Semantic Embeddings for Medical Providers and Fraud Detection

Justin M. Johnson and Taghi M. Khoshgoftaar
College of Engineering and Computer Science
Florida Atlantic University
Boca Raton, Florida 33431
jjohn273@fau.edu, khoshgof@fau.edu

***Abstract*—A medical provider's specialty is a significant predictor for detecting fraudulent providers with machine learning algorithms. When the specialty variable is encoded using a one-hot representation, however, models are subjected to sparse and uninformative feature vectors. We explore three techniques for representing medical provider types with dense, semantic embeddings that capture specialty similarities. The first two methods (GloVe and Med-Word2Vec) use pre-trained word embeddings to convert provider specialty descriptions to short phrase embeddings. Next, we propose a method for constructing semantic provider type embeddings from the procedure-level activity within each specialty group. For each embedding technique, we use Principal Component Analysis to compare the performance of embedding sizes between 32–128. Each embedding technique is evaluated on a highly imbalanced Medicare fraud prediction task using Logistic Regression (LR), Random Forest (RF), Gradient Boosted Tree (GBT), and Multilayer Perceptron (MLP) learners. Experiments are repeated 30 times and confidence intervals show that all three semantic embeddings significantly outperform one-hot representations when using RF and GBT learners. Our contributions include a novel method for embedding medical specialties from procedure codes and a comparison of three semantic embedding techniques for Medicare fraud detection.**

***Keywords*—Semantic Embeddings, Clinical Concept Extraction, Class Imbalance, Big Data, Medicare, Fraud Detection, Deep Learning**

1. Introduction

The United States (US) Medicare program provides affordable health insurance to individuals 65 years and older, and other individuals with permanent disabilities [1]. There are currently more than 62.2 million U.S citizens enrolled in Medicare [2], and 2019 expenditures exceeded $796 billion [3]. The Federal Bureau of Investigation estimates that fraud accounts for 3–10% of all billings [4], i.e. $23 to $79 billion per year in the Medicare program. Examples of Medicare fraud include billing for appointments that the patients did not keep, billing for services more complex than those performed, or billing for services not provided. In an effort to reduce fraud, the Centers for Medicare and Medicaid Services (CMS) makes Medicare data sets publicly available for analysis [5].

This study uses the 2012–2016 Medicare Provider Utilization and Payment Data: Physician and Other Supplier Public Use File (Part B) data sets made available by CMS. Fraudulent and non-fraudulent class labels are mapped to the Part B data claims data using the List of Excluded Individuals and Entities (LEIE) [6]. The LEIE is maintained by the Office of Inspector General (OIG), and it lists providers that are prohibited from participating in Federal healthcare programs. The Part B data's provider type attribute describes a provider's specialty and contains 123 distinct values, e.g. Internal Medicine, Anesthesiology, and Cardiology. In all previous Medicare fraud works, the provider type has either been excluded from the feature set or represented as a one-hot vector. Related works show, however, that there is a significant amount of overlap between provider types, and equidistant one-hot vectors are unable to capture these provider type similarities. We address this by exploring three semantic embedding techniques for the provider type variable and comparing results to traditional one-hot vector results.

The concept of learning low-rank semantic representations for provider types is largely inspired by word embeddings. Unlike traditional bag-of-words models, word embeddings represent each word using a dense real-valued vector. Mikolov et al. [7] proposed two Word2Vec models for efficiently learning high-quality representations of words. Pennington et al. [8] later proposed Global Vectors for Word Representation (GloVe), a method for generating word embeddings from a global word-word co-occurrence matrix. Both Word2Vec and GloVe embeddings are word-level representations, i.e. they do not take into account the order of words. ELMo (Embeddings from Language Models) uses a bidirectional Long Short-Term Memory (LSTM) algorithm to encode context-aware word embeddings [9]. Similarly, BERT (Bidirectional Encoder Representations from Transformers) achieves contextualized word embeddings by pre-training masked language models with a multi-layer bidirectional Transformer encoder [10]. These advances have transformed natural language processing with state-of-the-art results in a wide range of tasks, e.g. question and answering [11]. In this paper, we use pre-trained word embeddings to capture provider type similarities and

978-1-7281-1054-7/20/$31.00 ©2020 IEEE
DOI 10.1109/IRI49571.2020.00039

improve Medicare fraud prediction performance and leave more advanced language model embeddings open to future work. Additionally, we propose a technique that encodes provider type variables using the procedure occurrences within each specialty group.

