





Cost-Sensitive Learning for Medical Insurance Fraud Detection With Temporal Information

Haolun Shi , Mohammad A. Tayebi , Jian Pei , *Fellow, IEEE*, and Jiguo Cao 

Abstract—Fraudulent activities within the U.S. healthcare system cost billions of dollars each year and harm the wellbeing of many qualifying beneficiaries. The implementation of an effective fraud detection method has become imperative to secure the welfare of the general public. In this article, we focus on the problem of fraud detection using the current year’s Medicare claims data from the perspective of utilizing temporal information from the previous years. We group the data into temporal trajectories of the key covariates and base our feature engineering around these trajectories. For effective feature engineering on the temporal data, we propose to use the functional principal component analysis (FPCA) method for analyzing the temporal covariates’ trajectory as well as the distributional FPCA for extracting features from the empirical probability density curve of the covariates. Moreover, we introduce the framework of cost-sensitive learning for analyzing the Medicare database to allow for asymmetrical losses in the confusion matrix, such that the classification rule reflects the realistic tradeoff between the fixed cost and the fraud cost. The issue of class imbalance in the database is tackled through the random undersampling scheme. Our results confirm that the trained classifier has a reasonably good prediction performance and a significant percentage of cost savings can be achieved by taking into account the financial cost.

Index Terms—Centers for medicare & medicaid services, cost-sensitive learning, functional principal component analysis, functional data analysis.

I. INTRODUCTION

AS TECHNOLOGY in medical research advances and healthcare services continue to improve in the United States, increasing costs follow as a result. Fair access to healthcare services has thus become a pressing issue that impacts the general population in the United States. To help alleviate the financial strain on people to purchase their medical services, the US government created a national healthcare insurance program named Medicare, which covers parts or even all of

the expenditures of medical procedures, prescription drugs, and equipment. According to data released by the Centers for Medicare & Medicaid Services (CMS), funding for Medicare accounts for 20% of the annual US healthcare budget with possible expense recovery of \$4–13 billion [1]. One underlying issue that lurks within the Medicare system is the fraudulent activities that waste billions of dollars each year at the cost of the wellbeing of many qualifying beneficiaries. Fraud causes a tremendous amount of financial strain on the annual US healthcare budget. It is important to note that this loss implies not only a large amount of money going into wrong people’s pockets but also the unavailability of services to many who require a constant supply of medical service. Fraud accounts for an estimated spending of \$700 billion out of the total budget of \$2.7 trillion in healthcare in 2013 [1]. Therefore, the implementation of an effective healthcare delivery system and fraud detection method has become imperative to secure the welfare of the general public, especially the elderly population, who is in dire need of affordable healthcare services.

Though a considerable amount of effort has been put into reducing fraudulent activities, we have not seen a significant relief on the financial strain. The primary fraud detection method involves investigators searching through a great number of files and records to discover suspicious activities [2]. Unfortunately, this method has already become outdated and less effective, as a massive amount of data regarding healthcare transactions are generated each year. As an increasing volume of healthcare-related data become storable and accessible, more sophisticated data manipulation tools and machine learning methods need to be implemented in order to extract useful information from a pool of data and improve the detection of healthcare fraudulent activities. Many government programs and organizations have declared the importance of finding an effective fraud detection method; in particular, the CMS joined the cause of improving fraud detection and made available online a series of datasets named “Medicare Provider Utilization and Payment Data” [1].

These datasets released by CMS consist of primarily three parts: Physician and Other Supplier (Part B), Part D Prescriber (Part D), and Referring Durable Medical Equipment, Prosthetics, Orthotics, and Supplies (DMEPOS). These datasets comprise a large range of claims submitted by healthcare providers to Medicare, thereby providing a complete and thorough representation of the cost-related activities in the Medicare program. The Part B dataset is related to the average cost amount for the procedures performed, the Part D is related to the prescribed medication, and DMEPOS related to the issued supplies.

Manuscript received 11 January 2022; revised 30 November 2022; accepted 14 January 2023. Date of publication 30 January 2023; date of current version 15 September 2023. The work of Jian Pei and Jiguo Cao was supported by the Strategic Partnership Grant of the Natural Sciences and Engineering Research Council of Canada (NSERC). Recommended for acceptance by L. Zou. (Corresponding author: Jiguo Cao.)

Haolun Shi and Jiguo Cao are with the Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC V5A1S6, Canada (e-mail: haolun_shi@sfu.ca; jca76@sfu.ca).

Mohammad A. Tayebi and Jian Pei are with the School of Computing Science, Simon Fraser University, Burnaby, BC V5A1S6, Canada (e-mail: tayebi@cs.sfu.ca; jpei@cs.sfu.ca).

This article has supplementary downloadable material available at <https://doi.org/10.1109/TKDE.2023.3240431>, provided by the authors.

Digital Object Identifier 10.1109/TKDE.2023.3240431

The focus of this paper is on fraud detection using data of the Fee-For-Service payment of Medicare that involves physicians submitting claims for medical services directly to Medicare. Using the Medicare claim data, we have considered several strategies for feature engineering: first, we group the data into temporal trajectories of the key covariates and create numerical summaries as the trajectory-specific features; second, we introduce the methodology of functional principal component analysis (FPCA), a useful tool for analysis of temporal data, for effective feature extraction from the trajectories; third, we conduct a distributional FPCA on the empirical probability density curves of the drug-level features, and obtain predictive features from it. Based on the engineered features, our modeling is different from the previous studies on Medicare data. We specifically adopt the framework of cost-sensitive learning to reflect the tradeoff between the fixed cost associated with fraud investigation and the amount of fraud. Moreover, one issue in training the classifier using the data set is the class imbalance. In our constructed data set, only 0.025% of the cases are fraudulent/positive. This creates challenges for the training of the machine learning algorithm. We use the random undersampling approach with the varying class ratio to handle the issue of class rarity.

We experiment with four machine learning algorithms to train the classifier. Our results show that based on the undersampling dataset, the trained classifier has a reasonably good predictive performance. More importantly, under the cost-sensitive approach, a significant percentage of cost saving can be achieved by taking into account the financial cost. In terms of the cost-saving percentage, the best algorithm achieves the largest saving percentage of around 55%, i.e., when compared with the case where no fraud investigation is conducted, using such a classifier may save the cost by more than 50%. In contrast, the traditional non cost-sensitive approach tends to have a much smaller cost saving percentage. In fact, under certain algorithm and undersampling class ratio, the non cost-sensitive approach could lead to an even higher cost than doing no fraud detection at all.

A major novel aspect that distinguishes our work from existing methods is the use of functional principal component analysis to extract useful information from the data. Functional principal component analysis is a highly efficient and effective technique for detecting the primary direction of variations in longitudinal trajectories and extract predictive information from them. It has been applied across various scientific fields such as medicine, finance, genetics, behavioral science and ecology [4]. To the best of our knowledge, no prior work has ever applied the functional principal component analysis to fraud detection in the Medicare data set, and thus our work fills an important research gap. Most conventional approaches only use standard numeric function (e.g., minimum, mean and maximum) and one-hot encoding for constructing numerical covariates from the raw data. We utilize two classes of functional principal component analysis to extract information from (a) the temporal trajectories of a raw numerical feature and (b) the empirical distribution of the numerical feature. These covariates may extract new insight from the data.

Another major novel aspect of our approach is cost-sensitive learning, which to the best of our knowledge, has not been previously applied to fraud detection in the Medicare data set. By incorporating the cost measure, our method is able to guide the fraud detection towards the suspicious case that would lead to the highest monetary recovery of the fraud cost. The cost-sensitive learning method is meaningful and pragmatic for healthcare fraud detection and has important policy implications.

The rest of the paper is organized as follows. Section II provides a discussion on existing works related to Medicare data and cost-sensitive fraud detection methods. The content and structure of the data set are described in the Section III. Section IV explains the three strategies for creating predictive features from the data and discusses the approaches for handling class imbalance as well as the framework of cost-sensitive learning. Section V elaborates the performance metrics and presents the analysis of the Medicare data set. Section VI discusses the policy implication of the proposed methods and finally, Section VII summarizes the paper with a short discussion.

II. RELATED WORKS

We review the existing studies on healthcare fraud detection that have been conducted using the Medicare datasets released by CMS. The common objective shared by all these studies is to detect fraudulent activities using a machine learning approach. Depending on the types of techniques used, we classify the related works as follows.

▷ *Supervised Models* define healthcare fraud detection as a binary classification to distinguish fraudulent behavior from non-fraudulent behavior. Bauder et al. [7] explored how abnormalities in physicians' activities may point to possible fraudulent activities using the Naive Bayes algorithm, such that physicians' suspicious actions such as submitting claims data outside their specialties can be detected. Bauder and Khoshgoftaar [7] proposed to estimate the expected cost amount of different types of medical services and then compute its discrepancy to the actual amount paid to mark possible fraudulence. The multivariate adaptive regression splines are identified to be the best-performing model for estimating the expected cost amount. Herland et al. [10] validated the performance of the model constructed in the previous studies with Part B and LEIE datasets and strived to improve the model performance through feature selection, removal of certain specialties, and specialty grouping. They concluded that the strategy of removing certain specialties that involved several procedures could significantly improve the model performance. Herland et al. [5] created a combined dataset from Part B, Part D, and DEMPOS datasets, and used various machine learning methods to detect fraud in Medicare. The authors evaluated the performances of random forest, gradient boosting and logistic regression on each of these three datasets and also the dataset obtained by grouping all the parts. The results presented in this paper show that the performance of all classifiers improves significantly using the integrated dataset and logistic regression outperforms all other studied models.

