

UTILIZING A BLEND OF CLINICAL CONCEPTS TO IDENTIFY FRAUDULENT HEALTH INSURANCE CLAIMS

¹Mr.P.Murthuja,²Gajjala Pavan Kumar Reddy

¹Assistant Professor,²MCA Student

Department of Master of Computer Application,

Rajeev Gandhi memorial College of Engineering and Technology,Nandyal,Andhra Pradesh, 518501,India.

ABSTRACT

For patients to pay for the expensive medical bills, they rely on health insurance offered by either private, public, or both systems. Some healthcare practitioners conduct insurance fraud as a result of their reliance on health insurance. Despite the tiny number of these service providers, it is said that fraud costs insurance companies billions of dollars annually. In this work, we pose the issue of fraud detection over a basic, definite claim data set that consists of operation codes and medical diagnosis. Using a unique representation learning technique, we provide a solution to the fraudulent claim detection issue by converting procedure and diagnostic codes into Mixtures of Clinical Codes (MCC). We also look on ways to extend MCC using Robust Principal Component Analysis and Long Short Term Memory networks. Our test findings show encouraging results in the detection of false records.

I. INTRODUCTION

DATA analytics has progressively become crucial to almost any economic development area. Since healthcare is one of the largest financial sectors in the US economy, the massive amount of data, including health records, clinical data, prescriptions, insurance claims, provider information, and patient information “potentially” presents incredible opportunities for data analysts. Health insurance agencies process billions of claims every year and healthcare expenses is over three trillion dollars in the United States [1]. Figure 1 presents a concise flow of a typical healthcare

reconciliation process by using different entities involved. First, the service provider’s office ensures that the patient has adequate coverage through his/her insurance plan or other funds before getting any service. Next, the service provider identifies relevant diagnoses based on the initial examinations performed on the patient. The service provider then runs tests on the patient using one or more medical interventions such as further diagnostics and surgical procedures. These diagnoses and procedures are usually tagged with the patient’s report along with other information such as personal, demographic, and past/present visit information. At this point, the patient typically pays a copay defined in his/her insurance plan and checks out. Then, the patient’s report is sent to a medical coder who abstracts the information and creates a “superbill” containing all information about the provider, Given the economic volume of the healthcare industry, it is natural to observe fraudulent and fabricated claims submitted to insurance companies. The National Health Care Anti- Fraud Association (NHCAA) defines healthcare fraud as “An intentional deception or misrepresentation made by a person, or an entity, with the knowledge that the deception could result in some unauthorized benefit to him or some other entities” [3]. Those fabricated claims bear a very high cost, albeit they constitute a small fraction. According to NHCAA the fraud related financial loss is in the orders of tens of billions of dollars in the United States [3]. Although there are strict policies regarding

fraud and abuse control in healthcare industries, studies show that a very small portion of the losses are recovered annually [4].

Most typical fraudulent activities committed by dishonest providers in the healthcare domain include the following.

- _ Making false diagnoses to justify procedures that are not medically necessary.
- _ Billing for high priced procedures or services instead of the actual procedures, also called “upcoding”.
- _ Fabricating claims for unperformed procedures.
- _ Performing medically unnecessary procedures to claim insurance payments.
- _ Billing for each step of a procedure as if it is a separate procedure, also called “unbundling”.
- _ Misrepresenting non-covered treatments as medically necessary to receive insurance payments, especially for cosmetic procedures.

It is not feasible or practical to apply only domain knowledge to solve all or a subset of the issues listed above. Automated data analytics can be employed to detect fraudulent claims at an early stage and immensely help domain experts to manage the fraudulent activities much better.

