Embedding Representations of Diagnosis Codes for Outlier Payment Detection

Anvesh Matta*
Deloitte & Touche Assurance & Enterprise Risk Services India Private Limited
India
manveshreddy@deloitte.com

Michael Suesserman*
Deloitte & Touche LLP
United States
msuesserman@deloitte.com

David McNamee
SFL Scientific, a Deloitte Business
Canada
damcnamee@deloitte.com

Daniel Lasaga
Deloitte Transactions & Business Analytics LLP
United States
dlasaga@deloitte.com

Dan Olson
Deloitte Transactions & Business Analytics LLP
United States
danolson@deloitte.com

Edward Bowen
Deloitte & Touche LLP
United States
edbowen@deloitte.com

Sanmitra Bhattacharya
Deloitte & Touche LLP
United States
sanmbhattacharya@deloitte.com

* These authors contributed equally

***Abstract*—Models for detecting payment outliers from healthcare claims often rely on sparse, high-dimensional feature vector encodings of diagnosis codes. These encodings tend to lose inherent relationships between the diagnosis codes, and lead to more complex and less efficient models. In this paper, we propose a novel approach that leverages word and graph embeddings to represent diagnosis codes, particularly when predicting healthcare claim payment amounts within an outlier detection model. Word embeddings are generated using BioSentVec, utilizing medical descriptions of diagnosis codes extracted from the Unified Medical Language System (UMLS) database. Graph embeddings are created using node2vec applied to a claims graph network that connects claims information with the diagnosis code hierarchy. On a dataset of 36 million claims, the graph embeddings outperformed other feature representations, improving R^2 by over 99% compared to sparse encodings. Embedding representations produce significantly smaller dense vectors that encapsulate more information than large, sparse multi-hot encoded diagnosis code vectors. These dense embedded vectors provide meaningful representations of diagnosis codes, significantly improving payment prediction and outlier detection capabilities.**

Keywords—Healthcare Fraud Waste and Abuse, FWA, Payment Outlier Detection, Word Embedding, Graph Embedding

I. Introduction

Healthcare providers and insurance companies rely on payment models to determine the reimbursement amount for services rendered. In the evolving landscape of healthcare, the accurate detection of outlier payments plays a pivotal role in ensuring the integrity and sustainability of healthcare reimbursement systems. While identification of outlier payments may not always imply fraudulent activities, it does serve as an essential mechanism for flagging cases that require further scrutiny and review. Prior research [1], [2] has demonstrated that outliers in payment or billing can serve as indicators of potential fraudulent activities within healthcare claims.

Broadly, healthcare fraud occurs when unscrupulous individuals or entities attempt to exploit vulnerabilities and loopholes in payment systems through deception to achieve unlawful financial gain. To counteract these activities, advanced techniques, including artificial intelligence and machine learning, are being employed to analyze medical claims data and detect potential instances of fraud, waste, and abuse (FWA), as well as improper payments [3], [4]. These methods scrutinize the intricacies of the data, aiming to uncover previously unrecognized trends and patterns, effectively thwarting emerging fraudulent schemes. This countermeasure promotes a preventive posture and helps payors mitigate the adverse consequences that a new fraud scheme may inflict.

Most FWA from purposeful or erroneous claims result in augmentation of billing through overutilization, upcoding, and unbundling [5]. One approach to identifying FWA could be to build a model for each of the individual drivers of FWA, such as overutilization, upcoding, and unbundling. This research chose to focus on identifying where treatment costs of diagnosis are above expected peer group averages allowing investigators to capture multiple drivers of FWA overpayments under a single model.

A model that accurately predicts the typical range of treatment costs for specific diagnoses can assist FWA investigators in identifying treatment costs that deviate significantly from the norm, where there might be a risk of inflated billing. However, primary diagnoses alone cannot account for ranges in costs. There is natural variation in the cost of treating a diagnosis [6]. Different factors go into what procedures a healthcare provider may need to apply to treat a condition (for example, cast, splint, or surgery for a broken leg). Older patients may require more complex and longer treatment. Patients with autoimmune disorders or heart conditions may add different complexities to the procedures used and the overall cost of treatment. Differences in the place of service or network of providers may add additional variation to the overall costs. We can see examples of natural variation among some example diagnoses in Table I. Natural variations in treatment costs,

1946-0759/23/$31.00 ©2023 IEEE
DOI 10.1109/ICMLA58977.2023.00325

combined with the overall complexity of healthcare billing systems, allow providers to conceal errors and abuses in plain sight.

