Insurance Claim Data Analysis

Pubudu Cooray, and **Nagulanathan Lavalojan,** and **Ragunaathan Sugirjan**
Faculty of Computer Science and Engineering
University of Moratuwa
Colombo, Sri Lanka
pubudu.20@cse.mrt.ac.lk lavalojan.14@cse.mrt.ac.lk ragunaathan.20@cse.mrt.ac.lk

***Abstract*—Health Insurance companies do not always make payments for their client's treatments and they try to deny the claims of them. In the health Provider view, they are getting half of the claims rejected by the patients' insurance companies which impacts their revenue cycle. By analysing the past submitted claim details, a health provider can forecast the acceptance of the new claim and do further actions on it. Previous research has primarily relied on the Insurance industry, like claim fraud detection, risk analytic and claim triage. No research happens regarding getting claim acceptance from a health provider point of view. We use data of one healthcare provider's past claim details and do descriptive analysis, diagnosis analysis and predictive analysis on it. From these analyses, we have created a machine learning prediction model which will predict the acceptance of the new claim with more than 90 percent accuracy.**

I. Introduction

According to USA government policies, healthcare providers should provide services to patients and get claimed for the cost of treatments from insurances. However, these claims can get denied by the insurances for various other reasons such as missing/incomplete details and incorrect information etc. Therefore healthcare providers have to manually reprocess the claims by doing the appropriate fixes and resubmit to the insurance with the required specifications. This process takes much time and affects the revenue cycle. With the descriptive analysis of the past claim data, we could find much useful information about data and used it for diagnosis analysis. From the diagnosis analysis, we could find why the claims are rejected or accepted. For predicting the new claim is rejected or not we could use our diagnosis analysis. By doing this descriptive, diagnosis, the predictive analysis in future we can create recommendations to make claims accepted.

II. Methodology

A. Date Set

For the analysis, we used actual claims data of a Health Provider in the United States of America. Because of The Health Insurance Portability and Accountability Act of 1996(HIPAA) compliance, we couldn't use any patients related data in our analysis. The data set contains Payment details, treatments and insurance company details.

B. Descriptive Analysis

Here we categorized the variables into quantitative and qualitative and check their distributions and check whether there are any correlations between each variable and see the relationships.

TABLE I
Data Set columns

Table Columns	Table Column Head		
	Description	***Categorical***	***Metric***
Ticket Id	Unique value for a claim	Nominal	Descrete
DateOfService	Service is provided date	Ordinal	Descrete
CarrierCode	Insurance provider unique code	Nominal	Descrete
Charges	Treatment cost	Ordinal	Continuous
Payments	Total payments	Ordinal	Continuous
Insurance Payment	Payments from Insurance	Ordinal	Continuous
PatientPayment	Payments from patients	Ordinal	Continuous
Writeoffs	Amount cannot pay	Ordinal	Continuous
Insurance Balance	Will be paid by Insurance	Ordinal	Continuous
Patient Balance	Will be paid by the patient	Ordinal	Continuous
Allowed Amount	Allowed amount for Treatment	Ordinal	Continuous
FinancialClass	Insurance category	Nominal	Descrete
CPT	Treatment codes	Nominal	Descrete
NPI	Physician's registration number	Nominal	Descrete
Ordering Clinic	Clinic name	Nominal	Descrete
DeniedCode	The reason for denial	Nominal	Descrete
FirstBilledDate	Date of submitted claim	Odinal	Descrete

*Value in US dollars

1) Quantitative: We found Charges, Payments, Insurance Payment, patient Payment, WriteOffs, Insurance Balance, Patient Balance, Allowed Amount under quantitative variables. For this analysis we need to calculate central tendency such as mean, median, mode and see the dispersion of data by calculating standard deviation, variance and interquartile ranges.

