HEMClust: An Improved Fraud Detection Model for Health Insurance using Heterogeneous Ensemble and K-prototype Clustering

Shamitha S Kotekani[1], V Ilango[2]
Research Scholar (VTU Belgavi), CMR Institute of Technology, Bangalore[1]
Professor, CMR Institute of Technology, Bangalore[2]

***Abstract*—Health insurance plays an integral part of society's economic well-being; the existence of fraud creates innumerable challenges in providing affordable health care support for the people. In order to reduce the losses incurred due to fraud, there is a need for a powerful model to predict fraud on the data accurately. The purpose of the paper is to implement a more sophisticated technique for fraud detection using machine learning: HEM*Clust* (Heterogeneous Ensemble Model with Clustering). The first phase of the model aims in improving the quality of claims data by providing effective preprocessing. The second stage addresses the overlapping instances in provider specialties by grouping them using k-prototype clustering. The final stage includes building the model using a heterogeneous stacking ensemble that performs classification on multiple levels, with four base learners in level 0 and a meta learner in level 1. The results were assessed using evaluation metrics and statistical tests such as Friedman and Nememyi to compare the performance of base classifiers against the proposed HEM*Clust*. The empirical results show that the HEM*Clust* produced 94% and 96% overall precision-recall rates on the dataset, which was an increase of 45% to 50% in the fraud detection rate for each class in the data.**

***Keywords*—Fraud detection; health insurance; ensemble learners; meta-level learning; clustering; classification algorithms**

I. INTRODUCTION

Health insurance has become a rapidly growing industry that plays a vital role in ensuring country's economic well-being. It provides us with much-needed cover during a financial crisis; it has benefitted many by reducing their healthcare expenditure burden, which otherwise jeopardizes their financial stability. The services provided by insurance industry can broadly be divided into two parts: life insurance and non-life insurance. This study considers life insurance, particularly health insurance. Many researchers have primarily administered claims data to be used extensively for healthcare data analytics[1]–[3]. The claims data include information related to medical examinations, diagnosis, drug-related information, doctor prescriptions along with medical diagnosis, it also contains financial data such as reimbursement amount, billing information etc. [4]. Claims are usually submitted from the patient's end or provider's end. A claim is processed when a policyholder submits a demand covering a particular treatment. Claims submitted to the facility will be validated further, and the request will be approved and reimbursed either to the practitioner (doctor / hospital) or the patient directly.

According to the study conducted by III (Insurance Information Institute); the overall net income (in billions) of the insurance industry over the last three years (2017-2019) was approximately 36.1, 59.6 and 61.4 billion [5]. This shows the growth of demand and dependence of people on the insurance sector. The rising demand for health cards has also increased the risk of fraudulent transactions. Health insurance fraud can be defined as intentional deception in which an insurance or medical provider provides false, misleading information to an insurer to obtain improper benefits from the policyholder's policy[6][7]. The studies state that around 10 per cent of healthcare expenditure is wasted on fraudulent transactions [8]. The fraudsters use several techniques to perpetrate fraud, such as altering the bills, forgery of documents or using powerful technologies for illegitimately collecting money from the consumers and the health providers. The main offenders of the health insurance sector can be broadly categorized into two providers and consumers [7]. Among the offenders, the study will concentrate on detecting providers fraud. The fraudulent activities involved by the providers include the following types of fraud:

- Phantom billing – This is a way of fabricating claims, it basically includes applying charges for treatment that has never been performed.
- Upcoding – This includes charging higher billed services when the patient might have received basic or recommending unnecessary procedures or test which is not required.
- Unbundling - This includes dividing a single procedure into many and providing multiple bills for each.
- Kickback fraud – This is a kind of bribery given to the provider for an improper service, for example, an inflated bill will be presented for reimbursement and the party of the difference amount will be paid to the provider as a reward.

There are two different ways to combat fraud, one way is detecting fraud (Fraud Detection) and the other one is to prevent fraud (Fraud Prevention) [9]. Fraud prevention involves stopping the fraud before it occurs by setting new rules or protocols. In health insurance, fraud prevention could be achieved in various ways, like denying policies for people by checking the risk possibilities or excluding providers from the authorized list of providers having malicious records. On

the other hand, fraud detection is applied when all the rules for preventing fraud fail, and a fraud transaction has already been committed. When any fraudulent transaction is detected, the aim will be to reverse it. To identify fraud for every incoming data, a fraud detection system will check every transaction to find any possibility of fraud. This will help the organization monitor or identify the fraudulent transactions quickly despite any change in the strategies adopted by the fraudsters.

As explained, claims data is a major source to retrieve information related to healthcare utilization and expenses, making it an appropriate database for our study in detecting fraud on health insurance. Though claims data have proved to be an attractive source of data for research, few challenges are associated while analyzing it like [10]–[12]:

1) Billers or coders with a lack of knowledge in medical terminologies tend to misinterpret the terminologies and end up entering incorrect information. Also, the way a particular data is fed into the system will vary from place to place.

2) Overlapping of codes, providers of different specialties will be referring to single code. This later causes code variability and sparsity in the data.

The above challenges demand that claims database needs efficient data preprocessing and a proper technique for dealing with misrepresented procedures or specialties.

There was a noted evolution seen in applications of analytical approaches in claims data, from simple record-based calculation to the use of machine learning (ML) techniques. Traditional fraud investigation on the data was time-consuming and also costly. Manually investigating fraud is not advisable in this era as fraudsters keep changing their way of committing fraud. Machine learning algorithms apply artificial intelligence techniques to help the fraud detection system learn from experience and improve its ability to see any fraud patterns [13]. Using ML techniques for fraud detection helps in executing entire data in a shorter period of time. All kinds of fraud could be detected more accurately, also with a slight variation applied in the analysis, the system could be used to anticipate new patterns of fraud. Since the insurance sector accumulates a large amount of data in the form of claims, the use of big data analytics will help in revealing complex claim patterns since more data leads to improving the predictive power of the model [14][15][16].

The study proposes a novel heterogenous ensemble model with clustering to overcome the above challenges. The fundamental idea was to incorporate k-prototype clustering and an efficient preprocessing procedure to detect fraud from health insurance claims data. Finally, preprocessed and clustered subsets are obtained for training the base classifiers. The major contributions of the study are listed as below:

- To improve the data quality by applying effective preprocessing techniques.
- To group similar providers based on their specialties to reduce the overlapping procedures while detecting providers fraud.
- To provide a novel combination of heterogeneous ensembles through stacking framework for detecting fraud in health insurance using preprocessed and clustered data.

Paper uses distributed computing for handling the volume of data. Using the distributed environment, complex big data analysis can be performed in a second of time without any computational overheads. Traditionally, when the size of the data increase only solution lies was to upgrading or scaling up the machine, which is expensive as it doubles the cost. Instead, modern approaches have started focusing on scaling out. Scaling out increases the computational power by adding more machines to the network. Spark is one such distributed opensource framework where, big data applications could be easily distributed by its data structure called RDD (Resilient distributed dataset)[17], [18],[19]. Since spark is built on top of the Hadoop framework, it can perform the computations faster using in-memory primitives[19]. Distributed computing and scalability using spark is achieved using the following information in RDD:

- Data are partitioned into several sets; each partition contains an atomic piece of the database.
- The location of each partition will be included for providing faster access.
- Total number of dependencies are on the parent RDD.

The entire proceedings of the work are distributed as follows; Section 2 performs a detailed background review of the works and explains the gaps found in the research; Section 3 explains the proposed HEM*Clust* model. Section 4 details the dataset and experimental setup used for implementation. Section 5 discusses the results obtained from the model finally, Section 6 concludes the work.

II. Literature Review

Many researchers used several complex learning algorithms in fraud detection, such as deep learning and ensemble learning, mainly because of their capability to learn complex relationships between the patterns. X. Zhou et al [20] developed a fraud detection model for online banking based on convolutional neural networks. The network consisted of six layers, including a feature sequencing layer, four convolutional layers, and a pooling layer. The model was used to verify and detect fraud on the online transactions that are performed in the bank. The model produced a good precision and recall rate.

Other than conventional classification algorithms, bagging and voting ensembles have been used as a state of the art techniques for fraud detection in several articles[9], [21], [22]. Most of the articles focused on bagging ensembles, which takes bootstrap samples, and training was concentrated on each chosen sample [23]. M. Zareapoor and P. Shamsolmoali [24] applied bagging classifier for detecting fraudulent transactions using credit card. The author combined three different learners such as Naïve Bayes (NB), kNearest Neighbours (knn), and a Bagging ensemble with a 10-cross validation for building a model.