We compare sparse one-hot provider type vectors to dense real-valued representations, i.e. entity embeddings [12], using four machine learning algorithms. The first approach (GloVe) converts the provider type variable to its equivalent word embedding by replacing the tokens in the provider type textual description with their respective GloVe word embeddings. Most provider types are short phrases, e.g. 2–4 words, so we combine them to a single vector by taking their unweighted average [13]. In a similar manner, the second approach (Med-Word2Vec) converts provider type variables to word embeddings using publicly available word embeddings that have been pre-trained on clinical notes from PubMed and Pubmed Central Open Access (PMC OA) [14]. We then propose HcpcsVec, an embedding technique that constructs provider type embeddings from procedure code occurrences, as outlined in Section 3. The Healthcare Common Procedures Coding System (HCPCS) procedure codes describe provider billing behavior, and related works in Section 2 suggest significant procedure overlap between similar provider types. We expect these procedure embeddings to capture semantic relations between providers types. For all three techniques, we use Principal Component Analysis (PCA) to reduce dimensionality and compare results using embedding sizes between 32 and 128. Our results show that the proposed embedding techniques outperform one-hot vectors when using Logistic Regression (LR), Gradient Boosted Tree (GBT), and Random Forest (RF) learners. To the best of our knowledge, this is the first work to explore semantic provider type embedding techniques for Medicare fraud detection. Furthermore, we show that these new embedding techniques outperform previous works [15] based on the area under the Receiver Operating Characteristic curve (ROC AUC).

The remainder of this paper is structured as follows. In Section 2, we review previous work related to Medicare fraud prediction and embedding methods for medical concept extraction. Section 3 describes the data set, preprocessing, and experiment design. Section 4 presents results, and Section 5 concludes with areas for future work.

2. Related Work

2.1. Medicare Fraud Prediction

Several studies related to Medicare fraud prediction either do not use the provider type attribute, or they model each provider type independently. Bauder and Khoshgoftaar [16] estimate Medicare payments using five regression models. Actual payment amounts are compared to estimated payments, and the deviations are used to flag potentially fraudulent providers. Ko et al. [17] use a linear regression model to analyze the variability of service utilization and payments in 2012 CMS Part B data. Both studies use only a subset of medical specialties and model each specialty separately. Branting et al. [18] use graph-based features and a decision tree learner to predict Medicare fraud using the 2012–2014 CMS Part B and Part D data sets. Advanced features are constructed from behavioral similarity between providers and risk propagation through geospatial collocation, but the provider type predictor is not included.

Another set of related Medicare studies considers the relationship between HCPCS procedures and provider types. In [19], Bauder et al. use a Naive Bayes learner to predict Medicare provider specialties from HCPCS procedure occurrences. Results show that 7 of 82 provider types scored very highly (F1-score > 0.90), and 18 provider types scored reasonably ($0.5 <$ F1-score < 0.90). This suggests that the majority of the provider types have overlapping procedure activity that prevents accurate provider type prediction. Herland et al. [20] expand on this by incorporating 2014 CMS Part B data and real-world fraud labels defined by the LEIE data set. A Naive Bayes learner is used to predict a provider's specialty from their respective HCPCS frequencies, and misclassified provider types are assigned a fraudulent class label. Herland et al. discover that grouping similar provider types improves overall classification performance. Chandola et al. [21] use healthcare claims and fraudulent provider labels provided by the Texas OIG exclusion database to detect anomalies and bad actors. The authors model providers as documents and use Latent Dirichlet Allocation to identify 20 hidden topics from a provider-diagnosis matrix. Chandola et al. show how some topics are dominated by diagnoses belonging to the same area of medicine, e.g. Oncology and Ophthalmology, and suggest that the topic distributions can be used as features for downstream learning. Experimental results showed that the inclusion of the provider type attribute increases ROC AUC score from 0.716 to 0.814. These related works stress the importance of the provider type feature in predicting Medicare fraud, and allude that more semantic representations for provider types can improve the performance of Medicare fraud classification.

