Medicare Fraud Detection using Machine Learning Methods

Richard A. Bauder
College of Engineering & Computer Science
Florida Atlantic University
Boca Raton, Florida, USA
rbauder2014@fau.edu

Taghi M. Khoshgoftaar
College of Engineering & Computer Science
Florida Atlantic University
Boca Raton, Florida, USA
khoshgof@fau.edu

***Abstract*—**

Healthcare is an integral component in people's lives, especially for the rising elderly population, and must be affordable. Medicare is one such healthcare program. Claims fraud is a major contributor to increased healthcare costs, but its impact can be lessened through fraud detection. In this paper, we compare several machine learning methods to detect Medicare fraud. We perform a comparative study with supervised, unsupervised, and hybrid machine learning approaches using four performance metrics and class imbalance reduction via oversampling and an 80-20 undersampling method. We group the 2015 Medicare data into provider types, with fraud labels from the List of Excluded Individuals/Entities database. Our results show that the successful detection of fraudulent providers is possible, with the 80-20 sampling method demonstrating the best performance across the learners. Furthermore, supervised methods performed better than unsupervised or hybrid methods, but these results varied based on the class imbalance sampling technique and provider type.

Keywords: *Supervised and Unsupervised Classification; Class Imbalance; Fraud Detection; Medicare; Healthcare*

I. Introduction

Healthcare is a critical component in most people's lives and as such, should be affordable. The need for healthcare is particularly important for the rising elderly population. They require increased healthcare and therefore, appropriate insurance coverage for various medical drugs and services. The number of elderly individuals rose 28% from 2004 to 2015, versus an increase of just 6.5% for those under 65 years of age [3]. Thus, the upkeep and improvement in the health of this population becomes more important to the elderly and their family and friends. This increased healthcare need comes at a price and is usually managed by a healthcare insurance program. Clearly, these programs need to be economical for the general populace, but program costs, along with the elderly population, continue to increase, which can financially cripple individuals and families. In 2015, U.S. healthcare spending increased 5.8%, totaling over $3.2 trillion [2]. Even given such social prominence and financial stakes, fraud, waste, and abuse (FWA) reduction efforts are doing little to diminish these costs [1].

Given the gravitas of elderly healthcare, we focus our research on the Medicare insurance program. Medicare is a U.S. government program, with over 54.3 million beneficiaries, providing insurance for people over the age of 65 or younger individuals with specific medical conditions and disabilities [7]. Medicare contributes to 20% of the overall U.S. healthcare spending, at $646 billion in 2015 [2]. With such a large financial capacity, Medicare is a prime target for current and future FWA activities. The Federal Bureau of Investigations (FBI) estimates that fraud accounts for 3-10% of all medical claims [35], which is $19 to $65 billion in potential losses due to FWA. Healthcare is an attractive area for perpetrators of fraud necessitating the novel application of fraud detection methods to mitigate potential fraudulent activities in Medicare.

The Centers for Medicare and Medicaid Services (CMS) has recently released 2012 to 2015 Medicare-related data for public consumption [6], to help combat healthcare fraud. For our case study, we use the 2015 *Medicare Provider Utilization and Payment Data: Physician and Other Supplier*, hereby called Medicare PUF, which provides information on services provided to Medicare beneficiaries grouped by physicians and other healthcare professionals, such as nurses. The Medicare PUF dataset does not include labels indicating fraud. However, a list of physicians and other healthcare entities who are currently excluded from participation in Medicare, for a certain period of time, can be obtained from the Office of Inspector General's (OIG) List of Excluded Individuals/Entities (LEIE) database [33]. The OIG holds authority to exclude individuals and entities from federally funded healthcare programs in accordance with Sections 1128 and 1156 of the Social Security Act [33]. It is important to note that even though providers are listed in the LEIE database, 38% with fraud convictions continue to practice medicine and 21% were not suspended from medical practice despite their convictions [37], thus this database is not all inclusive. Even so, the use of such publicly available large-scale data repositories, and accessible data labels, with the application of machine learning to detect fraud can lead to substantial cost recovery for Medicare. More information on Medicare fraud can be found in [15].

In our experiment, we present an exploratory analysis comparing several supervised and unsupervised classification methods to detect known fraudulent activities. We use the 2015 Medicare PUF data with corresponding LEIE exclusion labels, and put the data into eight different prov

0-7695-6321-X/17/31.00 ©2017 IEEE
DOI 10.1109/ICMLA.2017.00-48

a reasonable percent of exclusions. The availability of such fraud labels enables supervised fraud detection methods, as well as validation of unsupervised, anomaly-based, detection methods. We use 10 techniques (also referred to as learners or methods) and divide them into three groups: supervised, unsupervised, and hybrid. The supervised learners are made up of Gradient Boosted Machine (GBM), Random Forest (RF), Deep Neural Network (DNN), and Naive Bayes (NB). The unsupervised methods include: Autoencoder, Mahalanobis distance, k-Nearest Neighbors (kNN), and Local Outlier Factor (LOF). The final group is what we call the hybrid learners, which includes a neural network model that is pre-trained using the unsupervised autoencoder and another method using a combination of multivariate regression and Bayesian probability. Because the labeled data is scant, we use two methods to mitigate the extreme class imbalance between the fraud and normal data labels. The first technique applies oversampling to help balance the classes, whereas the second forces an 80% normal and 20% fraud data split using undersampling. Fraud detection performance is measured with four different metrics: balanced accuracy [16], F-measure, G-measure, and Matthew's Correlation Coefficient (MCC) [34]. Lastly, statistical analysis is provided to demonstrate the significance of the experiment results using ANOVA (ANalysis Of VAriance) [26] and post hoc analysis via Tukey's Honestly Significant Different (HSD) test [45]. Our results show that supervised learners can be significantly better than other types of learners, especially with more reasonable class balance. Furthermore, when data are severely imbalanced, unsupervised methods performed similarly to most supervised learners.