David W.Fan et al. [25] had compared Stacked Generalization with other combiners to analyze using multiple algorithms for prediction. The results proved that stacked generalization had given impressive results than other base classifiers. Kerwin et al. [26] used stacking to deal with imbalanced class distribution for detecting fraud from the dataset. The author used different classification techniques and sampling techniques as a base learner and meta learner to improve the performance. Meta learner with Gradient Boosting Ensemble classifier produced a more excellent f1 score. It was observed that multiple algorithms have always been proved efficient in fraud detection from all these works.

The studies discussed above reveals that there is no comprehensive work related to health insurance fraud detection, mainly because of the lack of data availability. As far as our knowledge, CMS Medicare data [27] is the only available open-source data. CMS database consists of details regarding procedures and drug descriptions. There were few studies based on CMS Medicare data conducted by Mathew Herald et al. [28]–[31] using multiple data sets from CMS Medicare. Herald et al. [28] constructed a fraud detection model for big data by combining four datasets from CMS, resulting in 37,147,213 records. They applied neural networks and tree based ensembles such as random forest and gradient descent trees. The results were validated using cross-validation during learning, and the results showed that MLP learners outperformed GBT and RF with a ROC Score of 0.816. They further expanded their work by applying the deep learning model on big data sets with 4,692,370 instances, which improved the model's performance further.

The author also addressed the problem of imbalanced data learning in fraud detection. Data level sampling and algorithmic level techniques were applied on a given range of class ratios. The results showed that deep learning with oversampling and an ensemble of over and under sampling outperformed the baseline algorithmic models with an AUC score of 0.8505 and 0.8509, respectively. A similar dataset was used by L. K. Branting and F. Reeder [32] to calculate the fraud risk for 2012-2014. The author proposed a graph-based model for calculating the risk that appears on the dataset after combining the Part B and LEIE data. The whole aggregation was based on the NPI's, since the exclusions database contained missing entities, the author used the NPPES registry, which maintains the list of providers under Medicare. V. Chandola et al. [33] used Medicare claims data and LEIE exclusions database to find the hidden anomalies inside. The techniques used were social network analysis, spatial-temporal analysis and text mining. Later, weighted MLP was used to classify bad actors and produced an accuracy score from 71% to 81.4%.

In their paper, Mathew Herland et al. [29] concentrated on detecting upcoding fraud by finding providers who had procedural code other than one. It was also found that grouping the providers practicing on similar area had produced an improved prediction result.

A. Research Gap

Considering the above review, it was observed that Cart[34], RF [35], MLP [36] had been widely used in detecting fraud in health insurance. It was observed that RF and Cart were good in classifying normal transactions, and MLP performed well in classifying fraudulent transactions. Studies conducted by M.Paz Sesmero et al. [37] Saurabh Tewari[38] proved that the hypothesis generated from varied classifiers on a space using stacking or voting would boost the overall predictions reduce the bias or variance than using homogeneous classifiers. Though several works reveal the dominance of ensemble learners over single learners for fraud detection on various domains[21], [24], [26], in health insurance, its implementation is minuscule. Overlapping of procedures between the specialties was also a major issue discussed in the literature, and the authors have grouped the classes manually considering the similarities [30], [39]. Since claims data contains hundreds of provider specialties with thousands of procedural codes, the current manual grouping to reduce the overlapping could not be considered as a feasible solution.

III. Heterogenous Ensemble Classifier with Clustering (HEMClust) : Proposed Technique

As emphasized, HEM*Clust* incorporate stacking ensemble with an extension to the existing work of M. Herland et al. [30] by applying clustering to similar group providers based on their specialties. Data quality was also a significant concern during processing as the database contained lot of missing values. To overcome that, Feature-Wise Imputation (FWI) is applied. Using FWI, missing values are imputed using mean/mode/knn on each attribute by looking the severity and type of data. The proposed HEM*Clust* works in three phases:

A. Phase 1: Data Pre-processing

As emphasized in Section 1, claims data contains many ambiguities caused by automated data entries, data redundancies, missing values and incorrect entries [12]. Enhancing data quality is inevitable as only perfect data could lead to a better model. The choice of methods was entirely dependent on the nature of our data. Following are the steps carried out to improve the data quality:

- To Identify and remove single-valued predictors as those attributes will not give any information for modelling.
- Cleaning incorrect data entry errors through fuzzy matching. Fuzzy matching finds the text that is very similar to the search given. It also lists the matches along with the matching ratio.
- Feature-wise imputation of missing values.
- Data normalization using a min-max scaler.

B. Phase 2 : Clustering Provider Specialties

The second phase of the model aims in reducing the overlapping instances and model variance. Here, clustering techniques are used to group provider specialties instead of manual grouping. Clustering finds groups in the data that are similar to each other. It divides the data into similar groups, such that the distance between two instances is identical if they belong to one cluster and far if they are from different clusters.

Before applying the clustering algorithm, the clustering tendency was measured using Hopkins's test.

The hypothesis generated from the test was used to find whether the data inherently contains any clusters. The statistic's null hypothesis(h0) will state that the data has no meaningful clusters and is distributed uniformly. If the value of the results (H) is greater than 0.5 h_0 will be rejected, and an alternate hypothesis (data is not uniformly distributed and it contains meaningful clusters) is accepted. Later, K-prototype (kproto) clustering will be applied for creating the groupings since it can efficiently handle large and heterogeneous data types[47], [48]. kproto clustering defines prototypes as centroids which is built from mean values of numerical and mode of categorical variables[49]. The whole procedure works similar to K-means clustering. It iteratively relocates the data based on partitioning to minimize the distance between a cluster and its prototype (similar to the centroid in K-means). Here, the distance between two points A and B is defined as[48].

$$f_n(A,B) = \sum_{j=1}^{r}(a_j - b_j)^2) + \gamma l \sum_{j=r+1}^{s}\delta(a_j, b_j) \quad (1)$$

Where r is the Euclidean distance applicable to numerical data, followed by hamming distance for dealing with categorical variable s. The variables γl *and* δ are user-defined values, which will be used avoid the influence of numerical and categorical variables when applied to the model.

C. Phase 3 : Heterogenous Ensemble Framework based on Stacking

Heterogenous ensembles possess the capability to generate varied results in a single space using different base classifiers. Individual classifiers used here will solve both binary class and multi-class classification problems. Following criteria was considered for the construction of base classifiers,

- Algorithms should be scalable for both large and small data sets.
- Algorithms should be able to provide quick predictions after training.

Multi-Layer Perceptron (MLP), Logistic Regression (LR), Cart and Random Forest (RF) were considered as the base pool of classifiers as it satisfies the above said criteria. Optimal set of parameters for all the base model was found by applying grid search optimization. Table I lists the parameters adopted throughout the study. A detailed explanation of these algorithms is out of the scope of this paper. Its explanation and implementation could be referred from the following articles [41], [23], [42], [43]. There are basically two types of ensembles Stacking ensemble and Voting ensemble. Though our model will be using stacking as the base classifier, it was evident to give a brief on voting ensemble. The voting ensemble combines predictions from different learners intending to attain the highest possible prediction accuracy. It uses majority voting or average voting techniques to combine the predictions generated from the base classifiers. During majority voting, the result of the final prediction of a sample will be based on the total number of times a class label predicted. The classifiers which get more than half of the vote against the test labels will be considered for final predictions. Whereas in average voting, every base classifier will be assigned a weight. During the validation phase, prediction probabilities will be generated for each sample from all the classes. Finally, a product of weights assigned and their likelihood will be averaged. The class that scores the highest average will be considered [38], [44], [45].

TABLE I. PARAMETER LIST FOR THE BASE CLASSIFIERS

Acronym	Parameters
LR	Penalty: L2 (Ridge Regression), Solver: lbfgs, maxIter=150, regParam=0.3, elasticNetParam=0.2
Cart	criterion of split = gini, splitter = best, max_dept = 30 min_samples_leaf = 1, maxBins = 5000
RF	numTrees=100, maxBins = 5000
MLP	Learning_rate=0.1, No of epochs = 50.Momentum = 0.6, Batch_size = 256, No of hidden nodes = 5, Optimizer = adam

The stacked generalized model uses a meta learner on top of the base learners using stacking. Meta learners optimize the output or boost the predictions generated from the base learners. Stacking operates on multiple levels (Level 0 and 1). Level 0 learns with multiple classifiers, and these learners' weights (w_1,w_2,….w_n) will be fed into a meta learner. Predictions made by each learning algorithm in the first phase become training data for the level 1 meta learner. The equation for stacking(stack) predictions from set of classifiers (x_1, x_2,,…,x_n) with a linear combination of weights (w_1,w_2,….w_n) is expressed in equation 1[37] [46].

$$f_{stack}(y) = \sum_{i=1}^{n} w_i f_i\,(x) \quad (2)$$

Algorithm 1: Procedure for building a stacking ensemble

Input: Preprocessed data T = $\{a_j, b_j\}_{j=1}^{n}$

Output: Ensemble model H

1. Learn level-0 classifier models
2. for d=1 to D do
   learn h_d based on T
   end for
3. Create new set of predictions from set T
   for j=1 to n do
   T_h ={ a'_j , b_j }, where a'_j={$h_1(a_i), h_2(a_i), \dots, h_D(a_i)$}
   end for
4. Learn meta classifier
   learn H according to T_h
5. return H

So, for understanding the behaviour of the transaction, the level 0 classifiers will first classify the new data. Then the prediction results will be passed to the meta learner for making the final decision on a transaction to be fraud or non-fraud.