There are several related works that use CMS Medicare Part B data with one-hot encoded provider type predictors to classify Medicare fraud. Herland et al. [22] explore fraud prediction using three 2012–2015 CMS Medicare data sets, i.e. Part B, Part D, and DMEPOS. Fraudulent providers are identified using the LEIE data set and the feature set includes provider activity statistics, e.g. the average amount billed, average amount paid, and average number of beneficiaries treated. Part B, Part D, and DMEPOS claims data are used independently to perform cross-validation with LR, RF, and GBT learners. The LR learner performed best on the Part B data set with a maximum AUC score of 0.805. In another study [15], we reuse the CMS Part B data set from Herland et al. to evaluate deep neural networks and various techniques for addressing class imbalance. Data-level and algorithm-level methods are used to treat class imbalance, and results show that balancing training data with random over-sampling (ROS) and random under-sampling

TABLE 1. DESCRIPTION OF PART B FEATURES

Feature	Description	Type
NPI	Unique provider identification number	Categorical
HCPCS	Procedure/service code	Categorical
Provider_type	Medical provider's specialty (or practice)	Categorical
Nppes_provider_gender	Provider's gender	Categorical
Line_srvc_cnt	Number of procedures/services the provider performed	Numeric
Bene_unique_cnt	Number of distinct beneficiaries receiving the service	Numeric
Bene_day_srvc_cnt	Number of distinct beneficiary/per day services	Numeric
Average_submitted_chrg_amt	Average of charges that provider submitted for the HCPCS	Numeric
Average_medicare_payment_amt	Average payment made to a provider per claim for the HCPCS	Numeric
Exclusion	Fraud labels from the LEIE data set	Categorical

(RUS) maximizes performance with an average ROC AUC of 0.8506. We believe that the one-hot encoding of provider types in these previous works is insufficient, and that meaningful relations between provider types are lost in equidistant one-hot vectors. To address this information loss, this paper explores multiple encoding techniques that capture meaningful relations between similar provider types.

2.2. Clinical Concept Extraction

Breakthroughs in word embeddings and natural language processing have inspired authors to explore similar techniques for the purpose of clinical concept extraction and downstream prediction tasks. Clinical concept extraction is the process of identifying medical concepts from clinical notes. Khattak et al. [23] survey methods for embedding clinical concepts and cover a range of topics, e.g. word representation, clinical text corpora, pre-trained clinical embeddings, applications, and limitations. Choi et al. [24] use medical journals, medical claims, and clinical notes to learn low-dimensional embeddings for medical concepts. Results show that clinical embeddings learned on different sources of data capture different semantics, suggesting different embedding sources may perform better on specific downstream tasks. In another work, Choi et al. [25] propose Med2Vec for learning embeddings for medical codes and patient visits using procedure code co-occurrences and sequential patient data. Med2Vec outperforms Skip-Gram, GloVe, one-hot encoding, and stacked autoencoder embeddings on the task of predicting future medical codes. Yuqi et al. [26] evaluate the impact of word-level and context-sensitive word representations on the task of clinical concept extraction. The authors compare Word2Vec, GloVe, fastText [27], ELMo, and BERT on the i2b2 and SemEval data sets. Both open-domain and MIMIC-III [28] (clinical corpora) are used for pre-training word embeddings, and results show that pre-training on clinical corpora generally performs better. Huang et al. [29] propose ClinicalBERT for generating and evaluating representations of clinical notes. ClinicalBERT is trained on clinical notes from a patient's intensive care notes and discharge summary, and is evaluated on a 30-day hospital readmission prediction task. The proposed model outperforms a bag-of-words model and a bidirectional LSTM network that is trained using Word2Vec embeddings trained on the MIMIC-III clinical data set.

In summary, related work shows that semantic embeddings for medical concepts improve performance on downstream classification tasks. We extend existing Medicare fraud detection work by leveraging embedding ideas from these related works on clinical concept extraction. More specifically, we represent Medicare provider types using open-domain word embeddings, clinical word embeddings, and a new HcpcsVec representation that is derived from procedure occurences. Given the success of word embeddings for representing Medicare provider types, we plan to explore more advanced language model embeddings (e.g. ELMo, BERT, and ClinicalBERT) in future work.

3. Methods

3.1. Medicare Part B Data

Embedding techniques for Medicare providers are evaluated using the CMS 2012–2016 Medicare Part B data sets. These data sets use both provider-level and procedure-level attributes to describe Medicare providers and their billing activity, as illustrated by Table 1. The National Provider Identifier (NPI) is only used to map fraud labels from the LEIE data set, and is removed prior to training. HCPCS codes identify the specific procedures and services provided by healthcare professionals. HCPCS codes are not used for model training, but they are used in Section 3.2 to learn provider type embeddings. Five numeric attributes describe each provider's activity relative to a HCPCS code for a given year and place of service, e.g. the number of times the procedure was performed and the number of receiving beneficiaries. The provider type attribute consists of 123 unique values that describe a provider's specialty, e.g. Internal Medicine, Nurse Practitioner, and Urology.