	Charges	Payments	InsurancePayment	PatientPayment	Writeoffs	InsuranceBal	PatientBal	AllowedAmount	timedelta
count	22562.000000	22562.000000	22562.000000	22562.000000	22562.000000	22562.000000	22562.000000	22562.000000	22562.000000
mean	1883.571669	73.384245	70.593954	2.790291	870.662427	404.176953	532.703014	533.018205	16.490648
std	1860.330346	272.607128	270.589905	38.596924	1491.963427	1144.770922	1234.355380	1189.371360	6.111933
min	10.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	6.000000
25%	185.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	11.000000
50%	500.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	15.000000
75%	3900.000000	0.000000	0.000000	0.000000	500.000000	38.380000	249.000000	651.000000	21.000000
max	6600.000000	4600.000000	4600.000000	1590.000000	6600.000000	6600.000000	6600.000000	10110.420000	43.000000

Fig. 1. Descriptive statistics for quantitative data

Now let's go through some variables and check their distribution.

From the histogram and probability density graph, we can see the variable Charges' distribution has 2 peaks one is around dollar 100 and other is around dollar 4000. Here for central tendency mean is the best choice.

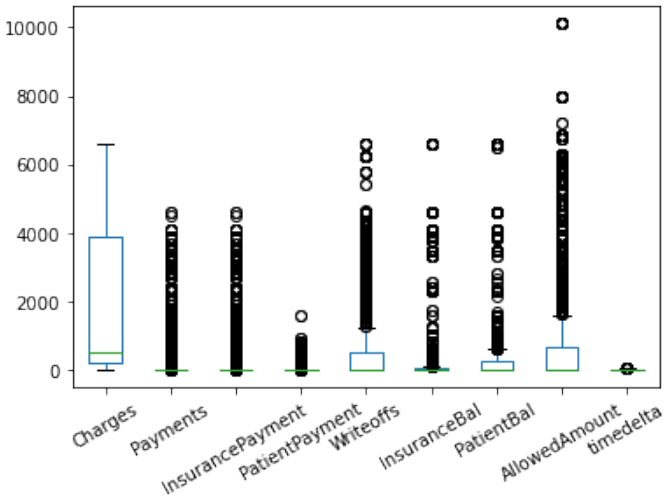

Fig. 2. Box plot for quantitative data

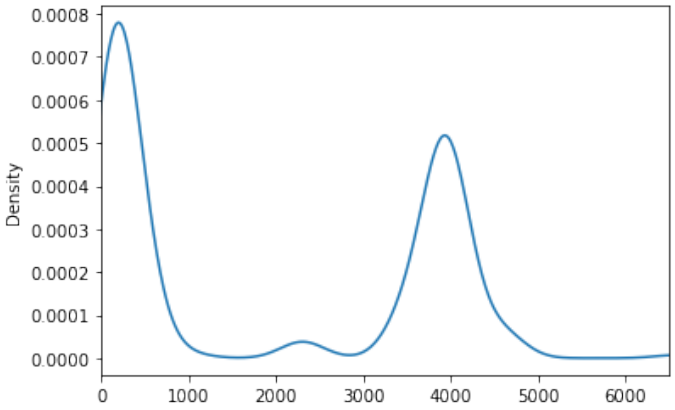

Fig. 3. Probability density for Charges

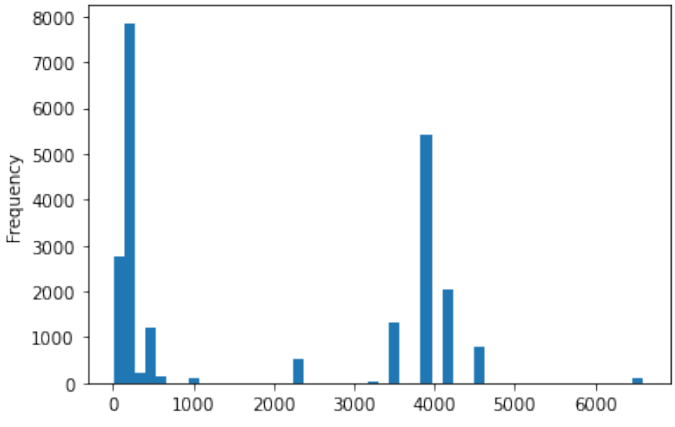

Fig. 4. Histogram for Charges

level of accepted	Mean	Standard Error	Standard Deviation	Min	First Quartile	Median	Third Quartile	Max	n used	n missing
False	1819.30	15.97	1867.14	10.00	175.00	500.00	3900.00	6600.00	13661	0
True	1982.22	19.56	1845.58	10.00	200.00	600.00	3900.00	6600.00	8901	0
Overall	1883.57	12.39	1860.33	10.00	185.00	500.00	3900.00	6600.00	22562	0

Fig. 5. Charges statistics with accepted data

When we observe the variable Charge, it is evenly distributed with variable accepted tag true and false. Like this we plot and see each variables and analyze.