The basic structure of the stacking process used in this study is shown in Fig. 1. The model is applied to train and test data. k-cross validation(cv) is applied on the training data on level 0 to avoid chances of overfitting. Using cv a set of data is generated from each fold and creates a new portion of dataset for each of the four learners. In level 1, that particular dataset generated from the first level prediction is trained by the Random Forest, the meta classifier, and the final prediction

results will be generated. One more significant reason was that the time for prediction in RF is significantly faster than training the model as trees generated during the training are for future reference.

The conceptual architecture of the framework is explained in Fig. 2.

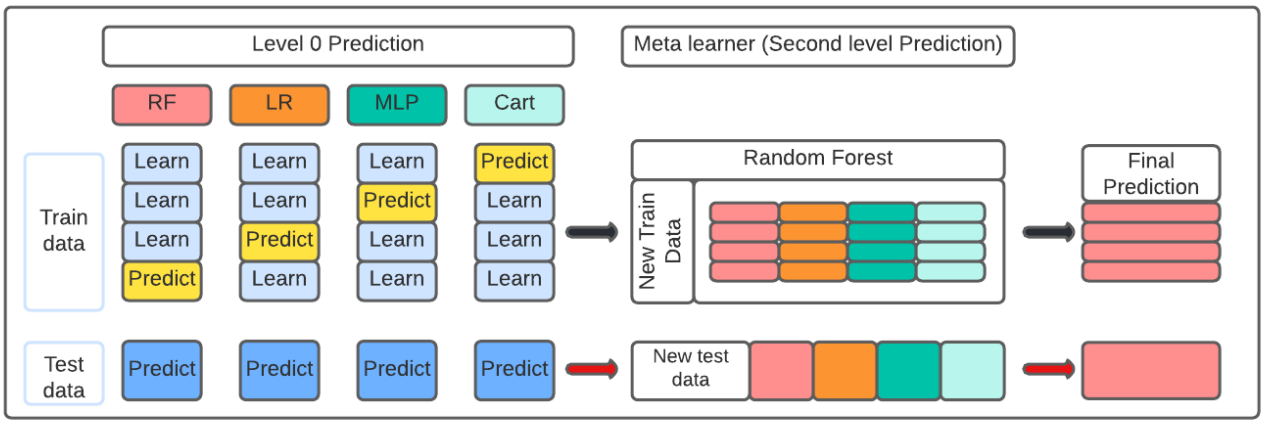


Fig. 1. Structure of Stacking Ensemble used in the Study.

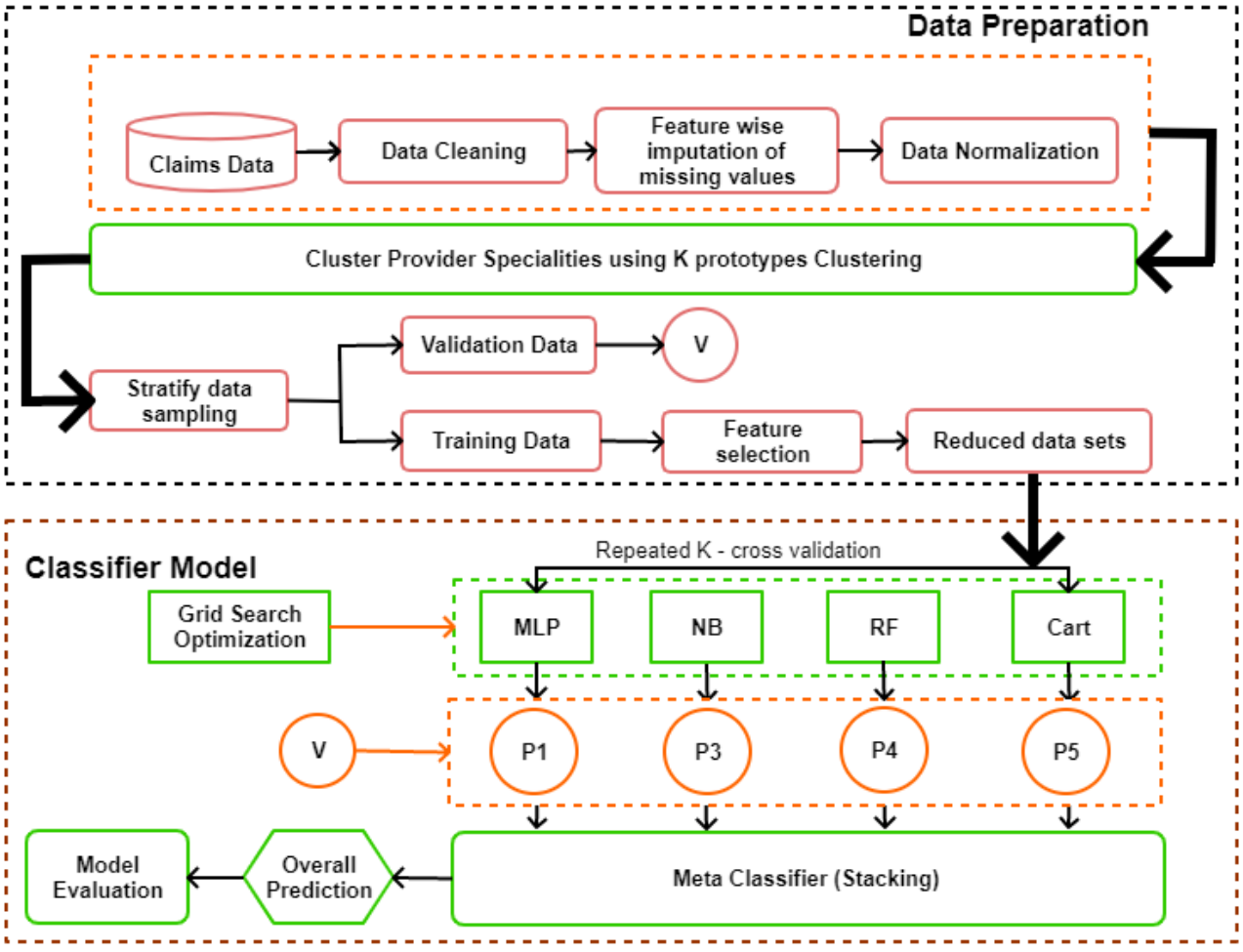


Fig. 2. Conceptual Framework of HEMClust Model.

IV. Experimental Design

A. Experimental Data

The study uses Medicare Part B Providers data published on the CMS Medicare website for years 2014-2017[27]. CMS is a wing in the United States that manages national health care services. CMS collects all claims related data such as prescription, drug-related data, etc.- and analyses it to find and reduce fraud that occurs within the healthcare system. Study uses two datasets from Medicare healthcare for implementation of the model; the first is provider claims data. Providers claim data set which will be mentioned to as Part B, which provides information on all procedures performed by a physician in a particular year. Each physician has been given a unique identifier named NPI. NPI is used to represent a specific physician and procedure he performs for a particular disease. The procedures he performs against the details of the actual procedure could be found by matching the HCPCS code (Health care Common Procedure Code System). This database also provides necessary information about the total number of services performed by the physician, billed, submitted, and allowed charges for a particular service, place of service etc. The nature of the procedure also varies based on the location of the service.

The Second database used for the study is LEIE (List of Excluded Individuals and Entities), generally referred to as the LEIE database[40]. This database contains the list of providers who have been exempted from their service due to some reason. The exemption criteria are based on the crime they have performed, which matches the sections from the Social Security Act. The LEIE database is updated and maintained by the OIG (Office of Inspector General). OIG categorized exclusions into two types Mandatory exclusions and Permissive exclusions. So, example, Section 1128(a)(2) explains "Conviction based on doctor's behaviour towards the patient". Say, abuse or Neglect and the period of conviction is 5 years. Section 1128(b)(4) will be convicted if the provider has not renewed his license or if he is under suspension or surrender. the period of exclusions varies based on the kind of prohibitions. There are many kinds of coded reasons for exclusions, for Section 1128(b)(7), Providers will be convicted for kickback fraud etc. After combining Part B for 4 years (2013-2017), the total number of instances was 2740138. Overall dataset descriptions are available in Table II.