The class label (Exclusion) is derived by joining the Part B data set with a subset of the LEIE data set using the NPI column. The LEIE data set is a list of providers that have been excluded from practice due to fraudulent behaviors, as determined by the OIG. Following previous works [22], we use a subset of exclusion types representative of fraud to label Part B providers as fraudulent and non-fraudulent.

For each year of data, we group records by NPI and provider type, and then aggregate the numeric attributes into a single record of summary statistics, i.e. minimum, maximum, sum, median, mean, and standard deviation. After converting the five numeric attributes to their summary

statistics and one-hot encoding the provider gender, there are 32 numeric attributes, a categorical provider type attribute, and a class label for model training. For more information on the Medicare Part B data set and the pre-processing steps outlined here, we refer readers to the original work by Herland et al. [22].

We create training and test sets for model evaluation by randomly partitioning the data 80-20%. The same test set is used for all experiments. All features are normalized to the range $[0, 1]$ using a min-max scaler. Finally, we replace the categorical provider type variable with a new set of attributes obtained from one of four embedding techniques at training time. The size and class imbalance levels of training and test partitions are described in Table 2.

TABLE 2. TRAINING AND TEST PARTITION DETAILS

Data Set	Total Samples	Fraudulent Samples	% Fraudulent
Training Data	3,752,690	1206	0.032%
Test Data	938,172	302	0.032%

One of the most challenging components of this Medicare fraud prediction task is its severe class imbalance. The fraudulent class makes up just 0.032% of the entire data set, roughly three fraudulent providers for every 10,000 non-fraudulent providers. This combination of high class imbalance and big data has proven detrimental to machine learning performance [30]. Since the focus of this study is not on addressing class imbalance, we reuse a hybrid ROS-RUS sampling technique that proved to be effective in recent work [31]. This hybrid sampling technique combines over-sampling and under-sampling to simultaneously improve efficiency and address class imbalance. The specific implementation used in this study first under-samples the majority class by randomly selecting 50% of the non-fraudulent class from the training data. The minority group, or fraudulent class, is then oversampled by randomly duplicating fraudulent samples until both classes are equal in size. The end result is a class-balanced (50:50) training set with approximately 1.87 million samples from each class. The test set is not sampled and can be considered representative of real-world Medicare fraud.

3.2. Provider Type Embeddings

We employ three provider type embedding techniques that were inspired by advances in natural language processing, and we compare results to those achieved with traditional one-hot vectors. One-hot vector encodings of provider types are sparse, atomic, and uninformative representations that fail to capture any form of provider similarity. Consider, for example, the provider types Certified Clinical Nurse Specialist, Certified Nurse Midwife, and Orthopedic Surgeon. With one-hot encoding, all specialties are equidistant in vector space and no pair of specialties is geometrically closer than another. In this study, we capture relationships between provider types using word embeddings and a specialty-HCPCS occurrence matrix.