There are very limited studies using Medicare data in conjunction with the list of LEIE database fraud providers for the detection of Medicare fraud. Thus, the contribution of our exploratory analysis is the application and comparison of various supervised and unsupervised machine learning methods to detect Medicare fraud by provider type, leveraging the LEIE exclusion data for fraud labels, using features that encompass only procedures performed, submitted charges, and payment amounts. We feel our study provides meaningful comparisons not readily available in the literature, such as which provider types benefit most from which methods for Medicare fraud detection. To the best of our knowledge, no other work provides comprehensive comparative analysis exploring supervised, unsupervised, and hybrid machine learning methods for the detection of fraud, using the latest 2015 Medicare dataset released in June 2017.

The rest of the paper is organized as follows. Section II discusses works related to the current research, focusing primarily on Medicare-related fraud. In Section III, we detail our research methodology to include data, learners, class imbalance, and model performance. In Section IV, the results of our research are discussed. Finally, Section V summarizes our conclusions and ideas for future work.

II. Related Works

Because Medicare utilization and payment data is relatively new, there are limited studies using or referencing this particular dataset in fraud detection research. Three earlier studies look for patterns in the Medicare utilization and payment data employing descriptive statistics and correlations, but do not apply machine learning techniques. Feldman et al. [23] use 2012 Medicare data to attempt to find correlations between a physician's educational background and practices performed to detect possible misuse or insurance inefficiencies. In doing this, the authors analyzed medical-related variables, such as charges, number of procedures, and payments, to find anomalous behaviors by comparing results with the national distribution of payments and charges. A study by Ko et al. [32] also uses the 2012 Medicare data with focus on the Urology specialty only. They calculated the variability among Urologists and determined a possible savings of 9% from high utilization variability. Additionally, the authors found that the number of patient visits was strongly correlated with the reimbursements made by Medicare. Pande et al. [37] take older Medicare data and exclusions from the LEIE database to assess who the Medicare fraud perpetrators are and what happens to them after they get caught. The authors only use descriptive statistics to find patterns and make recommendations on Medicare fraud. One of the recommendations made is to use predictive models to detect claims fraud.

The use of descriptive statistics, correlations, etc., are extremely useful, but they tend to rely heavily on humans extracting patterns from the data. Machine learning methods can be employed to lessen this dependence by automatically extracting patterns to produce meaningful results, such as the detection of possibly fraudulent behaviors. One such study that uses machine learning is by Thornton et al. [44]. The authors explore several outlier-based detection techniques using Medicaid claims data for dental providers. Medicaid is another, distinct U.S. healthcare program that provides health coverage to low-income people [5]. They employ three univariate methods which include linear regression, box plots, and time series plots, as well as one multivariate method via clustering. The authors provide a case study of 500 dentists for which they claim the successful identification of 17 possibly fraudulent activities detected out of 360 records. A general coverage paper by Chandola et al. [21] explores several methods to detect healthcare fraud. They do not specify the use of Medicare data, but they do use a provider exclusion list, from the Texas Office of Inspector General's exclusion database, for fraudulent provider labels. The authors use several techniques including social network analysis, text mining, and temporal analysis in order to translate the problem of healthcare data analysis into some well-known data mining methods. The authors discussed the use of normal treatment profiles in order to compare providers and detect possible issues. A recent paper, by Branting et al. [18] creates a graph of providers, drug prescriptions, and procedures. The authors use two algorithms where one calculates the similarity to known fraudulent and non-fraudulent providers, and the other estimates fraud risk via shared practice locations or addresses. Medicare data from 2012 to 2014 was used, as was the LEIE database for fraud labels. They used 11 graph-based features, such as the similarity to the closest k members and the number of colocated excluded providers, and 4 additional features and the J48 decision tree implemented in the Weka framework. This resulted in an F-measure of 0.919 and ROC area of 0.960.

Additional research by Bauder et al. explores Medicare fraud detection through several different studies. In one study [11], the authors use multivariate regression to establish a baseline for expected Medicare payments, per provider type. This baseline is then used as the normative case in which to compare the actual payment amounts, with deviations flagged as outliers. Another study [13] incorporates a two-step approach in detecting Medicare fraud, per provider type. The first step involves a multivariate regression model returning model residuals. These residuals are passed into a Bayesian probability model that produces the final probabilities indicating how likely it is that a particular value is fraudulent. They compared their method versus other common outlier detection methods, and found their method performed favorably. Additionally, works [10], [12] continue to explore the use of Bayesian models to detect Medicare fraud. The final studies [14], [29] are exploratory studies that look to predict fraudulent providers by using only the number of procedures performed. The authors employ Multinomial Naive Bayes to predict the provider type. If the predicted provider type does not match what is expected, then this provider is performing outside of normal practice patterns and should be investigated.