Labels for CMS Part B database were generated by joining with LEIE on NPI as a primary key, and the matching records were marked as fraud. While analyzing the LEIE database, around 93.7 percentage of NPI values were found missing, i.e., labelled "0". Out of 93.7 percentage of missing values, seven percentage had UPIN (Unique Physician Identifier Number). While matching this database, only 465 fraud classes could be found initially. This was quite disappointing that the proportion of fraud occurred and working data was contrastingly low. So, it became inevitable to find the NPI for the missing records. NPPES NPI Registry was used further in the study to refill the missing NPI's. Matching 72k records manually was a tedious task. To speed up the task, an "NPI Matching Algorithm" was developed and used further for matching the NPI's from the registry. Where NPI was not present, UPIN was used to compare and match the records. After applying the algorithm, 9862 fraudulent records were matched.

B. Runtime Environment

The whole experiment was conducted in UBUNTU Linux Environment. The experiment setup was run on 2.8 GHz Intel Core i7-7700HQ, Quad-core CPU with 8 logical cores. NVIDIA beForce GTX 1050 with 4 GB dedicated GPU was also used along with 32GB RAM. Both Python and Spark was used for implementing the whole model. The spark ecosystem contains 5 significant components: Spark Core, Spark SQL, Spark Streaming, Spark MLib, and GraphX. The Spark Core component serves as a basis for distributed processing of big data sets. Resilient Distributed Datasets (RDD) from spark core was applied, which helped save the execution time while loading and reusing the data because it provides distributed and in-memory computations. The machine learning model was implemented using Python Sklearn and Spark ML Library[50].

C. Post Processing or Validation of Results

For evaluating the efficiency of the framework on the fraud detection environment performance metrics such as Precision, Recall, f1 Score will be used. The metrics will check for each provider specialties in detecting upcoding fraud using multi-class classification and overall fraud detection using binary classification.

The model will also be evaluated using stratified repeated k fold cross-validation to ensure that it does not overfit the testing data. Table III explains the formulae for calculating the metrics used for evaluation.

TABLE II. Description of Dataset used for the Study

Name of Data	Description of data	Features
Providers Data (Part B)	Information regarding claims a provider performs for a given procedure Oriented by Fields 1) National Provider Indicator (NPI) 2) Provider Speciality 3) Drug Description Code (HCPCs) 4) Place of Service	29
Exclusion Data (LEIE)	Information regarding providers that are exempted for committing fraud Oriented by Fields 1) National Provider Indicator (NPI) 2) Reason of Exclusion.	17

TABLE III. Evaluation Metrics used for the Study

EM	Equation	Description
ACC	$ACC = \frac{\mathrm{TP} + \mathrm{TN}}{TP + FP + TN + FN}$	Explains the ratio of correctly predicted instances against the total number of instances
Pr	$Pr = \frac{\mathrm{TP}}{TP + FP}$	Percentage of positive samples that are actually predicted correctly from the positive samples.
Re	$Re = \frac{\mathrm{TP}}{TP + TN}$	Percentage of positive samples that are actually predicted from total number of samples
f1	$\mathrm{f1} = 2 * \frac{\mathrm{Pr} * \mathrm{Re}}{\mathrm{Pr} + \mathrm{Re}}$	Evaluate the balanced performance of classes in a model.

*__EM__,Evaluation Metrics
*__ACC__,Accuracy; __Pr__,Precision;__Re__,Recall;__f1__,F1 score

V. Results and Discussion

A. Results

The section explains the outcomes from the experiments performed by using the proposed model. For a better understanding, the results are bifurcated into three parts.

1) Performance evaluation of heterogenous ensemble learners over individual classifiers.

2) Validation of results on HEM model after applying the improvement strategies.

3) Validating the performance of proposed framework (HEMClust) over the baseline ensemble model (HEM model) using Friedman and Nemenyi tests.

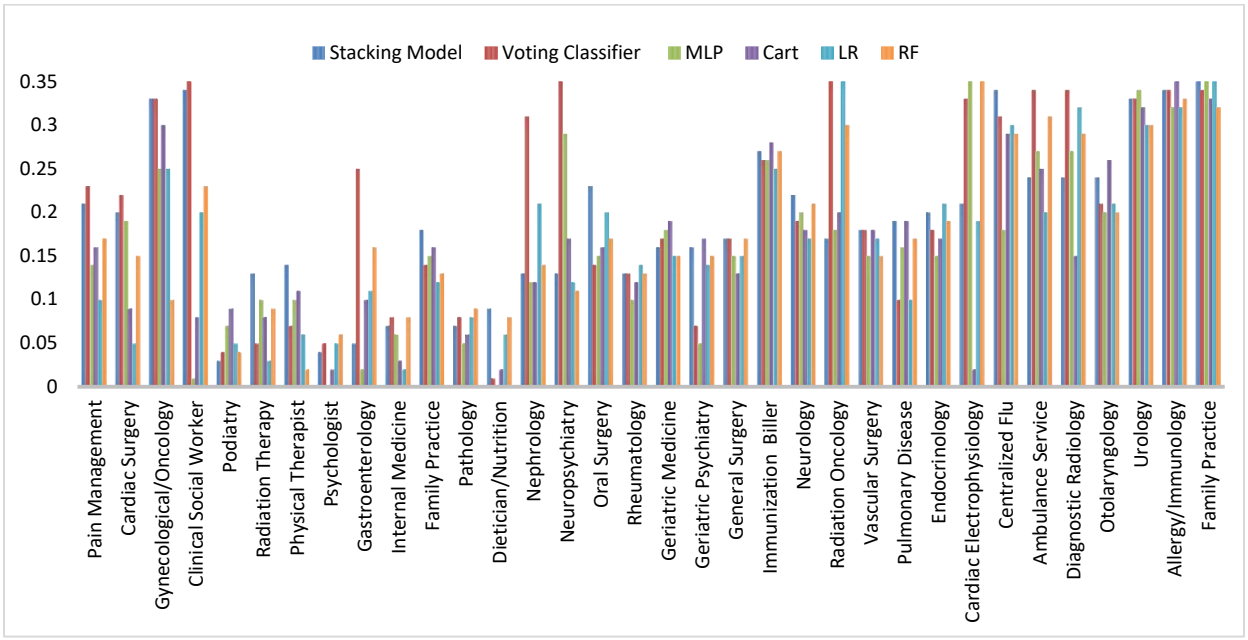
The individual learners and HEM models were evaluated by comparing each model based on their performance criteria. The data was evaluated on the model in two different ways, binary classification (LEIE labelled data set which contains two classes, fraud and Non-fraud) and Multi-class classification (considering each provider specialties as class labels). Considering provider specialties as class labels were necessary to detect upcoding fraud because it could be detected by finding misclassified provider labels against their given specialties. Forty percent fragment of the data was kept aside for validation to see the generalization of the model on the unseen data. The result of the performance of each classifier on the data is shown in Table IV for fraud and Non-fraud class. It was evident that the heterogeneous stacking ensemble outperformed the individual classifiers and voting ensemble. Fig. 3(a) and 3(b) explains the learners' performance on each provider type. Displaying the results of all the specialities on one single plot was not feasible. To improve the readability, the entire plot was divided into two sections. The first section explains the lower performing specialities classes with an f1-score less than 35%, and the second describes the f1-score greater than 35% on the baseline model.

As emphasized earlier two improvement strategies were adopted in the study preprocessing and clustering. To begin with preprocessing, cleaning was performed on the data by identifying the variables which returns the variance zero, especially like single valued attributes as it can no way be influential for the predictor. As a result, attribute "CountryCode" was removed from the dataset. Columns HCPCS code and HCPCS description could be called duplicate columns because HCPCS code itself describes the drug code and its purpose. drug description feature was not found important as it just details the description provided in the feature HCPCS code. Cleaning the values inside the data was also mandatory. While observing "Provider Credentials" column, it was found that a single value is interpreted several ways. For example, credentials "MD", on some places it is said as "M.D." and in some other places it is referred as "M D" and so on. So, necessary actions was taken to clean those values and make them similar. Data normalization is also considered an important part of designing a model. Normalization is used to bring down the features with varying scales to a similar scale [0-1 or $^{-}1$-$^{+}1$]. Paper used min-max normalization which takes values of a feature and transforms it into a predefined interval between 0 and 1. It also tries to preserve the outlier relationship with scaling the data. In the second phase preprocessing was to deal with missing values, it was handled feature-wise by considering the rates of missing values on each attribute. Different approaches were applied to each attribute based on the severity of missing values on it. Following the work of Esra'a Alshdaifat [51], features were categorized based on certain categories. If 1-5% of data is missing in a column, it comes to the category of 'Manageable' or if 5-15% is missing, it will be categorized as 'Sophisticated' and anything above 15% will be categorized as 'Severe'. When the data sample falls under the category of manageable or sophisticated, missing values were handled by imputing it with the mean for numerical data. The values were replaced by any global constant or mode for categorical data. Normally imputing the missing values with mean or mode leads to bias by changing the correlations of the data. Since the amount of data that fall under this category is very small, it wouldn't affect much on the performance. If the category is above 15%, Knn was applied as a method of imputation. Using knn, missing values are filled with similar occurring instances or by finding its distance measure.