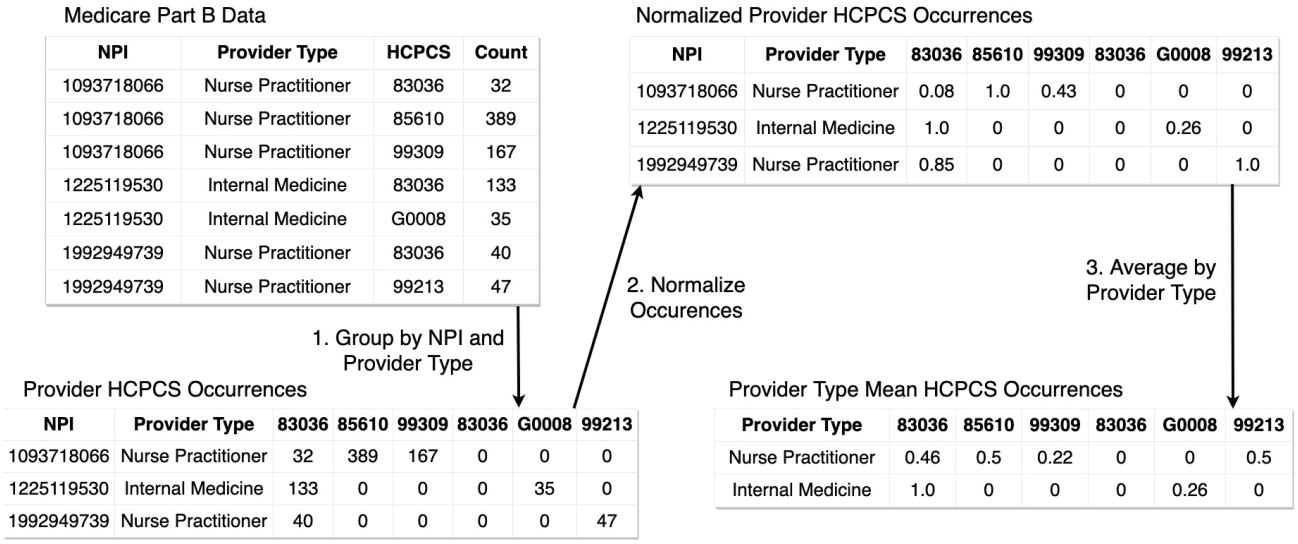
Provider type attributes are nothing more than short phrases, e.g. 1–4 word textual descriptions. Therefore, we can capture semantic relationships between provider types by replacing their textual descriptions with their word embeddings. Our first word embedding technique uses GloVe embeddings to represent each provider type using an unweighted average of word embeddings. For example, we encode Orthopedic Surgeon by adding the GloVe embedding for orthopedic and the GloVe embedding for surgeon, and then dividing by two. We chose to use unweighted averages because they are relatively simple and effective for short phrases [13]. Similarly, we repeat this process using domain-specific word embeddings. We refer to these embeddings as Med-Word2Vec embeddings in this study, because they were induced from large biomedical corpora using a Word2Vec model [14]. For both word embedding techniques, we use the *P+PCA* algorithm [32] to reduce the dimensionality of the word embeddings to lengths of 32, 64, and 128 based on preliminary results.

The third embedding technique (HcpcsVec) constructs dense provider type representations from a specialty-HCPCS occurrence matrix using training data. We accomplish this by aggregating HCPCS occurrences over all providers within a single specialty. First, we group the Medicare data on (NPI, provider type) and sum the total number of times a given HCPCS was performed. Since there are 7527 unique HCPCS values and most providers only use a small subset of procedures, this produces a very sparse HCPCS occurrence vector for each (NPI, provider type) pair. Next, we normalize each row of the occurrence matrix by dividing by the maximum occurrence in each row. We apply this normalization step because some providers are naturally more active than others, but the relative HCPCS frequencies should be similar for providers within the same specialty group. Finally, we group the normalized HCPCS occurrence matrix on provider type and take the average for each HCPCS column. The resulting specialty-HCPCS occurrence matrix has a size of 123×7527, and each row summarizes the average procedure activity for a particular specialty. We summarize this process visually using a small subset of both data and HCPCS codes in Figure 1. Unlike the word embedding techniques, we use plain PCA to reduce the dimensionality of the occurrence vectors to lengths of 32, 64, and 123. Note that we were not able to achieve an embedding size of 128 when using HcpcsVec because there are only 123 provider types.

3.3. Performance Evaluation

We evaluate embedding techniques for predicting Medicare fraud using four popular machine learning algorithms. First, we use the scikit-learn Python package [33] to implement the LR, RF, and GBT learners. Hyperparameters for each learner are seleted by maximizing performance on the training partition using five-fold cross-validation. For the LR learner, we set the maximum number of iterations to 200. For the RF and GBT ensembles, we restrict trees to a maximum depth of eight, and for the GBT learner we use

Figure 1. HcpcsVec Embedding Technique



the exponential loss. The fourth learner is a MLP neural network with two hidden layers and 32 neurons per layer. The MLP is implemented using the Keras deep learning library [34], and following our previous work [35] we employ ReLU activations, batch normalization, dropout, and the Adam optimizer. Finally, scikit-learn's Nearest Neighbor algorithm is used to interpret embeddings and list a subset of most similar specialties. Unless stated otherwise here, the remaining hyperparameters are left at their default values per scikit-learn v0.21.1 and Keras v2.1.6-tf.

Each machine learning algorithm is trained using all embedding techniques and is evaluated on the 20% holdout test set. To account for random error that is caused by data sampling, each experiment is repeated 30 times. This repetition enables us to report the mean AUC and margin of error with a significance level of $\alpha = 0.05$. These confidence intervals are used to identify sample means that are statistically significant with 95% confidence. We use the ROC AUC performance metric to evaluate results because the threshold-agnostic score is well suited for class-imbalanced data and previous works have reported ROC AUC results for comparison.