Our exploratory, comparative study takes into account many of the methods and data sources found in the related works. We apply these methods in a comprehensive experiment to assess the efficacy of different learners in predicting Medicare fraud. In addition to the learners and data, we incorporate the very real issue of class imbalance and assess detection using four performance metrics.

III. Methodology

In this section, we outline the datasets, learners, and training and evaluation. We also discuss the methods used to mitigate class imbalance. Parameters for each machine learning algorithm are provided (otherwise the default values are used) in addition to the training and evaluation framework to ensure our results are reproducible, given the same data.

A. Data

The data in our experiment is from the Centers for Medicare and Medicaid Services which, at the point of this publication, encompass the 2012 to 2015 calendar years. Specifically, we use the *Physician and Other Supplier Data 2015* dataset, which describes payment and utilization claims data, with information on services and procedures provided to Medicare beneficiaries. The Medicare dataset contains values that are recorded after claims payments were made and with that, we assume that the Medicare dataset was appropriately recorded and cleansed by CMS [22], thus are not concerned with bad values adversely affecting the models. We filter the Medicare dataset for non-prescription data. The non-prescription data are those codes that are not for specific services listed on the Medicare Part B Drug Average Sales Price file [6], thus are actual provider services versus drug-specific activities.

The entire 2015 Medicare PUF dataset has 8,904,316 instances and 30 features, covering 91 provider types, e.g. Cardiology or Podiatry, for 968,276 distinct providers. Out of the 30 features, we focus on detecting fraud by only using the procedures performed, charges, and payments, with the additional features being used for filtering and identification purposes. The use of any remaining variables, along with applying different feature engineering approaches, is left as future work. Table I describes the subset of features chosen for our study.

TABLE I: Description of Medicare Features

Feature	Description
npi	Unique provider identification number
last_name	Provider's last name
first_name	Provider's first name
zip	Provider's 5-digit zip code
provider_type	Medical provider's specialty (or practice)
line_srvc_cnt	Number of procedures performed per provider
bene_unique_cnt	Number of distinct beneficiary per day services
average_submitted_chrg_amt	Average of the charges that the provider submitted
average_medicare_payment_amt	Amount paid to the provider for services performed

In order to obtain labels indicating fraudulent providers, we incorporate excluded providers from the List of Excluded Individuals/Entities (LEIE) database [33]. The exclusions are categorized by various rule numbers, which indicate severity as well as the length of time of each exclusion. We selected the providers excluded for more severe reasons, as seen in Table II. Additionally, because these exclusions last from 5 to 10 years and we are using the 2015 Medicare PUF data, we subset the exclusions further by restricting the date range from 2010 to 2015 to account for these exclusion time frames. Note that we assume these exclusion labels are the only fraudulent labels in the data and those not on the list are normal, for training and testing of the learners. Unfortunately, the LEIE does not contain National Provider Identification (NPI) number for most of the providers, so along with matching directly to NPI, we performed simple string matching using last name, first name, and zip code. Fuzzy string matching or other methods to match and extract additional exclusions is left for future work.

TABLE II: LEIE Exclusion Rules

Rule Number	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 or suspension.
1128(b)(7)	Fraud, kickbacks, and other prohibited activities.

From the 2015 Medicare dataset and the LEIE database, we selected certain provider types based on the diversity of procedures performed, submitted charges, and payments, as well as the number of known matched provider exclusions. Table III lists the provider types chosen and the number of matched exclusions.

TABLE III: 2015 Medicare Provider Types

Provider Type	Instances	Features	Exclusions	% Exclusions
Internal Medicine	1,066,572	9	270	0.025%
Family Practice	845,812	9	126	0.015%
Podiatry	183,944	9	100	0.054%
Neurology	129,789	9	45	0.035%
General Practice	41,187	9	44	0.107%
Radiation Oncology	75,479	9	41	0.054%
Pain Management	31,121	9	28	0.090%
Geriatric Medicine	19,058	9	17	0.089%

B. Supervised Learners

In deep learning networks, each layer of nodes trains on a distinct set of features based on the previous layer's output.

These networks perform automated extraction and abstraction of complex features. A hierarchical architecture of learning and representing data is created from these networks, where more abstract features are defined in terms of less abstract ones. In this experiment, the deep learning algorithm is based on a multi-layer feedforward artificial neural network that is trained with stochastic gradient descent using back-propagation [9]. This feedforward artificial neural network is also known as deep neural network (DNN) or multi-layer perceptron (MLP) and is one of the most common types of deep neural networks. This type of DNN works well on tabular data, such as Medicare, but other DNNs like Convolutional Neural Networks work better with image or text data [38]. For our study, we create a 3 layer DNN with 50 nodes per layer. We use a hyperbolic tangent (Tanh) which is a nonlinear activation function used to determines a node's output based on the input. We run the model through 1,000 epochs, i.e. the number of passes over the training data.

Gradient Boosting Machines (GBM) [25] is a boosting ensemble method, where new models are added to the ensemble sequentially. At each iteration, a new base learner is trained with respect to the errors learned in the prior iterations, taking steps to reduce the overall model error. In the case of GBM, the prediction error is minimized using a gradient-descent based technique, finding the gradient of the error with respect to the prediction. Each learner is created to be maximally correlated with the negative gradient of the loss function, which is associated with the whole ensemble (thus reducing the overall model error).