TABLE I. Cost Examples for Common Diagnoses

Medical Condition	Cost Range
Vaginal Delivery	\$5,923 to \$9,535 (TX)[1]
Broken Leg	\$15,000 to \$40,000 (nationally)[2]
Sepsis	\$18,000 to \$50,000 (nationally)[3]

For a model to account for variation among common diagnoses it should condition on additional factors such as the location where the service was rendered, patient demographics, secondary diagnosis, and other preexisting health conditions. Creating a model that can accurately predict treatment cost ranges, considering the intricacies of primary and secondary diagnoses as well as any preexisting conditions, is challenging due to the presence of over 68,000 distinct diagnosis codes [7]. Typically, one to three diagnosis codes are associated with a claim. Attempting to train a model on sparse diagnosis data can produce an over specified model with potentially spurious predictions. Additionally, diagnosis codes have inherent interrelated structure among each other that is lost if each diagnosis is treated discretely. Additional disadvantages associated with sparse feature representations of diagnosis codes include heightened model complexity, a general decline in model performance due to the *curse of dimensionality* [8], and an expanded memory footprint. The need to address these challenges related to sparsity prompted the exploration of dense embedding representations of diagnosis codes, with the aim of effectively capturing relevant information and establishing meaningful relationships among commonly used diagnosis codes.

Our key contributions in this paper are: 1) we propose two novel techniques of representation learning of diagnosis codes in healthcare claims using word and graph embeddings techniques, and 2) to the extent of our knowledge, this study is the first one to show the effectiveness of these embeddings over sparse encodings of diagnosis codes in predicting healthcare claim payment amounts, which is a key step in our outlier payment detection model.

The paper covers related research in section II, data sources in section III, methods in section IV, and presents and discusses results in sections V and VI, respectively.

II. Related research

Bauder and Khoshgoftaar [1] proposed a multivariate outlier payment detection method by combining multivariate regression with probabilistic programming. Using Medicare Provider Utilization and Payment Data: Physician and Other Supplier Public Use File CY 2012-2014 [9], the authors implement the regression model with features such as zip code, year, number of services performed per provider, total sum of procedures performed across providers, number of distinct Medicare beneficiaries served per day, average amount allowed for the services, and the average amount Medicare paid for the services performed. They show the viability of their methods in detecting potential payment fraud activities. In a prior research [10] Bauder and Khoshgoftaar demonstrated that it was possible to flag potential fraud by using a regression model to establish a baseline for expected payments and when actual payments deviate beyond a certain threshold they can be seen as outliers. Khurjekar et al. [11] proposed unsupervised approaches to detecting fraud using a multivariate regression model with average payment as the dependent variable. Claims which exceeded a residual threshold of \$500 were clustered to identify fraudulent cases based on average cluster distances.

Thornton et al. [12] applied outlier detection methods on state Medicaid claim provider and patient data, specifically for dental claims, to identify fraudulent activities. In their study they define metrics such as visit length, patient retention, and frequency of visits from known fraud cases and experts, which are treated as features in their study. While they developed an initial set of 100 behavioral metrics, this list was refined to only fifteen that could be applied to the dental provider pool. Out of 360 dental providers, 17 were identified as outliers on three or more metrics. Hu et al. [13] presented a framework based on regression models and Grubb's outlier detection test for utilization analysis including hot spotting and anomaly detection. Specifically for anomaly detection they built models that map patients' demographics and clinical characteristics such as International Classification of Disease (ICD) codes and Hierarchical Condition Categories codes to utilization levels, which can then be used to study deviations between expected and actual utilization levels. In [14] researchers focused on prediction of claim amounts and large spending anomalies from Medicare claims data. Using general linear regression and multivariate outlier detection on the residuals they show feasibility of studying outlier Medicare claim payments. They found that diagnostic radiation services in larger population states typically contributed to extreme outlier in terms of cost of services.

Dense embeddings approaches have been previously used to train models for detecting healthcare FWA. Kumar et al. [15] combined claim information, such as diagnosis codes, procedure codes, and provider information, to create various provider embeddings. The provider embeddings are concatenated and passed to a meta embedding generator that produces a single meta embedding for a provider. The provider meta embedding is then passed to a classifier to determine if the provider is engaged in fraudulent activity. Sun et al. [16] evaluated medical knowledge graphs for detecting healthcare FWA. Specifically, the knowledge graph identifies three variations of inappropriate combinations of diagnosis and medication: a drug does not have the disease as an indication, the disease is a contraindication of the drug, and no suitable drugs for treating the disease appear in this claim. The authors constructed a medical knowledge graph by extracting entities and relationships from unstructured knowledge sources, including medical textbooks and medical examination records. Word embeddings were used to transform

[1] https://www.mdsave.com/procedures/vaginal-delivery/d785fdca/texas

[2] https://enhancehealth.com/how-much-does-a-broken-bone-cost-without-insurance/

[3] https://www.wolterskluwer.com/en/expert-insights/the-true-cost-of-sepsis-how-performance-improvement-programs-are-missing-patients

the unstructured medical data into a format that can more effectively be used to build the medical knowledge graph. In [17] the authors propose Hcpcs2Vec for embedding procedure codes and show that for provider fraud classification these embeddings outperform one-hot encodings and other pre-trained embeddings. While various approaches to encoding diagnosis codes [18], [19] have been explored within the context of healthcare records, their adoption for detecting healthcare FWA from claims data remains limited.

