation for this approach is to drastically reduce the size and dimensionality of each data set. While there is some risk of information loss that is caused by removing the service codes, i.e., the HCPCS code and drug name, we make up for this by introducing up to 36

new predictors that explain each provider's overall activity over a given year.

The aggregated data sets consolidate all records for a given provider and year by dropping the high-dimensional service variables and converting numeric attributes to a series of summary statistics. As an example, let us consider the subset of Part B records listed in Table 1 that contains all five records from the provider with NPI 1003000142. In the Part B data set, we aggregate over the NPI, year, provider type, and gender attributes, and then drop the HCPCS attribute. When we combine all rows for this provider, we convert each numeric attribute to six summary statistics: minimum, maximum, median, mean, sum, and standard deviation. The aggregation removes the HCPCS column, and transforms the `line_srv_count` attribute to `line_srv_count_min`, `line_srv_count_max`, `line_srv_count_median`, `line_srv_count_mean`, `line_srv_count_sum`, and `line_srv_count_std` values. This transformation is applied to all numeric attributes, and when there exists only one row for a given provider, then the standard deviation is imputed with 0. These summary statistics describe the provider's overall billing activity relative to all HCPCS procedure codes claimed that year. For the Part B data set, this increases the total number of numeric attributes from 5 to 30.

The same procedure is applied equally to the Part D and DMEPOS data sets. For the Part D data set, we aggregate over the NPI, year, and provider type attributes, remove the drug name attribute, and convert the remaining numeric attributes to their summary statistics. In the DMEPOS data set, we aggregate over the NPI, year, gender, and provider type attributes, remove the HCPCS attribute, and compute summary statistics for remaining numeric attributes. For both data sets, this aggregation removes the high-dimensional categorical attributes of HCPCS code and drug name. The total number of numeric attributes in the Part D and DMEPOS data sets is increased to 30 and 36, respectively. This feature engineering step effectively reduces the size of each data set by an order of magnitude, and introduces new features that capture each provider's overall billing behaviors. For example, the Part D data set is reduced from 173 million rows and 7 features to 6 million rows and 31 features, and reduces its memory footprint by more than 80%.

## Enriching Medicare Data Sets

We introduce three new data sets by leveraging the CMS Medicare Summary by Provider data sets for the Part B [46], Part D [47], and DMEPOS [48] Medicare programs. Similar to the Summary by Provider and Service data, the CMS makes the Summary by Provider data publicly available on their website for years 2013–2019. These data sets do not include statistics at the procedure (HCPCS) or drug level. Instead, they include summary data that describes all of a

provider's services and beneficiaries over a given year. As such, it is similar to our Aggregated data sets that we have manually derived, i.e., with one record for each NPI and year combination. We use the CMS Summary by Provider data sets to enrich our Aggregated data sets with new features and create the Aggregated-Enriched Medicare data sets.

The Part B Summary by Provider data enables us to enrich our Part B Aggregated data sets with 47 new features for analysis and modeling. The Summary by Provider data includes 7 features which describe the provider's overall billing activity across all procedure codes. These include totals for the number of HCPCS codes billed, the number of beneficiaries seen, the number of services performed, the dollar amount submitted to Medicare for reimbursement, the Medicare allowed amount, the Medicare payment amount, and a standardized total Medicare payment amount. The standardized total Medicare payment amount is calculated by removing the differences in payments based on geographic region to make prices more comparable across regions. In addition to these 7 features, the CMS provides another 14 features by breaking these service statistics down into medical and drug-related services. The totals within the drug category capture the statistics related to a subset of drug services that are covered within the Medicare Part B program. Next, the Part B Summary by Provider data introduces features related to the beneficiaries that are receiving the services. The total number of beneficiaries seen are broken down into four categories: less than 65, 65–74, 75–84, and greater than 84. These features, plus their average, account for 5 new features. The total number of male and female beneficiaries and the racial frequency of beneficiaries account for 8 new features. The total number of beneficiaries covered by both Medicare and Medicaid, and the total number of beneficiaries covered by Medicare only provide 2 new features. The next 18 features are defined as the percentage of patients with a particular chronic condition (CC). Some examples of chronic conditions included Alzheimer's, asthma, cancer, and kidney disease. Finally, the CMS provides an average beneficiary risk score using a risk-adjustment model based on hierarchical condition categories (HCC). According to the CMS methodology, beneficiaries with above-average risk scores ( $> 1.08$ ) are estimated to have above-average Medicare spending. We exclude the beneficiary-race statistics, as the majority of the records have missing values, yielding a total of 47 new numeric predictors.

The Part D Summary by Provider data introduces 51 new features to our Part D Aggregated data set. Similar to the Part B Summary by Provider data, the first 10 features capture various beneficiary summary statistics. These include the total number of beneficiaries prescribed for, broken down by their age group, gender, and Medicare/Medicaid membership. The average HCC risk score is also included for each provider. The Part D Summary by Provider data does not

include the chronic condition statistics that are included in the Part B Summary by Provider data, but it does provide more features describing claims subtotals within categories. Claims data are summarized for each provider using the total number of claims, total number of 30-day prescription orders, total drug cost, total day's supply dispensed, and the total number of beneficiaries seen. These summary statistics are then broken down to provide subtotals within the categories: Medicare advantage prescription drug plan (MAPD) coverage, Medicare prescription drug plan (PDP) coverage, and low income subsidy (LIS). The statistics are also broken down by several drug categories: opiate drug claims, long-acting opiate drug claims, antibiotic drug claims, and anti-psychotic drug claims.