Existing solutions in this category only apply classic machine learning applications to classify normal and fraudulent samples, and what differentiates these works from each other is not about the methodical part but how they preprocess and integrate the data and extract learning features from it. The work presented by Herland et al. [5] outperforms the other approaches because of fusing information from different sources. The advantage of the supervised learning models mainly lies in its accuracy (in terms of AUC).

▷ *Unsupervised Models* generally aim to identify and highlight data points that deviates from the overall pattern of the data. Khurjekar et al. [15] presented an unsupervised learning approach based on a multivariate regression model. They set a residual threshold and applied clustering to residuals that are above the threshold. Sadiq et al. [16] introduced the patient rule induction method, an unsupervised learning method to detect fraudulence by marking anomalies indicated by higher modes in the datasets. Bauder and Khoshgoftaar [7] presented an outlier detection model which is based on Bayesian inference using a sub-dataset derived from the 2012–2014 Part B Medicare dataset specifically focusing on dermatology and optometry claims. In their work, probabilistic programming is used to produce probability distribution and creates credibility intervals to evaluate the precision of outlier prediction. One of the major challenges that unsupervised solutions need to address is the class imbalance issue. Johnson and Khoshgoftaar [6] studied different existing resampling techniques for imbalanced classes using the CMS data. The authors concluded that not only maintaining sufficient representation of the majority class plays more important role than reducing the level of class imbalance but also downsampling the majority class to reach balanced proportions can degrade classification performance. While the unsupervised models can point to abnormalities in the data, its disadvantage lie in the relatively low accuracy in comparison with the supervised solutions.

▷ *Deep Learning Models* have yielded outstanding results in different fields, and has become an indispensable part of data-driven solutions. Deep learning models can be potentially used both in supervised and unsupervised ways to solve the fraud detection problem. While modern deep-learning-based solutions show promising solutions for different applications, fraud detection has not received enough attention from the scholars. As the only work in this domain, Johnson and Khoshgoftaar [6] applied various deep learning models to the combined CMS dataset for fraud detection with a focus on addressing class imbalance issues. They evaluated the significance of identifying optimal decision thresholds in case of having imbalanced training data. The authors of this work noted improvement over the existing methods used by Herland et al. [5]. Applications of deep learning in healthcare fraud detection is in its infancy, and offers many interesting research directions to pursue.

For example, to address fraud detection as an anomaly detection problem, one can employ deep learning models in two ways: to learn feature representation of normality, and to develop an end-to-end anomaly scoring approach, as discussed by Pang et al. [29]. In the first approach, general-purpose deep learning models such as autoencoders and generative adversarial

networks can be used to learn a representation of given data, and by capturing the essential underlying data regularities, these models are capable of detecting anomalies. Learning feature representation can be optimized based on an anomaly measure, such as the distance of anomalous samples from normal samples, or it can be formulated as a one-class classification problem. In the second approach, the goal is to learn an anomaly scoring approach directly. These approaches aim at devising a novel loss function to learn anomaly scores. Moreover, different deep learning models such as Recurrent Neural Networks (RNNs), Long Short-term Memory (LSTMs), and Gated Recurrent Units (GRUs) can be used as a supervised solution to classify fraudulent and normal data samples.

▷ *Other Perspectives* In addition to the previously studied models, researchers have explored healthcare fraud detection problem from other perspectives. Liu et al. [30] used providers and clients geo-location information for healthcare fraud detection. The underlying hypothesis in this work was that medicare clients prefer to use health service providers located in a relatively short distance specifically when they are senior or disabled, and the long distance between the service provider and client locations may indicate a fraudulent activity. Chandola et al. [12] elaborated on the challenges in analyzing the healthcare claims datasets from Texas and identifying fraudulent physicians. They also discussed the potential to utilize text mining and temporal analysis for detecting fraud from big healthcare datasets. Ko et al. [11] applied linear regression to examine the correlation between patient visits and utilization payment, with a concentration on urology. Branting et al. [13] proposed graph-based methods for fraud detection. Two types of algorithms were applied: one estimates the similarity between fraudulent and non-fraudulent providers' activities; the other estimates the risk propagation from physicians according to geospatial collocation. Another highlight of the work by Branting et al. [13] is that they refined the fraud labels by filling in the missing NPI from the National Plan & Provider Enumeration System registry website.

III. DATA

We focus our study on the Part D dataset, which provides information on the prescription drugs the physicians entered into an electronic medical record system in a certain year. Five years of data are available from 2013 to 2018 on the CMS website. Each row in the data set corresponds to the information related to a certain drug administered by a certain physician under a certain specialty type, i.e., the three columns which together uniquely define a row are the physician, the specialty type, and the brand name of the drug. The unique identifier for each physician is the NPI, and each physician may have more than one specialty type. Under a certain combination of physician and specialty types, multiple rows pertaining to the information of a certain drug are available in the data set. In addition to a drug's brand name and generic name, the drug-related information in the data set includes its total cost, the total number of claim count, the total number of beneficiaries, total 30-day fill count, and total daily supply under the physician and specialty in that given year. We

refer to these numeric features as “key covariates”, which are used for constructing predictive features in the model.

For the binary labeling of fraud, we obtain the list of fraudulent physicians and their NPIs from the List of Excluded Individuals and Entities (LEIE) on the website of the Office of Inspector General’s (OIG). The database is updated monthly to provide a list of physicians whose exclusions are currently in effect (as of March 2020). We use the NPI as the unique identifier in the LEIE data to link and map back to the Medicare Part D database, such that the fraudulent physicians can be identified.

IV. MODELING

A. Feature Engineering

1) *Trajectory-Specific Feature of Key Covariates:* We perform the analysis on the level of physician and specialty, i.e., given a physician and his/her specialty, our goal is to predict whether the physician committed fraud, using the numerical features related to all the drugs administered by the physician and under that specialty type. Five years of Part D data are available from 2013 to 2018. For a given year, we group the data by physician’s NPI and specialty type and compute the group-level minimum, maximum, mean, median, standard deviation, and summation of all the key covariates. The 5-year trajectories of these group-level numerical summary quantities can then be constructed. Fig. 1 shows the yearly trajectories of a fraud physician and a non-fraud physician. It is worth noting that our method is based on the entire temporal trajectories of a physician, whereas most existing fraud detection applications related to Medicare data focus on a snapshot of the data, and do not distinguish the identity of the physician [5], [6], [7]. In this sense, our method offers a more sound and novel perspective to the analysis of the Medicare database.

We create trajectory specific features for each trajectory, such as the trajectory mean, median, maximum, minimum, and standard deviation. Moreover, to capture the trend or slope information from the trajectory, a linear regression model is fitted on the trajectory. The model uses the trajectory value as the response and the year as the predictor and includes an intercept and a slope coefficient term. The fitted slope coefficient is used as a trajectory specific feature.

2) *FPCA for Physician-Level Trajectories of Key Covariates:* Functional principal component analysis (FPCA) is a widely used tool in statistics for analyzing temporal data [23]. It achieves dimensionality reduction by summarizing the information in the temporal trajectories into a series of functional principal component (FPC) score. In our constructed data set, the temporal trajectories of key quantities over the past 5 years are used as the target for performing FPCA. The FPC scores extracted from such an analysis are then used as the predictors in the model.

We model the trajectories of a particular key quantity as independent realizations from a stochastic process $X(t)$ and let $X_i(t)$ denote the trajectory realization of the i th subject. Let $\mu(t) = E(X(t))$ and $K(s, t) = \text{Cov}(X(s) - \mu(s), X(t) - \mu(t))$ denote the mean function and the covariance function, respectively. Based on the Karhunen-Lovève decomposition,

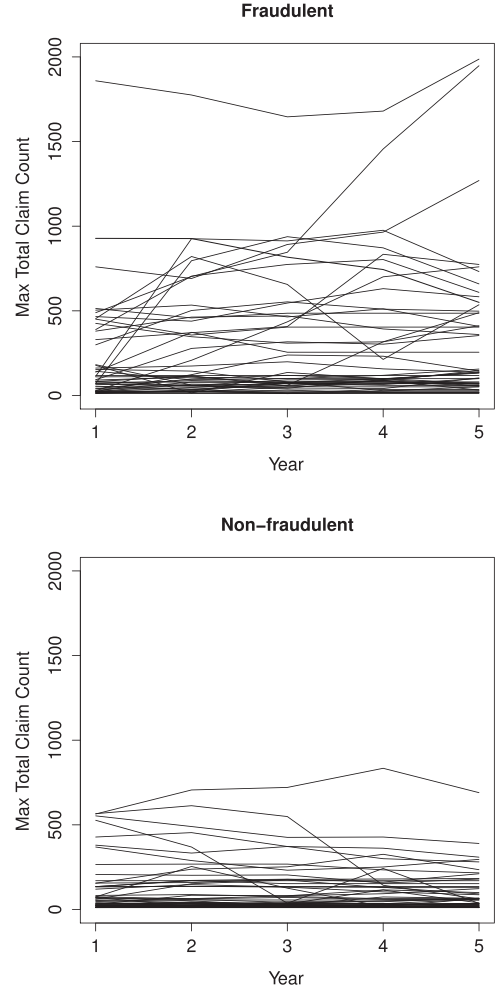


Fig. 1. Yearly trajectories of the maximum total claim count under fraud and non-fraud physician. Compared with the non-fraudulent cases, the trajectories of the fraudulent cases tend to have more extreme outliers that have much higher values than the population.