In contrast to prior studies which used procedure codes as key predictors of payment amounts and identification of outlier payments, we focus on diagnosis codes as key predictors of payment amount. The advantages of this approach have been discussed earlier. Most prior studies on payment outlier detection focus on traditional sparse feature encoding techniques such as one-hot or multi-hot encoding techniques to transform the claims data into feature vectors that can be used by machine learning models. Instead, we focus on using dense graph and word embeddings of diagnosis codes to predict paid amounts, which are in turn used to detect claims with outlier payments.

III. Dataset

A redacted and anonymized dataset of outpatient medical claims from state Medicare programs is used in this study. Outpatient claims are billed when patients visit healthcare providers for treatment but do not get admitted to a hospital. This dataset comprises of over 36 million distinct claims, over 2.8 million patients and around 370,000 healthcare providers.

Claims submitted to Medicare generally contain the following information:

- Claim number: a distinct identifier for each claim.
- Diagnosis codes: represent patient diagnosis and typically encoded using a standardized coding system for clinical terms, ICD-10[4]. In this dataset each claim contains up to three ICD-10 codes – primary, secondary, and tertiary. ICD-10 codes consist of up to seven characters with the first three characters representing the general diagnosis, with remaining characters representing more specific categories. For example, the ICD-10 code S86.011D can be resolved as:
    - S86: injury of muscle, fascia, and tendon at lower leg. This is referred to as the general category code.
    - S86.011: strain of right Achilles tendon
    - D: represents a subsequent encounter

  There are over 68,000 distinct ICD-10 codes.
- Procedure codes: capture procedures performed by healthcare provider and represented as Current Procedural Terminology (CPT) or Healthcare Common Procedure Coding System (HCPCS) codes[5]. These codes are made up of 5 characters. For example, the CPT code 73615 represents "Review X-ray to determine if ankle is broken" and the HCPCS code E0112 represents "Prescribe underarm crutches". There are over 10,000 CPT/HCPCS codes. A claim contains at least one procedure but can consist of multiple procedures.
- Provider ID: a distinct identifier for each provider typically following the National Provider Identifier (NPI) registry.
- Patient demographics: patient age and gender.
- Business practice state: primary state where the provider's practice is enrolled.
- Paid amount: total amount paid to a healthcare provider for all the procedures provided to a patient for a single claim.

During data preprocessing we remove claims that violate data integrity checks such as claims with payment amount of zero United States dollars (USD), voided claims (claims not processed for payment due to user, patient, or payor errors), and claims where providers cannot be identified.

IV. Methods

A. Featurization of Data for Machine Learning Models

To build machine learning models for payment outlier detection we perform various transformations on the raw claims data. Encodings for variables with low dimensionality such as age and gender, as well as those with high dimensionality such as diagnoses codes are discussed below:

- Age: Age is bucketed into three categories: under 18, 18 to 54, and 55 and older. These categories are selected to align with the differences in procedures and medical costs that occur at different stages of life (Table II). The three age categories are one-hot encoded.
- Gender: Gender consists of a one-hot encoded vector with 2 categories: male and female (Table II).
- Business practice state: The physician's business practice state is one of a finite set of US states and is one-hot encoded.
- Payment amount: In this dataset the mean payment amount was 229.34 USD, and the standard deviation was 1913.80 USD. To reduce the skewness of the data the payment amounts are log transformed.
- Diagnosis codes: All claims are required to contain a primary diagnosis (ICD-10) code. Claims may also contain optional secondary and tertiary ICD-10 codes. Three variations of diagnosis code features are tested: sparse, word embedded, and graph embedded. The embedding approaches are elaborated in further details in the following

[4] https://www.cms.gov/medicare/coordination-benefits-recovery-overview/icd-code-lists

[5] https://www.cms.gov/medicare/fraud-and-abuse/physicianselfreferral/list_of_codes

sections. To reduce the length of the sparse encodings, the general category codes for the primary, secondary, and tertiary diagnosis are multi-hot encoded for each claim. Results from features with sparse, multi-hot encoded diagnosis codes are compared with word and graph embedded diagnosis codes. Our dataset consisted of 40,353 individual full ICD-10 codes and 1882 general category codes.

TABLE II. Summary of Gender and Age Category Distributions

	Female	Male	
Less than 18	10.64%	11.72%	22.36%
18 to 54	31.02%	15.20%	46.22%
55 and older	19.29%	12.13%	31.42%
	60.95%	39.05%	100%

B. Word Embedding of Diagnosis Codes

Word2vec is a natural language processing technique that uses a neural network trained on a large text corpus to learn associations between different words [20], [21]. It does this by representing words as multi-dimensional vectors that can then be processed using vector mathematics. BioSentVec [22] is an extension of word2vec that is trained on biomedical content from PubMed articles and clinical notes from Medical Information Mart for Intensive Care-III (MIMIC-III) Clinical Database[6]. It can generate embeddings for biomedical words and phrases and has been shown to achieve state-of-the-art results in sentence pair similarity tasks. Compared to more recent Transformer-based models, such as BioBERT [23] or ClinicalBERT [24], BioSentVec has demonstrated an average inference time that is 50 times faster (a critical consideration given the scale of our dataset) while also exhibiting superior performance in benchmarking experiments [25].

