Show Me Your Claims and I'll Tell You Your Offenses: Machine Learning-Based Decision Support for Fraud Detection on Medical Claim Data

Tizian Matschak
Göttingen University
Göttingen, Germany
tizian.matschak@uni-goettingen.de

Christoph Prinz
Göttingen University
Göttingen, Germany
christoph.prinz@uni-goettingen.de

Florian Rampold
Göttingen University
Göttingen, Germany
florian.rampold@uni-goettingen.de

Simon Trang
Göttingen University
Göttingen, Germany
strang@uni-goettingen.de

Abstract

Health insurance claim fraud is a serious issue for the healthcare industry as it drives up costs and inefficiency. Therefore, claim fraud must be effectively detected to provide economical and high-quality healthcare. In practice, however, fraud detection is mainly performed by domain experts resulting in significant cost and resource consumption. This paper presents a novel Convolutional Neural Network-based fraud detection approach that was developed, implemented, and evaluated on Medicare Part B records. The model aids manual fraud detection by classifying potential types of fraud, which can then be specifically analyzed. Our model is the first of its kind for Medicare data, yields an AUC of 0.7 for selected fraud types and provides an applicable method for medical claim fraud detection.

1. Introduction

Healthcare has become a major expenditure for social and financial systems in recent years, while annually increasing global spending in health is expected to reach $18.28 trillion by 2040 [1, 2]. Due to its complexity, analogous processes, and value of monetary transactions, healthcare emerged as an attractive fraud target [3, 4, 5]. Consequently, the healthcare domain faces an increasing number of fraud incidents every year [6]. In this context, fraud is defined as the misuse of a system by obtaining financial advantage or causing loss by implicit or explicit deception [7, 8]. The Federal Bureau of Investigation (FBI) estimated that 3–10% of all billings are fraudulent [9]. In parallel, limited healthcare resources are challenged by an increasing population, a rising number of elderly people, and a generally expanding health insurance coverage [1, 10].

Faced with these challenges, any fraudulent use of the healthcare system manifests as a huge bulk driving up costs for insurers, premiums for policyholders, expenses for providers, and consequently weaken the backbone of the healthcare system [4]. Thus, it is of public interest to effectively and efficiently detect fraudulent claims to provide economical and quality healthcare.

Focusing on this goal, the American Center for Medicare and Medicaid Service (CMS) [11] publicly releases datasets containing information about the types of Medicare services, requested charges, and payments issued by providers across the country [12]. By doing so, CMS provides a valuable contribution shedding light on *Medicare fraud, waste, and abuse* [12] and empowers the development of innovative Machine Learning (ML)-based fraud detection approaches that require significant amounts of data for effective training.

However, in the healthcare domain, detecting fraud is mainly done by domain experts who can only review a subset of total claims and only detect a few suspicious claims [5]. Consequently, healthcare fraud detection is time- and resource-consuming in practice since it depends on the knowledge and decision-making of these experts [5]. This raises the need for more efficient fraud detection approaches.

A potential solution to this problem could be a ML-based fraud detection model that pre-classifies claims and aids experts' decision-making. While the latest studies agree that human involvement is still required when it comes to the final classification of fraudulent claims, the implementation of automated fraud detection has the potential to reduce the number of observations to be reviewed [13, 14]. Moreover, the output of a fraud detection model can provide additional information on a potentially fraudulent claim supporting the analysis an expert performs on each claim.

Recent research mainly focused on the detection of fraud in general. The objective was to classify whether a claim is potentially fraudulent and requires further investigation. Therefore, several unsupervised

URI: https://hdl.handle.net/10125/79790
978-0-9981331-5-7

and supervised ML models have been implemented and evaluated. With regard to this, the latest research suggests that neural networks provide superior performance to this specific problem domain. While Johnson and Khoshgoftaar [15], Bauder et al. [1], and Bauder and Khoshgoftaar [16] analyzed Dense Neural Networks, research on other Neural Network architectures, especially Convolutional Neural Networks (CNN), for fraud detection within the Medicare claim domain is limited. However, CNNs are already successfully applied in other research domains like intrusion detection (e.g., Wu et al. [17]) and financial fraud detection (e.g., Fu et al. [18] and Zhang et al. [19]). Here, their potential of especially dealing with high dimensional data and efficiently detecting patterns in this data has been indicated. Since healthcare incorporates many stakeholders, the resulting data are also diverse and extensive and challenge current fraud detection approaches.

Against this background, our paper aims to design, implement, and evaluate an innovative CNN-based fraud detection approach that supports claim auditing experts to analyze medical claim data for fraud. Consequently, our approach allows them to perform efficient prescription claim fraud detection and contributes to improved economic and high-quality healthcare. To do so, we use publicly available Medicare Part B Data from 2018 in combination with the List of Excluded Individuals/Entities published by the Office of the Inspector General to create labeled training data.