$X_i(t)$ can be expressed as

$$X_i(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_{ik} \phi_k(t), \quad (1)$$

where $\phi_k(t)$ is the k th eigenfunction, and ξ_{ik} is the associated FPC score for the i th subject. The eigenfunctions should satisfy

$$\int_{\mathcal{T}} \phi_k(t) \phi_j(t) dt = \delta_{kj}, \quad (2)$$

where $\delta_{kj} = 1$ if $k = j$ and 0 otherwise.

The FPC score is defined as

$$\xi_{ik} = \int_{\mathcal{T}} (X_i(t) - \mu(t)) \phi_k(t) dt. \quad (3)$$

The magnitude of ξ_{ik} represents the degree of similarity between the $X_i(t) - \mu(t)$ and the eigenfunction $\phi_k(t)$. The mean and variance of the distribution of ξ_{ik} are $E(\xi_{ik}) = 0$ and $\text{Var}(\xi_{ik}) = \lambda_k$, where $\lambda_1 \geq \lambda_2 \geq \dots \geq 0$.

To obtain the FPC estimates and FPC scores, we perform the following procedures.

- 1) Do a local linear regression to obtain a smoothed estimate of $X_i(t)$, denoted as $\hat{X}_i(t)$.
- 2) Calculate $\hat{\mu}(t) = \frac{1}{n} \sum_{i=1}^n \hat{X}_i(t)$. The sample covariance function is given by

$$\begin{aligned} \hat{K}(t, t') &= n^{-1} \sum_{i=1}^n \left\{ \hat{X}_i(t) - \hat{\mu}(t) \right\} \left\{ \hat{X}_i(t') - \hat{\mu}(t') \right\} \\ &= \sum_{j=1}^{\infty} \hat{\lambda}_j \hat{\phi}_j(t) \hat{\phi}_j(t'), \end{aligned} \quad (4)$$

$\{\hat{\lambda}_j, j \geq 1\}$ are the estimated eigenvalues, and $\{\hat{\phi}_j(\cdot), j \geq 1\}$ the estimated eigenfunctions. Both are obtained by spectral decomposition on the $\hat{K}(t, t')$.

- 3) Finally, the FPC scores ξ_{ik} are obtained by $\hat{\xi}_{ik} = \int_{\mathcal{T}} \{X_i(t) - \hat{\mu}(t)\} \hat{\phi}_k(t) dt$

The obtained FPC scores can be used as covariates or predictors in the machine learning model. As an example, we consider the following functional logistic regression model.

$$\Pr(Y = 1|X_i) = \Psi \left(\int_0^T \beta(t) X_i(t) dt \right), \quad (5)$$

where $\Psi(x) = \exp(x)/(1 + \exp(x))$ is the logistic function.

Based on the basis $\{\phi_k(t) : 1 \leq k < \infty\}$, we can expand $\beta(t)$ as

$$\beta(t) = \sum_{k=1}^{\infty} \phi_k(t) \beta_k, \quad (6)$$

where $\beta_k = \int_0^T \beta(t) \phi_k(t) dt$ is the basis coefficient.

Through functional principal component analysis, rewrite

$$\begin{aligned} \int_0^T X_i(t) \beta(t) dt &= \int_0^T \left\{ \mu(t) + \sum_{k=1}^{\infty} \xi_{ik} \phi_k(t) \right\} \beta(t) dt \\ &= \int_0^T \mu(t) \beta(t) dt + \sum_{k=1}^{\infty} \xi_{ik} \int_0^T \phi_k(t) \beta(t) dt \\ &= \int_0^T \mu(t) \beta(t) dt + \sum_{k=1}^{\infty} \xi_{ik} \beta_k \\ &\approx \int_0^T \mu(t) \beta(t) dt + \sum_{k=1}^K \xi_{ik} \beta_k. \end{aligned} \quad (7)$$

Thus the functional logistic regression model can be rewritten into a usual logistic regression model using the FPC scores as the predictors.

$$\Pr(Y = 1|X_i) = \Psi \left(\sum_{k=1}^K \xi_{ik} \beta_k \right). \quad (8)$$

Example 4.1: To illustrate the computation and interpretation of the FPC and FPC scores, we consider the trajectories of the sum of the total claim count. Fig. 2 plots the first three FPCs $\phi_1(\cdot)$ to $\phi_3(\cdot)$ in (1). The first FPC $\phi_1(\cdot)$ is flat and above zero, representing the main degree of variation around the mean function. The second FPC $\phi_2(\cdot)$ is downward sloping and crosses the zero axis once. Negative before year 3 and positive after year

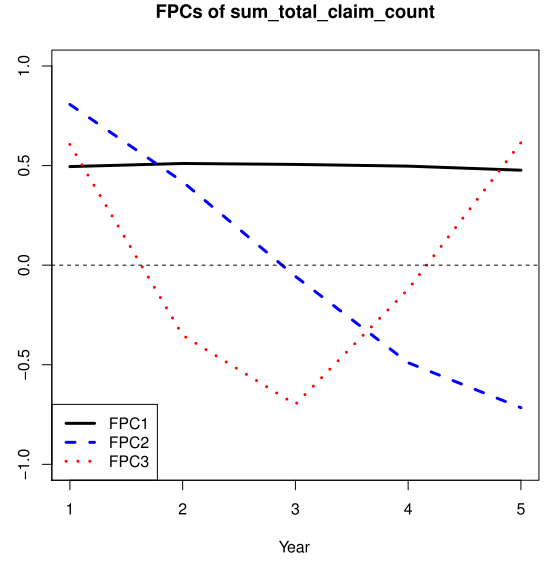


Fig. 2. The first three FPCs of the yearly trajectories of the sum of the total claim count in the Medicare database.

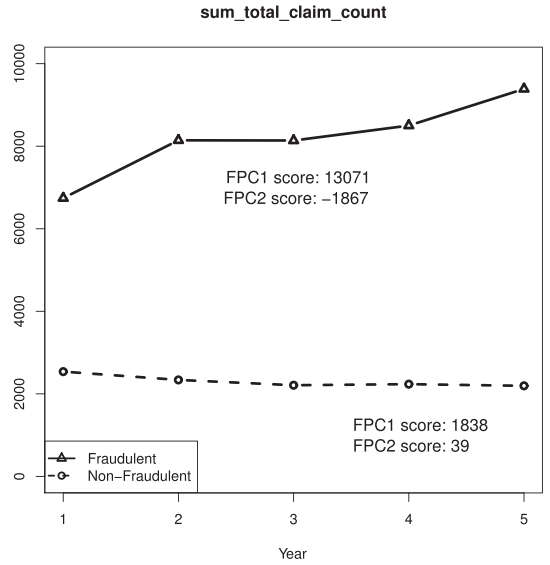


Fig. 3. Example of the yearly trajectories of the sum of the total claim count of a fraudulent case and a non-fraudulent case in the Medicare database.

3, the $\phi_2(\cdot)$ represents the degree of the change in the trajectory after year 3. The third FPC $\phi_3(\cdot)$ crosses the zero axis twice at year 2 and 4, i.e., it is negative in $[2, 4]$ and positive in the other two intervals, which can be interpreted as the difference in values during $[2, 4]$ and those in the other time intervals.

The computed FPC scores varies from subject to subject, and Fig. 3 shows two examples of the yearly trajectories of the features and their respective first and second FPC scores, one fraudulent case and one non-fraudulent case. As the first FPC score represents the main degree of variation, trajectories with a higher overall value tend to have a larger first FPC score. Since its overall level is higher, the fraudulent case has a larger first FPC score than the non-fraudulent case (13,071 versus 1,838). The second FPC score represents the degree of change, trajectories with a sharper change tend to have a larger (absolute) second

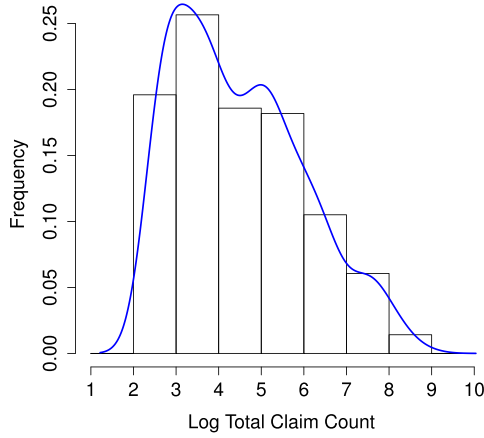


Fig. 4. Histogram and empirical probability density curve of the log of total claim count under a physician.

FPC score. The fraudulent case has a more evident upward trend while the non-fraudulent case is almost flat, and thus the absolute value of the second FPC score of the fraudulent case is larger than that of the non-fraudulent case (1,867 versus 39). It is worth noting that the sign of the FPC score does not matter because we may invert the FPC functions to obtain oppositely signed FPC scores.

3) *Distributional FPCA for Drug-Level Covariates*: The drug-related information in the data set includes the drug's total cost, the total number of claim count, the total number of beneficiaries, total 30-day fill count, and total daily supply under the physician and specialty in that given year. For each combination of physician and specialty type, multiple rows related to the information of a specific drug are available in the data set. The key covariates under each combination of physician and specialty type thus have a unique probability distribution across the past 5 years. For example, Fig. 4 shows a histogram of the distribution of the value of log total claim count under a specific physician. It is of particular interest to derive useful predictive features from the shape of such a distribution.