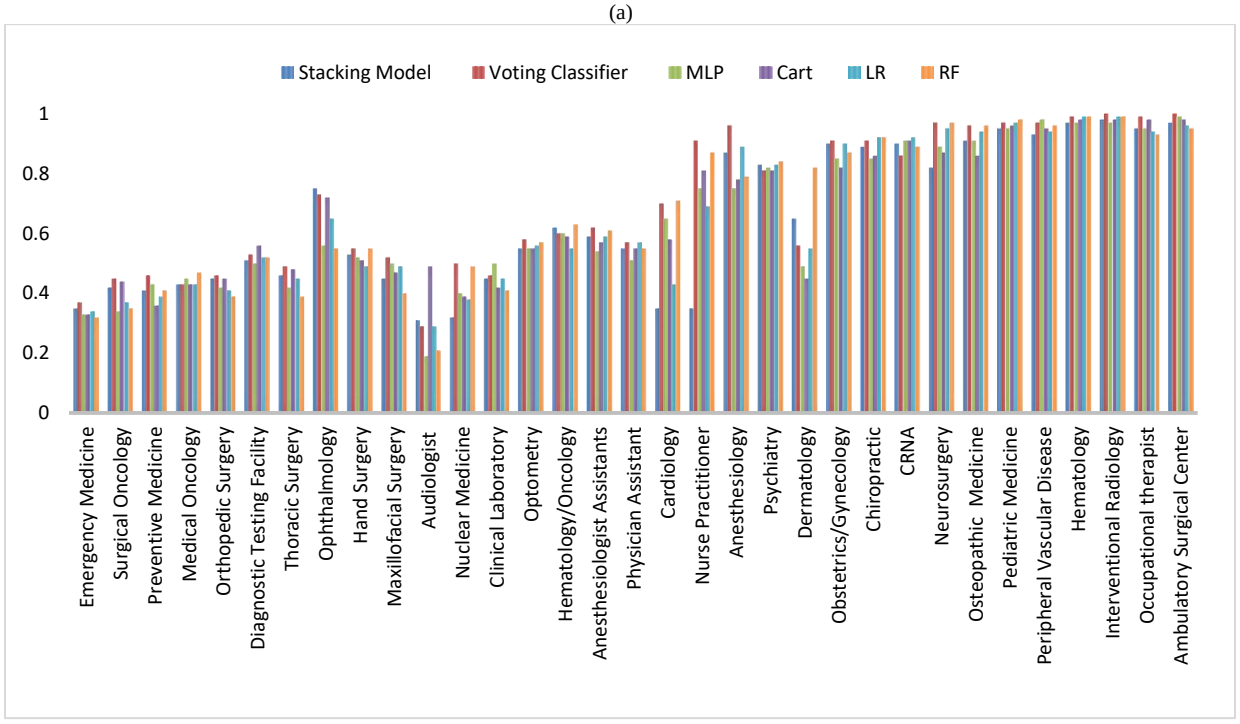
The feature selection was performed using the Extra tree classifier, a decision tree ensemble. Extra tree classifier is more reliable as it randomly selects the split. It is also computationally faster than any other classifier. Following are the discussion and conclusion on each feature's behaviour and importance after analyzing the results.

- From 26 features, 16 features have a value of importance greater than 0.
- Out of the 16 essential features, six numerical features hold 44.0% of the feature importance and 7 categorical features hold 49.0% of the feature importance and 3 features holds value less than 0.

It was found that the aggregated service, Billing information of the procedures, Provider Indicator and City are the most relevant features. Provider's Year of Service and Drug Indicator are less important features. After feature selection, the final number of attributes selected for further study was 15.



(a)



(b)

Fig. 3. (a). Performance of Dataset on each Classifier in Classifying Provider Specialities (f1 < 50), (b). Performance of Dataset on each Classifier in Classifying Provider Specialities (f1 > 50).

TABLE IV. FRAUD-NONFRAUD CLASSIFICATION RESULTS ON HEM MODEL AND INDIVIDUAL LEARNERS

	MLP		**LR**		**Cart**		**RF**		**Heterogenous Voting**		**Heterogenous Stacking**	
Class	NF	F	NF	F	NF	F	NF	F	NF	F	NF	F
Pr	1.0	0.50	0.99	0.50	1.0	0.73	1.0	0.92	0.99	1.0	1.00	0.96
Re	0.93	0.73	0.89	0.55	1.0	0.72	1.0	0.64	1.00	0.68	1.00	0.82
f1	0.96	0.68	0.94	0.53	1.0	0.72	1.0	0.75	0.99	**0.77**	1.00	**0.88**
Acc	0.92		0.88		0.99		0.99		0.99		0.99	

The most important step to be followed prior to clustering process is finding optimal number of clusters (i.e., the value of k). Determining appropriate value for k is important as different values lead to different conclusions and characteristics in the clusters. Also, it is important to find that the resultant value of k has the tendency to produce good clusters. Here, Hopkins's test is applied to measure cluster tendency and optimal k value. Basically, Hopkins value greater than 0.5 consisting of larger value of k shows the probability of grouping data into larger clusters[52]. The results revealed a higher degree of 0.99 value of clustering tendency in the data. Since, value obtained is (~0.99) which is greater than 0.50, null hypothesis is rejected and alternate hypothesis is concluded that the dataset is significantly clusterable. While determining the optimum value of k, there was a considerable decline in the value of statistic with the increase of parameters. The optimal value of k clusters against Hopkins's statistic is shown in Fig. 4. With 12 clusters good cluster tendency of 0.65 was achieved. So, with k value as 12 kproto clustering was applied on the PCA subspace. Partitioning clustering methods have proved to produce better results when applied with pca[53], [54]. Four principal components that explained a total variation of 98% was selected further for clustering. Overall mean accuracy of cluster wise provider specialties are cross-validated, its characteristics and how each provider specialties are distributed in each cluster are shown in Fig. 5.

From the results on the boxplot of each cluster for provider specialties ranging from 0-76, Cluster 3 and 6 hold a maximum number of specialties. It was observed that cluster 3 contained specialties related to surgical procedures like, Vascular Surgery, Anesthesiology, Internal medicine and so on. Those procedures are formed in one group because of the common procedural code shared by these specialties for a particular treatment. The contents of clusters 1, 2 and 3 contained a smaller number of providers, and that group was dedicated to specialties with similar behaviour for example, cluster 4 had only 5 members such as Cardiology, Cardiac Surgery, Diagnostic Radiology, Anesthesiology and Internal Medicine. It can be said that these groups are related to cardiac surgery. It was also found that certain groups of providers like Anesthesiology, Ambulance providers, Internal medicine, Nurse Practitioner are included in more than 1 cluster. The reason might be that these providers are commonly included in many procedures. Further, the HEM model was applied on the clustered data, by considering each cluster as classes. Fig. 6 plots the confusion matrix using a color-encoded heatmap obtained from multi-class classification here, each class represents a cluster with grouping specialties.

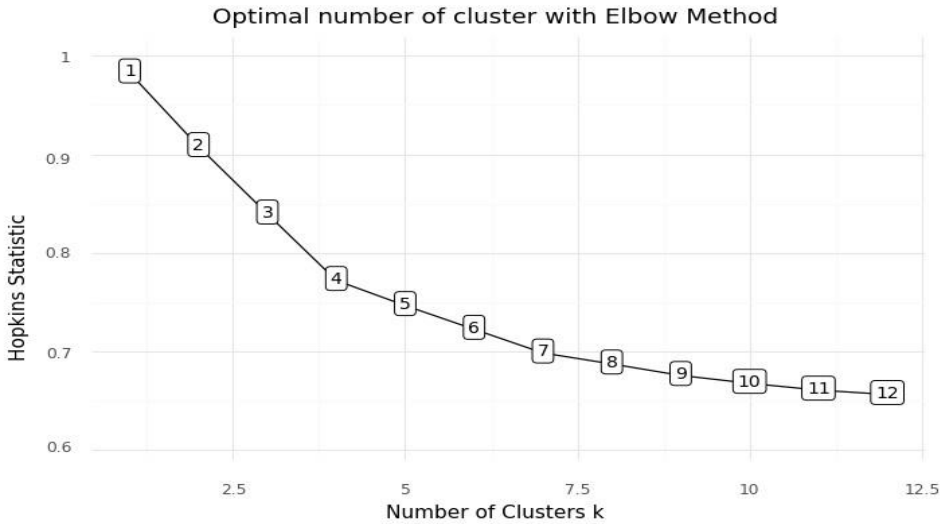


Fig. 4. Results of Hopkins Test Statistic for Measuring Clustering Tendency using k nearest Neighbour Distances.