In this paper, we focus on the problem of healthcare fraud detection from health insurance providers’ viewpoint. We answer the question of how to classify a procedure as legitimate or fraudulent from a claim when we only have limited data available, i.e. diagnosis and procedure codes. The problem of fraud detection in medical domain has been identified using different approaches such as data mining [5], classification methods [6], [7], Bayesian analysis [8], statistical surveys [9], non-parametric approaches [10], and expert analysis. Existing methods use physicians profile, background history, claim amount, service quality, services performed per provider, and related metrics from a claim database to create models for claim

status prediction. Although these methods are successful, they often employ datasets that are not publicly available. Furthermore, the variables featured in those datasets are diverse and generally incompatible, which makes the solutions very difficult to transfer. In this study we limit our available data to diagnosis and procedure codes, because obtaining third-party access to richer datasets is often prohibited by Health Insurance Portability and Accountability Act (HIPAA) in the US, General Data Protection Regulation (GDPR) in Europe or similar law in other regions. Besides, the healthcare industry is more apprehensive to share data compared to other sectors. Moreover, different software systems report different patient variables, which prohibits transferring solutions from one system to another. As a result, we confine our problem formulation to diagnosis and procedure codes which can always be handled in the same way whether they are country-specific or international. Our solution approach assumes the claim data as a mixture of medical concepts with respect to clinical codes of diagnoses and procedures in International Classification of Diseases (ICD) coding format. Moreover, the proposed approach works on other coding formats, e.g., Current Procedural Terminology (CPT) and Healthcare Common Procedure Coding System (HCPCS), or their combinations without any modification.

We represent an insurance claim as a Mixture of latent Clinical Concepts (MCC) using probabilistic topic modeling. To the best of our knowledge this is the first work representing insurance claims as mixtures of clinical concepts in a latent space. We assume that every claim is a representation of latent or obvious mixtures of clinical concepts such as pain, mental or infectious diseases. Moreover, each clinical concept is a mixture of clinical codes, i.e., diagnosis and procedure codes. The intuition behind our model comes from the services provided by doctor’s offices, clinics,

and hospitals. In general, a patient gets services based on specific issues consisting of one or more diagnoses. Next, the service provider performs necessary procedures to treat the patient. Therefore, the diagnoses and procedures in a claim can be represented as a mixture of clinical concepts such as pain, mental, infectious diseases and/or their treatments. Note that, we do not explicitly label or interpret these concepts, as they are often not obvious, complex or require domain knowledge.

We extend the MCC model using Long-Short Term Memory networks and Robust Principal Component Analysis. Our goal in extending MCC is to filter the significant concepts from claims and classify them as fraudulent or non-fraudulent. We extend MCC by using the concept weights of a claim as a sequence representation within a Long-Short Term Memory (LSTM) network. This network allows us to represent the claims as sequences of dependent concepts to be classified by the LSTM. Similarly, we apply Robust Principal Component Analysis (RPCA) to filter significant concept weights by decomposing claims into a low-rank and sparse vector representations. The low-rank matrix ideally captures the noise-free weights.

Our unique contributions in this study can be summarized as follows.

- _ We formulate the fraudulent claim detection problem over a minimal, definitive claim data consisting of procedure and diagnosis codes.
- _ We introduce clinical concepts over procedure and diagnosis codes as a new representation learning approach.
- _ We extend the mixtures of clinical concepts using LSTM and RPCA for classification.

We compare our approaches to the Multivariate Outlier Detection (MOD) [11] and a baseline method and report improved performance. Multivariate Outlier Detection method consists of two steps which are used to detect anomalous provider payments within

Medicare claims data. In the first step, a multivariate regression model is built on 13 hand picked features to generate corresponding residuals. Next, the residuals are used as inputs to a generalized univariate probability model. Specifically, they used probabilistic programming methods in Stan [12] to identify possible outliers in the claim data. The authors use the same CMS (Centers for Medicare and Medicaid Services) dataset that we use in our experiments with a different problem formulation. Their study incorporates providers and beneficiary data that was related to Medicare beneficiaries within the state of Florida, while we employ MOD on MCC features. On the other hand, the baseline classifier assigns a test claim as the majority label present in the training claim data.

Our experimental results show that MCC + LSTM reaches an accuracy, precision, and recall scores of 59%, 61%, and 50%, respectively on the inpatient dataset obtained from CMS. In addition, it demonstrates 78%, 83%, and 72% accuracy, precision, and recall scores, respectively on the outpatient dataset. We believe that the proposed problem formulation, representation learning and solution will initiate new research on fraudulent claim detection using minimal, but definitive data.