2) Qualitative: To see the trend of Qualitative data, we need to find the frequencies of each unique category of variable. For example variable FinancialClass has 8 unique values. We need to calculate the occurrences of category and plot a bar chart or pie chart to visualize the data. We are having 2860 unique

	TicketNumber	CarrierCode	FinancialClass	CPT	NPI	OrderingClinic	DeniedCode	ProviderProfile	accepted
count	22562	22562	22562	22562	22562	22562	22562	22562	22562
unique	22562	392	8	39	2860	1721	80	2	2
top	263979	AETLIF-E	CT - CONTRACTED	81420	1669532347	MOORE OB/GYN	-	NTINC	False
freq	1	2511	12375	5019	310	310	8901	22223	13661

Fig. 6. frequency statistic for qualitative data

values for NPI and 1721 unique values for OrderingClinic. There is not enough data to describe each category of these. We can plot the bar chart or pie chart for each variables and

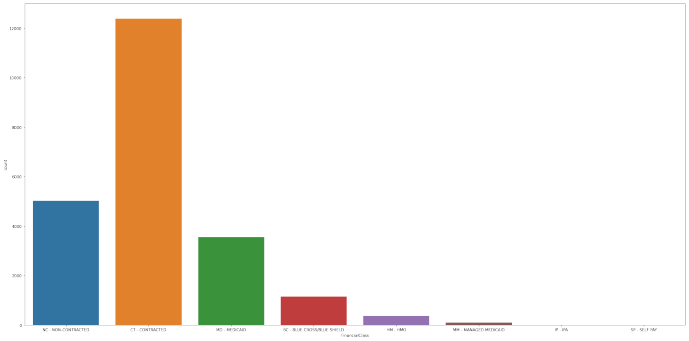

Fig. 7. Frequency Bar chart for Financial class

see the distributions.

3) Correlation: In the analysis, we can find the relations of each variable. From the scatter plot, we can see how one variable correlated with other. From the heat map, we can find how each variables related with other variables. We can categorized coefficient less than 2 as weakly correlated, coefficient 2 to 5 as medium correlated, 5 to 8 as strongly correlated, and above 8 as very strongly correlated. By observing the data we can come up with the relationship between some variables as below.

Payments = PatientPayment + InsurancePayment

Charges = Payment + writeoffs + patientBal+insuranceBal

C. Diagnosis Analysis

If the claim is rejected by Insurances then they send a denial Code to the clients else if insurance companies accept the