Similar to the other Summary by Provider data sets, the DMEPOS Summary by Provider data include beneficiary demographics data, Medicare versus Medicaid coverage of beneficiaries, and the average beneficiary HCC risk score. The DMEPOS also includes 16 CC features, similar to the Part B Summary by Provider data, denoting the percentage of beneficiaries treated that have been diagnosed with the CC. The 8 claims data statistics include totals for the number of suppliers, number of HCPCS codes, number of beneficiaries, number of claims, number of services, dollar amount submitted to Medicare, dollar amount paid by Medicare, and the allowed Medicare payment. The claims totals are then broken down into categories to provide a more granular view: durable medical equipment (DME), prosthetic and orthotic specific (POS), and drug-related services. In total, the DMEPOS Summary by Provider data introduces 58 new numeric features that we can use to enrich the DMEPOS Aggregated data set.

Minimal preprocessing is required for the Summary by Provider data sets. Like the Summary by Provider and Service data sets, several dollar-amount columns are cleaned of non-numeric characters and cast to floating point numbers. Missing values are imputed with 0 and rows missing NPI values are removed. The beneficiary-race statistics and the suppression flag features are not included in the final data sets, because they are missing values for more than 50% of the records. All attributes introduced by the supplementary data sources are numeric and did not require feature normalization.

We perform an inner join on the NPI and year attributes for the Aggregated data sets and the processed Summary by Provider data sets for each Medicare program. This feature engineering step more than doubles the total number of features in each data set. The Part B and DMEPOS data sets provided clean joins with no data loss, but the Part D data set lost approximately 1 million records during this join, because the Summary by Provider data set did not include data for all providers and years. We refer to these new Medicare data sets as the aggregated-enriched data sets. To the

**Table 5** Sample of LEIE data

Specialty	NPI	Excltype	Excldate
Podiatry practice	1598041998	1128a1	20190320
Pharmacy	1275750374	1128a1	20190320
Internal medicine	1477537496	1128b4	20140520

**Table 6** Fraud-related LEIE rules [49]

Social security act	Description
1128(a)(1)	Conviction of program-related crimes
1128(a)(2)	Conviction relating to patient abuse or neglect
1128(a)(3)	Felony conviction relating to health care fraud
1128(b)(4)	License revocation, suspension, or surrender
1128(b)(7)	Fraud, kickbacks, and other prohibited activities
1128(c)(3)(g)(i)	Conviction of second mandatory exclusion offenses
1128(c)(3)(g)(ii)	Conviction of third mandatory exclusion offenses

best of our knowledge, this is the first time that these additional data sets have been used to enrich Medicare claims data for fraud detection. We run classification experiments with the Aggregated and Aggregated-Enriched data sets to determine if the new predictors are able to significantly improve fraud detection rates.

## Fraud Labeling

The Medicare data sets from the CMS do not include fraudulent or non-fraudulent labels. Instead, we identify real-world fraudulent providers using the publicly available LEIE data set [23] and use these to label the CMS data. The LEIE is maintained by the Office of Inspector General (OIG) in accordance with Sections 1128 and 1156 of the Social Security Act [49] and is updated on a monthly basis. The OIG has the authority to exclude providers from Federally funded health care programs for a variety of reasons. Excluded individuals are unable to receive payment from Federal healthcare programs for any services, and must apply for reinstatement once their exclusion period has been satisfied. The current LEIE data format contains 18 attributes that describe the provider and the reason for their exclusion. Table 5 provides a sample of the LEIE data set. Some additional attributes not listed include first and last name, date of birth, address, and the provider's reinstatement date.

The LEIE exclusion type attribute is a categorical value that describes the offense and its severity. Following the work by Bauder and Khoshgoftaar [50], a subset of exclusion rules that are indicative of fraud are selected for labeling Medicare providers. Table 6 lists the exclusion rules used in this study. We use the NPI numbers of excluded individuals



**Table 7** Aggregated medicare data summary

Dataset	Records	Positive count	Positive ratio	Feature count
Part B	8,669,497	3954	0.0456%	33
Part D	6,262,115	4064	0.0649%	31
DMEPOS	2,056,075	1809	0.0880%	38

**Table 8** Aggregated-enriched medicare data summary

Dataset	Records	Positive count	Positive ratio	Feature count
Part B	8,669,497	3954	0.0456%	80
Part D	5,344,106	3700	0.0692%	82
DMEPOS	2,056,075	1809	0.0880%	96

that have been convicted under one of these rules to identify fraudulent providers within each of the Medicare data sets. For each excluded provider, we compute their last exclusion year to be the exclusion year plus the minimum exclusion period, rounded to the nearest year. For the providers within the Medicare data sets whose NPI number matches those of the LEIE data set, claims that are dated prior to the provider's last exclusion year are labeled as fraudulent. This works under the assumption that the claims made prior to the exclusion data, or during the exclusion period, are representative of fraudulent claims activity.