Our research contributes to the discussion of innovative fraud detection approaches in the healthcare domain in at least three ways. First, we provide a methodology to enrich and transform the data to suit CNN-specific requirements and confirm and extend previous research by developing a novel CNN-based model for fraud detection on medical claim data. By showcasing this instantiation, we provide a blueprint for further studies that follow this path. Second, for the first time, a ML model to classify different types of healthcare fraud is evaluated on Medicare data and is able to archive reliable results with an AUC of 0.7 for selected fraud types. Third, we provide evidence about the potentials of CNN-based models in this context. These first insights are the foundation for comparing different ML techniques and their specific performance in the medical prescription domain.

The paper is structured as follows. We first provide an overview of previous research works dealing with Medicare fraud detection. Subsequently, we present the data and methods used and report our results. Finally, we discuss our results in light of practical and literature-based implications.

2. Literature Review

Initialized by CMS releasing the first "Medicare Provider Utilization and Payment Data: Physician and Other Supplier" (PUF) in 2014, a number of research works relating to Medicare anomaly and fraud detection have been published. We have selected this open-source data set to develop and evaluate a novel CNN-based fraud detection approach since it is well recognized in the research community and previous work has left an opportunity for improvement.

In [20], Ko et al. used the 2012 CMS data to analyze the dependency between service utilization and paid reimbursements. The authors found that the number of patient visits is strongly correlated with the Medicare payments and that utilization variability of services performed per visit offers a possible 9% savings within the field of Urology. Feldman and Chawla examined the impact of physicians' medical school education on their practicing decisions in [12] using the 2012 Medicare Part B data. They enriched the Medicare data with provider-level medical school data and significant school procedures that were further used to evaluate school similarities and present a geographical analysis of procedure charges and payment distributions. Branting et al. [21] performed fraud risk estimation based on a graph analytics approach. The authors deployed two types of algorithms: one for behavioral similarity calculation and one for estimation of fraud risk propagation through geospatial colocation. PUF data of 2012, 2013, and 2014 was labeled using the *List of Excluded Individuals/Entities* and the *National Plan and Provider Enumeration System* (NPPES) [22]. However, the obtained AUC score of 0.96 by tenfold cross-validation is recognized as misleading by related research [15].

Following research activities were mainly focused on fraud and anomaly detection driven by supervised and unsupervised approaches. Bauder and Khoshgoftaar [23] designed a probabilistic programming approach for anomaly detection on a small subset of the 2012–2014 Medicare Part B data limited to dermatology and optometry claims from Florida office clinics. The authors validated the Bayesian inference approach using claims data from a known fraudulent provider. Another study by Bauder and Khoshgoftaar [24] aims to identify potential fraud by analyzing actual payment amount deviations from the expected payment amounts. Five different regression models were used on Medicare Part B data of 2012 and 2013 from Florida. The authors reported superior performance for their multivariate adaptive regression splines model. In [3], Bauder et al. applied a Naïve Bayes classifier to predict provider specialty

types building on the idea that providers practicing outside their specialty may be fraudulent and are worthy of further investigation. Based on a Florida-only subset of 2013 Medicare Part B claims data the classification of 7 of 82 provider types scored very high (F1-score > 0.90), and 18 types resulted in an F1-score between 0.5 and 0.90. A similar approach was developed by Herland et al. [25] using 2014 Medicare Part B data enriched by real-world fraud labels derived from the LEIE data set. The implemented Naïve Bayes model obtained an overall accuracy 0.67. Rather than focusing on a specific ML model, Bauder and Khoshgoftaar [26] use 2012–2015 Medicare Part B data sets mapped with fraud labels from the LEIE dataset to compare multiple learners. The authors report a superior performance of C4.5 decision tree and logistic regression learners against the support vector machine with average AUC scores of 0.883 and 0.882, respectively. Moreover, they showed that a random undersampling (RUS) with a ratio of 80:20 yields better results than alternative sampling strategies. In [15], 2012-2016 Medicare Part B combined with fraud labels from the LEIE dataset is used to implement and evaluate six deep learning methods for addressing class imbalance. They show that Neural Networks combined with a hybrid random under-oversampling (RUS-ROS) outperform baseline models with an average AUC score of 0.8509.

The related works listed here provide evidence that the LEIE dataset can be reliably used for deriving a ground truth of fraud labels. Furthermore, previous research mainly focuses on detecting fraud and anomalies in general but does not consider different fraud types. Thereby, the possible potentials of innovative CNN approaches have not yet been taken into account. Consequently, we aim at closing this research gap and extend these related works by providing a CNN-based fraud detection model that classifies different types of fraud to support the manual auditing of health insurance claims.