Random forest (RF) is an ensemble method in which multiple unpruned decision trees are built and a final classification is made by combining the results from the individual trees [19]. The algorithm creates random datasets using sampling with replacement to train each of the decision trees. At each node within a tree, RF chooses the most discriminating feature between the classes using information entropy. Additionally, RF performs random feature subspace selection, at each node of a tree, where a subset of *m* features are considered for the decision at that node. In our experiment, we use 100 trees.

The Naive Bayes algorithm [40] uses Bayes' theorem of conditional probability to determine the probability that an instance belongs in a particular class. It is considered "naive" due to its assumption that the values of the various features are independent from one another in contributing to the decision of the class label. The formula for this classifier is $\hat{y} = p(C_k)\prod_{i=1}^{n} p(x_i|C_k)$, where $\hat{y}$ is the predicted class, $p(C_k)$ is the probability that the instance belongs to class *k*, and $p(x_i|C_k)$ is the product of the conditional probabilities.

C. Unsupervised

Mahalanobis and k-Nearest Neighbors are typically considered global, distance-based outlier detection methods. The Mahalanobis distance gives the distance from a case to the centroid of all cases for the predictor variables [8]. A large distance indicates an observation that is an outlier in the space defined by the predictors [42]. In our study, we use the so-called robust Mahalanobis distance for deviations from multivariate normal, where the center and covariance are estimated via the Minimum Covariance Determinant estimator [24]. The k-Nearest Neighbors (kNN) [17] algorithm is a relatively simple method that looks at the *k*-nearest neighbors around some particular value to determine which neighbors are most similar, or closest in distance. Conversely, points having large kNN distances are seen as outliers. More specifically, to detect outliers, we use a *k* of one giving us a 1NN approach that utilizes the cover tree algorithm for the nearest neighbor search. In order to create the threshold from which to differentiate outliers from inliers for 1NN, we use 1.5 times the interquartile range (IQR) of the resulting 1NN distances. This method is typical in outlier detection indicating that about 99% of the points are normal, with the other 1% considered as outlying values. For Mahalanobis, we use a 99% threshold based on generating p-values from a Chi-Square cumulative distribution function and using these p-values to check against some threshold value [4]. Choosing outlier detection thresholds is another topic unto itself and there are various methods that exist, in practice, for selecting outliers based on scores with no single method used exclusively in application. We do not discuss this further in this paper.

Local outlier factor (LOF) is considered a local, density-based outlier detection method. With LOF, the local density of a particular point is compared with the density of its neighbors [20]. If the density of the point is substantially lower than its neighbors, this point is in a sparse area, thus is possibly an outlier. LOF requires that *k* neighbors be defined apriori, for which we followed [36] and used 20 for all experiments. We are primarily concerned with general algorithm performance compared to others, not necessarily the best *k* for each method. We use the same method, as used with 1NN, to create the outlier detection threshold.

Autoencoders are neural networks trained to reconstruct their input [27], [28]. They reduce dimensionality in a way similar to principle component analysis, but with non-linear activation functions. An autoencoder is composed of an encoder and decoder in order to learn the patterns from the data to create representative features of that data. These features are then used to reconstruct the original data patterns, with the reconstruction error indicating the divergence in the model's prediction versus the original input. In this study, we incorporate "bottleneck" training creating a middle hidden layer that is very small. We use a hidden layer of just 2 nodes for which the autoencoder will have to reduce the dimensionality of the input data. The autoencoder model will then learn the input data patterns. To determine the threshold to indicate an outlying value, we use the average model reconstruction error with anything greater than this value being considered an outlier. We use 50 nodes for the input and output layers, 2 nodes for the hidden middle layer, the Tanh activation function, and 1,000 epochs.

D. Hybrid Learners

For the purposes of our research, we consider the last two learners hybrid methods, because they build off other methods to increase model performance. In this paper, the pre-training autoencoder is what we are calling the model that uses the original unsupervised autoencoder as pre-training input to a supervised model which, in this case, is a neural network. The learner uses the weights from the inputted autoencoder for model fitting. We use the same parameters as with the unsupervised autoencoder, including the threshold for detecting outliers. This learner is more supervised, with unsupervised inputs used only for weighting.

The model by Bauder et al. [13], known herein as *hybrid*, takes a twofold approach for the detection of outliers. A Multivariate Adaptive Regression Splines (MARS) model fits multiple predictor variables with a single response variable. The MARS model residuals are used as inputs to a generalizable, fully Bayesian probability model, using the Stan language, to detect anomalous values (or outliers). In general, probabilistic programming allows for the creation of a probability model to fully represent uncertainty, or variability, about any underlying information explaining observed data, for probabilistic inference. MARS is a non-parametric regression model that accounts for the non-linearities between variables and their interactions. For the probability model, the posterior distributions are drawn from the full conditional of each unknown parameter. The model fitting is done by specifying the full likelihood function and the prior distributions of unknown parameters. We use a 99% outlier probability threshold to indicate outlying values, with anything less than this probability being considered normal.