Previous works have always used the latest LEIE data set to obtain a list of fraudulent providers. This risks introducing false negatives, however, because the LEIE database only contains exclusions that are currently in effect. If a provider was fraudulent in previous years and then reinstated, this approach would label the provider's claims as non-fraudulent. This adds noise to the non-fraudulent class and reduces the total number of fraudulent providers. As such, we expect this shortcoming to hurt classification performance. We refer to this fraud labeling approach as the Active Exclusion List, and propose a new approach for labeling that mitigates this class label noise and increases the total number of fraudulent providers.

We leverage an Internet Archive tool [51] to obtain the active LEIE data set for each year of Medicare data and create a Cumulative Exclusion List. More specifically, we obtain the active LEIE exclusion lists for the first month of each year 2013–2019. We then concatenate each year of provider exclusions and use the exclusion end year from the most recent record to identify fraudulent labels within the Medicare data. This increases the total number of known fraudulent providers from 4355 to 5046, i.e., a 15% increase, and improves data quality by reducing the mislabeled instances.

Table 7 and Table 8 provide a summary of the Aggregated and Aggregated-Enriched data sets. All data sets are

characterized by high levels of class imbalance, with an overwhelming number of non-fraudulent samples and a rare fraudulent class sizes ranging from 0.0456 to 0.0880%. The Aggregated data sets have between 31 and 38 features, and the Aggregated-Enriched data sets have between 80 and 96 features.

## Experiment Design

All data exploration, preparation, and experiments are executed using the Python data science stack [52] on high-performance compute machines with sufficient memory and CPU to avoid the complexities of distributed systems. Data sets are evaluated on the fraud classification task using the RF and XGB learners. The RF algorithm is trained using the scikit-learn package [53] and the XGB algorithm is trained using its Python implementation [54]. Hyperparameters are identified for each learner during preliminary experiments. A maximum depth of 4 and 16 is used for the XGB and RF learners, respectively, and all remaining hyperparameters are left as their default values.

Six runs of fivefold cross-validation are performed to produce a total of 30 results for each learner and data set combination. We use an adjusted k-fold cross-validation technique to provide a fair and realistic performance evaluation. We refer to this as k-fold-by-npi cross-validation, because it ensures that providers do not co-exist in both the train and test folds. We find this necessary, because learners can overfit to specific providers and memorize specific characteristics of providers, instead of learning a general fraud distribution.

We report performance using the AUC, AUPRC, TPR, and TNR metrics over 30 repetitions. The AUC metric summarizes the trade-off between the TPR and the false-positive rate (FPR), and it is a popular metric for class-imbalanced problems that is often used in related works [55]. When reporting TPR and TNR scores, we use the prior probability of the positive class as a decision threshold to assign class labels to model output probabilities [56, 57]. Finally, we analyze feature contributions using the average of the absolute Shapley additive explanations (SHAP) values for each feature. The SHAP values measure the marginal contribution that each feature makes towards the model's output [58]. We use SHAP values, instead of each learner's built-in feature importance metrics, because they directly quantify the contribution of each feature [59].

## Experiment Results

We evaluate the aggregated, summary by provider, and aggregated-enriched data sets on the Medicare fraud classification task. This section begins with a discussion on fair

model evaluation. Next, we review the performance of each data set using the XGB and RF learners. Finally, we analyze the feature contributions of each data source using SHAP values.

### Fair Evaluation with K-Fold-by-NPI Cross-validation

During preliminary experiments, we explored the *max-depth* hyperparameter of the XGB learner to maximize classification performance. These results show that increasing the max-depth of the XGB learner's decision trees provides consistently better AUC scores. After evaluating the XGB learner with an unlimited tree depth and obtaining near-perfect AUC > 0.99, we decided to reconsider the Medicare data and evaluation strategy.

Recall that the Aggregated Medicare data sets contain one or more rows for each provider. If a provider has submitted claims for years 2013–2017, and they were flagged as fraudulent in 2017 or later, then the data set will include five records for this provider, i.e., one record for each year. Using the standard k-fold cross-validation, or train-test split, as has been done in related works, it is possible for a provider to exist in both the training and evaluation data sets. Furthermore, intuition suggests a provider is likely to have somewhat similar billing patterns over these years, e.g., the same provider type, similar billing frequencies, and similar beneficiary populations. Given the rich feature set that we have introduced in this study, it is possible that we are not modeling general claims fraud, but instead are modeling specific providers through de-anonymization. We acknowledge that this is an unfair evaluation, and most importantly, it is not appropriate for real-world applications that need to make inferences on new providers that were not observed during model training. We address this using a k-fold-by-npi cross-validation technique that selects random providers for the training and evaluation folds. This ensures that for a given iteration of cross-validation, a provider will not exist in both the train and evaluation folds.