3. Data and Methodology

3.1 Datasets

Our study is grounded on two open-source datasets. First, we use a sample of the latest Medicare Part B PUF data of 2018 to draw our predictive variables. The PUF data summarizes each provider's annual charges of drugs, services, and procedures provided to Medicare's Fee-For-Service beneficiaries. Accordingly, records within the dataset contain several provider- (e.g., National Provider Identifier (NPI), first and last name, gender, credentials, and address) and claim-related attributes. Claim-related attributes provide summarized information about a provider's Medicare-related operations within a year. This includes the claimed subject, the average charge amount submitted to Medicare, the average amount paid by Medicare, and the place of service treatment. Thereby, claim subjects are indicated by a Healthcare Common Procedure Coding System (HCPCS) [27] code. Previous research has provided evidence that fraudulent providers can be distinguished from others based on their distinctive billing patterns. Thus, we believe that PUF data will be useful within our research context as well.

The second dataset is the publicly available "List of Excluded Individuals/Entities" (LEIE) published by the Office of the Inspector General (OIG) [28]. Since we investigate supervised ML learners, predictive variables have to be enriched by labels of the respective target class (fraudulent or not). These labels can be obtained from LEIE as it includes healthcare providers excluded from Medicare eligibility. The list is updated monthly but unfortunately does not include all fraudulent providers since some fraudulent actions are relinquished without any public acknowledgment (e.g., overcharging) [21]. The recognized fraud is documented as one record per provider, including provider's metadata (e.g., name, address) and healthcare fraud-related attributes (e.g., exclusion type, provider's reinstation date). In addition, the LEIE dataset is only available in its latest version (April 2021), with updated logs of the last twelve months. This is a good problem, since the relevant exclusion times are five to ten years (see Table 1). Thus, all relevant exclusions should be included.

Despite the limitations of these datasets, we purposefully decided to ground our research on open-access data to provide a testable and adaptable fraud detection approach that can be built on by future research.

3.2 Data Labeling

Based on the described datasets and with reference to [15], the general assumption behind our labeling strategy is that a fraudulent provider's claim activities, before their date of exclusion from Medicare, are decisive for their exclusion soon after. Consequently, we map the listed providers in the LEIE to the PUF data.

To do so, we followed a four-step process. First, we follow the work by Bauder and Khoshgoftaar [24] and filter the LEIE on exclusion types most indicative of healthcare fraud. The resulting subset of exclusion types is listed in Table 1.

Table 1: Healthcare fraud-related exclusion types [28]

Social Security Act	Description	Minimum exclusion period	Defined Fraud Class
1128(a)(1)	Conviction of program-related crimes	5 years	**1**
1128(a)(2)	Conviction relating to patient abuse or neglect	5 years	**2**
1128(a)(3)	Felony conviction relating to health care fraud	5 years	**3**
1128(b)(4)	License revocation, suspension, or surrender	State-dependent	**4**
1128(b)(7)	Fraud, kickbacks, and other prohibited activities	None	**5**
1128(c)(3)(G)(i)	Conviction of second mandatory exclusion offenses	10 years	**6**
1128(c)(3)(g)(ii)	Conviction of third mandatory exclusion offenses	Permanent exclusion	**7**

Second, excluded providers are filtered based on their date of exclusion. Following our previously mentioned labeling assumption, the provider's claim records data submitted before the exclusion date have to be labeled as fraudulent. In practice, we preserve only providers excluded later than the end of 2018 since we use cumulated PUF data of 2018 that does not provide a dedicated attribute for the date of claim submission. By doing so, we differ in the definition of exclusion (fraud) applied by [25] and [15] since they round the exclusion-related dates. This does not seem to be practical in our case since despite them, dealing with PUF data of multiple years, we only use the latest PUF data of 2018. Second, providers are matched by their NPI provided in each dataset as a common attribute. Unfortunately, the LEIE data suffers from a significant amount of missing NPI values leading to the need for another matching method. Third, we followed the approach by Branting et al. [21] and applied name matching using the first name and last name attributes of the remaining LEIE and PUF data. Moreover, we extended name matching by zip code mapping to further mitigate false-positive labeling. As a result, each record in our PUF data is now assigned whether and what type of fraud has been detected.

3.3 Data Preprocessing

PUF data provided by CMS is published as a text file in comma-separated format. In order to transform this data into a useful representation for our ML model, the data has to be processed with a focus on several topics.