The National Institutes of Health maintains a Unified Medical Language System (UMLS) database[7] that contains text descriptions for all ICD-10 codes. Descriptions are available for both the general category and full ICD-10 codes.

To embed the ICD-10 codes, the UMLS description of each code is pre-processed to remove common stop words, and then it is embedded using BioSentVec. The result is a vector with 200 dimensions that represents an ICD-10 code description. This is done for both the general category code description and the full ICD-10 code description.

Since we use up to three ICD-10 codes (primary, secondary and tertiary) in this study, the final word embeddings used as model features are generated using two approaches: averaging and concatenation. For averaging, the three ICD-10 codes are averaged to create a single, dense vector with 200 dimensions. For concatenation, the embedding for the primary ICD-10 code is concatenated with the average for the secondary and tertiary ICD-10 codes, creating a single, dense vector with 400 dimensions. Averaging and concatenation is done with both the embedded general category codes and full ICD-10 codes, resulting in four variations of word embedded ICD-10 codes that are tested.

C. Graph Embedding of Diagnosis Codes

Graphs serve as a means of representing network structure information, where nodes represent individual entities interconnected by edges signifying relationships. Healthcare claims data stands out as an excellent candidate for effective graph modeling. In this context, we present a novel approach that combines claim-level details with the ICD-10 hierarchy to create diagnosis embeddings.

In the realm of downstream machine learning applications, node embeddings play a crucial role [26]. In this study, our focus lies solely on the nodes associated with ICD-10 diagnosis codes. To create node embeddings from graphs, three key steps are required: (1) generating a graph structure from the claims data, (2) learning dense vector representation for individual nodes, and (3) extracting the diagnosis code node embeddings for downstream application. For the purposes of this study, we opted for a simple graph topology (the rule set determining how nodes and edges are arranged in a network) consisting of diagnosis (ICD-10) codes, procedure (CPT) codes, and claim numbers as nodes [27]. Three distinct types of edges connect these nodes to create the graph structure:

(1) Edges are formed between claim numbers and any CPT codes reported in reference to the claim.

(2) Edges are formed between claim numbers and any ICD-10 codes reported in reference to the claim.

(3) ICD-10 codes have a hierarchical tree structure [28]. For example, diabetes mellitus (child) belongs to the “endocrine, nutritional and metabolic diseases” (parent) category. Edges are created between parent and child ICD-10 codes.

This approach results in a highly connected graph, with claims numbers serving as hubs linking ICD-10 and CPT codes with similar attributes. To maintain consistent graph properties, we utilized more than 200,000 distinct claims, extracted from the larger dataset of 36 million claims. We also experimented with constructing graphs using larger and smaller claim samples from the dataset. The results demonstrated that embeddings derived from graphs with larger claim samples remained consistent with those using 200,000 claims, while smaller claim samples exhibited slight variations. The constructed graph is then processed using node2vec [29] to generate embeddings for all general category and full ICD-10 codes. This produced embedded vectors with 32 dimensions for representing the ICD-10 codes.

As indicated for BioSentVec embeddings, we use up to three diagnosis codes for this study. The final embeddings used for model features are generated using two approaches: averaging and concatenation. For averaging, the three diagnosis codes are averaged to create a single, dense vector with 32 dimensions. For concatenation, the embedding for the primary diagnosis code is concatenated with the average for the secondary and tertiary diagnosis codes, creating a single, dense vector with 64

[6] https://github.com/ncbi-nlp/BioSentVec

[7] https://www.nlm.nih.gov/research/umls/index.html

dimensions. Averaging and concatenation is done with both the embedded general category codes and full diagnosis codes, resulting in four variations of graph embedded diagnosis codes that are tested.

D. Outlier Payment Detection Model

The outlier payment detection model comprises of two parts: a supervised paid amount predictor model, followed by an unsupervised probabilistic outlier detector model.

The supervised model learns to predict the paid amount for a given claim. We experimented with random forest (RF) and gradient boosted tree (GBT) regression models. These models have been shown to outperform other regression models of varying complexities ranging from linear regression to neural networks across a wide range of applications [30]–[32], including healthcare fraud detection [33], [34].