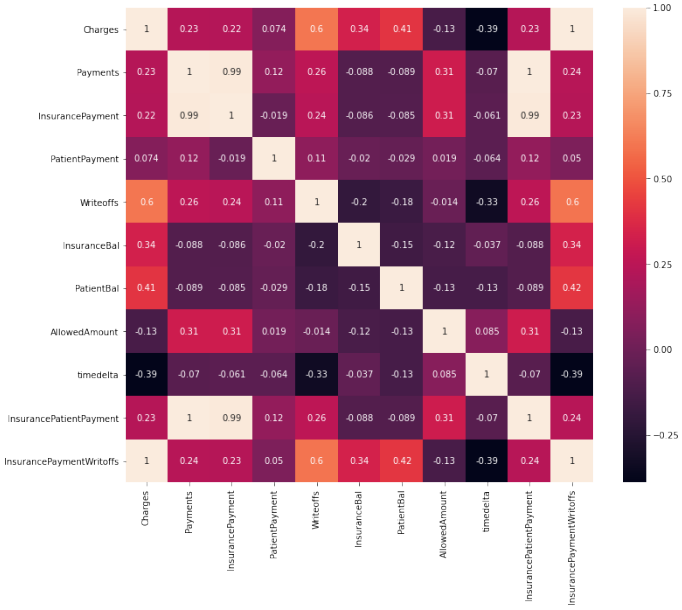

Fig. 8. Correlation between each column with a heat map

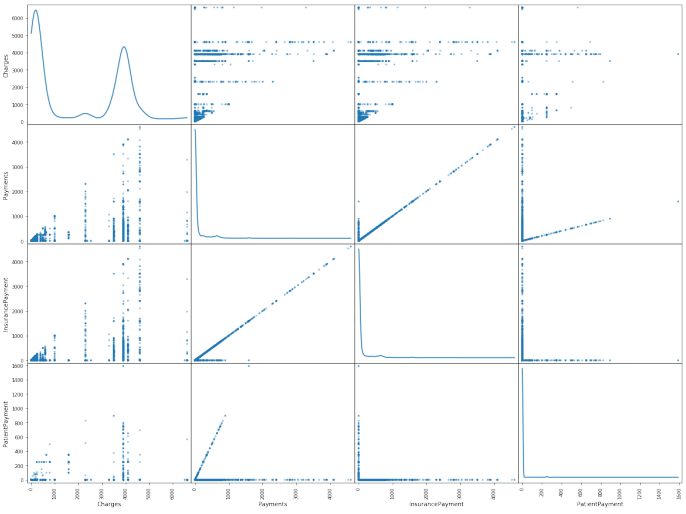

Fig. 9. Scatter plot between some variables

claims then they will do the payment according to the allowed amount.

People have different categories of Insurances like Medicaid and Medicare which is in the 'FinancialClass' column in Data Set. Fig. 10 illustrates how claims are accepted or rejected by insurances based on the Financial Class Column.

Self Pay means the patient has no insurance which covers the treatment he got and he has to pay by himself. So those claims not get rejected by any insurance companies and Healthcare Providers can get the payments from the patient. When the health provider gets the patient's details and without submitting to Insurance company, it can get the payments directly from the patient. There is another insurance [IP-IPA] company which does not have any rejection. But we come to

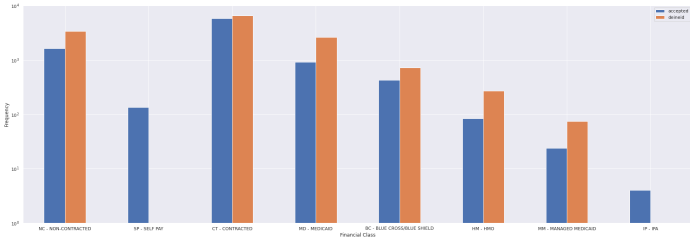

Fig. 10. Financial Class vs Acceptance.

any conclusion since the claims for the insurance company is smaller compared to other insurance company claims. For other categories of Financial Class, claims get rejected more than accepted.

When insurance companies get claims from a health provider it checks the patient's eligibility and does the claim processing. Fig. 11 shows how the claim status is changed with carrier codes(Unique Id of Insurance Company)

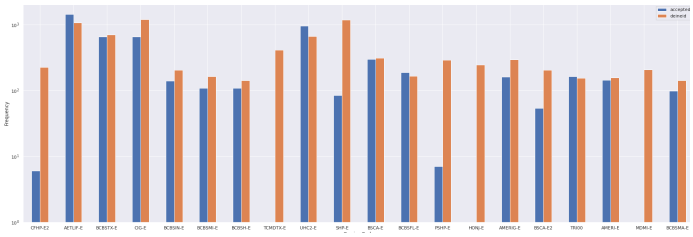

Fig. 11. Carrier Code vs Acceptance

Some insurance companies rejected all the claims they got. So we compared the financial class of the claims they rejected and accepting.