To achieve this goal, we resort to the distributional FPCA proposed by Petersen and Muller [24]. The idea is to conduct FPCA to the probability density curves of the distribution of a specific key covariate under a physician. The challenges of conducting FPCA on probability density curves are: first, a probability density function f has to satisfy the constraint $\int_{-\infty}^{\infty} f(t)dt = 1$; second, the probability density functions may have different support. To address these problems, a transformation approach that maps the probability density curves to a Hilbert space of functions through a continuous and invertible transformation can be used. The typical transformations include log quantile density and log hazard transformations.

After obtaining the kernel density estimate of the probability density curve, we may perform the FPCA on the transformed Hilbert space and use the FPC scores as the predictor variables. To summarize, the procedure for performing distributional FPCA for a certain drug-level key covariate is as follows.

- 1) For each physician, create a sub data set consisting of all the rows corresponding to the physician's NPI.

TABLE I
COST MATRIX FOR THE MEDICAL INSURANCE FRAUD PROBLEM

		True	
		Fraudulent	Legitimate
Predicted	Fraudulent	I	I
	Legitimate	$K \times DrugCost_i$	0

- 2) Compute the kernel density estimate of the key covariates in the sub data set, denoted as $f(t)$.
- 3) Map $f(t)$ to the quantile function $Q(t) = F^{-1}(t)$, where $F(t)$ is the cumulative distribution function; then, take the derivative of $Q(t)$, resulting in the quantile density function $q(t)$; lastly, take the log of $q(t)$ as our transformed function, denoted as $g(t)$. The function $g(t)$ is supported on $[0,1]$ and is unconstrained.
- 4) Repeat the above steps to obtain function g for all the physicians. Finally, conduct the regular FPCA on all the function g 's.

B. Cost-Sensitive Learning

Cost-sensitive learning models develop classification rules that are capable of reflecting the actual savings of detecting a fraudulent activity versus the actual cost of inspecting a suspicious activity. The detection of fraud inherently carries a financial tradeoff. If the potential saving in cost when the fraudulent activities are caught and stopped outweighs the cost of investigating the fraud, then it is beneficial to conduct the investigation. Otherwise, it might not be a worthwhile decision to start the investigation if the potential "gain" from a successful fraud intervention is too small. Thus, a framework of learning models called cost-sensitive learning is proposed to reflect the actual financial cost involved in binary classification problems. Various cost-sensitive learning methods are proposed, primarily in the area of credit card fraud detection [17].

For fraud detection problem, we define the confusion matrix of our binary classification model. A fraudulent case is defined as positive and non-fraudulent case as negative, and TP and FN respectively stand for the numbers of true positives and false negatives. The true positive (TP) case is where the predicted positive cases is truly fraudulent; the false positive (FP) case is where the predicted positive case is in fact legitimate. The true negative (TN) case is where the predicted negative case is legitimate; the false negative (FN) case is where the predicted legitimate case is actually fraudulent. Under the traditional framework, the false positive incurs the same cost as the false negative, which does not reflect the actual costing in the case of fraud investigation. For medical insurance, the administrative cost for fraudulent cases may consist of cost for investigation, evidence collection, litigation, etc. As shown in Table I, we assume that the administrative cost is fixed and denote it as I . The administrative cost is assigned to both TP and FP. On the other hand, the cost of TN is zero, and the cost of FN is defined as the total amount of fraudulently claimed drug cost. This amount is usually related to the total drug cost for under a specific physician, and we may assume that it equals to $K\%$

of the total drug cost of physician i . The value of K and I can be derived from historical data. While we do not possess such information, it is possible that reasonable estimates can be computed using the legal department's budgetary information and the litigation documents of the past fraudulent cases. Such a cost matrix reflects realistically the actual costs related to medical insurance fraud because when a fraudulent case goes undetected, its associated losses are the amount of the claimed drug cost.

Based on this cost matrix, we may define the following performance measure to reflect the total real cost associated with the fraud detection in each category.

$$C = \sum_{i=1}^n y_i \hat{y}_i I + K y_i (1 - \hat{y}_i) \text{Drug Cost}_i + (1 - y_i) \hat{y}_i I, \quad (9)$$

where y_i is the true binary fraud label, and \hat{y}_i is the predicted fraud label. This measure is the sum of the costs for all the physician and specialty types.

We consider the problem of fraud prediction under the Bayesian decision theoretic framework [17]. For each case, a choice needs to be made between two actions, predicting the case as fraudulent, denoted as a_f , or predicting it as legitimate, denoted as a_l . Let $L(a_f|f)$ (or $L(a_f|l)$) be the loss incurred for taking action a_f when the true state is f or l . Denote the data as D . The risk value for predicting a case as fraudulent is

$$R(a_f|D) = L(a_f|\text{fraud})P(\text{fraud}|D) + L(a_f|\text{legitimate})P(\text{legitimate}|D). \quad (10)$$

Similarly, let $L(a_l|f)$ (or $L(a_l|l)$) be the loss incurred for taking action a_l when the true state is f or l . The risk value for predicting a case as legitimate is

$$R(a_l|D) = L(a_l|\text{fraud})P(\text{fraud}|D) + L(a_l|\text{legitimate})P(\text{legitimate}|D). \quad (11)$$

Here, $P(\text{legitimate}|D)$ and $P(\text{fraud}|D)$ can be understood as the predicted probability of being fraud/legitimate under the uniform prior distribution. Under Bayesian decision theory, a case would be classified as fraud if $R(a_f|D) < R(a_l|D)$, i.e., the risk of predicting the case as fraud is lower than the risk of predicting it as legitimate.

By plugging in the values in the cost matrix into the computation of $R(a_f|D)$ and $R(a_l|D)$, we classify a case as fraud if

$$I < K \times \text{Drug Cost}_i P(\text{fraud}|D). \quad (12)$$

The calculation of $P(\text{fraud}|D)$ depends on the machine learning algorithm deployed for modeling and the ratio of undersampling. For example, in the case of logistic regression, let $P^*(\text{fraud}|D)$ denote the predicted probability from the logistic function of the product sum of the linear coefficients and the predictors trained in the undersampled data set. The predicted fraud probability is $P(\text{fraud}|D) = \frac{p_{full}}{p_{RUS}} P^*(\text{fraud}|D)$, where p_{full} is the proportion of fraud in the full data set, and p_{RUS} the proportion of fraud in the undersampled data set.

Example 4.2: To illustrate how the decision is made under the cost-sensitive learning framework, we revisit the example of the fraudulent versus non-fraudulent cases in Section IV-A2. For simplicity, we focus on the truly fraudulent case. Suppose that based on the feature engineering and the machine learning models, the probability estimate of the case being fraudulent is $P(\text{fraud}|D) = 0.873$, and its drug cost in the most recent year is $\text{Drug Cost}_i = 5000$. We assume the fixed administrative cost for investigating the fraud is $I = 4000$ and the total amount of fraud is $K \times \text{Drug Cost}_i$, where $K = 20$. Justification for setting these values can be found in Section V-C.

Under the cost-sensitive learning framework, we are concerned with the risk value associated with classifying a case as fraudulent or as legitimate. As both $L(a_f|\text{fraud})$ and $L(a_f|\text{legitimate})$ is equal to I . The risk value for predicting the case as fraud is simply $R(a_f|D) = I = 4000$, i.e., an investigation into the suspected fraudulent cases will incur a fixed cost of 4,000. On the other hand, as $L(a_l|\text{fraud}) = K \times \text{Drug Cost}_i$ and $L(a_l|\text{legitimate}) = 0$, the risk value for predicting the case as legitimate is $K \times \text{Drug Cost}_i \times P(\text{fraud}|D) = 20 \times 5000 \times 0.873 = 87300$, i.e., the potential expected loss of being oblivious to the fraudulent case would be equal to the probability of fraud times the total fraud amount. In this case, the risk value for predicting the case as fraud is smaller than predicting it as legitimate, and thus the case is classified as fraudulent and worthy of further investigation.

On the other hand, for example, if the total drug cost is very low, e.g., $\text{Drug Cost}_i = 100$. The risk value for predicting the case as legitimate would be $K \times \text{Drug Cost}_i \times P(\text{fraud}|D) = 20 \times 100 \times 0.873 = 1746$, and thus the expected loss due to the suspected fraudulent case is less than the investigational cost of 4,000. In this case, the cost-sensitive model would deem the case unworthy of further investigation and would classify the case as legitimate, despite that the predicted probability of the case being fraudulent is as high as $P(\text{fraud}|D) = 0.873$.

C. Class Rarity

The number of fraudulent cases is extremely rare in the Medicare data. For our data set, there are a total of 918,009 instances, among which only 227 cases are fraudulent. This implies that only 0.025% of the cases are positive.

The severe imbalance between the positive and negative classes may lead to the biased predictive performance of the commonly used machine learning algorithms because the classifier might be tilted towards predicting the cases as negative in order to achieve high accuracy. Moreover, in the case of fraud detection, the rare positive class incurs a much higher cost than the majority negative class. Thus, an algorithm's bias towards the majority class could lead to high costs.