E. Class Imbalance

We employ two sampling methods to mitigate issues arising from the class imbalance problem. The first method uses oversampling which is a method for balancing classes by adding instances to the minority class, rather than removing samples from the majority class as in undersampling. We use the default method provided in h2o, described in [31], which creates a balanced class distribution. The second is the so-called 80-20 method which is similar to what was done in [18] using an undersampling technique, except that we do not generate a balanced class distribution. We use an 80-20 ratio to retain more of the normal class and reduce any loss of information relative to the fraud class [30]. More specifically, we retain all fraud labeled instances and randomly sample without replacement from the remaining normal data to get to an 80% normal and 20% fraudulent class distribution. It is important to note that some of the available labeled instances are so small that undersampling alone would make the total dataset too small and most likely unrepresentative of the whole dataset. Further research will include more testing with additional sampling techniques.

F. Training and Evaluation

In our experiments, we use the h2o package [43] in R [39] to implement all supervised learning algorithms, as well as the autoencoder models. The remaining models were also implemented in R using standard packages. Default parameters were used unless otherwise stated.

We generate training and testing datasets, by randomly sampling without replacement, with 80% of instances used for training and the remaining 20% for model testing and validation. The supervised models were trained on this 80% and tested on the remaining 20% with all performance metrics derived from testing only. The exception to this is the training of the autoencoders. The unsupervised autoencoder is trained with 40% of the training data, with the remaining 40% used to create the pre-trained autoencoder model. Both, however, are tested on the same 20% test data. To better assess learner performance, we elected to use four different performance measures. Using more than one performance score can assist in either finding the best overall model, or indicate the difficulty in learning a particular dataset, e.g. if F-measure is a 0.5 but MCC is <= 0 then the F-measure score may be too optimistic. In this study, we use balanced accuracy, F-measure, G-measure, and Matthew's Correlation Coefficient.

Balanced accuracy (BACC) is a measure of performance that has been proposed as an approximation for the Area Under the Receiver Operator Curve (AUC) [16] for binary classification. This measure takes into account the true positive rate (TPR) and true negative rate (TNR) of the model, thus is more balanced than the standard accuracy score. Balanced accuracy is defined as $BACC = \frac{1}{2}(\frac{TP}{P} + \frac{TN}{N})$, where TP and TN are the counts of true positives and negatives, and P and N are the totals for the positive and negative classes, respectively.

F-measure, also known as F-Score, calculates the harmonic mean of precision and recall, while G-measure calculates the geometric mean of precision and recall. In F-measure, Beta can be varied to adjust the importance of precision and recall in computing the metric. We use the F1-measure, Beta = 1, as it gives equal weight to precision and recall. Both measures generate a number between 0 and 1, with the best score being 1 and the worst being 0 [41]. The F-measure, and G-measure, represent a more balanced view but could give a biased result since it does not include the true negatives. The formulas for both are defined as:

$$\textit{F-measure} = (1+\beta^2)(\frac{precision \times recall}{\beta^2 \times precision + recall})$$

$$\textit{G-measure} = \sqrt{(}precision * recall)$$

Unlike the other metrics discussed, Matthew's Correlation Coefficient (MCC) [34] takes into account all values in the confusion matrix in its formula. The equation below describes MCC which includes all true positives and negatives (TP, FP, TN, and FN).

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{((TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN))}}$$

MCC is generally regarded as a balanced measure, even if the classes are of very different sizes. Similar to a correlation coefficient between observed and predicted values, the range of MCC values lie between -1 to +1. A model with a score of +1 is a perfect model and -1 is a poor model, with 0 being essentially randomly guessing.

IV. Discussion and Results

As mentioned, we apply the learners to each provider type using only the variables associated with the procedures performed, submitted charges, and payments to classify fraudulent providers. Regardless of the percentage of labeled training data used, each learner was evaluated on the testing dataset only. In our experiment, we are not only looking at learner performance by provider type, but also performance based on class imbalance. Figure 1 depicts the average performance for each classifier across all provider types, grouped by oversampling or 80-20 class imbalance reduction method.

One observation is the substantial increase in model performance between the two class imbalance sampling methods.