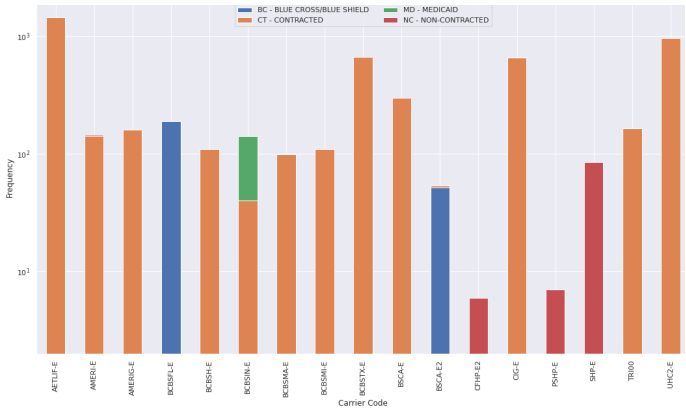

Fig. 12. Carrier Codes With Acceptance True

Some insurance companies rejected without regard to any financial class. Another insurance company has rejected one 'FinancialClass' related claim mostly than others.

CPT codes are the treatments performed by the health providers. We compared the CPT codes with Insurances and analyzed how they impact the claims hat get accepted and rejected.

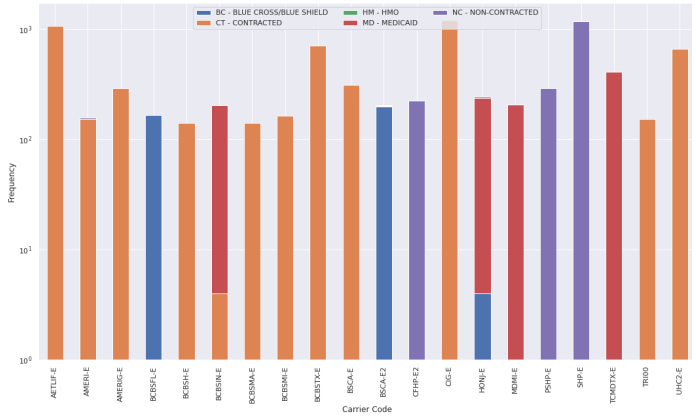

Fig. 13. Carrier Codes With Acceptance False

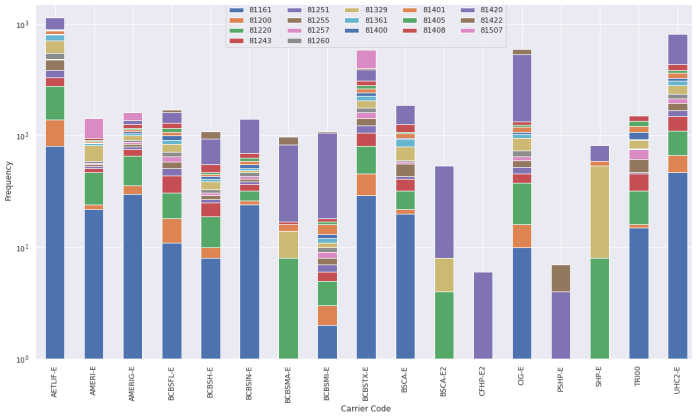

Fig. 14. Insurances vs CPT code for accepted claims

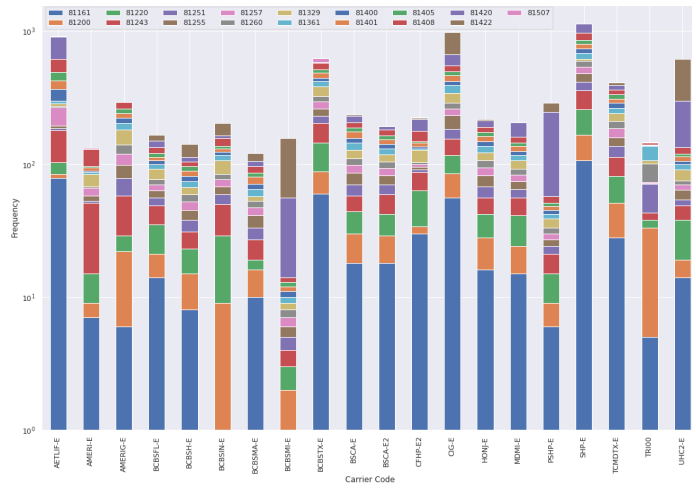

Fig. 15. Insurances vs CPT code for denied claims

CPT code 81161 is used in claims. CPT code 81251(purple) is mostly accepted by most of the carriers and CPT 81200(orange) is mostly denied by the carriers. Insurance companies have their own rules which CPT codes they can pay or not. IF we see carrier code BCBSIN-E has accepted the claims which have CPT 81161 than others.