The AUC scores obtained using both k-fold evaluation techniques are compared in Fig. 2 using the Part B

Aggregated-Enriched data set. These results show that a maximum depth of 8 is all that is required to fit the entire training set. The traditional k-fold cross-validation obtains the minimum validation AUC at a maximum depth of 2 and the maximum AUC of approximately 0.99 at a maximum depth of 32. In other words, the AUC performance increases monotonically with the maximum tree depth parameter when using the traditional k-fold cross-validation.

When we apply the k-fold-by-npi evaluation technique, we obtain a maximum AUC of 0.9536 using a maximum tree depth of 4. Unlike the traditional k-fold cross-validation, the k-fold-by-npi validation performance decreases consistently as the tree depths increase past 4. The k-fold-by-npi exhibits the expected bias-variance trade-off, where increasing the capacity of the learner beyond some threshold leads to overfitting and degrading validation performance. Therefore, we can conclude that the traditional k-fold cross-validation provides unreliable, optimistic classification performance on the validation set. We use the k-fold-by-npi throughout the remainder of our work and recommend that this technique be adopted by future works as it provides a fair evaluation that reflects fraud detection systems in practice.

The problem of over-optimistic cross-validation that we observe in this study is not unique to healthcare fraud detection. Fortunately, the k-fold-by-npi can easily be extended to other problem domains. In the more general sense, we refer to this as the k-fold-by-entity cross-validation technique, and its purpose is to ensure that entities do not co-exist within the training and validation folds.

### Data Set Performance

Table 9 compares the performance of the newly curated summary by provider (SbP) data, the original aggregated (A) data that is used in related works, and the new aggregated-enriched (AE) data set. The aggregated-enriched data set is the result of joining the new summary by provider features with the original Aggregated data set on the provider NPI. The average AUC and 95% confidence intervals (C.I.) are reported and the confidence intervals listed in bold font

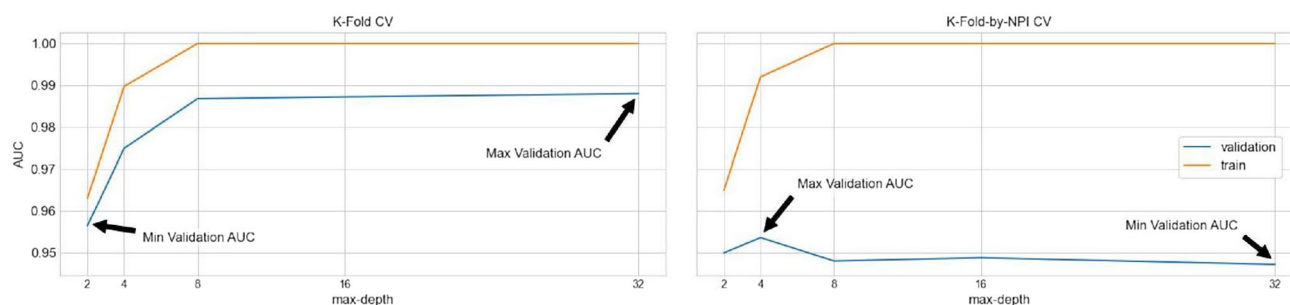


Fig. 2 Comparing cross-validation techniques

**Table 9** AUC performance results

Data set	XGB learner		RF learner	
	Mean	95% C.I	Mean	95% C.I
Part B SbP	0.8310	(0.8256, 0.8365)	0.7943	(0.7895, 0.7991)
Part B A	0.9392	(0.9345, 0.9439)	0.8245	(0.8196, 0.8295)
Part B AE	0.9536	<b>(0.9497, 0.9575)</b>	0.8717	<b>(0.8660, 0.8774)</b>
Part D SbP	0.8086	(0.8043, 0.8129)	0.7488	(0.7433, 0.7544)
Part D A	0.9376	(0.9343, 0.9409)	0.7548	(0.7383, 0.7532)
Part D AE	0.9580	<b>(0.9533, 0.9627)</b>	0.8022	<b>(0.7949, 0.8096)</b>
DMEPOS SbP	0.7856	(0.7799, 0.7892)	0.7435	(0.7373, 0.7497)
DMEPOS A	0.9400	(0.9341, 0.9447)	0.8213	(0.8130, 0.8296)
DMEPOS AE	0.9495	(0.9447, 0.9547)	0.8376	<b>(0.8306, 0.8446)</b>
Average SbP	0.8080	(0.7997, 0.8164)	0.7589	(0.7496, 0.7681)
Average A	0.9387	(0.9345, 0.9428)	0.7972	(0.7824, 0.8120)
Average AE	0.9528	<b>(0.9485, 0.9571)</b>	0.8372	<b>(0.8250, 0.8493)</b>

indicate performance that is significantly better according to non-overlapping intervals.

Across all Medicare programs and learners, the Summary by Provider data set obtains significantly lower AUC scores than the original Aggregated data set that is used in related works. For example, the average Summary by Provider AUC score for the XGB learner is 0.8080, while the Aggregated data set obtains an average AUC score of 0.9387. Similarly for the RF learner, the Summary by Provider data obtains an average AUC of 0.7589 that is significantly lower than the Aggregated data set's average AUC of 0.7972. Despite the larger feature space of the Summary by Provider data set, it is clear that it does not contain sufficient information to perform competitively with existing data sets on the fraud detection task.