When the data is initially loaded, it contains 26 attributes per record, and records of claims for drugs and services/procedures are mixed. Thus, with regard to [15], we filter the data for service claims at first since they differ in characteristics. So, dealing with drug claims is left for future research. Then, we perform feature reduction to reduce features and thus mitigate negative effects related to the *curse of dimensionality* [29]. As mentioned previously, PUF data contains provider- and claim-related attributes. In the process of feature reduction, we removed most provider-related features (e.g., name and address) to boost generalization. In addition, we perform feature engineering and provide several computed features based on claim-related data. This aims to improve model fitting because more informative features improve the effectiveness of the model to identify relationships and correlations in the training data [30]. In detail, we computed the numeric differences between *Average Submitted Charge Amount and Average Medicare Allowed Amount and Average Medicare Payment Amount each.* As a result, the following features were used as input for further preprocessing (see Table 2).

Table 2: Description (see [22]) of used features for provider grouping

Feature	Description	Type
National Provider Identifier	The provider NPI is the numeric identifier registered in NPPES	Categorical
Provider Type	Derived from the provider specialty code reported on the claim	Categorical
State Code of the Provider	The state where the provider is located	Categorical
Country Code of the Provider	The country where the provider is located	Categorical
Number of Services	Number of services provided	Numeric
Number of Medicare Beneficiaries	Number of distinct Medicare beneficiaries receiving the service	Numeric
Number of Distinct Medicare Beneficiary/Per Day Services	Number of distinct Medicare beneficiary/per day services	Numeric
HCPCS Code	HCPCS code used to identify the specific medical service furnished by the provider	Categorical

Average Submitted Charge Amount	Average of the charges that the provider submitted for the service	Numeric
Average Medicare Allowed Amount	Average of the Medicare allowed amount for the service	Numeric
Average Medicare Payment Amount	Average amount that Medicare paid	Numeric
Average Medicare Standardized Amount	Standard deviation of the Medicare allowed amounts	Numeric
Payment_diff	Difference between *Average Submitted Charge Amount* and *Average Medicare Payment Amount*	Numeric
Allowed_diff	Difference between *Average Submitted Charge Amount* and *Average Medicare Allowed Amount*	Numeric

We apply feature encoding to transform existing categorical features into a suitable format. In general, feature encoding describes the process of transforming the representation of feature values. This procedure is required since some ML algorithms cannot handle categorical data well [31]. Thus, this data has to be transformed to a more suitable format. To do so, we use the one-hot encoding method that is widely utilized in related research (e.g., Bauder et al. [1]). One-hot encoding describes a technique used to represent each categorical feature by a sparse vector representing the categories and their binary value, indicating whether the category is the value of the original feature or not [1]. A disadvantage coming along with one-hot encoding is that it drastically increases dimensionality and tends to fail to capture relevant relationships between similar providers [15].

Since feature values within our dataset differ in their scale as they increase without bound, they affect models that are smooth functions of the input [32]. Consequently, data normalization is a critical step that speeds up training and influences model performance by limiting the scale [33]. In this case, we decided to use Min-Max Scaling [34] because it is well-known and provided good results in multiple research works (e.g., Johnson and Khoshgoftaar [15]).

As mentioned before, PUF data is organized as records per HCPCS code and provider. In order to follow our primary goal of detecting fraud through claims activity per provider, we have to rearrange our data. With reference to the work by Johnson and Khoshgoftaar [15], we re-group the records by NPI so that the resulting groups contain a unique provider's annual claims data, with one row for every HCPCS code. In contrast to them, we not only implement provider-related and summary attributes but follow the one-hot encoding logic and create a sparse vector representing distinct service types (HCPCS codes) and their respective number of services. Thus, information loss should be minimized, and valuable patterns should be preserved. This approach is applied to state and country codes respectively. Lastly, additional summary attributes for the remaining numeric attributes are added. By doing so, each group is aggregated, converting the multiple rows into a single record. As a result, this vector contains 5,924 features.