Strategies for dealing with class imbalance usually involve creating a new data set with a class distribution less imbalanced than the original and trains the model based on such an adjusted data set. Two commonly used approaches are random undersampling (RUS) and random oversampling. Random undersampling retains all the minority class cases and randomly removes cases

TABLE II
CLASS DISTRIBUTION IN THE TRAINING, TESTING AND UNDERSAMPLED
DATA SET

Data	Cases	Frauds	Fraud percentage
Full	918,009	227	0.025%
Training	459,004	113	0.025%
Testing	459,005	114	0.025%
RUS-1	11,300	113	1%
RUS-5	2,260	113	5%
RUS-10	1,130	113	10%
RUS-20	565	113	20%
RUS-50	226	113	50%

from the majority class. Random oversampling creates duplicates from the minority class such that a particular class ratio is achieved.

We choose to use the random undersampling approach to handle the highly imbalanced data set for our application because random undersampling carries the least amount of computation burden compared with other approaches. We consider 5 different class ratio, 1%, 5%, 10%, 20%, and 50%, which are denoted as RUS-1, RUS-5, RUS-10, RUS-20, and RUS-50. The goal is to assess the predictive performance and cost saving measure under different class ratio. Table II shows the number of samples in each data set. It is worth noting that only the training data set is transformed with random undersampling while the testing data set still represents the true class distribution.

V. RESULTS

A. Experimental Setting

We experiment with four machine learning algorithms from the h2o library in Java/R, namely the logistic regression (LR), random forest (RF), gradient boosting machine (GBM), and neural network (NN). Unless explicitly stated otherwise, we use the default configuration of the algorithm. The optimal tuning parameters for the algorithms are chosen through 4-fold cross-validation on the undersampled training data set. The LR uses the logistic function as the link function between the probability and the product sum of the covariates and coefficients. We add an L1 penalty to the coefficients in order to select the covariates with predictive values. RF is an ensemble learning algorithm that predicts the outcome based on combining outcomes from numerous decision trees. The prediction results are computed by summarizing results from multiple trees through majority voting. Each decision tree is trained from a random subset sampled from the training data with replacement. The number of trees is chosen as 50, with a maximum tree depth of 20. GBM is an ensemble method that sequentially trains a decision tree to maximize the log-likelihood function. In each round of training, the predicted class from the model is compared with the actual class, allowing more weight to be assigned to the misclassified instances. We set the learning rate of the GBM algorithm to be 0.1 and the maximum number of trees to be 50. NN is a multi-layer model that is highly suited for complex and nonlinear classification problems. For the NN model, we choose the number of layers to be 2 and the number of neurons 200, with a learning rate of 0.005.

B. Performance Metrics

For binary classification problems, we consider two commonly used metrics, the F1-measure and the AUC, to evaluate the classification accuracy. The F1-measure is derived from precision and recall, which are defined as follows.

- Precision = $TP / (TP + FP)$
- Recall = $TP / (TP + FN)$
- F1 = $2 \cdot \text{Precision} \cdot \text{Recall} / (\text{Precision} + \text{Recall})$

In addition, another metric is the AUC, which summarizes the area under the receiver operating characteristic (ROC) curve. The ROC curve consists of measurements of performance metrics at various probability thresholds. The AUC is a summary of a classifier's capability of distinguishing between the classes and takes a value between 0 and 1. The larger the AUC, the better the predictive performance. A perfect classifier would have an AUC equal to 1.

For cost-related metrics, we consider the following measures. First, the total cost measure is defined as

$$C_{total} = \sum_1^n y_i \hat{y}_i I + K y_i (1 - \hat{y}_i) Drug Cost_i + (1 - y_i) \hat{y}_i I, \quad (13)$$

where y_i is the true binary fraud label, and \hat{y}_i is the predicted fraud label.

In addition, we may decompose the total cost measure into two categories, one consisting of the fixed investigational cost, and the other the cost related to fraud,

$$C_{fixed} = \sum_1^n \hat{y}_i I, \quad (14)$$

and

$$C_{fraud} = \sum_1^n K y_i (1 - \hat{y}_i) Drug Cost_i. \quad (15)$$

Moreover, we compute the cost saving percentage, which equals to $C_{total} / C_{null} - 1$, where C_{null} is the cost where no fraud investigation is carried out,

$$C_{null} = \sum_1^n K y_i Drug Cost_i. \quad (16)$$

C. Data Analysis

In this section, we present the results of our data analysis. As mentioned before, the model is trained on five undersampled data sets with class ratios equal to 99:1, 95:1, 90:10, 80:20, 50:50, respectively denoted as RUS-1, RUS-5, RUS-10, RUS-20, and RUS-50. As shown in Table II, there are only 227 positive cases in the whole data set. Because of such a small number of positive cases, the usual 5-fold cross-validation approach (i.e., 181 positive instances in the training set versus 46 positive instances in the testing set) would often result in unstable and unreliable performance in the testing data set. To obtain a relatively stable fit, we use a 1:1 training-versus-testing ratio and repeat the experiment 10 times to obtain the average performance metrics.

TABLE III

TOTAL COST MEASURE, FIXED COST AND FRAUD COST (IN MILLIONS) UNDER VARIOUS UNDERSAMPLING SCHEMES (RUS) AND MACHINE LEARNING ALGORITHMS (LOGISTIC REGRESSION (LR); GRADIENT BOOSTING MACHINE (GBM); NEURAL NETWORK (NN); AND RANDOM FOREST (RF)) WITH THE COST-SENSITIVE (CS) AND CONVENTIONAL NON COST-SENSITIVE (NCS) APPROACHES

Model	Learning	RUS-1	RUS-5	RUS-10	RUS-20	RUS-50
<i>Total Cost</i>						
LR	CS	335.30	312.68	269.09	326.93	297.61
	NCS	435.99	427.01	416.71	650.68	673.22
GBM	CS	282.12	237.45	256.37	390.80	366.01
	NCS	532.24	447.38	523.24	555.51	547.53
NN	CS	511.01	498.93	492.42	497.18	460.80
	NCS	510.39	549.69	687.68	758.88	601.06
RF	CS	289.40	239.94	261.56	303.67	293.65
	NCS	489.50	448.28	572.32	547.79	558.60
<i>Fixed Cost</i>						
LR	CS	61.26	85.20	70.34	62.25	46.68
	NCS	20.41	59.00	148.72	496.72	490.69
GBM	CS	42.39	57.73	62.70	58.62	46.74
	NCS	8.57	41.30	73.70	107.32	173.73
NN	CS	0.42	3.09	1.91	7.06	10.30
	NCS	3.96	35.23	210.72	335.30	477.02
RF	CS	115.64	120.72	105.34	84.92	51.64
	NCS	12.77	109.24	146.27	108.37	427.39
<i>Fraud Cost</i>						
LR	CS	274.03	227.49	198.75	264.68	250.93
	NCS	415.59	368.01	267.99	153.95	182.53
GBM	CS	239.73	179.72	193.67	332.19	319.27
	NCS	523.67	406.08	449.54	448.20	373.81
NN	CS	510.59	495.84	490.51	490.12	450.50
	NCS	506.43	514.46	476.96	423.59	124.04
RF	CS	173.76	119.22	156.22	218.74	242.01
	NCS	476.73	339.04	426.06	439.42	131.21

It is also worth noting that if the machine learning model is trained using the full training data set, without adjusting for the class imbalance, the model fit and the performance metrics in the testing data set is often very poor, e.g., precision equal to 0 and F1-measure undefined due to division of 0 error. Therefore, we only train the model on the undersampled data set, and evaluate its performance on the full testing data set.

Regarding the ratio between the fraud amount and the drug cost K , we have the following consideration. By summarizing the news of health care fraud published on the Federal Bureau of Investigation website, we note that the amount of the fraud usually ranges from 1 million to 50 million [27]. On the other hand, the mean value of the total drug cost column in our database is around 1 million. Thus, we take the ratio between the fraud amount and the drug cost K to be equal to 20. This value is reasonable because, first, when fraud is detected, oftentimes, fraudulent activities have already been going on for many years. Second, a physician may have means other than through drug costs to create artificial medical claims. Regarding the fixed cost I for fraud investigational, we set it to be 4,000.

For each of the four machine learning algorithms, we consider two approaches for predicting the fraud label. The first is cost-sensitive learning, as described in the previous sections. The second is a conventional non cost-sensitive approach based on maximizing the F1-measure.

Table III outlines the total cost measure along with fixed cost and fraud cost under different algorithms across the five undersampling ratios. We compare both the cost-sensitive and the

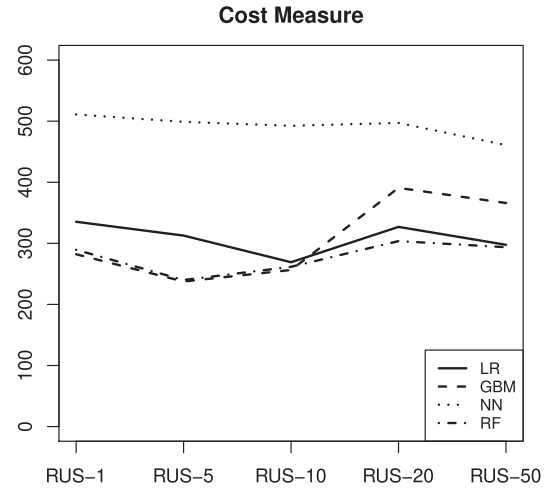


Fig. 5. Comparison of the cost measure (in millions) for different machine learning algorithms and undersampling schemes.

non-cost-sensitive approaches with each one of the four machine learning algorithms. Among the cost-sensitive approaches, the RF algorithm renders a steady good performance. It produces a total cost of mostly under 300 million for all ratios with the cost-sensitive approach. LR and GBM obtain total cost measures over a slightly bigger range: some can be as low as around 260 million, yet others can reach as high up to around 400 million depending on the class ratio. NN has the highest total cost as all of the total costs are sitting at around 500 million, while other algorithms attain total costs far below this number. In general, lower total costs are achieved at RUS-1 and RUS-10; higher cost measures are achieved at the ratio of RUS-20 and RUS-50. For the non-sensitive approaches, the total costs obtained are higher than those of the cost-sensitive approach. GBM and RF achieve relatively lower total costs than LR and NN, although none of the costs obtained are small enough to be deemed more favorable than the cost-sensitive approach.