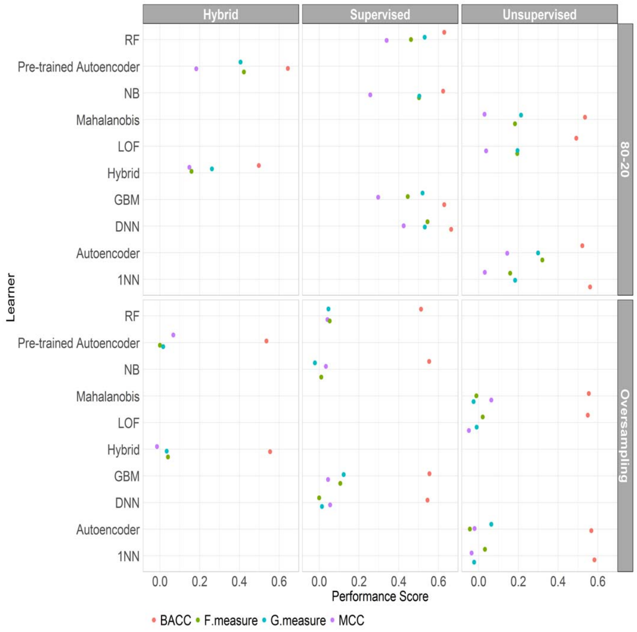

Fig. 1: Average Performance Scores by Classifier

It appears that oversampling alone does not provide the necessary data representation to discriminate between fraudulent exclusions and normal instances. This could be because the number of available exclusions is too small or the quality of these labeled exclusions is such that increasing these labeled instances, via oversampling, adds additional noise rather than discriminating samples. The 80-20 sampling method, which keeps all of the exclusion instances while downsampling the normal instances, does not appear to produce this additional noise. This could explain the large difference in performance across all learners. Another noticeable difference is the relative consistency of BACC versus the other metrics. The sampling methods did not appear to have a significant effect on BACC with an oversampling average of 0.55 and 80-20 average of 0.59, across all learners. The remaining metrics show substantial differences between the oversampling and 80-20 methods. Given the difficulty in classifying fraudulent Medicare cases using the exclusion database only, BACC does not appear to adequately capture true model performance.

The difference in performance between the supervised, unsupervised, and hybrid learners is less obvious within each sampling group. With oversampling, all methods perform poorly with MCC, F-measure, and G-measure scores which are well below 0.1, whereas with the 80-20 method, supervised, unsupervised, and hybrid methods perform adequately. Moreover, the performance difference between supervised learners and the hybrid and unsupervised learners is fairly distinct, with supervised learners having superior performance. Even so, given the limited labeled data, pre-training a model using an unsupervised autoencoder as the pre-training input shows promise with performance close to most of the supervised learners. Furthermore, the hybrid method, as with the pre-trained autoencoder, is comparable to some of the supervised methods and performs better than the unsupervised methods.

Figure 2 shows the performance of all learners across the provider types. Again, the BACC is fairly high and consistent between both sampling methods. The remaining performance metrics show similar patterns between sampling methods as seen in Figure 1. One point to take away from this figure is that some provider types are much more difficult to learn versus others. Radiation Oncology is easier to learn across the different models, whereas a more general field, like Family Practice, is more difficult. This is intuitive since Family Practice encompasses many different procedures, while Radiation Oncology is more specific in the types of procedures.

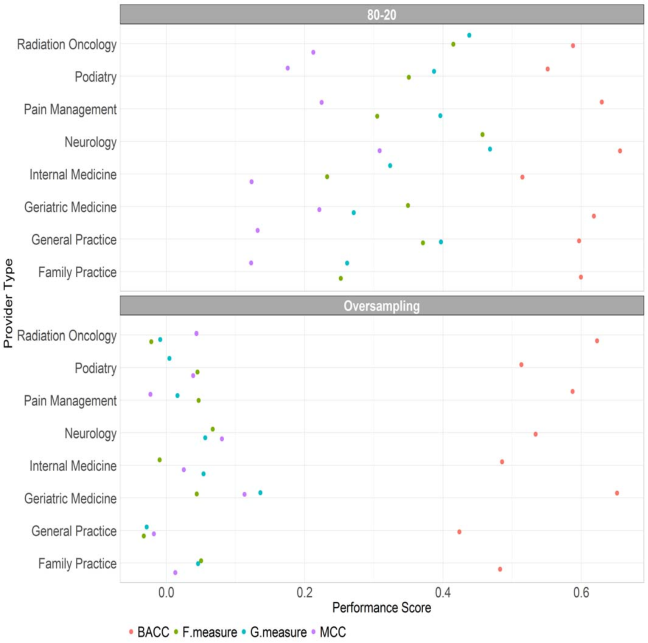

Fig. 2: Average Performance Scores by Provider Type

We evaluated the statistical significance of our results with a two-factor ANOVA and Tukey's HSD tests, at a 95% confidence level. The ANOVA and corresponding Tukey's HSD tests used learner and provider type as factors, across all performance metrics. We again grouped the results by sampling methods. Figure 3 shows the results of the hypothesis testing. Table 3a and 3b (in Figure 3) show the choice of learner and the provider type are both significant. Tables 3e and 3f show, for either sampling method, the majority of the supervised learners perform significantly better than the other learners. Moreover, DNN, GBM, and RF are generally the best learners for detecting fraudulent events. As for the unsupervised and hybrid learners, there is generally no significant difference between them meaning that any of these methods would work similarly with the Medicare data. Tables 3c and 3d indicate that the more specific provider types have better detection performance than the more general specialties.

V. Conclusion

Healthcare fraud significantly impacts the ability of insurance programs, such as Medicare, to provide effective and affordable care. Leveraging the power of machine learning can

Fig. 3: Two-factor ANOVA for Learners and Provider Types, with Tukey's HSD post hoc results, for oversampling and 80-20 sampling methods

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Learner	9	0.7612	0.08458	2.652	0.0113
Provider Type	7	1.7589	0.25126	7.877	7.8e-07
Residuals	63	2.0095	0.03190		

(a) Oversampling ANOVA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Learner	9	16.439	1.8266	8.928	1.56e-08
Provider Type	7	3.857	0.551	2.693	0.0166
Residuals	63	12.889	0.2046		

(b) 80-20 ANOVA

Group	Provider Type
a	Geriatric Medicine
ab	Neurology
abc	Radiation Oncology
bc	Pain Management
c	Family Practice
c	Internal Medicine
c	Podiatry
c	General Practice