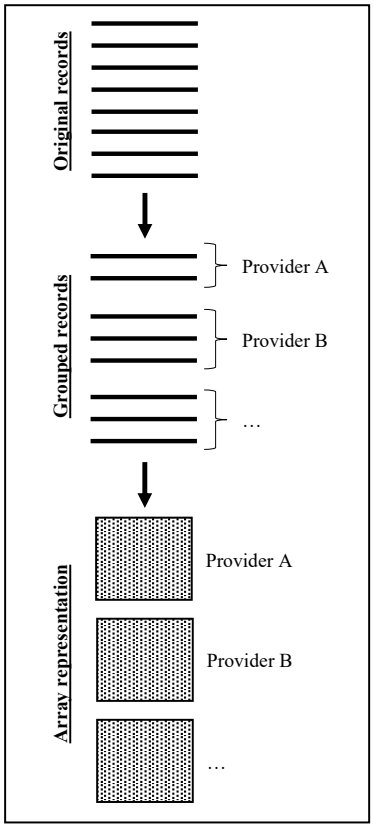

Figure 1: Data Transformation Steps

On the one hand, this number of features pushes baseline ML approaches to their limits. On the other hand, it provides potentials for applying innovative ML techniques to this research problem. As described in section 2, there is no prior research dealing with either the application of CNNs in the context of Medicare fraud detection or using CNNs for corresponding decision support. Therefore, the data is transformed in an image-like (multi-dimensional array) format, suiting requirements of CNN input, subsequently. We use a 78x78 pixel representation providing a potential capacity for 6,084 values. Since the vector of transformed PUF data does not occupy the space completely, the remaining pixels are filled with zeros. An exemplary segment of the image representation is presented in Figure 1.

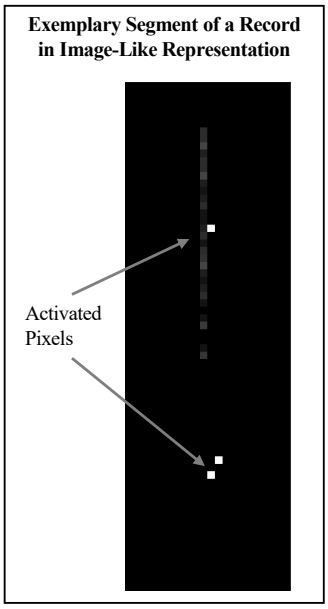

Figure 2: Image-like claim data representation

A relevant problem of most fraud detection tasks is class imbalance. ML algorithms may face degradation of classification performance caused by the class imbalance, minority class decomposition into sub-parts, and overlapping classes [35]. Considering a classification task, as most anomaly detection is, and a majority class (normal ones) partition of 99.5% (minority partition 0.5% respectively), an algorithm can trivially gain 99.5% accuracy by simply learning the rule f(x)=normal (always classify as normal; [36]). This makes the learning of a classifier quite challenging [37]. As a solution to this problem, relevant literature recommends the application of class balancing techniques. Here, two main methods are adopted: under-sampling and over-sampling. Since it is still unclear which sampling method performs best and what sampling ratio should be used, the choice is domain-specific [38].

In this study, we have to deal with a strong class imbalance since only 109 of 1,060,834 records are labeled as fraudulent (≈0.01%). To do so, we apply hybrid random over- and undersampling. First, minority classes are oversampled until they reach a 5% proportion of the original dataset each. Then the majority class is undersampled until the class balance is obtained. More advanced and deeper analysis of the impact and performance of different sampling strategies on this multi-class classification problem is declared as future research.

3.4 The CNN Model

In general, detection performance is closely related to selected features and classifiers, but traditional classification algorithms cannot perform well on massive data [17]. As mentioned before, data preprocessing led to a total number of 5,924 features, raising the need for a specialized classification algorithm.

A CNN is a classical deep learning algorithm and has been applied in many fields, particularly in visual recognition. It can be composed of several layers, including convolution, pooling, flatten, and fully connected layers (see Figure Figure 3). Compared to other deep learning algorithms, the greatest advantage of CNNs is their implementation of convolutional kernels, which reduce the number of parameters and calculation amount of training [17]. Moreover, CNNs are valued for their ability to reduce over-fitting and reveal hidden fraud patterns [17].

Based on that, we assess a CNN as a suitable ML model for classifying our high dimensional data. The CNN model comprises of the following structure:

- Three pairs of convolutional and max-pooling layers
- One flatten layer
- One dense layer
- One dense output layer

Stratified random sampling without replacement is used to create the 20% test set. Consequently, the model is trained on 80% of the data. To further increase model performance, relevant hyperparameters are tuned.

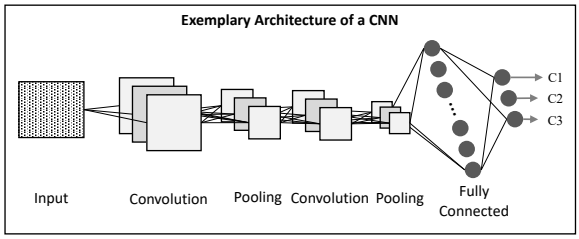

Figure 3: Exemplary CNN Architecture

4. Analysis and Evaluation

We analyze the performance of our fraud detection approach by applying the developed model to 1,849 preprocessed and unseen records. Therefore, the Area Under the Receiver Operator Curve (AUC), Precision, and Recall are used as evaluation metrics since they are well known and applied in multiple related publications (e.g., Kalid et al. [39] and Bauder et al. [1]). In addition, this enables a comparative assessment against previously developed and upcoming approaches.