The fixed cost measure is related to the number of predicted positive cases. Among the cost-sensitive methods, the RF algorithm results in higher fixed costs across all the class ratio, and NN renders the smallest fixed costs of around and below 10 million, indicating an insufficient number of fraud predictions. Regarding the fraud cost measure attained by the four algorithms under the cost-sensitive framework, RF achieves the best result, i.e., the lowest fraud cost of around 200 million consistent through all the class ratios. NN renders the highest fraud costs at around 500 million for all ratios, and LR and GBM render better results at around 200-300 million. The superiority of RF over other algorithms in terms of fraud cost is only present in a cost-sensitive approach. For non cost-sensitive approach, all four algorithms seem to have similar results.

Fig. 5 presents the comparison of the cost measure induced by the four algorithms at different class ratios. The cost measures in millions across all ratios for each algorithm are connected by a fragmented line and marked by a designated color. Among the four algorithms tested, NN achieves the highest 500 million cost measure for all ratios. For the other algorithms, it is evident that LR, GBM and RF come by the highest cost measure at class ratio

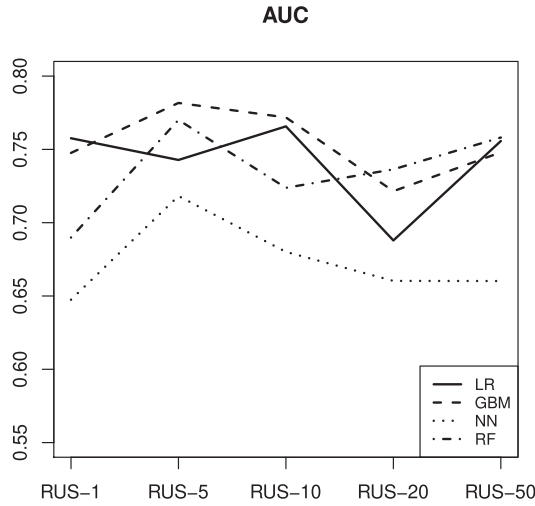


Fig. 6. Comparison of the AUC for various undersampling schemes (RUS) and machine learning algorithms (logistic regression (LR); gradient boosting machine (GBM); neural network (NN); and random forest (RF)).

TABLE IV
COST SAVING PERCENTAGE UNDER VARIOUS UNDERSAMPLING SCHEMES (RUS) AND MACHINE LEARNING ALGORITHMS (LOGISTIC REGRESSION (LR); GRADIENT BOOSTING MACHINE (GBM); NEURAL NETWORK (NN); AND RANDOM FOREST (RF)) WITH THE COST-SENSITIVE (CS) AND CONVENTIONAL NON COST-SENSITIVE (NCS) APPROACHES

Model	Learning	RUS-1	RUS-5	RUS-10	RUS-20	RUS-50
LR	CS	-37.0%	-41.3%	-49.5%	-38.6%	-44.1%
	NCS	-18.1%	-19.8%	-21.7%	22.2%	26.4%
GBM	CS	-47.0%	-55.4%	-51.9%	-26.6%	-31.3%
	NCS	-0.1%	-16.0%	-1.7%	4.3%	2.8%
NN	CS	-4.0%	-6.3%	-7.5%	-6.6%	-13.5%
	NCS	-4.2%	3.2%	29.1%	42.5%	12.9%
RF	CS	-45.7%	-54.9%	-50.9%	-43.0%	-44.9%
	NCS	-8.1%	-15.8%	7.5%	2.9%	4.9%

80:20; GBM produces the highest cost measure of about 400 million. Overall, RF achieves the best performance and produces a steadily low cost measure with the lowest of around 240 million at the ratio of 95:1. GBM and LR have moderate performance; GBM also produces the same lowest cost measure of around 240 million at 95:1, yet a high cost measure of 400 million at 80:20. LR produces the lowest measure of 290 million at 90:10 and the highest measure at 310 million at 80:20. In summary, NN has the worst performance in cost measure; GBM and LR have medium performance; RF has the best performance.

Table IV shows the cost saving percentage of the four machine learning algorithms for 5 class ratios under cost-sensitive and non cost-sensitive approaches. In terms of this performance metric, the more negative the percentage, the larger the saving in cost and hence the more preferable the model. A positive percentage indicates that carrying out fraud investigation according to the model cost even more than doing no investigation at all. Overall, we observe that the cost-sensitive methods achieve evidently better saving in cost than the non cost-sensitive ones. GBM achieves the best cost saving percentage of 55.4% at the class ratio of RUS-5, yet the performance of GBM is not as consistent for all the ratios as RF under cost-sensitive approach. The range of cost saving percentage achieved by GBM is 26% – 55.4%

TABLE V
F1-MEASURE ($\times 10^{-4}$) UNDER VARIOUS UNDERSAMPLING SCHEMES (RUS) AND MACHINE LEARNING ALGORITHMS (LOGISTIC REGRESSION (LR); GRADIENT BOOSTING MACHINE (GBM); NEURAL NETWORK (NN); AND RANDOM FOREST (RF)) WITH THE COST-SENSITIVE (CS) AND CONVENTIONAL NON COST-SENSITIVE (NCS) APPROACHES

Model	Learning	RUS-1	RUS-5	RUS-10	RUS-20	RUS-50
LR	CS	23.34	12.15	16.96	8.93	13.59
	NCS	72.95	37.69	25.21	10.30	11.24
GBM	CS	24.29	23.38	20.28	10.84	11.87
	NCS	44.46	32.59	25.90	24.50	20.21
NN	CS	377.36	68.26	171.23	42.74	14.91
	NCS	127.50	17.95	11.37	11.68	10.89
RF	CS	12.41	15.19	12.86	11.25	12.29
	NCS	66.67	24.80	20.18	26.47	12.53

Standard errors of all entries do not exceed 4×10^{-4} .

TABLE VI
AUC UNDER VARIOUS UNDERSAMPLING SCHEMES (RUS) AND MACHINE LEARNING ALGORITHMS (LOGISTIC REGRESSION (LR); GRADIENT BOOSTING MACHINE (GBM); NEURAL NETWORK (NN); AND RANDOM FOREST (RF))

Model	RUS-1	RUS-5	RUS-10	RUS-20	RUS-50
LR	0.7576	0.7428	0.7657	0.6879	0.7557
GBM	0.7476	0.7817	0.7718	0.7215	0.7480
NN	0.6475	0.7182	0.6801	0.6603	0.6602
RF	0.6899	0.7701	0.7238	0.7364	0.7581

while the range achieved by RF is 43% – 54.9%. Therefore, GBM has the best peak performance, yet RF has a steadily excellent performance. LR attains the medium performance, which yields cost saving percentages in the range of 37.0% to 49.5% under the cost-sensitive approach. NN yields the worst results: it achieves cost saving of less than 15% for cost-sensitive approach and even positive percentage for non-cost-sensitive approach. Under non cost-sensitive approaches, LR, GBM and RF exhibit more varied patterns: GBM and RF are no longer the preferable models for calculating cost saving percentages since there are unwanted positive percentages; LR generates negative percentages at some ratios, but they are not desirable percentages. The poor performance of NN might be attributed to the small size of undersampled data set, which makes the NN more prone to producing overfitted predicted probabilities and cost measures.

Table V outlines the F1-measure of the four algorithms of interest at five different class ratios under both cost-sensitive and non-sensitive approaches. Overall, the non cost-sensitive approach works better than cost-sensitive approach for all models at all ratios. Based on all the number present on the chart, NN is the best performing model as it achieves the top 3 highest scores: it achieves the highest F1-measure of 0.0377 at ratio RUS-1 under non cost-sensitive approach; it achieves the second-highest score of 0.171 at RUS-10; NN also achieves the third-highest score of 0.128 at RUS-1 under cost-sensitive approach. The other models all produce low F1 measures under both approaches, which are mostly under 0.0100.

Table VI and Fig. 6 compares the AUC under the four machine learning methods. LR achieves the highest AUC of 0.77 at RUS-10; GBM does of 0.78 at RUS-5; RF does of 0.76 at RUS-5, NN does of 0.72 at RUS-5. In general, GBM, RF and NN all have

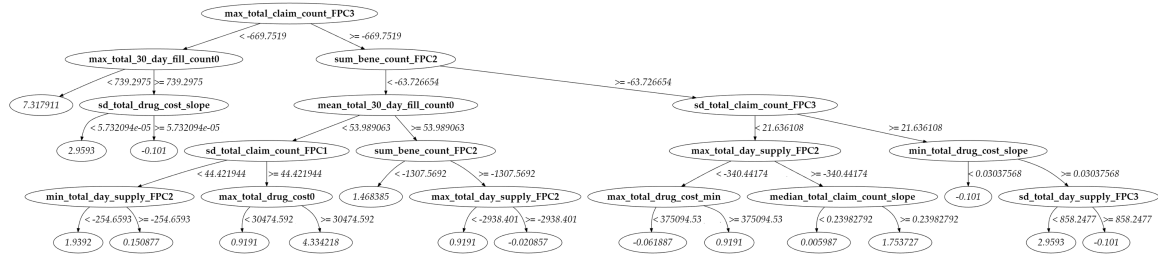


Fig. 7. The decision tree with the highest weighting under the gradient boosting machine with RUS-1. The values in leaf node are the logit values of the predicted probabilities of fraud.