(c) Oversampling HSD by Provider Type

Group	Provider Type
a	Neurology
ab	Radiation Oncology
ab	Podiatry
ab	Pain Management
ab	General Practice
ab	Geriatric Medicine
b	Family Practice
b	Internal Medicine

(d) 80-20 HSD by Provider Type

Group	Learner
a	GBM
a	RF
ab	Deep
ab	Mahalanobis
ab	Hybrid
ab	NB
ab	1NN
ab	Unsupervised Autoencoder
ab	Pre-trained Autoencoder
b	LOF

(e) Oversampling HSD by Learner

Group	Learner
a	Deep
ab	RF
ab	GBM
ab	NB
abc	Pre-trained Autoencoder
bc	Unsupervised Autoencoder
c	Hybrid
c	LOF
c	1NN
c	Mahalanobis

(f) 80-20 HSD by Learner

help in detecting more fraudulent events, thus removing the perpetrators and reducing healthcare costs. In this paper, we explore different machine learning methods to detect fraudulent Medicare providers and compare performance results and statistical significance. We use two different sampling methods and four performance metrics. Each of the ten models (supervised, unsupervised, and hybrid) is trained and tested on the 2015 Medicare PUF data with fraud labels obtained from the LEIE database.

Our results indicate that there is a large performance gap between the sampling methods. The 80-20 sampling technique has better learner performance versus oversampling. Overall, oversampling demonstrates poor performance for all learners. Additionally, BACC has been shown to be unreliable in measuring model performance, across all methods, unable to adequately reflect the more realistic differences seen in the other metrics. Learner performance is better using the 80-20 approach with the supervised methods being significantly better than the unsupervised and hybrid learners. Finally, the provider type contributes to the difficulty in detecting fraud, with relatively specialized provider types having better performance over more general specialties.

Ongoing research and future work will involve improving model performance with parameter and hyperparameter tuning. Adding more Medicare datasets, with additional LEIE exclusion labels, as well as using different sampling methods for class imbalance, will be considered for future research. Finally, using both unsupervised and supervised models to detect fraud with a limited number of labels, leveraging the strengths of both will be pursued.

Acknowledgment

We acknowledge partial support by the NSF (CNS-1427536). Opinions, findings, conclusions, or recommendations in this paper are the authors' and do not reflect the views of the NSF.

References

[1] "The facts about rising health care costs," 2015. [Online]. Available: http://www.aetna.com/health-reform-connection/aetnas-vision/facts-about-costs.html

[2] "National Health Expenditures 2015 Highlights," 2015. [Online]. Available: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/Downloads/highlights.pdf

[3] "Profile of older Americans: 2015," 2015. [Online]. Available: http://www.aoa.acl.gov/Aging_Statistics/Profile/2015/

[4] "Compute Mahalanobis Distance and flag multivariate outliers," Sep 2016. [Online]. Available: http://www-01.ibm.com/support/docview.wss?uid=swg21480128

[5] "US Medicaid Program," 2016. [Online]. Available: https://www.medicaid.gov

[6] "Centers for Medicare and Medicaid Services: Research, Statistics, Data, and Systems," 2017. [Online]. Available: https://www.cms.gov/research-statistics-data-and-systems/research-statistics-data-and-systems.html

[7] "US Medicare Program," 2017. [Online]. Available: https://www.medicare.gov

[8] C. C. Aggarwal, *Data Mining: The Textbook*. Springer, 2015.

[9] S. Aiello, E. Eckstrand, A. Fu, M. Landry, and P. Aboyoun, *Machine Learning with R and H2O*, 2016, sixth Edition. [Online]. Available: https://h2o2016.wpengine.com/wp-content/themes/h2o2016/images/resources/RBooklet.pdf

[10] R. A. Bauder and T. M. Khoshgoftaar, "A probabilistic programming approach for outlier detection in healthcare claims," in *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Dec 2016, pp. 347–354.

[11] R. A. Bauder and T. M. Khoshgoftaar, "A novel method for fraudulent medicare claims detection from expected payment deviations (application paper)," in *Information Reuse and Integration (IRI), 2016 IEEE 17th International Conference on*. IEEE, 2016, pp. 11–19.

[12] R. A. Bauder and T. M. Khoshgoftaar, "Multivariate anomaly detection in medicare using model residuals and probabilistic programming," in *FLAIRS Conference*, 2017, pp. 417–422.

[13] R. A. Bauder and T. M. Khoshgoftaar, "Multivariate outlier detection in medicare claims payments applying probabilistic programming methods," *Health Services and Outcomes Research Methodology*, pp. 1–34, Jun 2017. [Online]. Available: http:https://doi.org/10.1007/s10742-017-0172-1

[14] R. A. Bauder, T. M. Khoshgoftaar, A. Richter, and M. Herland, "Predicting medical provider specialties to detect anomalous insurance claims," in *Tools with Artificial Intelligence (ICTAI), 2016 IEEE 28th International Conference on*. IEEE, 2016, pp. 784–790.

[15] R. A. Bauder, T. M. Khoshgoftaar, and N. Seliya, "A survey on the state of healthcare upcoding fraud analysis and detection," *Health Services and Outcomes Research Methodology*, vol. 17, no. 1, pp. 31–55, 2017.