Originally, the mentioned metrics were developed for binary classification and are defined for the positive class [40], which is the minority class (fraud) in this case. In particular, Recall is defined as the fraction of positive cases classified as positive, while Precision is defined as the fraction of positive classified cases that are truly positive [41]. Based on this, AUC considers Recall and Precision, thus being more balanced than Accuracy [16]. It is defined as the area under the Receiver Operator Curve (ROC), representing the capability of a classifier in differentiating classes [39]. The ROC is created by plotting Recall against Precision. As a result, the closer the AUC value is to 1 (perfect classification), the better the classification performs [39]. In addition, the closer the AUC value is to 0.5, the more similar is the classification to random guessing. Values below 0.5 indicate that the classifier performance is worse than random guessing.

Since this study faces a multi-class classification problem, we need to adapt the evaluation method to this. Thus, we implemented a one-vs.-all logic (one class vs. all other classes) to transform our approach into a binary decision problem. By doing so, the model output, in the form of a sparse vector containing binary values of whether the specific class was predicted or not, is separated into two groups. The first group represents one specific class and the other group all other classes. These values are then used as input for AUC, Precision and Recall. Receptive results are shown in Figure 4.

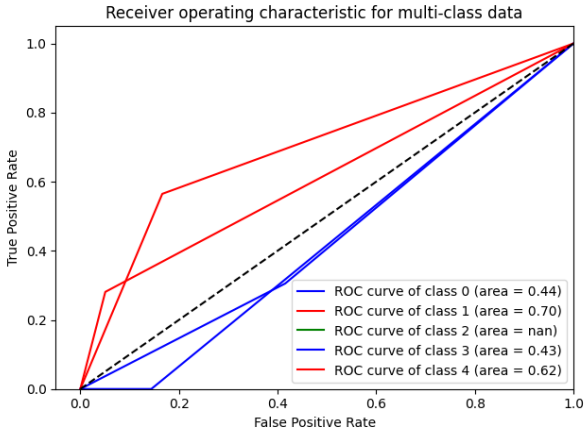

Figure 4: One-vs.-all AUC results per target class

It turned out that the model performance varies between target classes. While class 1 (exclusion type '1128a1') reaches an AUC value of 0.7, class 3 ('1128b4') is limited to 0.43. We assume that the difference in performance is correlated with the number of distinctive records for training. For class 1, there are more than six times as many records as for class 3 in our Medicare dataset. For class 4, there are more than five times as many records, respectively. Consequently, our approach should be validated on a broader dataset to provide each target class with a sufficient number of distinct training records.

In summary, the presented anomaly detection and classification approach provides evidence for the general practicability of our applied methodology including data labeling, data preprocessing and the CNN model. Moreover, it is the first study investigating the classification of different fraud types within Medicare data.

5. Discussion and Implications

This study proposes an innovative CNN-based classification model and a methodology for corresponding data combination and preprocessing to improve the detection of medical claim fraud. We designed a dedicated data transformation procedure based on previous research to enrich and transform publicly available PUF data into an image-like representation. The developed approach obtained comprehensible results for selected target classes (exclusion types). As a result, our work offers several implications and contributions to literature and the practice audience.

5.1 Implications to Literature

First, we transfer methodologies from related application domains and complement the status quo of fraud detection methods in the medical claim domain by an approach based on a CNN model and corresponding data preprocessing. Thereby, we designed a data labeling and preprocessing procedure to combine publicly available PUF data with LEIE data and subsequently use it as input for supervised ML models, in particular CNNs. Moreover, we challenged known problems such as high dimensionality data and extended the existing knowledge base by implementing and evaluating a CNN for fraud detection on Medicare claim data. This model addresses the detected research gap by aiming to classify different types of fraud rather than performing a simple binary classification (fraud or no fraud). Lastly, our novel approach provides a blueprint on how to challenge high dimensional fraud detection problems in the context of medical claim data and opens several leverage points for future research.

5.2 Implications for Practice

Regarding practice utility, we provide stakeholders of the healthcare domain, particularly health insurers, with an applicable method to analyze whether and what type of potential fraud a provider has committed. This is especially relevant to improving manual claim auditing efficiency since auditors review lots of claims to detect few fraudulent cases. Our approach can be used to build claim-bins, each coupled with a specific action that has to be performed for this particular fraud type. Thus, pre-filtering can enable auditors to act more efficiently by indicating which further claim assessment can be based on. Furthermore, investigators and managers can use the classification data about fraud types and assess the cost of further information gathering and investigation to realize a suspected case. Following this, limited resources can be used more efficiently and effectively, and the overall throughput can be increased. As a result, each detected fraud case contributes to economic and high-quality healthcare. Finally, we assume that the further spread of e-prescribing will increase the quantity and quality of available data and result in a future increase in performance potential.