II. LITERATURE SURVEY

"Identification of Fraudulent Healthcare Claims Using Fuzzy Bipartite Knowledge Graphs"

Md Enamul Haque, Mehmet Engin Tozal,

Health insurance is one of the most important services that people depend on for paying the bills related to hospital and clinical services. This dependency on health insurance lures some healthcare service providers to commit insurance frauds which has become a grave concern. The majority of healthcare fraud is committed by a very small number of untrustworthy providers. Yet, such fraudulent actions damage the reputation of the health service providers and

cost the system billions of dollars. In this article, we specifically focus on the fraudulent claim identification problem and develop different solution schemes to identify the fraudulent cases in healthcare claims with minimal data. We present a solution to the fraudulent claim identification problem that translates diagnoses and procedure code's relations into Bipartite Graphs with Fuzzy Edges (BiGFuzzE). We also investigate the extension of BiGFuzzE using vector representations of clinical codes instead of non-negative matrix factorization (NMF). Our experimental evaluations demonstrate significant outcomes.

"Overview of Digital Finance Anti-fraud",

Cheng Wang,

The development of digital financial technology and its penetration into the traditional financial industry have become an irreversible trend. At present, the applications of digital financial technology have significantly reduced the information asymmetry in the financial field and made great contributions to improving the financial market. However, everything has two sides, especially new things. Digital financial technology is on the ascendant, but new fraud means based on it are growing and financial fraud risks are escalating. Faced with this situation, the importance of financial supervision and risk prevention has been raised to an unprecedented height under the macro background of holding the bottom line of no systematic financial risk. Due to the marriage between digitization and finance, financial fraud has taken on new features such as specialization, industrialization, concealment and cross region, which poses great challenges to traditional anti fraud methods. Therefore, the anti fraud technology should also be constantly innovated. It is not only necessary to accurately combat the existing risks, but also to take the lead to prevent problems before they occur. The behavior-based method is recognized as an effective paradigm for anti-fraud in digital finance. It can be used

cooperatively as a second security line, rather than replaced with other types of existing anti-fraud methods. The behavior-based method is a highly-anticipated solution to pursue a non-intrusive and continuous authentication for online services. The efficacy of behavior-based methods significantly depends on the sufficiency and quality of behavior data. In this book, we propose anti-fraud engineering based on the behavioral modeling paradigm, which focuses on behavior associations to enhance behavior data.

"Design and development of big data-based model for detecting fraud in healthcare insurance industry"

A. Jenita Mary, S. P. Angelin Claret,

The advancement in healthcare services has been increasing widely to extend several services with intense quality. One of the important issues affecting the effective use of public funds is the detection of health insurance fraud. Previous techniques of detecting fraud pay close attention to characteristics of a single visit rather than many patient visits. Due to a higher false positive rate and poor profile construction, the common traits have reduced detection performance. This paper introduces a novel and intelligent Provider Fraud_Anomaly Detection System (PF_ADS) by combining big data and deep learning approaches for the healthcare insurance industry. The proposed framework contributes to improvising the preprocessing and classification phases to detect provider fraud at an untimely phase. Initially, the collected datasets are preprocessed using a Relative Risk-based MapReduce framework that builds an organized set of relationships between diseases, patients, and claiming variables. The classification phase is improvised using a proposed Recurrent Neural Network (RNN). It consists of sophisticated steps to consider the significant attributes using hyperparameter optimization. Recalling ability is one of the best parts of RNNs that defines the past and present

states of the networks. Therefore, the ability of network state predictions and the tuning of parameters is studied by improved Decisional Score-based Bayesian Optimization (DS_BO). Finally, the best attributes with the selective hyperparameters are fed into the input layer of the Recurrent Neural Networks (RNNs) to classify the anomalies from the provider's end. The proposed PF_ADS framework is experimented with and validated on the public repositories. The experimental results state that the proposed framework outperforms better than the other methods in terms of accuracy (88.09%), precision (14.15%), recall (32.80%), and 92.30 s computational time.