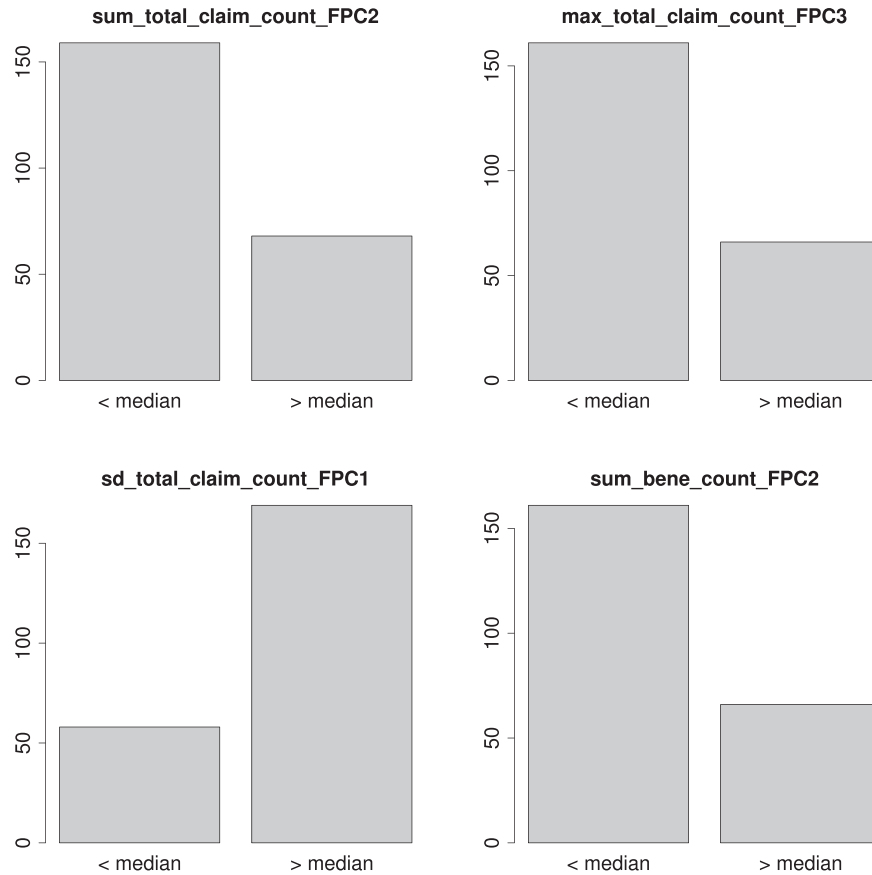


Fig. 8. Distribution of the fraudulent cases when the FPC score is less/greater than the median, for the top 4 FPCA-based features with the highest variable importance measure in the gradient boosting machine.

their own lowest AUC at the class ratio of 99:1, denoted by RUS-1. On the other hand, RUS-5 and RUS-10 appear to be the better-performing ratios for all four learner algorithms. As NN produces the smallest AUC at all ratios compared with other methods, therefore it is the worst-performing model in terms of AUC. GBM achieves the highest AUC of 0.78 at RUS-5 and also has steadily good performance across all ratios. RF and LR may achieve a good AUC at a particular ratio, yet their AUCs have a large degree of fluctuation across all the ratios. The most favorable model with steady performance in terms of high AUC is the GBM.

To illustrate the role of the FPCA-based features in the model, we examine the most important decision tree from the tree-based gradient boosting machine algorithm. Fig. 7

shows the plot of the decision tree with the highest weighting under the gradient boosting machine with RUS-1 scheme. We observe that the FPCA-based features play a crucial role in determining the label prediction. For example, the third FPC score of the maximum total claim count is situated at the root node of the tree, where the tree are divided based on whether or not the score is greater/smaller than -669.7519 . Beyond that, the FPCA-based features appear across various nodes of the tree. Moreover, we consider plotting the label distributions of the fraudulent cases when the FPC score is less than the median and when score is greater than the median. We choose the top 4 FPCA-features with the highest variable importance measure from the gradient boosting machine, namely the second FPC score of the sum of the total claim count, the third FPC

score of the maximum of the total claim count, the first FPC score of the standard deviation of the total claim count, and the second FPC score of the sum of the beneficiary count. As shown in Fig. 8, these FPCA-based features evidently distinguish between the fraudulent and non-fraudulent cases, indicating the predictiveness inherent in these FPCA-based features.

D. Computational Scalability

Next, we examine the computational scalability of our method. As there are 62.2 million people enrolled in Medicare in 2020 and new data are being created on a yearly basis, the total size of the database could potentially increase tremendously, both in terms of the number of rows in a year and also in terms of the number of years available. It is thus worthwhile to examine the scalability of our method on a large dataset. To do this, we conduct an experiment on a synthetic dataset with 100 million records and examine the computational time. Our focus is on the computational time of feature creation and conducting FPCA, as it is well known that the machine learning algorithms adopted in our methods have scalable and ready-to-use software implementation. We generate 100 million random curves from a stochastic process with three underlying orthogonal components. From each curve, we sample 20 data points, which correspond to 20 years of data. Using the R package “fdapace”, we found that on a 12-cores Intel machine, the creation of the yearly numerical features (max, min, sum, etc) only takes around 20 minutes, while the steps for conducting FPCA take around 4 hours. It is also worth noting that the memory requirement of the program is manageable. This is because the covariance matrix in (4) is only of size 20×20 and the largest matrix involved in the FPC score computation is the raw data matrix ($100 \text{ million} \times 20$).

VI. POLICY IMPLICATION

The CMS has partnered with different law enforcement agencies and initiated a series of anti-fraud programs (The Centers for Medicare and Medicaid Services, 2020). Investigations have been carried out on suspicious cases through, e.g., beneficiary/provider interview. The scale of the fraudulent activities and, most importantly, the potential amount of money that can be recovered are naturally taken into consideration when it comes to resource and personnel allocation of the CMS. However, there lacks a systematic and data-driven approach that can guide decision-making and answer the question “who should be investigated”. The results in this section indicate that significant improvement (more than 50%) in the cost-efficiency can be achieved if existing information in the database can be intelligently utilized, and thus call for a change in the policy implementation of the fraud prevention system. Currently, most fraud cases have been exposed by whistleblowers [28]. Instead of passively awaiting the exposure of fraudulent activities by whistleblowers, it would be better if a data-driven and cost-effective approach that guides the direction of investigation can be adopted at the policy implementation level.

VII. DISCUSSION

In this article, we tackle the problem of medical insurance fraud detection by utilizing information from previous years. We construct temporal trajectories of the key covariates and base our feature engineering around these trajectories. We introduce the framework of cost-sensitive learning for analyzing the Medicare database. It is important to allow for asymmetrical losses that are associated with the FP and FN cases, such that the classification rule reflects the realistic tradeoff between the fixed cost and the fraud cost.

While there have been many developments in the area of fraud detection, the use of complex statistical models and methods is relatively uncommon. One major highlight of our approach lies in the use of functional principal component analysis to extract key information from the data. By proposing a solution to the healthcare fraud detection problem, we hope that our method can serve as a small step towards the utilization of statistical methods such as functional principal component analysis in solving real problems. We introduce the FPCA methods for analyzing the temporal covariates’ trajectory as well as the distributional FPCA for extracting features from the empirical probability density curve of the covariates. Our results indicate that the trained classifier has a reasonably good predictive performance. A significant percentage of cost saving can be achieved by taking into account the financial cost. In terms of the cost-saving percentage, the RF or GBM algorithm with RUS-5 achieves the largest saving percentage of more than 50%. In terms of predictive capability, the GBM model has the largest AUC when trained with the RUS-5 data set. Future work in the area of medical fraud detection using the Medicare database includes incorporating additional key covariates from the Part B and DEMPOS database, expanding the fraud labels by finding NPIs through name-matching in the National Plan and Provider Enumeration System (NPPES) for the records whose NPIs are NPIs in the LEIE database and exploring more cost-sensitive learning algorithms where the financial cost are incorporated not only in the decision rule but also in the training stage of the model.

We adopt the random undersampling method for the class imbalance problem. Recently, Hasanin et al. [25] experimented with various approaches of their model to address the class imbalance problem on the Medicare dataset. They concluded that the RUS with a 50:50 ratio leads to the best performance in many cases. A useful extension of our method would be to experiment with other sampling techniques, e.g., the Synthetic Minority Over-sampling TEchnique (SMOTE), which oversamples the data set by generating new instances between minority instances close to one another [26].