4. Results and Discussion

Three semantic embedding techniques for medical provider types are compared to one-hot encodings using four popular machine learning algorithms. Table 3 lists the mean AUC score for each embedding technique and learner combination. A sample of specialty embeddings is interpreted by comparing their nearest neighbors in Table 4.

To establish a baseline, we first compare model performance without the provider type attribute to model performance with one-hot encoded provider types. All four models perform significantly better when the provider type predictor is included in the model. We see the largest increases in performance using the LR and MLP learners, with an average AUC gain of 0.063 each. The RF and GBT learners improve by 0.013 and 0.027, respectively. When using one-hot encoded vectors, the MLP model performs best with an average AUC score of 0.852.

Next, we compare the performance of two word embedding techniques, GloVe and Med-Word2Vec. Both techniques convert medical provider types to phrase embeddings by taking the unweighted average of the individual words in the specialty description. The LR model performs approximately the same with GloVe and Med-Word2Vec embeddings, and it consistently performs better with larger embeddings of 64 and 128. For the RF, GBT, and MLP learners, no single embedding size or word embedding technique consistently outperforms another, and all differences in average AUC scores are minimal. The RF and GBT learners both significantly outperform their one-hot encoded baselines on average by as much as 0.024 and 0.020, respectively. The LR and MLP learners perform worse with word embeddings, and they see a decrease in AUC of 0.003 and 0.001, respectively. Overall, GloVe and Med-Word2Vec perform approximately the same for each learner, and the GBT learner with the GloVe-64 embedding maximizes performance with an average AUC of 0.870.

The HcpcsVec embedding scored the highest AUC score overall for the LR, RF and GBT learners. The RF and GBT learners saw a maximum AUC gain of 0.026 and 0.023 compared to one-hot encodings, respectively. The LR performed significantly better with HcpcsVec-123, but the performance gain was marginal. The MLP learner performs best with the HcpcsVec-32 embedding, and while the MLP's highest AUC score is recorded using a one-hot encoding, the difference in mean AUC scores is not statistically significant. The GBT model trained with the HcpcsVec-32 encoding scored the highest average AUC (0.873) on the Medicare fraud classification task and outperforms the best-known score of

TABLE 3. AVERAGE EMBEDDING PERFORMANCE (30 RUNS)

Embedding Method	Embedding Size	Mean AUC ± 95% MOE LR	RF	GBT	MLP
None	0	0.750 ± 7.6e−4	0.817 ± 4.2e−4	0.823 ± 1.1e−3	0.788 ± 6.4e−3
One-hot	123	0.813 ± 1.8e−4	0.830 ± 3.7e−4	0.850 ± 1.1e−3	**0.852** ± 1.2e−3
GloVe	32	0.794 ± 2.5e−4	0.854 ± 9.8e−4	0.867 ± 9.8e−4	0.847 ± 2.1e−3
	64	0.810 ± 2.0e−4	0.851 ± 7.0e−4	0.870 ± 6.9e−4	0.847 ± 3.6e−3
	128	0.811 ± 2.2e−4	0.851 ± 7.3e−4	0.866 ± 8.3e−4	0.843 ± 2.2e−3
Med-Word2Vec	32	0.801 ± 3.4e−4	0.853 ± 4.0e−4	0.868 ± 8.6e−4	0.848 ± 2.0e−3
	64	0.810 ± 2.8e−4	0.853 ± 2.9e−4	0.869 ± 8.9e−4	0.851 ± 9.7e−4
	128	0.811 ± 1.4e−4	0.850± 3.4e−4	0.863 ± 7.7e−4	0.844 ± 3.0e−3
HcpcsVec	32	0.786 ± 2.0e−4	**0.856** ± 2.7e−4	**0.873** ± 6.3e−4	0.849 ± 4.1e−3
	64	0.810 ± 3.6e−4	0.855 ± 3.3e−4	0.864 ± 6.1e−4	0.849 ± 1.2e−3
	123	**0.814** ± 2.8e−4	0.852 ± 3.4e−4	0.865 ± 9.3e−4	0.846 ± 1.3e−3