**"Comprehensive Review of Multimodal Medical Data Analysis: Open Issues and Future Research Directions",
Shashank Shetty, Ananthanarayana V S, Ajit Mahale,**

Over the past few decades, the enormous expansion of medical data has led to searching for ways of data analysis in smart healthcare systems. Acquisition of data from pictures, archives, communication systems, electronic health records, online documents, radiology reports and clinical records of different styles with specific numerical information has given rise to the concept of multimodality and the need for machine learning and deep learning techniques in the analysis of the healthcare system. Medical data play a vital role in medical education and diagnosis; determining dependency between distinct modalities is essential. This paper gives a gist of current radiology medical data analysis techniques and their various approaches and frameworks for representation and classification. A brief outline of the existing medical multimodal data processing work is presented. The main objective of this study is to spot gaps in the surveyed area and list future tasks and challenges in radiology. The Preferred Reporting Items for Systematic Reviews and Meta-

Analysis (or PRISMA) guidelines were incorporated in this study for effective article search and to investigate several relevant scientific publications. The systematic review was carried out on multimodal medical data analysis and highlighted advantages, limitations and strategies. The inherent benefit of multimodality in the medical domain powered with artificial intelligence has a significant impact on the performance of the disease diagnosis frameworks.

**"Predictive and adaptive Drift Analysis on Decomposed Healthcare Claims using ART based Topological Clustering",
Lavanya Settipalli, G.R. Gangadharan, Ugo Fiore,**

Fraud in healthcare services dissipates funds that are important for improving the quality of life of people, thus enhancing the interest in predictive fraud analysis. The predictive analysis of fraudulent activity can be done by looking for unusual patterns in healthcare claims. However, unusual patterns may also occur due to sudden changes, isolated events, or concept drifts that frequently happen in healthcare which should not be considered fraud. Furthermore, analyzing drifts also supports predicting future trends and behaviors. In this study, we propose a novel approach, Drift Analysis on Decomposed Healthcare Claims (DADHC), to analyze the hidden patterns that hinder the performance of fraud prediction and detection. Our proposed model decomposes the series of healthcare claims into regular and irregular patterns using Pseudo Additive Decomposition (PAD) integrated with Simple Moving Average (SMA) smoothing technique. Then ART (Adaptive Resonance Theory) based Topological Clustering (TC) is used to analyze unusual patterns and identify the actual fraudulent activities. Our proposed model also incorporates correntropy based vigilance testing in ART to enhance adaptivity. Empirical evaluation on CMS Part B claims shows that our proposed

approach has significantly improved detection accuracy compared to existing models due to the drift analysis.

III. SYSTEM ANALYSIS EXISTING SYSTEM

Yang and Hwang developed a fraud detection model using the clinical pathways concept and process-mining framework that can detect frauds in the healthcare domain [13]. The method uses a module that works by discovering structural patterns from input positive and negative clinical instances. The most frequent patterns are extracted from every clinical instance using the module. Next, a feature-selection module is used to create a filtered dataset with labeled features. Finally, an inductive model is built on the feature set for evaluating new claims. Their method uses clustering, association analysis, and principal component analysis. The technique was applied on a real-world data set collected from National Health Insurance (NHI) program in Taiwan. Although the authors constructed different features to generate patterns for both normal and abusive claims, the significance of those features is not discussed.

Bayerstadler et al. [14] presented a predictive model to detect fraud and abuse using manually labeled claims as training data. The method is designed to predict the fraud and abuse score using a probability distribution for new claim invoices. Specifically, the authors proposed a Bayesian network to summarize medical claims' representation patterns using latent variables. In the prediction step, a multinomial variable modeling predicts the probability scores for various fraud events. Additionally, they estimated the model parameters using Markov Chain Monte Carlo (MCMC) [15].

Zhang et al. [16] proposed a Medicare fraud detection framework using the concept of anomaly detection [17]. First part of the proposed method consists of a spatial density based algorithm which is claimed to be more suitable compared to local outlier factors in

medical insurance data. The second part of the method uses regression analysis to identify the linear dependencies among different variables. Additionally, the authors mentioned that the method has limited application on new incoming data.

Kose et al. [18] used interactive unsupervised machine learning where expert knowledge is used as an input to the system to identify fraud and abuse related legal cases in healthcare. The authors used a pairwise comparison method of analytic hierarchical process (AHP) to incorporate weights between actors (patients) and attributes. Expectation maximization (EM) is used to cluster similar actors. They had domain experts involved at different levels of the study and produced storyboard based abnormal behavior traits. The proposed framework is evaluated based on the behavior traits found using the storyboard and later used for prescriptions by including all related persons and commodities such as drugs.