6. Conclusion

In this paper, we have investigated the potential of CNNs in the context of medical claim fraud detection. Medicare Part B PUF data and LEIE were combined to create labeled data with a strong class imbalance. Based on this, we designed dedicated preprocessing that transforms this data into a suitable input format for CNN-based fraud detection and classification model. The model is the first of its kind for classifying fraud types on Medicare claim data and proved practice utility by obtaining an AUC value of 0.7 for selected fraud types. Therefore, we were able to extend the literature by a novel fraud detection approach and provide confirmation and an extension of previously developed data preprocessing and fraud detection strategies. Based on that, it can be concluded for practice that our model has the potential to improve medical prescription fraud detection and serve economic and high-quality healthcare.

Furthermore, this study has several limitations and unveils leverage points for future research. Our research is limited only to the data of one American health insurer. Thus, national laws and regulations could affect our results. This should be considered when interpreting or transferring our results to different healthcare programs. However, we believe that most of our results may be easily transferrable to similar application domains with no or minor adaptations. Thus, our approach should be tested in other application environments and other healthcare programs. In this paper, the composed set of data features proved useful, though it is not excluded that additional significant fraud indicators were overlooked. Moreover, we only considered PUF data of 2018. In order to obtain more training data related to each fraud type and thus improve classification performance, PUF data of additional years should be integrated. Lastly, the described CNN model provides several configuration options so that the possibility remains that there is a better configuration. Therefore, based on our initial work, future research should investigate model configuration and class balancing in more detail.

7. References

[1] Bauder, R., R. da Rosa, and T. Khoshgoftaar, "Identifying Medicare Provider Fraud with Unsupervised Machine Learning", 2018 IEEE International Conference on Information Reuse and Integration (IRI), IEEE (2018), 285–292.

[2] Dieleman, J.L., T. Templin, N. Sadat, et al., "National spending on health by source for 184 countries between 2013 and 2040", The Lancet 387(10037), 2016, pp. 2521–2535.

[3] Bauder, R.A., T.M. Khoshgoftaar, A. Richter, and M. Herland, "Predicting Medical Provider Specialties to Detect Anomalous Insurance Claims", 2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI), IEEE (2016), 784–790.

[4] Dora, P., and D.G.H. Sekharan, "Healthcare Insurance Fraud Detection Leveraging Big Data Analytics", IJSR 4(4), 2015, pp. 2073–2076.

[5] Waghade, S.S., and A.M. Karandikar, "A Comprehensive Study of Healthcare Fraud Detection based on Machine Learning", 13(6), 2018, pp. 4.

[6] Rawte, V., and G. Anuradha, "Fraud detection in health insurance using data mining techniques", 2015 International Conference on Communication, Information & Computing Technology (ICCICT), IEEE (2015), 1–5.

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

[8] Levi, M., and J. Burrows, "Measuring the Impact of Fraud in the UK: A Conceptual and Empirical Journey", British Journal of Criminology 48(3), 2007, pp. 293–318.

[9] Morris, L., "Combating fraud in health care: an essential component of any cost containment strategy.", Health Affairs 28(5), 2009, pp. 1351–1356.

[10] Guo, X., Y. Sun, N. Wang, Z. Peng, and Z. Yan, "The dark side of elderly acceptance of preventive mobile health services in China", Electronic Markets 23(1), 2013, pp. 49–61.

[11] "Center for Medicare & Medicaid Services", Center for Medicare & Medicaid Services, 2021. https://www.cms.gov

[12] Feldman, K., and N.V. Chawla, "Does Medical School Training Relate to Practice? Evidence from Big Data", Big Data 3(2), 2015, pp. 103–113.

[13] Bauder, R.A., and T.M. Khoshgoftaar, "Multivariate outlier detection in medicare claims payments applying probabilistic programming methods", Health Services and Outcomes Research Methodology 17(3–4), 2017, pp. 256–289.

[14] Power, D.J., and M.L. Power, "Sharing and Analyzing Data to Reduce Insurance Fraud", Proceedings of the Tenth Midwest Association for Information Systems Conference, (2015), 1–6.

[15] Johnson, J.M., and T.M. Khoshgoftaar, "Medicare fraud detection using neural networks", Journal of Big Data 6(1), 2019, pp. 63.

[16] Bauder, R.A., and T.M. Khoshgoftaar, "Medicare Fraud Detection Using Machine Learning Methods", 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE (2017), 858–865.