TABLE 4. INTERPRETING PROVIDER TYPE EMBEDDINGS

Provider Type	GloVe Neighbor	Med-Word2Vec Neighbor	HcpcsVec Neighbor
Hematology/Oncology	Hematology	Hematology	Medical Oncology
Anesthesiology	Anesthesiology Assistant	Anesthesiology Assistant	CRNA
Interventional Radiology	Interventional Cardiology	Interventional Cardiology	Diagnostic Radiology
Preventive Medicine	Geriatric Medicine	Geriatric Medicine	Nurse Practitioner
Sports Medicine	Physical Medicine and Rehabilitation	Physical Medicine and Rehabilitation	Orthopedic Surgery
Radiation Oncology	Radiation Therapy	Radiation Therapy Center	Radiation Therapy
Ophthalmology	Gastroenterology	Dermatology	Optometry
General Surgery	Hand Surgery	Oral Surgery (dentists only)	Neurosurgery
Maxillofacial Surgery	Orthopedic Surgery	Orthopedic Surgery	Oral Surgery (Dentist Only)
Hospitalist	Critical Care (Intensivists)	Hospice and Palliative Care	Undefined Physician type

0.851 from [35]. We believe that the HcpcsVec performs best overall on this Medicare task because it uses historical procedure-level data to capture provider type relationships instead of auxiliary natural language modeling tasks.

Table 4 lists ten provider specialties along with their nearest neighbor in each embedding space. As expected, the GloVe and Med-Word2Vec nearest neighbor results have the most overlap, with six total nearest neighbors matching. The HcpcsVec has the most unique results, with only one of ten nearest neighbors matching specialties from the word embedding group. Overall, each encoding technique succeeds in capturing semantic relationships between provider types. For example, the nearest neighbors for the General Surgery specialty are all surgical provider types. GloVe and Med-Word2Vec encode Interventional Cardiology closest to the Interventional Radiology provider type, while the HcpcsVec encoding has a nearest neighbor of Diagnostic Radiology. In another example, HcpcsVec embeds CRNA (Certified Registered Nurse Anesthetists) closest to the Anesthesiology specialty, while the word embedding methods encode the Anesthesiology Assistant specialty as its nearest neighbor. While the HcpcsVec embedding performed best overall in this study, all three embedding techniques significantly improved AUC performance on RF and GBT learners and performed similarly on all learners.

5. Conclusion

This study explored three new techniques for encoding medical specialty types for the purpose of classifying Medicare fraud with machine learning algorithms. Traditionally, the medical specialty, or provider type, has been represented using one-hot encodings. These sparse representations are unfavorable, however, because they treat each provider type atomically and fail to capture common similarities among related provider types. We addressed this by evaluating three techniques for constructing dense, semantic representations of medical specialties. The first two methods, GloVe and Med-Word2Vec, convert provider type variables to phrase embeddings using their textual descriptions and an unweighted average of pre-trained word embeddings. We then proposed a novel provider type embedding, HcpcsVec, that creates a specialty-HCPCS occurrence matrix by aggregating historical procedure-level data. The HcpcsVec embedding technique enabled the highest average AUC for the LR, RF, and GBT learners. When compared to one-hot encoded baselines, all three semantic embedding techniques enabled significant performance gains for the RF and GBT learners. In future works, we plan to explore new encoding and dimension reduction techniques, apply these embeddings to alternative medical benchmarks, and investigate feature importance.

Acknowledgments

The authors would like to thank the reviewers in the Data Mining and Machine Learning Laboratory at Florida Atlantic University.

References

[1] U.S. Government, U.S. Centers for Medicare & Medicaid Services.

The official u.s. government site for medicare. [Online]. Available: https://www.medicare.gov/

[2] Centers for Medicare & Medicaid Services. (2019) Medicare enrollment dashboard. [Online]. Available: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Dashboard/Medicare-Enrollment/Enrollment%20Dashboard.html

[3] Centers For Medicare & Medicaid Services. (2020) Trustees report & trust funds. [Online]. Available: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/ReportsTrustFunds/index.html

[4] L. Morris, "Combating fraud in health care: An essential component of any cost containment strategy," *Health affairs (Project Hope)*, vol. 28, pp. 1351–6, 09 2009.

[5] Centers For Medicare & Medicaid Services. (2019) Medicare provider utilization and payment data. [Online]. Available: https://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/medicare-provider-charge-data

[6] Office of Inspector General. (2019) Leie downloadable databases. [Online]. Available: https://oig.hhs.gov/exclusions/exclusions_list.asp

[7] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013.