Next, we consider the effect of enriching the Aggregated data set with the Summary by Provider features. On average, the Aggregated-Enriched data sets obtain an AUC interval of 0.9485–0.9571 using the XGB learner and an AUC interval of 0.8250–0.8493 using the RF learner. These average Aggregated-Enriched AUC results are significantly greater than the AUC results of the original Aggregated data set. While the summary by provider data sets did not perform well independently, these results show that they are very effective when used to enrich existing data sources. We can conclude that the newly introduced provider-level, claims-level, and beneficiary-level features consistently improve the AUC performance across three Medicare programs and two machine learning algorithms.

Table 10 lists the XGB classification performance of the aggregated-enriched data set using the TPR, TNR, and G-Mean performance metrics. The percentages listed in parenthesis include the performance lift compared to the original Aggregated data from related works. We focus specifically on the XGB learner, because it consistently

**Table 10** Aggregated-enriched classification improvements with XGB learner

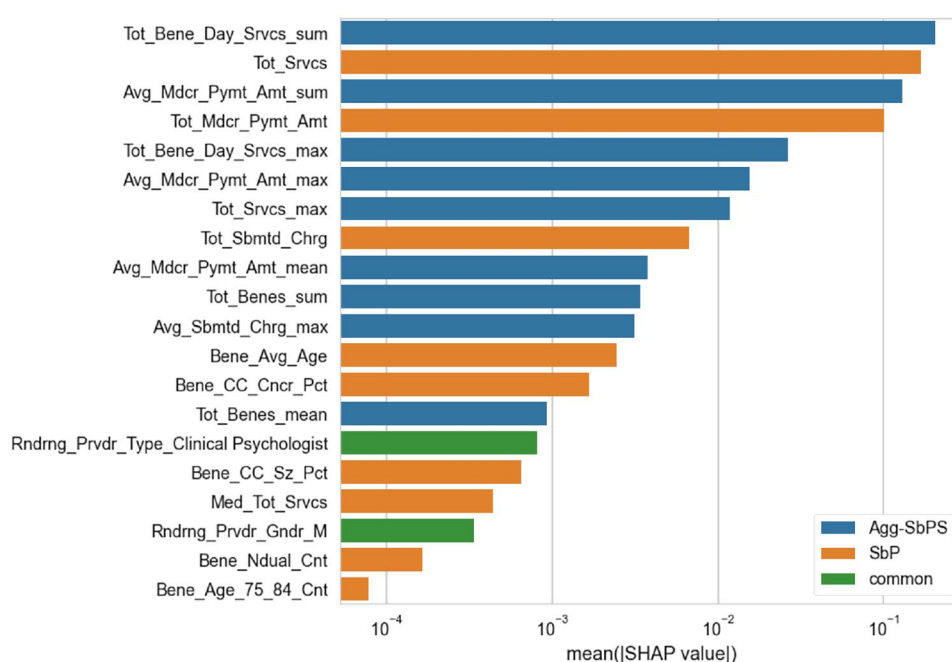
	TPR	TNR	G-mean
Part B	0.8484 (+2.23%)	0.9572 (+1.48%)	0.9011 (+1.88%)
Part D	0.8619 (+4.69%)	0.9678 (+2.77%)	0.9132 (+3.73%)
DMEPOS	0.8551 (+0.75%)	0.9687 (+0.60%)	0.9099 (+0.68%)
Average	0.8551 (+2.56%)	0.9646 (+1.63%)	0.9081 (+2.09%)

outperforms the RF learner according to AUC results. Due to the high class imbalance, the prior probability of the positive class is used as the decision threshold for assigning class predictions to model probability estimates based on related works [60].

In all cases, the Aggregated-Enriched data sets perform better than the Aggregated data sets according to the TPR, TNR, and G-Mean scores. The Part D data set sees the greatest increases to classification performance, with a 4.69% and 2.77% increase to the TPR and TNR scores, respectively. The DMEPOS data set sees the smallest increase to classification performance, with less than a 1% increase to the TPR and TNR scores. When comparing class-wise performance, the TPR performance is improved more than the TNR performance when using the enriched data sets. Overall, we can conclude that enriching the Medicare fraud data sets with new provider-level, claims-level, and beneficiary-level predictors provides meaningful improvements to all classification metrics reported in this study.

## Feature Contributions

To better understand which features contribute to the XGB performance, we illustrate the average SHAP value for the top 20 features of the combined Aggregated-Enriched data set in Fig. 3. Both the original Aggregated and the new Summary by Provider data set contribute 9 features to the top 20 most important features. The remaining two features are common across both data sets, i.e., single attributes from the gender and provider type one-hot vectors. Key predictors of Medicare fraud, with SHAP values  $\geq 0.1$ , include the total number of beneficiary services per day, the total number of services provided over the year, the average Medicare payment amount, and the total Medicare payment amount. Other interesting predictors that are introduced by the new Summary by Provider data set include beneficiary age and chronic condition statistics. These results show that the new Summary By Provider features make an approximately equal contribution to the model outputs and positively influence prediction results.