Bauder and Khoshgoftaar [19] proposed a general outlier detection model using Bayesian inference to screen healthcare claims. They used Stan model which is similar to [20] in their experiments. Note that, they consider only provider level-fraud detection without considering clinical code based relations. Many of those methods use private datasets or different datasets with incompatible feature lists. Therefore, it is very difficult to directly compare these studies. In addition, HIPAA, GDPR and similar law enforce serious penalties for violations of the privacy and security of healthcare information, which make healthcare providers and insurance companies very reluctant to share rich datasets if not at all. For these reasons, we formulate the problem over a minimal, definitive claim data consisting of diagnosis and procedure codes. Under this setting we tackle the problem of flagging a procedure as legitimate or fraudulent using

mixtures of clinical codes along with RNN and RPCA based encodings.

DISADVANTAGES

- Making false diagnoses to justify procedures that are not medically necessary.
- Fabricating claims for unperformed procedures.
- Performing medically unnecessary procedures to claim insurance payments.
- Billing for each step of a procedure as if it is a separate procedure, also called “unbundling”.
- Misrepresenting non-covered treatments as medically necessary to receive insurance payments, especially for cosmetic procedures.

PROPOSED SYSTEM

We extend the MCC model using Long-Short Term Memory networks and Robust Principal Component Analysis. Our goal in extending MCC is to filter the significant concepts from claims and classify them as fraudulent or non-fraudulent. We extend MCC by using the concept weights of a claim as a sequence representation within a Long-Short Term Memory (LSTM) network. This network allows us to represent the claims as sequences of dependent concepts to be classified by the LSTM. Similarly, we apply Robust Principal Component Analysis (RPCA) to filter significant concept weights by decomposing claims into a low-rank and sparse vector representations. The low-rank matrix ideally captures the noise-free weights.

Our unique contributions in this study can be summarized as follows.

The system formulates the fraudulent claim detection problem over a minimal, definitive claim data consisting of procedure and diagnosis codes.

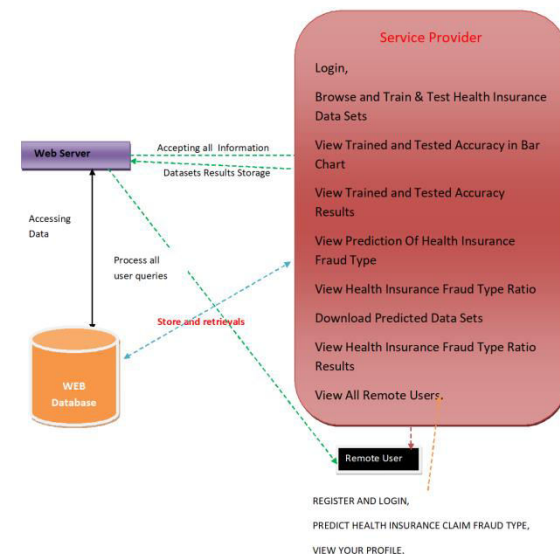
The system introduces clinical concepts over procedure and diagnosis codes as a new representation learning approach.

The system extends the mixtures of clinical concepts using LSTM and RPCA for classification.

Advantages

- The proposed system uses Support Vector Machine (SVM) for classification with MCC.
- Multivariate Outlier Detection method is an effective method which is used to detect anomalous provider payments within Medicare claims data.

IV. SYSTEM DESIGN SYSTEM ARCHITECTURE



V. ALGORITHMS DECISION TREE CLASSIFIERS

Decision tree classifiers are used successfully in many diverse areas. Their most important feature is the capability of capturing descriptive decision making knowledge from the supplied data. Decision tree can be generated from training sets. The procedure for such generation based on the set of objects (S), each belonging to one of the classes C_1, C_2, \dots, C_k is as follows:

Step 1. If all the objects in S belong to the same class, for example C_i , the decision tree for S consists of a leaf labeled with this class

Step 2. Otherwise, let T be some test with possible outcomes O_1, O_2, \dots, O_n . Each object in S has one outcome for T so the test partitions

S into subsets S_1, S_2, \dots, S_n where each object in S_i has outcome O_i for T . T becomes the root of the decision tree and for each outcome O_i we build a subsidiary decision tree by invoking the same procedure recursively on the set S_i .