By comparing the FinancialClass and CPT codes with Carrier Codes, both the patient's insurance category and treatment or test performed by health providers are impacted claim acceptance from Insurance companies.

D. Predictive Analysis

Prediction of claims acceptance or rejection is very important to enhance the operational workforce efficiency when collecting revenue. Predictions will help to cut off unnecessary expenditures that can happen in the claiming process. Further, it will help the operation employees to manage their work stress free individually while achieving their daily and weekly targets. Time consumption for an individual claims processing will be reduced.

In this section, prediction of the acceptance or rejection of claims is discussed using machine learning models like decision tree, random forest tree and gradient boosting.

And also when the claims get rejected, there will be a denied code for the rejection. Predicting the denied code will add value to the claiming process as the operations employees can take the necessary actions before rejecting the claims which is time effective. As the second step of the predictive analysis, denied code is predicted using a given data set and it leads to many future works for improving the accuracy of prediction.

Problem - Using the given data set, predict the claim processing results: accepted or denied. For the denied claims, the denied code will be predicted.

Process - The process of predictive analysis involves the steps such as pre-processing the given data set, identifying the importance of the features, splitting data into training and test sets, model selection and training, evaluating the trained models using accuracy, using recursive feature elimination to compare the accuracy and picking up the models with highest scores with respect to selected number of features.

Before creating classification model we did some preprocessing

1) Acceptance vs denied: The data has more than 20 thousand records and it contains accepted and denied claim records.First we did data cleaning by removing currency codes and changing the data types into correct one. To distinguish the accepted and denied ones we used Denied Code column in data set if it empty we categorized as accepted and if it has a denial code then categorized as denied status

2) timedelta: From the domain knowledge, it is known that claims should be billed to the insurance for the first time within a certain time period from the day that the patient has been taken treatments. Based on that a new derived feature is added to the data set called 'timedelta' which gives the days gap between the date of service (when the patient received treatments) and the first billed date.

3) Categorical data preprocessing: Some values were missing for the first billed date and the number of records were removed rather than processed with Imputations (e.g: mean/median, most frequent, zero/constant, k-NN) as very few records missed it. (134 records out of 22696 records).

4) Denied codes preprocessing: Though NPI field format was integer, it contains codes to uniquely identify doctors. It is converted as categorical data during preprocessing. sklearn.preprocessing. LabelEncoder is used for categorical data conversions before training.

When predicting the denied codes, there were 79 unique denied codes having frequencies in the range of 4091 to 1.

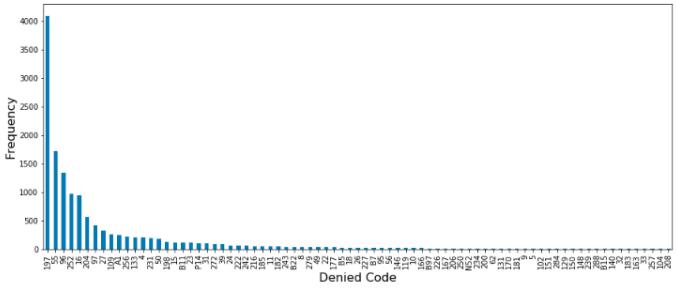

Fig. 16. Denied code frequencies

This data set is highly imbalanced and multi valued. Categorizing into many groups will reduce the business value to the organization. E.g: when a claim is predicted as denied and returns many denied codes as results will reduce it's value. And this data set is not enough for such analysis as the frequencies vary in large scale. Normalizing this data set removes some denied codes considering them as outliers.

Though transforming data with sigmoid function has little improvement, still it has neglected some values as outliers.