**Fig. 3** Source of top 20 features

## Conclusion

This study presents a data-centric approach to improve healthcare fraud classification results using publicly available claims data from the CMS. We utilize publicly available data from the CMS to introduce six new labeled Medicare data sets, i.e., the Summary by Provider and the Aggregated-Enriched data sets for each Medicare program. These data sets include new provider-level, claims-level, and beneficiary-level statistics for each provider participating in the Medicare program. We begin with a comprehensive review of existing data sources and data preparation techniques, e.g., data preprocessing, aggregation, enrichment, and labeling. Next, we leverage Internet archives to obtain historical lists of fraudulent Medicare providers through the LEIE website and construct cumulative exclusion lists of providers that increase the size of the fraudulent class by 15%. We demonstrate how these new feature-rich data sets can yield misleading classification results, and correct for this using an improved cross-validation technique that better reflects real-world fraud detection systems.

Multiple performance metrics and 95% confidence intervals show that the newly proposed Summary by Provider features are good indicators of healthcare fraud, especially when they are used to enrich existing data sets. AUC, TPR, TNR, and G-mean results show that the Aggregated-Enriched data set performs significantly better than the Aggregated data set used frequently in related works. Feature contribution measures are used to support these findings, indicating that the new Summary by Provider feature set contributes to 9 of the top 20 most

important features in the Aggregated-Enriched data set. In future works, we plan to explore additional feature engineering techniques for the provider summary data set, e.g., feature interactions. A comprehensive analysis of feature importance is also recommended for future works, as model interpretability is a critical component of healthcare fraud detection.

**Author Contributions** JMJ performed the literature review, executed the experiment design, and drafted the manuscript. TMK worked with JMJ to develop the article's framework and focus. All authors have read and approved the final manuscript.

**Funding** Not applicable.

## Declarations

**Conflict of Interest** The authors declare that they have no competing interests.

## References

1. Morris L. Combating fraud in health care: an essential component of any cost containment strategy. *Health Aff (Project Hope)*. 2009;28:1351–6. <https://doi.org/10.1377/hlthaff.28.5.1351>.
2. Bauder RA, Khoshgoftaar TM, Seliya N. A survey on the state of healthcare upcoding fraud analysis and detection. *Health Serv Outcomes Res Method*. 2017;17(1):31–55. <https://doi.org/10.1007/s10742-016-0154-8>.
3. The United States Department of Justice: Health Care Fraud Unit. <https://www.justice.gov/criminal-fraud/health-care-fraud-unit>. Accessed 1 Oct 2022.