Gradient boosting

Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees.^{[1][2]} When a decision tree is the weak learner, the resulting algorithm is called gradient-boosted trees; it usually outperforms random forest. A gradient-boosted trees model is built in a stage-wise fashion as in other boosting methods, but it generalizes the other methods by allowing optimization of an arbitrary differentiable loss function.

K-Nearest Neighbors (KNN)

- Simple, but a very powerful classification algorithm
- Classifies based on a similarity measure
- Non-parametric
- Lazy learning
- Does not “learn” until the test example is given
- Whenever we have a new data to classify, we find its K-nearest neighbors from the training data

Example

- Training dataset consists of k-closest examples in feature space
- Feature space means, space with categorization variables (non-metric variables)
- Learning based on instances, and thus also works lazily because instance close to the input vector for test or prediction may take time to occur in the training dataset

Logistic regression Classifiers

Logistic regression analysis studies the association between a categorical dependent

variable and a set of independent (explanatory) variables. The name *logistic regression* is used when the dependent variable has only two values, such as 0 and 1 or Yes and No. The name *multinomial logistic regression* is usually reserved for the case when the dependent variable has three or more unique values, such as Married, Single, Divorced, or Widowed. Although the type of data used for the dependent variable is different from that of multiple regression, the practical use of the procedure is similar.

Logistic regression competes with discriminant analysis as a method for analyzing categorical-response variables. Many statisticians feel that logistic regression is more versatile and better suited for modeling most situations than is discriminant analysis. This is because logistic regression does not assume that the independent variables are normally distributed, as discriminant analysis does.

This program computes binary logistic regression and multinomial logistic regression on both numeric and categorical independent variables. It reports on the regression equation as well as the goodness of fit, odds ratios, confidence limits, likelihood, and deviance. It performs a comprehensive residual analysis including diagnostic residual reports and plots. It can perform an independent variable subset selection search, looking for the best regression model with the fewest independent variables. It provides confidence intervals on predicted values and provides ROC curves to help determine the best cutoff point for classification. It allows you to validate your results by automatically classifying rows that are not used during the analysis.

Naïve Bayes

The naive bayes approach is a supervised learning method which is based on a simplistic hypothesis: it assumes that the presence (or absence) of a particular feature of a class is

unrelated to the presence (or absence) of any other feature .

Yet, despite this, it appears robust and efficient. Its performance is comparable to other supervised learning techniques. Various reasons have been advanced in the literature. In this tutorial, we highlight an explanation based on the representation bias. The naive bayes classifier is a linear classifier, as well as linear discriminant analysis, logistic regression or linear SVM (support vector machine). The difference lies on the method of estimating the parameters of the classifier (the learning bias).

While the Naive Bayes classifier is widely used in the research world, it is not widespread among practitioners which want to obtain usable results. On the one hand, the researchers found especially it is very easy to program and implement it, its parameters are easy to estimate, learning is very fast even on very large databases, its accuracy is reasonably good in comparison to the other approaches. On the other hand, the final users do not obtain a model easy to interpret and deploy, they does not understand the interest of such a technique.

Thus, we introduce in a new presentation of the results of the learning process. The classifier is easier to understand, and its deployment is also made easier. In the first part of this tutorial, we present some theoretical aspects of the naive bayes classifier. Then, we implement the approach on a dataset with Tanagra. We compare the obtained results (the parameters of the model) to those obtained with other linear approaches such as the logistic regression, the linear discriminant analysis and the linear SVM. We note that the results are highly consistent. This largely explains the good performance of the method in comparison to others. In the second part, we use various tools on the same dataset (Weka 3.6.0, R 2.9.2, Knime 2.1.1, Orange 2.0b and RapidMiner 4.6.0). We try above all to understand the obtained results.

Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.

The first algorithm for random decision forests was created in 1995 by Tin Kam Ho[1] using the random subspace method, which, in Ho's formulation, is a way to implement the "stochastic discrimination" approach to classification proposed by Eugene Kleinberg.