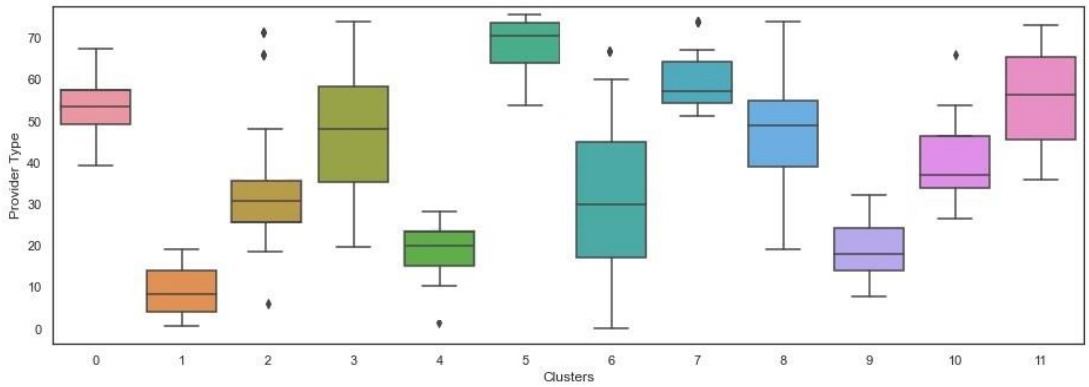


Fig. 5. Characteristics of each Cluster based on Provider Type.

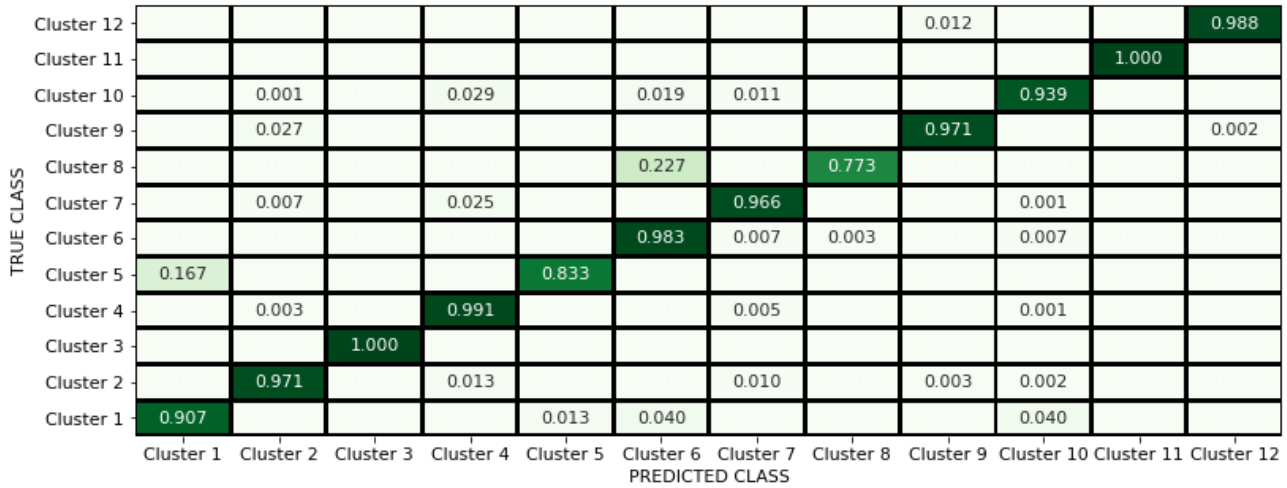


Fig. 6. Heatmap for Confusion Matrix Depicting the Multi Class Classification Results for each Cluster. Each Rows represents the True Class and Columns represents the Prediction done by the Classifier.

It was noticed that clusters 3 and 11 are detected with a higher TPR of 100% and cluster 8 is the least detected class compared to others with a TPR of 77.3%. There are also a few places where misclassified clusters were found from their original classes, although their percentage was tiny. A detailed description of results is plotted in Fig. 7. Heatmap is used to plot the overall key metrics such as precision-recall and f1-score for each class. The model produced an overall accuracy of 98%. Considering the weighted average precision, It could be noticed that almost 98% of data has been correctly classified only 2% was misclassified to wrong classes. All the above results prove that grouping strategies significantly increased the fraud detection ratio to at least 45-50%.

Further, for a more precise evaluation of the impact on the HEM model and the proposed improvement strategies, a comparison is made using two statistical tests Friedman and Nememyi. Initial steps were to find whether there existed any significant difference between the mean models. A Friedman test is applied on all base classifiers, heterogeneous ensembles and HEM*Clust* to determine whether or not these groups are statistically significant. The test statistic(X^2) and corresponding p_value(p) from Friedman test was 11.04 & 0.026 respectively. Since obtained p_value is lesser than the default 0.05 here, null hypothesis can be rejected and the post-hoc Nemenyi test could be performed for finding an exact model that is different in performance from others. Results from Fig. 8, shows that LR, MLP, Cart, RF and Voting classifiers belong to one group. Also, LR performed significantly worse than other models, and Cart and RF seem to have similar performances. Though it is difficult to conclude a comparison concerning the Stacking ensemble because it belongs to two groups. Although it can be affirmed that HEM*Clust* is significantly different from other groups, since HEM*Clust* is built using a stacking ensemble, a few similarities in their performances could be seen.

B. Discussion

Basic aim for building HEM*Clust* model is to identify provider fraud. So, the idea here was to detect misclassified provider specialties based on their respective procedural codes. Suppose a provider is classified into different group of class which it does not belongs to, that particular transaction could be alerted or further rechecked for fraud.

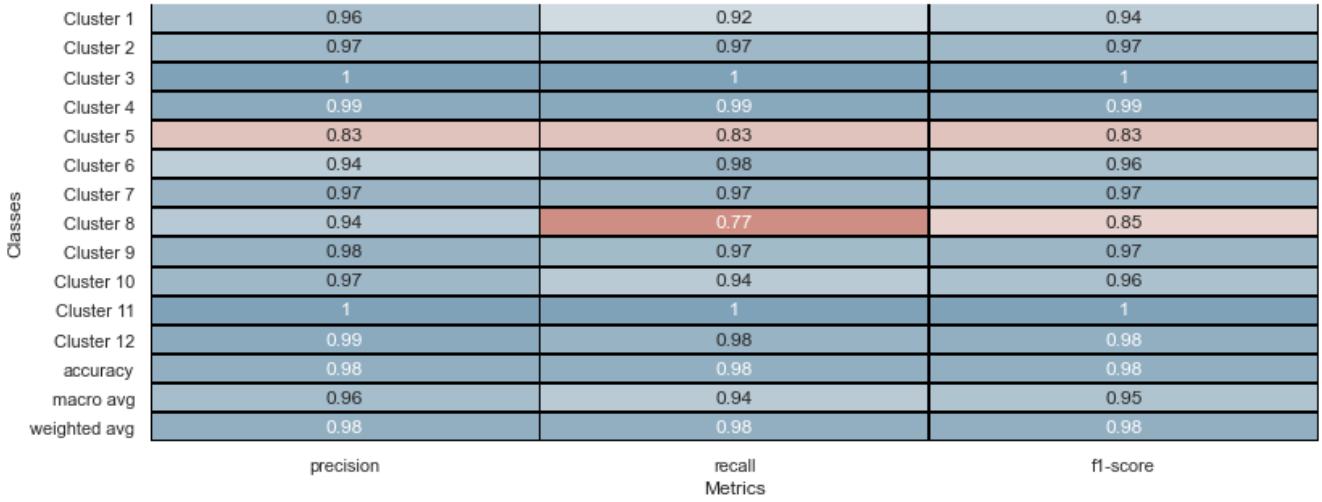


Fig. 7. Heatmap on Overall Performance Measures of HEMClustmodel on each Class.