4. Leevy JL, Khoshgoftaar TM, Bauder RA, Seliya N. A survey on addressing high-class imbalance in big data. *J Big Data*. 2018;5(1):42. <https://doi.org/10.1186/s40537-018-0151-6>.
5. Johnson JM, Khoshgoftaar TM. Encoding techniques for high-cardinality features and ensemble learners. In: 2021 IEEE 22nd international conference on information reuse and integration for data science (IRI). 2021. p. 355–361. <https://doi.org/10.1109/IRI51335.2021.00055>.
6. Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C, Wirth R. CRISP-DM 1.0: step-by-step data mining guide. CRISP-DM consortium: NCR Systems Engineering Copenhagen (USA and Denmark) DaimlerChrysler AG (Germany), SPSS Inc. (USA) and OHRA Verzekeringen en Bank Groep B.V (The Netherlands), 2000; CRISP-DM consortium: NCR Systems Engineering Copenhagen (USA and Denmark) DaimlerChrysler AG (Germany), SPSS Inc. (USA) and OHRA Verzekeringen en Bank Groep B.V (The Netherlands).
7. Chandola V, Sukumar SR, Schryver JC. Knowledge discovery from massive healthcare claims data. In: KD. 2013.
8. Branting LK, Reeder F, Gold J, Champney T. Graph analytics for healthcare fraud risk estimation. In: 2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM), 2016. p. 845–851. <https://doi.org/10.1109/ASONAM.2016.7752336>.
9. Ko J, Chalfin H, Trock B, Feng Z, Humphreys E, Park S-W, Carter B, D Frick K, Han M. Variability in medicare utilization and payment among urologists. *Urology* 2015;85. <https://doi.org/10.1016/j.urology.2014.11.054>.
10. Frye M, Schmitt RH. Structured data preparation pipeline for machine learning-applications in pro-duction. In: 17th IMEKO TC 10. 2020. p. 241–246.
11. Kotsiantis SB, Kanellopoulos D, Pintelas PE. Data preprocessing for supervised learning. *Int J Comput Inf Eng*. 2007;1(12):4104–9.
12. Data-Centric AI: Data-Centric AI Resource Hub. <https://datacentricai.org/>. Accessed 1 Mar 2022.
13. Goel N, Faltings B. Crowdsourcing with fairness, diversity and budget constraints. In: Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society, AIES '19. New York: Association for Computing Machinery; 2019. p. 297–304. <https://doi.org/10.1145/3306618.3314282>.
14. Pineau J, Vincent-Lamarre P, Sinha K, Larivière V, Beygelzimer A, d'Alché-Buc F, Fox E, Larochelle H. Improving reproducibility in machine learning research: a report from the neurips 2019 reproducibility program. *J Mach Learn Res*. 2021;22.
15. McCausland T. The bad data problem. *Res Technol Manag*. 2021;64(1):68–71. <https://doi.org/10.1080/08956308.2021.1844540>.
16. Chandrasekar P, Qian K, Shahriar H, Bhattacharya P. Improving the prediction accuracy of decision tree mining with data preprocessing. In: 2017 IEEE 41st annual computer software and applications conference (COMPSAC), vol 2. IEEE; 2017. p. 481–484.
17. Benhar H, Idri A, Fernández-Alemán J. Data preprocessing for heart disease classification: a systematic literature review. *Comput Methods Programs Biomed*. 2020;195: 105635.
18. HaCohen-Kerner Y, Miller D, Yigal Y. The influence of preprocessing on text classification using a bag-of-words representation. *PLoS One*. 2020;15(5):0232525.
19. U.S. Government, U.S. Centers for Medicare & Medicaid Services: The Official U.S. Government Site for Medicare. <https://www.medicare.gov/>. Accessed 1 Feb 2021.
20. Centers For Medicare & Medicaid Services: Trustees Report & Trust Funds. <https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/ReportsTrustFunds/index.html>. Accessed 15 Dec 2021.
21. Van Capelleveen G, Poel M, Mueller RM, Thornton D, Van Hillegersberg J. Outlier detection in healthcare fraud: a case study in the medicaid dental domain. *Int J Account Inf Syst*. 2016;21:18–31. <https://doi.org/10.1016/j.accinf.2016.04.001>.
22. Centers For Medicare & Medicaid Services: Provider Summary by Type of Service. <https://data.cms.gov/provider-summary-by-type-of-service>. Accessed 1 Jan 2022.
23. U.S. Department of Health and Human Services Office of Inspector General: LEIE Downloadable Databases. [https://oig.hhs.gov/exclusions/exclusions\\_list.asp](https://oig.hhs.gov/exclusions/exclusions_list.asp). Accessed 1 Jan 2022.
24. Chen T, Guestrin C. Xgboost. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016. <https://doi.org/10.1145/2939672.2939785>.
25. Witten IH, Frank E, Hall MA. Data mining: practical machine learning tools and techniques, 3rd edn. Morgan Kaufmann Series in Data Management Systems. Amsterdam: Morgan Kaufmann; 2011. <http://www.sciencedirect.com/science/book/9780123748560>.
26. Shmueli G, Bruce PC, Patel NR. Data Mining for business analytics: concepts, techniques, and applications with XLMiner. 3rd ed. Hoboken: Wiley Publishing; 2016.
27. Zhou L, Pan S, Wang J, Vasilakos AV. Machine learning on big data: opportunities and challenges. *Neurocomputing*. 2017;237:350–61. <https://doi.org/10.1016/j.neucom.2017.01.026>.
28. Hariharakrishnan J, Mohanavalli S, Srividya Sundhara Kumar KB. Survey of pre-processing techniques for mining big data. In: 2017 international conference on computer, communication and signal processing (ICCCSP). 2017. p. 1–5. <https://doi.org/10.1109/ICCCSP.2017.7944072>.
29. Goodfellow IJ, Bengio Y, Courville A. Deep learning. Cambridge: MIT Press; 2016. <http://www.deeplearningbook.org>.
30. Geron A. Hands-on machine learning with Scikit-learn and TensorFlow : concepts, tools, and techniques to build intelligent systems. Sebastopol: O'Reilly Media; 2017.
31. Manning CD, Raghavan P, Schütze H. Introduction to Information Retrieval. Cambridge: Cambridge University Press; 2008. <http://nlp.stanford.edu/IR-book/information-retrieval-book.html>.
32. Kadhimi AI. An evaluation of preprocessing techniques for text classification. *Int J Comput Sci Inf Secur (IJCSIS)*. 2018;16(6):22–32.
33. Mikolov T, Chen K, Corrado GS, Dean J. Efficient estimation of word representations in vector space. 2013. [arXiv:1301.3781](https://arxiv.org/abs/1301.3781) [CoRR abs].
34. Barnouti NH, et al. Improve face recognition rate using different image pre-processing techniques. *Am J Eng Res (AJER)*. 2016;5(4):46–53.
35. Johnson JM, Khoshgoftaar TM. Medicare fraud detection using neural networks. *J Big Data*. 2019;6(1):63. <https://doi.org/10.1186/s40537-019-0225-0>.
36. Bauder RA, Khoshgoftaar TM. Medicare fraud detection using machine learning methods. In: 2017 16th IEEE international conference on machine learning and applications (ICMLA). 2017. p. 858–865. <https://doi.org/10.1109/ICMLA.2017.00-48>.
37. Ekin T, Frigau L, Conversano C. Health care fraud classifiers in practice. *Applied stochastic models in business and industry*. 2021. <https://doi.org/10.1002/asmb.2633>.
38. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? *Int J Methods Psychiatr Res*. 2011;20(1):40–9. <https://doi.org/10.1002/mpr.329>.
39. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. Smote: synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16(1):321–57.
40. Herland M, Khoshgoftaar TM, Bauder RA. Big data fraud detection using multiple medicare data sources. *J Big Data*. 2018;5(1):29. <https://doi.org/10.1186/s40537-018-0138-3>.
41. Centers For Medicare & Medicaid Services: Medicare Physician & Other Practitioners - by Provider and Service. <https://data.cms.gov/provider-summary-by-type-of-service/medicare-physi>