An extension of the algorithm was developed by Leo Breiman and Adele Cutler, who registered "Random Forests" as a trademark in 2006 (as of 2019, owned by Minitab, Inc.).The extension combines Breiman's "bagging" idea and random selection of features, introduced first by Ho[1] and later independently by Amit and Geman[13] in order to construct a collection of decision trees with controlled variance.

Random forests are frequently used as "blackbox" models in businesses, as they generate reasonable predictions across a wide range of data while requiring little configuration.

SVM

In classification tasks a discriminant machine learning technique aims at finding, based on an *independent and identically distributed (iid)* training dataset, a discriminant function that can correctly predict labels for newly acquired instances. Unlike generative machine learning approaches, which require computations of conditional probability distributions, a discriminant classification function takes a data

point x and assigns it to one of the different classes that are a part of the classification task. Less powerful than generative approaches, which are mostly used when prediction involves outlier detection, discriminant approaches require fewer computational resources and less training data, especially for a multidimensional feature space and when only posterior probabilities are needed. From a geometric perspective, learning a classifier is equivalent to finding the equation for a multidimensional surface that best separates the different classes in the feature space.

SVM is a discriminant technique, and, because it solves the convex optimization problem analytically, it always returns the same optimal hyperplane parameter—in contrast to *genetic algorithms (GAs)* or *perceptrons*, both of which are widely used for classification in machine learning. For perceptrons, solutions are highly dependent on the initialization and termination criteria. For a specific kernel that transforms the data from the input space to the feature space, training returns uniquely defined SVM model parameters for a given training set, whereas the perceptron and GA classifier models are different each time training is initialized. The aim of GAs and perceptrons is only to minimize error during training, which will translate into several hyperplanes' meeting this requirement.

VI. SYSTEM IMPLEMENTATIONS

Modules

Service Provider

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as

Login, Browse and Train & Test Health Insurance Data Sets, View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results, View Prediction Of Health Insurance Fraud Type, View Health Insurance Fraud Type Ratio, Download Predicted Data Sets, View Health Insurance

Fraud Type Ratio Results, View All Remote Users

View and Authorize Users

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

Remote User

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like REGISTER AND LOGIN, PREDICT HEALTH INSURANCE CLAIM FRAUD TYPE, VIEW YOUR PROFILE.

VII. SCREENSHOTS





VIII. CONCLUSION

The Issue of identifying fraudulent insurance claims as a feature generation and classification process is presented in this work. As a result of legal restrictions and software system discrepancies, we construct the issue over a minimum, definite claim data set that consists of procedure and diagnostic codes. As a novel representation learning technique, we offer clinical ideas over procedure and diagnostic codes. Every claim is interpreted as a latent or evident blend of clinical concepts, which are themselves combinations of procedure and diagnostic codes. We add to the MCC model by filtering the important ideas from claims and categorizing them as fraudulent or non-fraudulent using the Long-Short Term Memory

Network (MCC + LSTM) and Robust Principal Component Analysis (MCC + RPCA). Our findings show that there is room for improvement in the identification of fraudulent healthcare claims using limited data. In the negative claim generation process, both MCC and MCC + RPCA behave consistently for different concept sizes and replacement probabilities. Using the inpatient dataset, MCC + LSTM achieves accuracy, precision, and recall scores of 59%, 61%, and 50%, respectively. Additionally, on the outpatient dataset, it displays accuracy, precision, and recall scores of 78%, 83%, and 72%, respectively. The findings of MCC and MCC + RPCA exhibit similarities as they both make use of an SVM classifier. We think that the design of the suggested issue, representation learning, and solution will start fresh research on the use of minimum but conclusive data to identify fraudulent insurance claims.