As neglecting some denial codes, considering as outliers, will direct the prediction towards inaccurate results. Log data transformations also resulted in the same as above. Therefore here another approach is followed where less frequency denied codes are grouped together. Denied code categorization is done as follows. thresholds = [20,30,40,50,70,100]

TABLE II
DATA SET COLUMNS

Category	*Frequency*
Other-category-1	Denied code freq ¡= 20
Other-category-2	20 ¡ Denied code freq ¡= 30
other-category-3	30 ¡ Denied code freq ¡= 40
other-category-4	40 ¡ Denied code freq ¡= 50
other-category-5	50 ¡ Denied code freq ¡= 70
other-category-6	60 ¡ Denied code freq ¡= 100

For splitting the data set, sklearn library is used and randomly selected 20 percent of data is taken as test data and 80 percent as train data.

Most straight forward attributes are removed before starting analysis such as Ticket Number, CaseCount, DateOfService, FirstBilledDate, PaymentPostDate, and DateOfEntry.

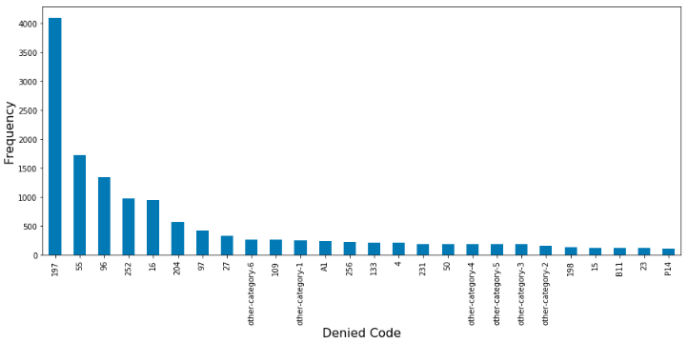

Fig. 17. categorized denied code frequencies

Mainly three machine learning algorithms are used to take higher accuracy training models: decision tree, random forest, gradient boosting.

Decision tree is a supervised machine learning algorithm where the data is continuously split according to a certain parameter. Less data cleaning required, data type is not a constraint, handling both numerical and categorical variables, being non-parametric method, and non-linear relationships between parameters which do not affect tree performance are some more reasons for the preferences. The ability to handle multi-output problems is another reason to be selected for denied code prediction.

Random Forest Algorithm gives higher accuracy and maintains the accuracy of a large proportion of data handling higher dimensionality. It has reduced over fitting trees in the model. Here higher accuracy is achieved with Random Forest Algorithms.

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. Rather than using a separated model for predicting acceptance or denial, it gives good results along with recursive feature elimination.

For the model training, 15 attributes are used after preprocessing steps. Though the trained models give good results, recursive feature elimination is used to pick the optimal number of attributes for the analysis. Selected number of features vs accuracy score is shown in the graph below. In Recursive feature elimination (RFE) which is a feature selection method, it eliminates that fits a model and removes the weakest feature (or features) until the specified number of features is reached. Results are shown in below For the acceptance and denied prediction, it is done with gradient boosting, random forest and decision tree models.

III. RESULTS

All trained models' accuracy values are plotted in the same graph as shown below.

When considering decision trees alone results as follows.

For random forests, two graphs can be drawn with different RFE algorithms.