- cian-other-practitioners/medicare-physician-other-practitioners-by-provider-and-service. Accessed 1 Jan 2022.
42. Centers For Medicare & Medicaid Services: Medicare Part D Prescribers - by Provider and Drug. <https://data.cms.gov/provider-summary-by-type-of-service/medicare-part-d-prescribers/medicare-part-d-prescribers-by-provider-and-drug>. Accessed 1 Jan 2022.
  43. Centers For Medicare & Medicaid Services: Medicare Durable Medical Equipment, Devices & Supplies - by Referring Provider and Service. <https://data.cms.gov/provider-summary-by-type-of-service/medicare-durable-medical-equipment-devices-supplies/medicare-durable-medical-equipment-devices-supplies-by-referring-provider-and-service>. Accessed 1 Jan 2022.
  44. Centers For Medicare & Medicaid Services: HCPCS General Information. <https://www.cms.gov/Medicare/Coding/MedHCPCSGenInfo/index.html>. Accessed 1 Jan 2022.
  45. Wang W, Zhang X, Gombault S, Knapiskog SJ. Attribute normalization in network intrusion detection. In: 2009 10th international symposium on pervasive systems, algorithms, and networks. 2009. p. 448–453. <https://doi.org/10.1109/I-SPAN.2009.49>.
  46. Centers For Medicare & Medicaid Services: Medicare Physician & Other Practitioners - by Provider. <https://data.cms.gov/provider-summary-by-type-of-service/medicare-physician-other-practitioners/medicare-physician-other-practitioners-by-provider>. Accessed 1 Jan 2022.
  47. Centers For Medicare & Medicaid Services: Medicare Part D Prescribers - by Provider. <https://data.cms.gov/provider-summary-by-type-of-service/medicare-part-d-prescribers/medicare-part-d-prescribers-by-provider>. Accessed 1 Jan 2022.
  48. Centers For Medicare & Medicaid Services: Medicare Durable Medical Equipment, Devices & Supplies - by Referring Provider. <https://data.cms.gov/provider-summary-by-type-of-service/medicare-durable-medical-equipment-devices-supplies/medicare-durable-medical-equipment-devices-supplies-by-referring-provider>. Accessed 1 Jan 2022.
  49. U.S. Department of Health and Human Services Office of Inspector General: Exclusion Authorities. <https://oig.hhs.gov/exclusions/authorities.asp>. Accessed 1 Jan 2022.
  50. Bauder RA, Khoshgoftaar TM. A novel method for fraudulent medicare claims detection from expected payment deviations (application paper). In: 2016 IEEE 17th international conference on information reuse and integration (IRI). 2016. p. 11–19. <https://doi.org/10.1109/IRI.2016.11>.
  51. Internet Archive: Wayback Machine. <https://archive.org/web/>. Accessed 1 Jan 2022.
  52. Oliphant T. Python for scientific computing. *Comput Sci Eng*. 2007;9:10–20. <https://doi.org/10.1109/MCSE.2007.58>.
  53. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825–30.
  54. Chen T, Guestrin C. Xgboost. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016. <https://doi.org/10.1145/2939672.2939785>.
  55. Provost F, Fawcett T. Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions. In: Proceedings of the third international conference on knowledge discovery and data mining. 1999. p. 43–48.
  56. Johnson JM. In: Wani MA, Khoshgoftaar TM, Palade V editors, Thresholding strategies for deep learning with highly imbalanced big data. Singapore: Springer; 2021. p. 199–227. [https://doi.org/10.1007/978-981-15-6759-9\\_9](https://doi.org/10.1007/978-981-15-6759-9_9).
  57. Johnson JM, Khoshgoftaar TM. Output thresholding for ensemble learners and imbalanced big data. In: 2021 IEEE 33rd international conference on tools with artificial intelligence (ICTAI). 2021. p. 1449–1454. <https://doi.org/10.1109/ICTAI52525.2021.00230>.
  58. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R editors. Advances in neural information processing systems, vol 30. New York: Curran Associates, Inc.; 2017. p. 4765–4774. <http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf>.
  59. Meng Y, Yang N, Qian Z, Zhang G. What makes an online review more helpful: an interpretation framework using xgboost and shap values. *J Theor Appl Electron Commer Res*. 2021;16(3):466–90. <https://doi.org/10.3390/jtaer16030029>.
  60. Silva EJR, Zanchettin C. On the existence of a threshold in class imbalance problems. In: 2015 IEEE international conference on systems, man, and cybernetics. 2015. p. 2714–2719. <https://doi.org/10.1109/SMC.2015.474>.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.