REFERENCES

1. National Health Care Anti-Fraud Association, "The challenge of health care fraud," <https://www.nhcaa.org/resources/health-care-antifraud-resources/the-challenge-of-health-care-fraud.aspx>, 2020, accessed January, 2020.
2. Font Awesome, "Image generated by free icons," <https://fontawesome.com/license/free>, 2020, online.
3. National Health Care Anti-Fraud Association, "Consumer info and action," <https://www.nhcaa.org/resources/health-care-anti-fraudresources/consumer-info-action.aspx>, 2020, accessed January, 2020.
4. W. J. Rudman, J. S. Eberhardt, W. Pierce, and S. Hart-Hester, "Healthcare fraud and abuse," *Perspectives in Health Information Management/AHIMA*,

- American Health Information Management Association, vol. 6, no. Fall, 2009.
5. M. Kirlidog and C. Asuk, "A fraud detection approach with data mining in health insurance," *Procedia-Social and Behavioral Sciences*, vol. 62, pp. 989–994, 2012.
 6. V. Rawte and G. Anuradha, "Fraud detection in health insurance using data mining techniques," in *2015 International Conference on Communication, Information & Computing Technology (ICCICT)*. IEEE, 2015, pp. 1–5.
 7. C. Phua, D. Alahakoon, and V. Lee, "Minority report in fraud detection: classification of skewed data," *Acm sigkdd explorations newsletter*, vol. 6, no. 1, pp. 50–59, 2004.
 8. T. Ekina, F. Leva, F. Ruggeri, and R. Soyer, "Application of Bayesian methods in detection of healthcare fraud," *chemical engineering Transaction*, vol. 33, 2013.
 9. J. Li, K.-Y. Huang, J. Jin, and J. Shi, "A survey on statistical methods for health care fraud detection," *Health care management science*, vol. 11, no. 3, pp. 275–287, 2008.
 10. R. J. Freese, A. P. Jost, B. K. Schulte, W. A. Klindworth, and S. T. Parente, "Healthcare claims fraud, waste and abuse detection system using non-parametric statistics and probability based scores," Jan. 19 2017, uS Patent App. 15/216,133.
 11. R. A. Bauder and T. M. Khoshgoftaar, "Multivariate anomaly detection in medicare using model residuals and probabilistic programming," in *The Thirtieth International Flairs Conference*, 2017.
 12. B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell, "Stan: A probabilistic programming language," *Journal of statistical software*, vol. 76, no. 1, 2017.
 13. W.-S. Yang and S.-Y. Hwang, "A process-mining framework for the detection of healthcare fraud and abuse," *Expert Systems with Applications*, vol. 31, no. 1, pp. 56–68, 2006.
 14. A. Bayerstadler, L. van Dijk, and F. Winter, "Bayesian multinomial latent variable modeling for fraud and abuse detection in health insurance," *Insurance: Mathematics and Economics*, vol. 71, pp. 244–252, 2016.
 15. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, "Introducing markov chain monte carlo," *Markov chain Monte Carlo in practice*, vol. 1, p. 19, 1996.
 16. W. Zhang and X. He, "An anomaly detection method for medicare fraud detection," in *Big Knowledge (ICBK), 2017 IEEE International Conference on*. IEEE, 2017, pp. 309–314.
 17. L. Zhang, J. Lin, and R. Karim, "Adaptive kernel density-based anomaly detection for nonlinear systems," *Knowledge-Based Systems*, vol. 139, pp. 50–63, 2018.
 18. I. Kose, M. Gokturk, and K. Kilic, "An interactive machine-learning based electronic fraud and abuse detection system in healthcare insurance," *Applied Soft Computing*, vol. 36, pp. 283–299, 2015.
 19. R. A. Bauder and T. M. Khoshgoftaar, "A probabilistic programming approach for outlier detection in healthcare claims," in *Machine Learning and Applications (ICMLA), 2016 15th IEEE*

- International Conference on. IEEE, 2016, pp. 347–354.
20. J. Wang and S. Luo, “Augmented beta rectangular regression models: A bayesian perspective,” *Biometrical Journal*, vol. 58, no. 1, pp. 206–221, 2016.
 21. Centers for Medicare and Medicaid Services, “ICD-10,” <https://www.cms.gov/Medicare/Coding/ICD10/>, 2020, accessed January, 2020.
 22. Medical Billing and Coding, “HCPCS codes,” <https://www.medicalbillingandcoding.org/hcpcs-codes/>, 2020, accessed January, 2020.
 23. American Academy of Professional Coders, “CPT codes,” <https://coder.aapc.com/cpt-codes>, 2020, accessed January, 2020.
 24. S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
 25. D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.