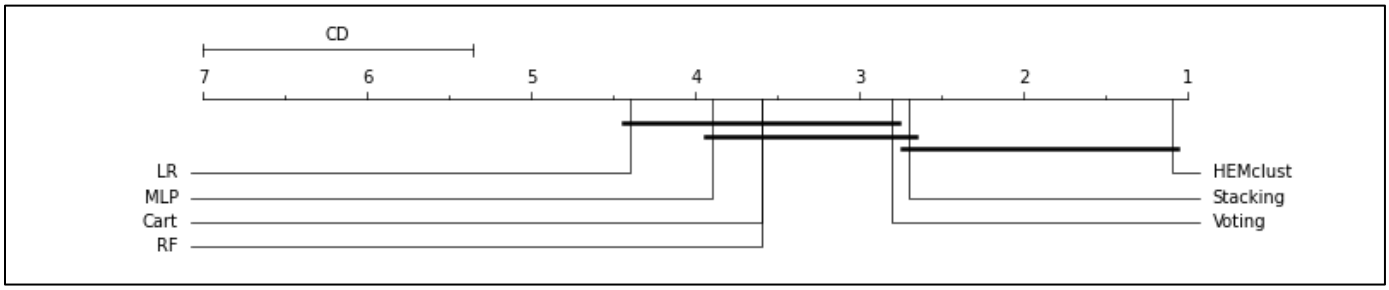


Fig. 8. Comparison of base Classifier against HEMClust with Nemenyi Test.

Attaining high accuracy was evident here as the risk of misclassification was very high. To build a better model all the areas related to claims data was studied in detail. Since the claims data is collected from various sources of healthcare sector it was evident to perform an appropriate preprocessing to improve the quality of data. Basically, detecting fraud is considered as a complex task as the boundary of separation between fraudulent and non-fraudulent classes is very noisy. Proposed model uses more sophisticated techniques for handling missing data to make it more convenient. feature engineering techniques were also used, which helped us select the essential feature that contributes in effective prediction. A varied performance result was observed from the initial experiment on each class when classified initially. It was found that the reason because of this was mainly due to the overlapping of procedures, which lead to the decrease in accuracy [30]. To improve the predictive accuracy, similar providers specialties were grouped using clustering. The results from clustering shows that providers performing similar kind of procedures were grouped in a single cluster. Further each cluster was considered as class labels and classified.

Following are the observations made from while implementing the classification model.

1) All the four learners, when individually applied, had an unstable performance. Though Cart and RF has good accuracy but there was high misclassification of classes.

2) Feature engineering and data cleaning had helped in improving the performance of the model also, the use of Spark Resilient distributed file system helped us in executing big data without time and memory overheads.

3) For selecting the meta learner, both RF and MLP was applied on base classifier separately. MLP as a Meta learner gave 83.9% precision score and 85.9% recall rate which states that the model could correctly classified only 83.9% of fraudulent samples. Random forest as a Meta learner gave 96% precision score and 94% recall rate and 98% of average f1 score, where the model could classify around 96% of fraudulent sample.

4) Statistical test like Friedman test and Nemenyi test was applied to know the differences in the performance of classifiers. Friend man test demonstrated a significant difference between the classifiers with the proposed method with a p_value of 0.02.

VI. Conclusion and Future Work

The paper proposes a Heterogenous ensemble model with clustering (HEM*Clust*) to detect fraud from claims data effectively. The model operates in three phases; first phase intends to apply preprocessing techniques to improve the data quality. The second phase aims to reduce the overlapping instances found in provider specialities using k-prototype clustering. The final step includes predicting fraudulent providers using a heterogenous ensemble model through stacking. The dataset used in the study was easily attributed to big data due to its voluminous nature. Spark framework was used on top of the Hadoop cluster to implement several model parts to avoid any computational overheads. Application of heterogeneous ensembles with meta learner helped in minimizing the error generated by these learners during prediction. It was found that the proposed HEM*Clust* model showed the best overall fraud detection performance with an Average Precision-Recall Rate of 98%. During the study it was also observed that the fraud detection domain keeps evolving with the changing patterns of fraud. The problem is mainly referred as concept drift. The proposed model could also be extended to address the particular problem as ensemble learning has been used as state of the art to detect concept drift. However, interpretation of concept drift detection in fraud detection is out of this work's scope and could be considered for future work.

References

[1] H. Joudaki et al., “Improving fraud and abuse detection in general physician claims: A data mining study,” Int. J. Heal. Policy Manag., vol. 5, no. 3, pp. 165–172, 2016, doi: 10.15171/ijhpm.2015.196.

[2] D. Thornton, M. Brinkhuis, C. Amrit, and R. Aly, “Categorizing and Describing the Types of Fraud in Healthcare,” Procedia Comput. Sci., vol. 64, pp. 713–720, 2015, doi: 10.1016/j.procs.2015.08.594.

[3] K. E. Mues et al., “Use of the Medicare database in epidemiologic and health services research: A valuable source of real-world evidence on the older and disabled populations in the US,” Clin. Epidemiol., vol. 9, pp. 267–277, 2017, doi: 10.2147/clep.s105613.

[4] E. Birman-Deych, A. D. Waterman, Y. Yan, D. S. Nilasena, M. J. Radford, and B. F. Gage, “Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors,” Med. Care, vol. 43, no. 5, pp. 480–485, 2005, doi: 10.1097/01.mlr.0000160417.39497.a9.

[5] "Insurance Information Institute” https://www.iii.org/fact-statistic/facts-and-statistics-insurance-fraud. .

[6] FICCI, “Health Insurance Fraud.”

[7] H. Joudaki et al., “Using data mining to detect health care fraud and abuse: a review of literature.,” Glob. J. Health Sci., 2015, doi: 10.5539/gjhs.v7n1p194.

[8] D. Erlangga, M. Suhrcke, S. Ali, and K. Bloor, "The impact of public health insurance on health care utilization, financial protection and health status in low: The middle-income countries: A systematic review(PLoS ONE (2019)14:8 (e0219731) DOI: 10.1371/journal.pone.0219731)," PLoS One, vol. 14, no. 11, pp. 1–20, 2019, doi: 10.1371/journal.pone.0225237.

[9] I. Sohony, R. Pratap, and U. Nambiar, "Ensemble learning for credit card fraud detection," ACM Int. Conf. Proceeding Ser., no. July, pp. 289–294, 2018, doi: 10.1145/3152494.3156815.

[10] R. Konrad, W. Zhang, M. Bjarndóttir, and R. Proaño, "Key considerations when using health insurance claims data in advanced data analyses: an experience report," Heal. Syst., vol. 9, no. 4, pp. 317–325, 2020, doi: 10.1080/20476965.2019.1581433.

[11] K. Ferver, B. Burton, and P. Jesilow, "The Use of Claims Data in Healthcare Research," Open Public Health J., vol. 2, no. 1, pp. 11–24, 2009, doi: 10.2174/1874944500902010011.

[12] S. R. Sukumar, N. Ramachandran, and R. K. Ferrell, "Data Quality Challenges in Healthcare Claims Data : Experiences and Remedies," no. April 2014, p. 15, 2014.

[13] S. K. Shamitha and V. Ilango, "A survey on machine learning techniques for fraud detection in healthcare," vol. 7, no. 4, pp. 5862–5868, 2018, doi: 10.14419/ijet.v7i4.15696.

[14] R. A. Derrig, "Insurance fraud," J. Risk Insur., vol. 69, no. 3, pp. 271–287, 2002.doi: https://doi.org/10.1111/1539-6975.00026.

[15] J. M. Johnson and T. M. Khoshgoftaar, Medicare fraud detection using neural networks, vol. 6, no. 1. Springer International Publishing, 2019.

[16] S. Dash, S. K. Shakyawar, M. Sharma, and S. Kaushik, "Big data in healthcare: management, analysis and future prospects," J. Big Data, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0217-0.

[17] Spark, "http://spark.apache.org/."

[18] S. Salloum, R. Dautov, X. Chen, P. X. Peng, and J. Z. Huang, "Big data analytics on Apache Spark," Int. J. Data Sci. Anal., vol. 1, no. 3–4, pp. 145–164, 2016, doi: 10.1007/s41060-016-0027-9.

[19] M. Zaharia et al., "Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing," Proc. NSDI 2012 9th USENIX Symp. Networked Syst. Des. Implement., pp. 15–28, 2012.

[20] X. Zhou, Z. Zhang, L. Wang, and P. Wang, "A Model Based on Siamese Neural Network for Online Transaction Fraud Detection," Proc. Int. Jt. Conf. Neural Networks, vol. 2019-July, 2019, doi: 10.1109/IJCNN.2019.8852295.

[21] S. Bagga, A. Goyal, N. Gupta, and A. Goyal, "Credit Card Fraud Detection using Pipeling and Ensemble Learning," Procedia Comput. Sci., vol. 173, no. 2019, pp. 104–112, 2020, doi: 10.1016/j.procs.2020.06.014.