For the unsupervised probabilistic outlier detection model, we convert the predicted payment amounts from the previous step into residuals by subtracting the predicted amount from the actual paid amount for a claim. For each primary diagnosis code, a distribution of residuals is generated, and *z-scores* are computed for individual claims. A probability (p) of each claim being an outlier is computed using a truncated normal distribution using the following formula:

$$p = \frac{\varphi(\mu,\delta;z_{score}) - \varphi(\mu,\delta;z_{cutoff})}{\varphi(\mu,\delta;\infty) - \varphi(\mu,\delta;z_{cutoff})} * I_{residual}$$

where $\varphi(\mu,\delta;x)$ is the cumulative density function (*cdf*) of the normal distribution. The *cdf* gives the probability that a random variable with a given mean μ and standard deviation δ will take a value less than or equal to x. z_{cutoff} is a threshold *z-score*, beyond which a data point is considered potentially anomalous. $I_{residual}$ is an identity function which is used to consider only one tail of the distribution. It is represented as:

$$I_{residual} = \begin{cases} 0 & residual < 0 \\ 1 & residual \geq 0 \end{cases}$$

In essence, this methodology applies a series of transformations and calculations to residuals from a regression model to identify and quantify data points that are potentially anomalous. This approach aids in prioritizing which claims might need further investigation due to their unusual nature.

For the supervised model, the dataset of 36 million claims is split into training and testing datasets in a 7:3 ratio, resulting in a training dataset with over 25 million distinct claims, and a testing dataset with almost 11 million distinct claims. Reliable detection of outlier payments depends on how accurately a model predicts paid amount from the input features. To evaluate the two models tested, we use Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R^2). These metrics are commonly used to identify how well a regression model fits a dataset. For MAE, MSE and RMSE, smaller values for these metrics indicate better fit. R^2 measures how well predictions approximate the actual data. R^2 ranges from zero to one with larger values indicating a better fit. Although we look at each of these metrics, R^2 generally is most informative for evaluating regression model performance [35].

For the unsupervised model that identifies payment outliers for a given claim, we do not have labeled data to validate model outputs. To evaluate model performance, FWA subject matter specialists (SMS) labeled a set of 58 distinct claims with the likelihood of being an outlier payment: 'not likely' (NO), 'low likelihood' (LL), 'medium likelihood' (ML), or 'high likelihood' (HL). We experiment with various class compositions of these likelihood labels to evaluate model performance (Table III).

TABLE III. Likelihood Label Combinations

Negative Class	Positive Class
NO	LL + ML + HL
NO + LL	ML + HL
NO + LL + ML	HL

For example, in NO vs (LL + ML + HL), cases labeled as unlikely to be outliers are categorized as the negative class, while cases likely to be outlier are categorized as the positive class. Other combinations explored are: Negative Class: (NO + LL) vs Positive Class: (ML + HL), and Negative Class: (NO + LL + ML) vs Positive Class: ML. These label combinations help us determine the likelihood levels at which our model outputs align best with SMS labels. As each label combination leads to the binarization of classes, we evaluate the model using the Area Under the Receiver Operating Characteristic (AU-ROC) curve, a commonly used metric for binary classification. AU-ROC scores range from 0 to 1 with higher values indicating better predictions.

V. RESULTS

A. Paid Amount Prediction

For paid amount prediction with RF and GBT regression models, results were generated with the following variations of the diagnosis code representations:

- Sparse encoded general category diagnosis codes
- Word embedded general category diagnosis codes
- Word embedded full diagnosis codes
- Graph embedded general category diagnosis codes
- Graph embedded full diagnosis codes

In every case, the GBT regression model outperformed the RF regression model. Since this study is focused on identification of optimal feature representation of diagnosis codes for payment outlier detection, we only present the results of the GBT regression model for various diagnosis code representations.

For word embedding, the embedded general category diagnosis codes outperformed embedded full diagnosis codes. For graph embedding, the embedded full diagnosis codes outperformed embedded general category diagnosis codes. The results presented in this section compare performance for sparse encoded diagnosis vectors, word embedded general category diagnosis codes, and graph embedded full diagnosis codes.

1) Sparse Encoded Versus Word Embedded Features

Results comparing sparse encoded and word embedded diagnosis codes as features in the model are presented in Table IV. The general category codes for diagnoses are used for both sparse encoded and word embedded features. Results for word embedded general category diagnosis codes are presented because they outperform full word embedded diagnosis codes on all evaluation metrics. Across all the metrics in Table IV, namely, MAE, MSE, RMSE, and R^2 we find that the model with concatenated word embeddings outperforms the other featurization techniques.

TABLE IV. Summary of Results on Test Data with GBT Model for Sparse General Category Codes and Word Embedded General Category Codes

Metric	Sparse Diagnosis Codes	Average Word Embedded Diagnosis Codes	Concatenated Word Embedded Diagnosis Codes
MAE	0.850	0.853	0.810
MSE	1.396	1.409	1.274
RMSE	1.182	1.187	1.129
R^2	0.127	0.168	**0.229**

2) *Sparse Encoded Versus Graph Embedded Features*

Results comparing sparse encoded, and graph embedded diagnosis codes are presented in Table V. Results for full graph embedded diagnosis codes are presented because they outperform general category graph embedded diagnosis codes on each of the evaluation metrics. As with the word embeddings we find that the concatenated graph embeddings outperform the other featurization approaches.

Overall, we find that the GBT regression model trained with graph embedded diagnosis codes outperforms the same model trained with word embedded codes.