[17] Wu, K., Z. Chen, and W. Li, "A Novel Intrusion Detection Model for a Massive Network Using Convolutional Neural Networks", IEEE Access 6, 2018, pp. 50850–50859.

[18] Fu, H., D. Cheng, Y. Tu, and L. Zhang, "Credit card fraud detection using convolutional neural networks.", Springer (2016), 483–490.

[19] Zhang, Z., X. Zhou, X. Zhang, L. Wang, and P. Wang, "A Model Based on Convolutional Neural Network for Online Transaction Fraud Detection", Security and Communication Networks 2018, 2018, pp. 1–9.

[20] Ko, J.S., H. Chalfin, B.J. Trock, et al., "Variability in Medicare Utilization and Payment Among Urologists", Urology 85(5), 2015, pp. 1045–1051.

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

[22] "National Plan & Provider Enumeration System.", NPPES NPI Registry. https://npiregistry.cms.hhs.gov/registry/

[23] Bauder, R.A., and T.M. Khoshgoftaar, "A Probabilistic Programming Approach for Outlier Detection in Healthcare Claims", (2016), 347–354.

[24] Bauder, R.A., and T.M. Khoshgoftaar, "A Novel Method for Fraudulent Medicare Claims Detection from Expected Payment Deviations (Application Paper)", IEEE (2016), 11–19.

[25] Herland, M., R.A. Bauder, and T.M. Khoshgoftaar, "Medical Provider Specialty Predictions for the Detection of Anomalous Medicare Insurance Claims", 2017 IEEE International Conference on Information Reuse and Integration (IRI), IEEE (2017), 579–588.

[26] Bauder, R.A., and T.M. Khoshgoftaar, "The Detection of Medicare Fraud Using Machine Learning Methods with Excluded Provider Labels", (2018), 6.

[27] "HCPCS", HCPCS - General Information. https://www.cms.gov/Medicare/Coding/MedHCPCSGenInfo

[28] "Office of Inspector General", LEIE downloadable databases, 2021. https://oig.hhs.gov/exclusions/exclusions_list.asp.

[29] Bellman, R., and S. Dreyfus, "Functional Approximations and Dynamic Programming", Mathematical Tables and Other Aids to Computation 13(68), 1959, pp. 247–251.

[30] Murdoch, W.J., C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Interpretable machine learning: definitions, methods, and applications", arXiv preprint arXiv:1901.04592, 2019.

[31] Matschak, T., C. Prinz, C. Masuch, and S. Trang, "Healthcare in Fraudster's Crosshairs: Designing, Implementing and Evaluating a Machine Learning Approach for Anomaly Detection on Medical Prescription Claim Data", (2021), 1–14.

[32] Zheng, A., and A. Casari, Feature engineering for machine learning: principles and techniques for data scientists., O'Reilly Media, Inc., Sebastopol, USA, 2018.

[33] Jayalakshmi, T., and A. Santhakumaran, "Statistical Normalization and Back Propagationfor Classification", International Journal of Computer Theory and Engineering, 2011, pp. 89–93.

[34] scikit-learn, "scikit-learn", sklearn.preprocessing.MinMaxScaler, 2021. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

[35] Dal Pozzolo, A., O. Caelen, and G. Bontempi, "When is Undersampling Effective in Unbalanced Classification Tasks?", In A. Appice, P.P. Rodrigues, V. Santos Costa, C. Soares, J. Gama and A. Jorge, eds., Machine Learning and Knowledge Discovery in Databases. Springer International Publishing, Cham, 2015, 200–215.

[36] Murphy, K.P., Machine learning: a probabilistic perspective, MIT Press, Cambridge, MA, 2012.

[37] Chandola, V., A. Banerjee, and V. Kumar, "Anomaly detection: A survey", ACM Computing Surveys 41(3), 2009, pp. 1–58.

[38] Weiss, G.M., "Mining with rarity: a unifying framework", ACM SIGKDD Explorations Newsletter 6(1), 2004, pp. 7–19.

[39] Kalid, S.N., K.-H. Ng, G.-K. Tong, and K.-C. Khor, "A Multiple Classifiers System for Anomaly Detection in Credit Card Data With Unbalanced and Overlapped Classes", IEEE Access 8, 2020, pp. 28210–28221.

[40] Fayzrakhmanov, R., A. Kulikov, and P. Repp, "The Difference Between Precision-recall and ROC Curves for Evaluating the Performance of Credit Card Fraud Detection Models", Proceedings of International Conference on Applied Innovation in IT 6(1), 2018, pp. 17–22.

[41] Davis, J., and M. Goadrich, "The relationship between Precision-Recall and ROC curves", Proceedings of the 23rd international conference on Machine learning - ICML '06, ACM Press (2006), 233–240.