[16] M. Bekkar, H. K. Djemaa, and T. A. Alitouche, "Evaluation measures for models assessment over imbalanced data sets," *Iournal Of Information Engineering and Applications*, vol. 3, no. 10, 2013.

[17] A. Beygelzimer, S. Kakadet, J. Langford, S. Arya, D. Mount, and S. Li, *FNN: Fast Nearest Neighbor Search Algorithms and Applications*, 2013, r package version 1.1. [Online]. Available: https://CRAN.R-project.org/package=FNN

[18] L. K. Branting, F. Reeder, J. Gold, and T. Champney, "Graph analytics for healthcare fraud risk estimation," in *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*. IEEE, 2016, pp. 845–851.

[19] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct 2001. [Online]. Available: http://dx.doi.org/10.1023/A:1010933404324

[20] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in *ACM sigmod record*, vol. 29, no. 2. ACM, 2000, pp. 93–104.

[21] V. Chandola, S. R. Sukumar, and J. C. Schryver, "Knowledge discovery from massive healthcare claims data," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, pp. 1312–1320.

[22] CMS Office of Enterprise Data and Analytics. (2017) Medicare Fee-For-Service Provider Utilization & Payment Data Physician and Other Supplier. [Online]. Available: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Downloads/Medicare-Physician-and-Other-Supplier-PUF-Methodology.pdf

[23] K. Feldman and N. V. Chawla, "Does medical school training relate to practice? evidence from big data," *Big Data*, vol. 3, no. 2, pp. 103–113, 2015.

[24] P. Filzmoser and K. Varmuza, *chemometrics: Multivariate Statistical Analysis in Chemometrics*, 2016, r package version 1.4.1. [Online]. Available: https://CRAN.R-project.org/package=chemometrics

[25] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001. [Online]. Available: http://www.jstor.org/stable/2699986

[26] A. Gelman *et al.*, "Analysis of variance: why it is more important than ever," *The annals of statistics*, vol. 33, no. 1, pp. 1–53, 2005.

[27] S. Glander. (2017) Autoencoders and anomaly detection with machine learning in fraud analytics. [Online]. Available: https://shiring.github.io/machine_learning/2017/05/01/fraud/

[28] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[29] M. Herland, R. A. Bauder, and T. M. Khoshgoftaar, "Medical provider specialty predictions for the detection of anomalous medicare insurance claims," in *Information Reuse and Integration (IRI), 2017 IEEE 18th International Conference*. IEEE, 2017, pp. 579–588.

[30] T. M. Khoshgoftaar, C. Seiffert, J. Van Hulse, A. Napolitano, and A. Folleco, "Learning with limited minority class data," in *Machine Learning and Applications, 2007. ICMLA 2007. Sixth International Conference on*. IEEE, 2007, pp. 348–353.

[31] G. King and L. Zeng, "Logistic regression in rare events data," *Political Analysis*, vol. 9, no. 2, pp. 137–163, 2001.

[32] J. S. Ko, H. Chalfin, B. J. Trock, Z. Feng, E. Humphreys, S.-W. Park, H. B. Carter, K. D. Frick, and M. Han, "Variability in medicare utilization and payment among urologists," *Urology*, vol. 85, no. 5, pp. 1045–1051, 2015.

[33] LEIE. (2017) Office of inspector general leie downloadable databases. [Online]. Available: https://oig.hhs.gov/exclusions/index.asp

[34] B. Matthews, "Comparison of the predicted and observed secondary structure of t4 phage lysozyme," *Biochimica et Biophysica Acta (BBA) - Protein Structure*, vol. 405, no. 2, pp. 442 – 451, 1975. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0005279575901099

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

[36] H. V. Nguyen, H. H. Ang, and V. Gopalkrishnan, *Mining Outliers with Ensemble of Heterogeneous Detectors on Random Subspaces*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 368–383. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-12026-8_29

[37] V. Pande and W. Maas, "Physician medicare fraud: characteristics and consequences," *International Journal of Pharmaceutical and Healthcare Marketing*, vol. 7, no. 1, pp. 8–33, 2013.

[38] J. D. Prusa and T. M. Khoshgoftaar, "Designing a better data representation for deep neural networks and text classification," in *2016 IEEE 17th International Conference on Information Reuse and Integration (IRI)*, July 2016, pp. 411–416.

[39] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2017. [Online]. Available: https://www.R-project.org

[40] I. Rish, "An empirical study of the naive bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22. IBM, 2001, pp. 41–46.

[41] Y. Sasaki *et al.*, "The truth of the F-measure," *Teach Tutor mater*, vol. 1, no. 5, 2007.

[42] J. P. Stevens, "Outliers and influential data points in regression analysis." *Psychological Bulletin*, vol. 95, no. 2, p. 334, 1984.

[43] The H2O.ai team, *h2o: R Interface for H2O*, 2017, r package version 3.10.4.6. [Online]. Available: https://CRAN.R-project.org/package=h2o

[44] D. Thornton, G. Capelleveen, M. Poel, J. Hillegersberg, and R. M. Müller, "Outlier-based health insurance fraud detection for us medicaid data," 2014.

[45] J. W. Tukey, "Comparing individual means in the analysis of variance," *Biometrics*, vol. 5, no. 2, pp. 99–114, 1949. [Online]. Available: http://www.jstor.org/stable/3001913