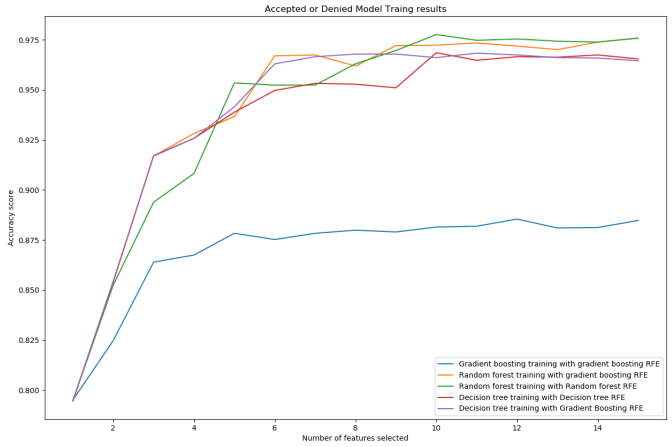

Fig. 18. Overall Results

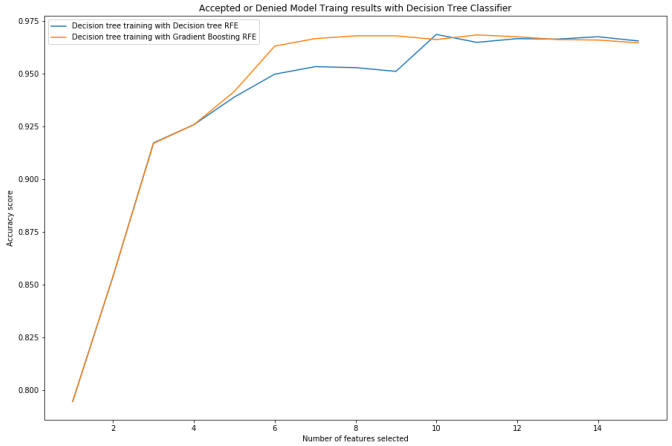

Fig. 19. Decision Tree Results

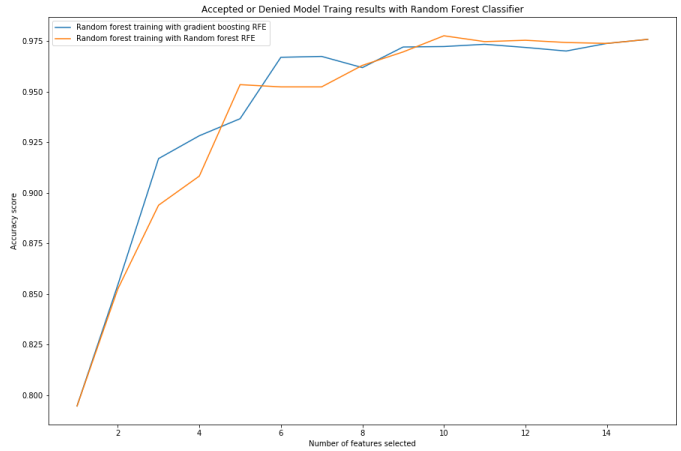

Fig. 20. Random Forest Results

When predicting denial codes using the decision tree, the accuracy score graph will be as below. After comparing the classification reports of trained models with optimal number of features, the best value is picked as 0.972 accuracy with random forest model with RFE using random forest classifiers. 12 features were selected there.

TABLE III
DATA SET COLUMNS

	Random forest	***Decision tree***
Random forest RFE	0.972	0.964
Decision tree RFE	0.971	0.965
Gradient boosting RFE	0.972	0.968

When predicting denied code, highest accuracy is achieved when 14 features are selected. The maximum accuracy score is 0.898. Precision and recall varied from 0.5 to 1.0 for different denied codes. As the data set is highly imbalanced and having only one record for many denied code, this analysis is still not directed to an end.

FUTURE WORKS

With predictive analysis, denied codes can be predicted using a denied claims set. As there are hundreds of common denied codes, choosing balanced dataset, multi value classification can be done using deep learning frameworks like keras in future works. Also we used the data without any patient details also. If one insurance denies the claim then the claim can be sent to secondary insurance of that patient. So there can be a recommendation model that can also be created for recommending the healthcare provider to get the claim accepted and quickly get revenue for it.

CONCLUSION

In this research, healthcare providers' claims data is analysed using descriptive analysis, reasons for denied claims are investigated using diagnosis analysis, finally for given claims acceptance or denial status and denied codes for denied claims are predicted using some machine learning models. Those findings will help for effective operations inside claims processing systems by enhancing organizations profits. In descriptive analysis, highly correlated features are identified and feature data distribution is analysed. Feature extractions for model training picked up based on the descriptive analysis. Moving to the diagnosis analysis, the relationship between CPTs and CarrierCodes are identified concluding some CPTs getting denied always by certain insurance(CarrierCodes). Finally in the predictive analysis random forest algorithm trained model gives 0.972 accuracy over predicting the denied or acceptance status of given claims. For predicting denied codes, though 0.898 accuracy is achieved, it can be trained using a balanced data set and trained using deep learning algorithms for better results.