[8] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. [Online]. Available: http://www.aclweb.org/anthology/D14-1162

[9] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *ArXiv*, vol. abs/1802.05365, 2018.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *ArXiv*, vol. abs/1810.04805, 2019.

[11] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100, 000+ questions for machine comprehension of text," in *EMNLP*, 2016.

[12] J. T. Hancock and T. M. Khoshgoftaar, "Survey on categorical data for neural networks," *Journal of Big Data*, vol. 7, no. 1, p. 28, 2020. [Online]. Available: https://doi.org/10.1186/s40537-020-00305-w

[13] S. Arora, Y. Liang, and T. Ma, "A simple but tough-to-beat baseline for sentence embeddings," in *ICLR*, 2017.

[14] S. Pyysalo, F. Ginter, H. Moen, T. Salakoski, and S. Ananiadou, "Distributional semantics resources for biomedical text processing," *Proceedings of Languages in Biology and Medicine*, 01 2013.

[15] J. M. Johnson and T. M. Khoshgoftaar, "Medicare fraud detection using neural networks," *Journal of Big Data*, vol. 6, no. 1, p. 63, 2019. [Online]. Available: https://doi.org/10.1186/s40537-019-0225-0

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

[17] J. Ko, H. Chalfin, B. Trock, Z. Feng, E. Humphreys, S.-W. Park, B. Carter, K. D Frick, and M. Han, "Variability in medicare utilization and payment among urologists," *Urology*, vol. 85, 03 2015.

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

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

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

[21] V. Chandola, S. R. Sukumar, and J. C. Schryver, "Knowledge discovery from massive healthcare claims data," in *KDD*, 2013.

[22] M. Herland, T. M. Khoshgoftaar, and R. A. Bauder, "Big data fraud detection using multiple medicare data sources," *Journal of Big Data*, vol. 5, no. 1, p. 29, Sep 2018. [Online]. Available: https://doi.org/10.1186/s40537-018-0138-3

[23] F. K. Khattak, S. Jeblee, C. Pou-Prom, M. Abdalla, C. Meaney, and F. Rudzicz, "A survey of word embeddings for clinical text," *Journal of Biomedical Informatics: X*, vol. 4, p. 100057, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2590177X19300563

[24] Y. Choi, C. Y.-I. Chiu, and D. A. Sontag, "Learning low-dimensional representations of medical concepts," *AMIA Summits on Translational Science Proceedings*, vol. 2016, pp. 41 – 50, 2016.

[25] E. Choi, T. Bahadori, E. Searles, C. Coffey, M. Thompson, J. Bost, J. Tejedor-Sojo, and J. Sun, "Multi-layer representation learning for medical concepts," in *22nd ACM SIGKDD International Conference*, 08 2016, pp. 1495–1504.

[26] Y. Si, J. Wang, H. Xu, and K. Roberts, "Enhancing clinical concept extraction with contextual embeddings," *Journal of the American Medical Informatics Association : JAMIA*, vol. 26, 07 2019.

[27] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *arXiv preprint arXiv:1607.04606*, 2016.

[28] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific Data*, vol. 3, no. 1, p. 160035, 2016. [Online]. Available: https://doi.org/10.1038/sdata.2016.35

[29] K. Huang, J. Altosaar, and R. Ranganath, "Clinicalbert: Modeling clinical notes and predicting hospital readmission," *ArXiv*, vol. abs/1904.05342, 2019.

[30] J. L. Leevy, T. M. Khoshgoftaar, R. A. Bauder, and N. Seliya, "A survey on addressing high-class imbalance in big data," *Journal of Big Data*, vol. 5, no. 1, p. 42, 2018. [Online]. Available: https://doi.org/10.1186/s40537-018-0151-6

[31] J. M. Johnson and T. M. Khoshgoftaar, "The effects of data sampling with deep learning and highly imbalanced big data," *Information Systems Frontiers*, 2020. [Online]. Available: https://doi.org/10.1007/s10796-020-10022-7

[32] V. Raunak, "Effective dimensionality reduction for word embeddings," *CoRR*, vol. abs/1708.03629, 2017. [Online]. Available: http://arxiv.org/abs/1708.03629

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[34] F. Chollet *et al.*, "Keras," https://keras.io, 2015. [Online]. Available: https://keras.io

[35] J. M. Johnson and T. M. Khoshgoftaar, "Deep learning and data sampling with imbalanced big data," in *2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)*, 2019, pp. 175–183.