TABLE V. Summary of Results on Test Data with GBT Model for Sparse General Category Codes and Graph Embedded Full Diagnosis Codes

Metric	Sparse Diagnosis Codes	Average Graph Embedded Diagnosis Codes	Concatenated Graph Embedded Diagnosis Codes
MAE	0.850	0.850	0.808
MSE	1.396	1.388	1.238
RMSE	1.182	1.178	1.113
R^2	0.127	0.177	**0.253**

B. *Probabilistic Outlier Detection*

As discussed in Section IV (D), prediction of paid amount for a given claim is key in determining outlier payments. Since we have already shown that for paid amount prediction the GBT regression model with diagnosis codes represented as graph embeddings outperform other approaches, in this section we only present evaluation of the unsupervised probabilistic outlier payment detector that uses this model.

AU-ROC scores of model predictions against FWA SMS labeled data with various combinations of likelihood are shown in Table VI.

TABLE VI. AU-ROC Scores of Unsupervised Outlier Detection Method on FWA SMS Labeled Data Across Various Class Compositions.

Negative Class	Positive Class	AU-ROC
NO	LL + ML + HL	0.715
NO + LL	ML + HL	0.760
NO + LL + ML	HL	**0.880**

We find that the model performs well across each of the class compositions with AU-ROC scores ranging between 0.715 and 0.88. Our model is most effective in identifying outliers that are considered highly likely by FWA SMS.

VI. Discussion

We conducted a feature importance analysis on the paid amount prediction model to gauge the influence of each feature on its performance. Models based on Classification and Regression Trees (CART), like the GBT model utilized in this study, can estimate feature importance [36]. The primary diagnosis code has by far the greatest impact on the model performance at 48 percent, which is twice as important as the secondary diagnosis code, and of much greater importance than the other features (business practice state: 13.5%, tertiary diagnosis code: 13.3%, age and gender combined: 0.4%). This result is expected since the primary diagnosis is directly related to the claim payment amount.

Tables IV and V compare performance for sparse and embedded ICD-10 codes. The MAE, MSE, RMSE, and R^2 metrics provide insight into how well the model performs with embedded ICD-10 versus the baseline sparse ICD-10. As discussed in Section IV (D), the ability of the model to reliably detect paid amount outliers for a specific primary diagnosis code depends on how accurately it predicts paid amount from input feature data, and R^2 is the most appropriate metric for evaluating model performance.

The results presented in Table IV compare performance for sparse ICD-10 codes with word embedded ICD-10 codes. For both the sparse and word embedded results, general category ICD-10 codes are used. Averaging the word embedded ICD-10 codes improved R^2 performance over the sparse ICD-10 codes by over 32 percent for the test data. Concatenating the word embedded ICD-10 codes improved R^2 performance over the sparse ICD-10 codes by over 80 percent for the test data.

The results presented in Table V compare performance for sparse ICD-10 codes with graph embedded ICD-10 codes. For both the sparse results, general category ICD-10 codes are used. For the graph embedding results, full ICD-10 codes are used. Averaging the graph embedded ICD-10 codes improved R^2 performance over the sparse ICD-10 codes by over 39 percent for the test data. Concatenating the graph embedded ICD-10 codes improved R^2 performance over the sparse ICD-10 codes by over 99 percent for the test data.

In addition to significantly improving model performance, embedding offers other benefits. With embeddings, the full ICD-10 code can be used. Since the data used in this study contains over 40 thousand ICD-10 codes, using the full diagnosis code for sparse features is not possible because the total feature vector would have a length of over 40 thousand of mostly zeros for just the diagnosis codes. Even the general category code produces a large, sparse ICD-10 vector with a length of over 1,800. Word embedding creates a dense vector with a total length of 200 for averaged embeddings or 400 for concatenated embeddings. Similarly, graph embedding the full diagnosis codes creates a dense vector with total length of 32 or 64 for averaged and concatenated, respectively. Also, the much smaller dense embedded vectors contain significantly more information than the large, sparse vectors. The combination of these benefits explains why embedded ICD-10 significantly outperform sparse, multi-hot encoded ICD-10.

When utilizing our payment outlier model to detect suspicious payment amounts from claims, we uncovered an instance in which a patient diagnosed with Autistic Disorder (ICD-10 code F84.0) was billed by a healthcare practitioner specializing in Applied Behavior Analysis (ABA) at rates ranging from $150 to $210 per 15-minute unit. This was a noticeable deviation from the peer group average of $30 per 15-minute unit found in our dataset. Additionally, our model flagged several claims from a Home Health Agency (HHA) which upon further investigation exhibited potential overcoding of billed units and discrepancies in the level of provider billing. These led to inflated reimbursements above peer group averages for various diagnosis codes, including Type II Diabetes Mellitus, without complications (ICD-10 code E11.9), Autistic Disorder (ICD-10 code F84.0), and Cerebral Palsy, unspecified (ICD-10 code G80.9).