[22] S. Liu, Y. Wang, J. Zhang, C. Chen, and Y. Xiang, "Addressing the class imbalance problem in Twitter spam detection using ensemble learning," Comput. Secur., vol. 69, pp. 35–49, 2017, doi: 10.1016/j.cose.2016.12.004.

[23] J. Novakovic and S. Markovic, "Classifier Ensembles for Credit Card Fraud Detection," 2020 24th Int. Conf. Inf. Technol. IT 2020, no. February, 2020, doi: 10.1109/IT48810.2020.9070534.

[24] M. Zareapoor and P. Shamsolmoali, "Application of credit card fraud detection: Based on bagging ensemble classifier," Procedia Computer Science, vol. 48, no. C. pp. 679–685, 2015, doi: 10.1016/j.procs.2015.04.201.

[25] D. W. Fan, P. K. Chan, and S. J. Stolfo, "A comparative evaluation of combiner and stacked generalization," Proc. AAAI-96 Work. Integr. Mult. Learn. Model., no. June, pp. 40–46, 1996.

[26] K. R. Kerwin and N. D. Bastian, "Stacked generalizations in imbalanced fraud data sets using resampling methods," J. Def. Model. Simul., vol. 2628, pp. 0–2, 2020, doi: 10.1177/1548512920962219.

[27] "Part B National Summary Data File (Previously known as BESS)," Data base _ Medicare C., p. 21244, 2018, [Online]. Available: https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/Part-B-National-Summary-Data-File.

[28] M. Herland, T. M. Khoshgoftaar, and R. A. Bauder, "Big Data fraud detection using multiple medicare data sources," J. Big Data, vol. 5, no. 1, pp. 1–21, 2018, doi: 10.1186/s40537-018-0138-3.

[29] R. A. Bauder, T. M. Khoshgoftaar, and T. Hasanin, "Data sampling approaches with severely imbalanced big data for medicare fraud detection," Proc. - Int. Conf. Tools with Artif. Intell. ICTAI, vol. 2018-Novem, pp. 137–142, 2018, doi: 10.1109/ICTAI.2018.00030.

[30] M. Herland, R. A. Bauder, and T. M. Khoshgoftaar, "Approaches for identifying U.S. medicare fraud in provider claims data," Health Care Manag. Sci., vol. 23, no. 1, pp. 2–19, 2018, doi: 10.1007/s10729-018-9460-8.

[31] R. Bauder, T. M. Khoshgoftaar, and N. Seliya, "A survey on the state of healthcare upcoding fraud analysis and detection," Heal. Serv. Outcomes Res. Methodol., vol. 17, no. 1, pp. 31–55, 2017, doi: 10.1007/s10742-016-0154-8.

[32] L. K. Branting, F. Reeder, J. Gold, and T. Champney, "Graph Analytics for Healthcare Fraud Risk Estimation," pp. 845–851, 2016.

[33] V. Chandola, S. R. Sukumar, and J. Schryver, "Knowledge discovery from massive healthcare claims data," Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., vol. Part F1288, no. August, pp. 1312–1320, 2013, doi: 10.1145/2487575.2488205.

[34] J. R. Gaikwad, A. B. Deshmane, H. V Somavanshi, S. V Patil, and R. A. Badgujar, "Credit Card Fraud Detection using Decision Tree Induction Algorithm," Int. J. Innov. Technol. Explor. Eng., no. 6, pp. 2278–3075, 2014.

[35] M. S. Kumar, V. Soundarya, S. Kavitha, E. S. Keerthika, and E. Aswini, "Credit Card Fraud Detection Using Random Forest Algorithm," 2019 Proc. 3rd Int. Conf. Comput. Commun. Technol. ICCCT 2019, no. March, pp. 149–153, 2019, doi: 10.1109/ICCCT2.2019.8824930.

[36] A. Gulati, P. Dubey, C. Mdfuzail, J. Norman, and R. Mangayarkarasi, "Credit card fraud detection using neural network and geolocation," IOP Conf. Ser. Mater. Sci. Eng., vol. 263, no. 4, 2017, doi: 10.1088/1757-899X/263/4/042039.

[37] I. K. Nti, A. F. Adekoya, and B. A. Weyori, "A comprehensive evaluation of ensemble learning for stock-market prediction," J. Big Data, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00299-5.

[38] S. Tewari and U. D. Dwivedi, "A comparative study of heterogeneous ensemble methods for the identification of geological lithofacies," J. Pet. Explor. Prod. Technol., vol. 10, no. 5, pp. 1849–1868, 2020, doi: 10.1007/s13202-020-00839-y.

[39] M. Herland, R. A. Bauder, and T. M. Khoshgoftaar, "Medical provider specialty predictions for the detection of anomalous medicare insurance claims," Proc. - 2017 IEEE Int. Conf. Inf. Reuse Integr. IRI 2017, vol. 2017-Janua, pp. 579–588, 2017, doi: 10.1109/IRI.2017.29.

[40] "LEIE : Office of Inspector General LEIE Downloadable Databases." https://www.oig.hhs.gov/exclusions/exclusions_list.asp.

[41] S. K. K. Asha RB, "Credit Card Fraud Detection using Artificial Neural Networks", Global Transitions Proceedings, pp. 35-41,Vol2, 2021, doi: 10.1016/j.gltp.2021.01.006.

[42] S. K. Shamitha and V. Ilango, "A hybrid technique for health insurance fraud detection on highly imbalanced dataset," Int. J. Innov. Technol. Explor. Eng., vol. 8, no. 11, 2019, doi: 10.35940/ijitee.K2489.0981119.

[43] A. Husejinović, "Credit card fraud detection using naive Bayesian and c4.5 decision tree classifiers," Period. Eng. Nat. Sci., vol. 8, no. 1, pp. 1–5, 2020, doi: 10.21533/pen.v.

[44] D. Gupta and R. Rani, "Improving malware detection using big data and ensemble learning," Comput. Electr. Eng., vol. 86, p. 106729, 2020, doi: 10.1016/j.compeleceng.2020.106729.

[45] N. Liu, H. Gao, Z. Zhao, Y. Hu, and L. Duan, "A stacked generalization ensemble model for optimization and prediction of the gas well rate of penetration: a case study in Xinjiang," J. Pet. Explor. Prod. Technol., 2021, doi: 10.1007/s13202-021-01402-z.

[46] R. Soleymanzadeh, M. Aljasim, M. W. Qadeer, and R. Kashef, "Cyberattack and Fraud Detection Using Ensemble Stacking," Ai, vol. 3, no. 1, pp. 22–36, 2022, doi: 10.3390/ai3010002.

[47] R. Nooraeni, M. I. Arsa, and N. W. Kusumo Projo, "Fuzzy Centroid and Genetic Algorithms: Solutions for Numeric and Categorical Mixed Data Clustering," Procedia Comput. Sci., vol. 179, no. 2020, pp. 677–684, 2021, doi: 10.1016/j.procs.2021.01.055.

[48] G. Preud'homme et al., "Head-to-head comparison of clustering methods for heterogeneous data: a simulation-driven benchmark," Sci. Rep., vol. 11, no. 1, pp. 1–14, 2021, doi: 10.1038/s41598-021-83340-8.

[49] Z. Huang, "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Mining and Knowledge Discovery 2, 283-304," Data Min. Knowl. Discov., vol. 2, no. 3, pp. 283–304, 1998.

[50] K. D. L. Library, "https://keras.io/."

[51] E. Alshdaifat, D. Alshdaifat, A. Alsarhan, F. Hussein, and S. M. F. S. El-Salhi, "The effect of preprocessing techniques, applied to numeric features, on classification algorithms' performance," Data, vol. 6, no. 2, pp. 1–23, 2021, doi: 10.3390/data6020011.

[52] R. F. Lachlan, L. Verhagen, S. Peters, and C. ten Cate, "Are There Species-Universal Categories in Bird Song Phonology and Syntax? A Comparative Study of Chaffinches (Fringilla coelebs), Zebra Finches (Taenopygia guttata), and Swamp Sparrows (Melospiza georgiana)," J. Comp. Psychol., vol. 124, no. 1, pp. 92–108, 2010, doi: 10.1037/a0016996.

[53] F. Afrin and M. Tabassum, "Comparative Performance Of Using PCA With K-Means And Fuzzy C Means Clustering For Customer Segmentation," Comp. Perform. Using PCA With K-Means Fuzzy C Means Clust. Cust. Segmentation, vol. 4, no. 10, pp. 70–74, 2015.

[54] C. Ding and X. He, "K-means clustering via principal component analysis," Proceedings, Twenty-First Int. Conf. Mach. Learn. ICML 2004, pp. 225–232, 2004, doi: 10.1145/1015330.1015408.