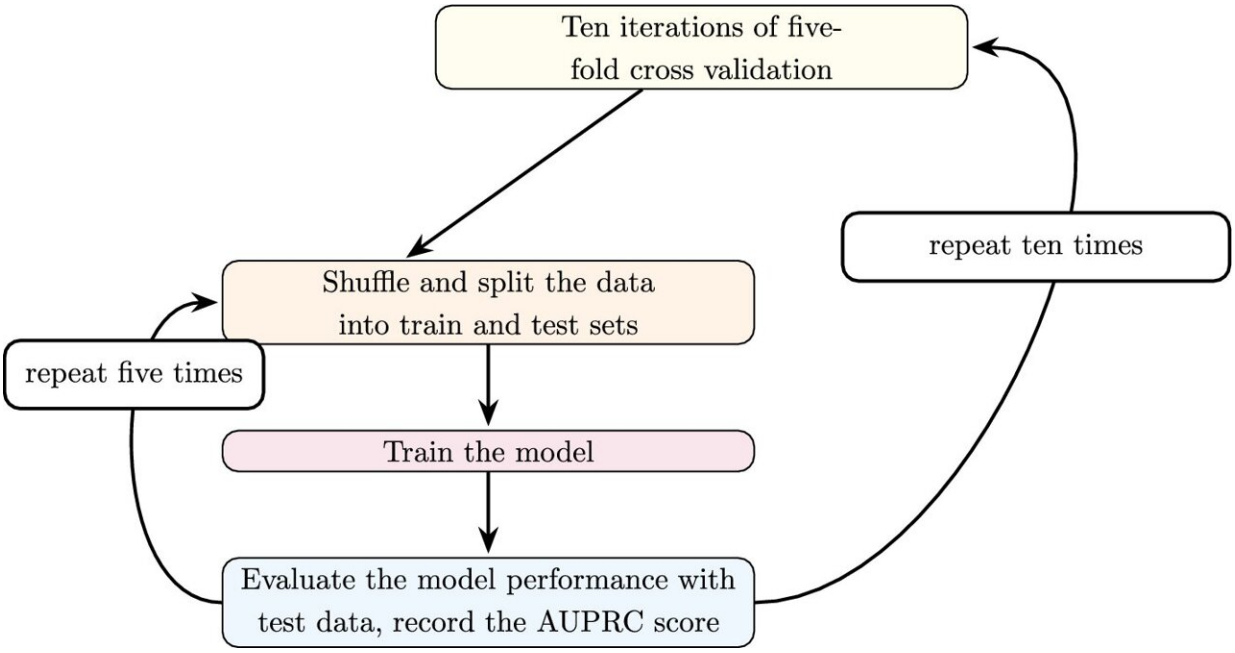
New AI technique significantly boosts Medicare fraud detection

January 31 2024



Five-fold cross validation. Credit: *Journal of Big Data* (2024). DOI: 10.1186/s40537-023-00869-3

Medicare is sporadically compromised by fraudulent insurance claims. These illicit activities often go undetected, allowing full-time criminals and unscrupulous health providers to exploit weaknesses in the system. Last year, the estimated annual fraud topped $100 billion, according to the National Health Care Anti-Fraud Association, but it is likely much higher.

Traditionally, to detect Medicare fraud, a limited number of auditors, or investigators, are responsible for manually inspecting thousands of claims, but only have enough time to look for very specific patterns indicating suspicious behaviors. Moreover, there are not enough investigators to keep up with the various Medicare fraud schemes.

Utilizing big data, such as from patient records and provider payments, often is considered the best way to produce effective machine learning models to detect fraud. However, in the domain of Medicare insurance fraud detection, handling imbalanced big data and high dimensionality—data in which the number of features is staggeringly high so that calculations become extremely difficult—remains a significant challenge.

New research from the College of Engineering and Computer Science at Florida Atlantic University addresses this challenge by pinpointing fraudulent activity in the "vast sea" of big Medicare data. Since identification of fraud is the first step in stopping it, this novel technique could conserve substantial resources for the Medicare system.

The study is published in the *Journal of Big Data*.

For the study, researchers systematically tested two imbalanced big Medicare datasets, Part B and Part D. Part B involves Medicare's coverage of medical services like doctor's visits, outpatient care, and other medical services not covered under hospitalization. Part D, on the other hand, relates to Medicare's prescription drug benefit and covers medication costs. These datasets were labeled with the List of Excluded Individuals and Entities (LEIE). The LEIE is provided by the United States Office of the Inspector General.

Researchers delved deep into the influence of Random Undersampling (RUS), a straightforward yet potent data sampling technique, and their

novel ensemble supervised feature selection technique. RUS works by randomly removing samples from the majority class until a specific balance between the minority and majority classes is met.

The experimental design investigated various scenarios, ranging from using each technique in isolation to employing them in combination. Following analyses of the individual scenarios, researchers again selected the techniques that yielded the best results and performed an analysis of results between all scenarios.

Results of the study demonstrate that intelligent data reduction techniques improve the classification of high imbalanced big Medicare data. The synergistic application of both techniques—RUS and supervised feature selection—outperformed models that utilize all available features and data. Findings showed that either combination of using the feature selection technique followed by RUS, or using RUS followed by the feature selection technique, yielded the best performance.

Consequently, in the classification of either dataset, researchers discovered that a technique with the largest amount of data reduction also yields the best performance, which is the technique of performing feature selection, then applying RUS. Reduction in the number of features leads to more explainable models and performance is significantly better than using all features.

"The performance of a classifier or algorithm can be swayed by multiple effects," said Taghi Khoshgoftaar, Ph.D., senior author and Motorola Professor, FAU Department of Electrical Engineering and Computer Science. "Two factors that can make data more difficult to classify are dimensionality and class imbalance. Class imbalance in labeled data happens when the overwhelming majority of instances in the dataset have one particular label. This imbalance presents obstacles, as it is

possible for a classifier optimized for a metric such as accuracy, which will mislabel fraudulent activities as non-fraudulent to boost overall scores in terms of the metric."

For feature selection, researchers incorporated a supervised feature selection method based on feature ranking lists. Subsequently, through the implementation of an innovative approach, these lists were combined to yield a conclusive feature ranking. To furnish a benchmark, models also were built utilizing all features of the datasets. Upon the derivation of this consolidated ranking, features were selected based on their position in the list.

"Our systematic approach provided a greater comprehension regarding the interplay between feature selection and model robustness within the context of multiple learning algorithms," said John T. Hancock, first author and a Ph.D. student in FAU's Department of Electrical Engineering and Computer Science. "It is easier to reason about how a model performs classifications when it is built with fewer features."

For both Medicare Part B and Part D datasets, researchers conducted experiments in five scenarios that exhausted the possible ways to utilize, or omit, the RUS and feature selection data reduction techniques. For both datasets, researchers found that data reduction techniques also improve classification results.

"Given the enormous financial implications of Medicare fraud, findings from this important study not only offer computational advantages but also significantly enhance the effectiveness of fraud detection systems," said Stella Batalama, Ph.D., dean, FAU College of Engineering and Computer Science. "These methods, if properly applied to detect and stop Medicare insurance fraud, could substantially elevate the standard of health care service by reducing costs related to fraud."

Study co-authors are Huanjing Wang, Ph.D., a professor of computer science, Western Kentucky University; and Qianxin Liang, a Ph.D. student in FAU's Department of Electrical Engineering and Computer Science.

More information: John T. Hancock et al, Data reduction techniques for highly imbalanced medicare Big Data, *Journal of Big Data* (2024). DOI: 10.1186/s40537-023-00869-3

Provided by Florida Atlantic University