VII. Conclusion

A GBT regression model, when combined with a probabilistic outlier detector and using embedded diagnosis codes as input features, performs effectively in detecting healthcare payment outliers. In all scenarios, dense word and graph embeddings significantly outperformed baseline results with sparse, multi-hot encoded diagnosis codes. Although it results in larger vectors, concatenating the embedded diagnosis codes yielded better results than averaging, likely due to the richer information in concatenated vectors. Notably, for paid amount prediction, graph embedding showed an improvement of over 99% in R^2 compared to sparse encodings. Due to the time-consuming and labor-intensive nature of manually labeling validation data by FWA SMS, the number of data points available for model testing was limited. Nonetheless, we intend to expand this dataset in the future as more cases are flagged by the model and prioritized for SMS review. Possible areas of future research include evaluating other ensemble approach for the regression model and incorporating more claims features, such as embedded procedure codes. While we present preliminary results of an unsupervised probabilistic method for outlier payment detection, we would like to explore other unsupervised approaches in future.

References

[1] R. A. Bauder and T. M. Khoshgoftaar, “Multivariate outlier detection in medicare claims payments applying probabilistic programming methods,” *Health Serv Outcomes Res Method*, vol. 17, no. 3–4, pp. 256–289, Dec. 2017, doi: 10.1007/s10742-017-0172-1.

[2] G. van Capelleveen, M. Poel, R. M. Mueller, D. Thornton, and J. van Hillegersberg, “Outlier detection in healthcare fraud: A case study in the Medicaid dental domain,” *International Journal of Accounting Information Systems*, vol. 21, pp. 18–31, Jun. 2016, doi: 10.1016/j.accinf.2016.04.001.

[3] N. Q. Tran *et al.*, “Leveraging deep survival models to predict quality of care risk in diverse hospital readmissions,” *Sci Rep*, vol. 13, no. 1, p. 10479, Jun. 2023, doi: 10.1038/s41598-023-37477-3.

[4] S. Aguinaga, D. Lasaga, S. Danthuri, J. Helms, E. Bowen, and S. Bhattacharya, “Identification of Providers with Similar Risk Profiles in Healthcare Claims Using Graphs,” in *2023 11th International Symposium on Digital Forensics and Security (ISDFS)*, Chattanooga, TN, USA: IEEE, May 2023, pp. 1–6. doi: 10.1109/ISDFS58141.2023.10131779.

[5] C. Buppert, “What Kind of Errors May Be Considered Fraud?,” *The Journal for Nurse Practitioners*, vol. 9, no. 3, pp. 180–181, Mar. 2013, doi: 10.1016/j.nurpra.2012.12.018.

[6] A. Chandra, D. Cutler, and Z. Song, “Who Ordered That? The Economics of Treatment Choices in Medical Care,” in *Handbook of Health Economics*, Elsevier, 2011, pp. 397–432. doi: 10.1016/B978-0-444-53592-4.00006-2.

[7] A. Barta, G. McNeill, P. Meli, K. Wall, and A. Zeisset, “ICD-10-cm primer.,” *Journal of AHIMA*, vol. 79, No. 5, 2008.

[8] L. Chen, “Curse of Dimensionality,” in *Encyclopedia of Database Systems*, L. Liu and M. T. Özsu, Eds., Boston, MA: Springer US, 2009, pp. 545–546. doi: 10.1007/978-0-387-39940-9_133.

[9] “CMS Research, Statistics, Data & Systems.” https://www.cms.gov/research-statistics-data-and-systems/research-statistics-data-and-systems

[10] 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)*, Pittsburgh, PA, USA: IEEE, Jul. 2016, pp. 11–19. doi: 10.1109/IRI.2016.11.

[11] N. Khurjekar, C.-A. Chou, and M. T. Khasawneh, “Detection of fraudulent claims using hierarchical cluster analysis,” in *IIE Annual Conference. Proceedings*, Institute of Industrial and Systems Engineers (IISE), 2015, p. 2388.

[12] D. Thornton, G. van Capelleveen, M. Poel, J. van Hillegersberg, and R. M. Mueller, “Outlier-based Health Insurance Fraud Detection for U.S. Medicaid Data,” *Proceedings of the 16th International Conference on Enterprise Information Systems - Volume 2*. SCITEPRESS - Science and Technology Publications, Lda, Lisbon, Portugal, pp. 684–694, 2014. [Online]. Available: https://doi.org/10.5220/0004986106840694

[13] J. Hu, F. Wang, J. Sun, R. Sorrentino, and S. Ebadollahi, “A healthcare utilization analysis framework for hot spotting and contextual anomaly detection,” *AMIA Annu Symp Proc*, vol. 2012, pp. 360–369, 2012.

[14] D. Chrimes, “Big data analytics of predicting annual US Medicare billing claims with health services,” in *2022 IEEE International Conference on Big Data (Big Data)*, Osaka, Japan: IEEE, Dec. 2022, pp. 2747–5754. doi: 10.1109/BigData55660.2022.10020524.