REFERENCES

- [1] Centers for Medicare & Medicaid Services, 2020. Accessed: 2020. [Online]. Available: <https://www.cms.gov/>
- [2] A. Rashidian, H. Joudaki, and T. Vian, “No evidence of the effect of the interventions to combat health care fraud and abuse: A systematic review of literature,” *PLoS One*, vol. 7, 2012, Art. no. 41988.
- [3] Centers for Medicare & Medicaid Services, “Medicare fraud and abuse: Prevention, detection, and report,” 2020. Accessed: 2020. [Online]. Available: <https://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNProducts/Downloads/Fraud-Abuse-MLN4649244.pdf>

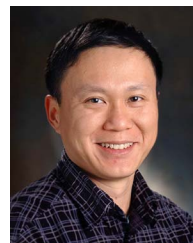
- [4] S. Ullah and C. F. Finch, "Applications of functional data analysis: A systematic review," *BMC Med. Res. Methodol.*, vol. 13, 2013, Art. no. 43.
- [5] M. Herland, T. M. Khoshgoftaar, and R. A. Bauder, "Big Data fraud detection using multiple medicare data sources," *J. Big Data*, vol. 5, pp. 29–35, 2018.
- [6] J. M. Johnson and T. M. Khoshgoftaar, "Medicare fraud detection using neural networks," *J. Big Data*, vol. 6, pp. 63–69, 2019.
- [7] R. A. Bauder and T. M. Khoshgoftaar, "A probabilistic programming approach for outlier detection in healthcare claims," in *Proc. IEEE 15th Int. Conf. Mach. Learn. Appl.*, 2016, pp. 347–354.
- [8] R. A. Bauder and T. M. Khoshgoftaar, "A novel method for fraudulent medicare claims detection from expected payment deviations," in *Proc. IEEE 17th Int. Conf. Reuse Integration*, 2016, pp. 11–19.
- [9] R. A. Bauder, T. M. Khoshgoftaar, A. Richter, and M. Herland, "Predicting medical provider specialties to detect anomalous insurance claims," in *Proc. IEEE 28th Int. Conf. Tools Artif. Intell.*, 2016, pp. 784–790.
- [10] M. Herland, R. A. Bauder, and T. M. Khoshgoftaar, "Medical provider specialty predictions for the detection of anomalous medicare insurance claims," in *Proc. IEEE Int. Conf. Inf. Reuse Integration*, 2017, pp. 579–588.
- [11] J. Ko et al., "Variability in medicare utilization and payment among urologists," *Urology*, vol. 85, pp. 1045–1051, 2015.
- [12] V. Chandola, S. R. Sukumar, and J. C. Schryver, "Knowledge discovery from massive health-care claims data," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2013, pp. 1312–1320.
- [13] L. K. Branting, F. Reeder, J. Gold, and T. Champney, "Graph analytics for healthcare fraud risk estimation," in *Proc. ACM Int. Conf. Adv. Social Netw. Anal. Mining*, 2016, pp. 845–851.
- [14] National Plan & Provider Enumeration System. NPPES NPI registry, 2020. Accessed: 2020. [Online]. Available: <https://nppesregistry.cms.hhs.gov/registry/>
- [15] N. Khurjekar, C. A. Chou, and M. T. Khasawneh, "Detection of fraudulent claims using hierarchical cluster analysis," in *Proc. IIE Annu. Conf. Inst. Ind. Syst. Engineers*, 2015, pp. 2388–2392.
- [16] S. Sadiq, Y. Tao, Y. Yan, and M. Shyu, "Mining anomalies in medicare Big Data using patient rule induction method," in *Proc. IEEE 3rd Int. Conf. Multimedia Big Data*, 2017, pp. 185–192.
- [17] A. C. Bahnsen, A. Stojanovic, D. Aouada, and B. Ottersten, "Cost sensitive credit card fraud detection using bayes minimum risk," in *Proc. 12th Int. Conf. Mach. Learn. Appl.*, 2013, pp. 333–338.
- [18] A. C. Bahnsen and A. Stojanovic, "Example-dependent cost-sensitive decision trees," *Expert Syst. Appl.*, vol. 42, pp. 6609–6619, 2015.
- [19] N. Sanaz and S. Mehdi, "Cost-sensitive payment card fraud detection based on dynamic random forest and k-nearest neighbors," *Expert Syst. Appl.*, vol. 50, pp. 6012–6020, 2018.
- [20] S. Akila and U. Reddy, "Cost sensitive risk induced Bayesian inference bagging (RIBIB) for credit card fraud detection," *J. Comput. Sci.*, vol. 27, pp. 247–254, 2018.
- [21] F. Feng, K.-C. Li, J. Shen, Q. Zhou, and X. Yang, "Using cost-sensitive learning and feature selection algorithms to improve the performance of imbalanced classification," *IEEE Access*, vol. 8, pp. 69979–69996, 2020.
- [22] M. Lzaro and A. R. Figueiras-Vidal, "A bayes risk minimization machine for example-dependent cost classification," *IEEE Trans. Cybern.*, vol. 51, no. 7, pp. 3524–3534, Jul. 2021, doi: [10.1109/TCYB.2019.2913572](https://doi.org/10.1109/TCYB.2019.2913572).
- [23] F. Yao, H. G. Muller, and J. L. Wang, "Functional linear regression analysis for longitudinal data," *Ann. Statist.*, vol. 33, pp. 2873–2903, 2005.
- [24] A. Petersen and H. G. Muller, "Functional data analysis for density functions by transformation to a hilbert space," *Ann. Statist.*, vol. 44, pp. 183–218, 2016.
- [25] T. Hasanin, T. M. Khoshgoftaar, and J. L. Leevy, "Severely imbalanced Big Data challenges: Investigating data sampling approaches," *J. Big Data*, vol. 6, 2019, Art. no. 107.
- [26] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.
- [27] Federal Bureau of Investigation, "Health care fraud news," 2020. Accessed: 2020. [Online]. Available: <https://www.fbi.gov/investigate/white-collar-crime/health-care-fraud/health-care-fraud-news>
- [28] L. Shi and D. A. Singh, *Delivering Health Care in America. A Systems Approach*. Sudbury, MA, USA: Jones and Bartlett Publishers, 2008.
- [29] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel, "Deep learning for anomaly detection: A review," *ACM Comput. Surv.*, vol. 54, no. 2, pp. 1–38, 2021.
- [30] Q. Liu and M. Vasarhelyi, "Healthcare fraud detection: A survey and a clustering model incorporating geo-location information," in *Proc. 29th World Continuous Auditing Reporting Symp.*, Brisbane, Australia, 2013, pp. 1–10.



Haolun Shi is an Assistant Professor with the Department of Statistics and Actuarial Science, Simon Fraser University. His research interests include Bayesian modeling, clinical trial design, functional data analysis, and statistics in sports.

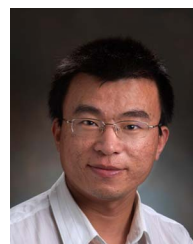


Mohammad A. Tayebi is a research associate with the School of Computing Science, Simon Fraser University. His research interests include social network analysis, social computing, machine learning for social good, and machine learning for cybersecurity.



Jian Pei (Fellow, IEEE) is a professor with the School of Computing Science, Simon Fraser University. He is a renowned leading researcher with the general areas of data science, Big Data, data mining, and database systems. He is recognized as a fellow of the Royal Society of Canada (Canada's national academy), the Canadian Academy of Engineering, ACM. He is one of the most cited authors in data mining, database systems, and information retrieval. Since 2000, he has published one textbook, two monographs and more than 300 research papers in refereed journals

and conferences, which have been cited extensively by others. His research has generated remarkable impact substantially beyond academia. For example, his algorithms have been adopted by industry in production and popular open source software suites. He also demonstrated outstanding professional leadership in many academic organizations and activities. He was the editor-in-chief of the *IEEE Transactions of Knowledge and Data Engineering (TKDE)* in 2013–2016, the chair of ACM SIGKDD, in 2017–2021, and a general co-chair or program committee co-chair of many premier conferences. He maintains a wide spectrum of industry relations with both global and local industry partners. He is an active consultant and coach for industry. He received many prestigious awards, including the 2017 ACM SIGKDD Innovation Award, the 2015 ACM SIGKDD Service Award, the 2014 IEEE ICDM Research Contributions Award, the British Columbia Innovation Council 2005 Young Innovator Award, an NSERC 2008 Discovery Accelerator Supplements Award, an IBM Faculty Award (2006), a KDD Best Application Paper Award (2008), an ICDE Influential Paper Award (2018), a PAKDD Best Paper Award (2014), and a PAKDD Most Influential Paper Award (2009).



Jiguo Cao is the Canada research chair in data science and professor with the Department of Statistics and Actuarial Science, Simon Fraser University. Over a short period of time, he has made inroads across an impressively diverse range of sub-disciplines of statistics and data science, he builds on the synergies between functional data analysis (FDA), differential equations and Big Data to unveil novel statistical methods that greatly enhance users' understanding of their data. His high-impact, statistical frameworks are applied to real-world problems across various disciplines, including neuroscience, pharmacology, genetics, ecology, environment, and engineering. His research is highly visible, with more than 80 refereed publications. He was awarded the prestigious CRM-SSC award, in 2021 jointly from the Statistical Society of Canada (SSC) and Centre de recherches mathématiques (CRM) to recognize his research excellence and accomplishments. He is currently associate editor for 4 prestigious journals: *Biometrics*, *Canadian Journal of Statistics*, *Journal of Agricultural, Biological, and Environmental Statistics*, and *Statistics and Probability Letters*.