[15] N. Kumar, D. Chaurasiya, A. Singh, S. Asthana, K. Agarwal, and A. Arora, “MeDML: Med-Dynamic Meta Learning-A multi-layered representation to identify provider fraud in healthcare.,” *The International FLAIRS Conference Proceedings*, vol. 34, 2021.

[16] H. Sun *et al.*, “Medical knowledge graph to enhance fraud, waste, and abuse detection on claim data: model

development and performance evaluation.," *JMIR Medical Informatics*, vol. 8, No. 7, p. e17653, 2020.

[17] J. M. Johnson and T. M. Khoshgoftaar, "Hcpcs2Vec: Healthcare Procedure Embeddings for Medicare Fraud Prediction," in *2020 IEEE 6th International Conference on Collaboration and Internet Computing (CIC)*, Atlanta, GA, USA: IEEE, Dec. 2020, pp. 145–152. doi: 10.1109/CIC50333.2020.00026.

[18] Y. C. Lee *et al.*, "ICD2Vec: Mathematical representation of diseases," *Journal of Biomedical Informatics*, vol. 141, p. 104361, May 2023, doi: 10.1016/j.jbi.2023.104361.

[19] M. J. Kane, C. King, D. Esserman, N. K. Latham, E. J. Greene, and D. A. Ganz, "A Compressed Language Model Embedding Dataset of ICD 10 CM Descriptions," Health Informatics, preprint, Apr. 2023. doi: 10.1101/2023.04.24.23289046.

[20] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space.," 2013.

[21] K. Ward, "Word2Vec," *Natural Lanuage Engineering*, vol. 23, No. 1, pp. 155–162, 2017.

[22] Q. Chen, Y. Peng, and Z. Lu, "BioSentVec: creating sentence embeddings for biomedical texts," in *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, Xi'an, China: IEEE, Jun. 2019, pp. 1–5. doi: 10.1109/ICHI.2019.8904728.

[23] J. Lee *et al.*, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, Art. no. 4, 2020.

[24] K. Huang, J. Altosaar, and R. Ranganath, "ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission," 2019, doi: 10.48550/ARXIV.1904.05342.

[25] Q. Chen, A. Rankine, Y. Peng, E. Aghaarabi, and Z. Lu, "Benchmarking Effectiveness and Efficiency of Deep Learning Models for Semantic Textual Similarity in the Clinical Domain: Validation Study," *JMIR Med Inform*, vol. 9, no. 12, p. e27386, Dec. 2021, doi: 10.2196/27386.

[26] J. Zhou, L. Liu, W. Wei, and J. Fan, "Network Representation Learning: From Preprocessing, Feature Extraction to Node Embedding," *ACM Comput. Surv.*, vol. 55, no. 2, pp. 1–35, Feb. 2023, doi: 10.1145/3491206.

[27] B. Bollobás, "Modern graph theory," in *Graduate Texts in Mathematics*, New York, NY: Springer, 1998.

[28] A. Blanco, A. Perez, and A. Casillas, "Exploiting ICD Hierarchy for Classification of EHRs in Spanish Through Multi-Task Transformers," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 3, pp. 1374–1383, Mar. 2022, doi: 10.1109/JBHI.2021.3112130.

[29] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," presented at the Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, 2016, pp. 855–864.

[30] M.-J. Jun, "A comparison of a gradient boosting decision tree, random forests, and artificial neural networks to model urban land use changes: the case of the Seoul metropolitan area," *International Journal of Geographical Information Science*, vol. 35, no. 11, pp. 2149–2167, Nov. 2021, doi: 10.1080/13658816.2021.1887490.

[31] R. Abedi, R. Costache, H. Shafizadeh-Moghadam, and Q. B. Pham, "Flash-flood susceptibility mapping based on XGBoost, random forest and boosted regression trees," *Geocarto International*, vol. 37, no. 19, pp. 5479–5496, Oct. 2022, doi: 10.1080/10106049.2021.1920636.

[32] A. Callens, D. Morichon, S. Abadie, M. Delpey, and B. Liquet, "Using Random forest and Gradient boosting trees to improve wave forecast at a specific location," *Applied Ocean Research*, vol. 104, p. 102339, Nov. 2020, doi: 10.1016/j.apor.2020.102339.

[33] J. Hancock and T. M. Khoshgoftaar, "Performance of CatBoost and XGBoost in Medicare Fraud Detection," in *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Miami, FL, USA: IEEE, Dec. 2020, pp. 572–579. doi: 10.1109/ICMLA51294.2020.00095.

[34] R. Bauder and T. Khoshgoftaar, "Medicare Fraud Detection Using Random Forest with Class Imbalanced Big Data," in *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, Salt Lake City, UT: IEEE, Jul. 2018, pp. 80–87. doi: 10.1109/IRI.2018.00019.

[35] D. Chicco, M. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation.," *PeerJ Computer Science*, vol. 7, p. e623, 2021.

[36] A. I. Adler and A. Painsky, "Feature importance in gradient boosting trees with cross-validation feature selection.," *Entropy*, vol. 24, No. 5, p. 687.