Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, No. 3, April : 2023

MEDICARE HOSPITAL FRAUD DETECTION USING SUPERVISED LEARNING

G. Kalpana Devi Assistant Professor,
Ch. Likitha Reddy, Y. Madhurvashini, Y. Anuhya UG Scholar,
Department of Computer Science and Engineering, CMR Technical Campus, Kandlakoya, Hyderabad, Telangana, India

ABSTRACT
With the overall increase in the elderly population come additional, necessary medical needs and costs. Medicare is a U.S. healthcare program that provides insurance, primarily to individuals 65 years or older, to offload some of the financial burden associated with medical care. Even so, healthcare costs are high and continue to increase. Fraud is a major contributor to these inflating healthcare expenses. Our paper provides a comprehensive study leveraging machine learning methods to detect fraudulent Medicare providers. We use publicly available Medicare data and provider exclusions for fraud labels to build and assess three different learners. In order to lessen the impact of class imbalance, given so few actual fraud labels, we employ Logistic Regression creating two class distributions. Our results show that the other algorithms have poor performance compared with Logistic Regression. Learners have the best fraud detection performance, particularly for the 80:20 class distributions with average AUC scores, respectively, and low false negative rates. We successfully demonstrate the efficacy of employing machine learning Models to detect Medicare fraud.

Keywords: Medicare Hospitals, fraud detection, regression, supervised learning.

1. INTRODUCTION
Health insurers receive millions of claims per year. Given that information asymmetries between the principal (insurer) and the agents (health care providers and the insured) can lead to moral hazard, insurance companies face the choice of either paying out insurance claims immediately without any adjustments or reviewing claims that are suspicious. The most common method for undertaking the latter involves manually auditing claims data, which is a time-consuming and expensive process. Machine learning models can greatly cut auditing costs by automatically screening incoming claims and flagging up those that are deemed to be suspicious – i.e., potentially incorrect – for subsequent manual auditing. Insurance fraud is a widespread and high-priced problem for each policyholder and insurance businesses in all sectors of the coverage industry [1]. India is one of the quickest developing economies in the international, has a burgeoning middle class, and has witnessed a giant upward push within the demand for medical insurance products [2]. Over the last 10 years, the medical health insurance industry has grown at a capital annual compounded boom rate of around 20%. But, with the exponential growth inside the industry, there has additionally been an extended prevalence of frauds within the us. Health insurance fraud contains a huge range of illicit practices and unlawful acts concerning intentional deception or misrepresentation. Statistics mining has an extraordinary effect in enhancing healthcare fraud detection system. Statistics mining has been implemented to fraud detection in both the way i.e., Supervised, and non-supervised way. Information mining strategies and its software for fraud detection in fitness care zones are defined below.
In latest years, systems for processing digital claims were increasingly carried out to mechanically perform audits and reviews of claims information. These systems are designed for figuring out regions requiring unique interest together with faulty or incomplete data entry, duplicate claims, and medically non-blanketed services [3]. Even though these structures may be used to locate sure varieties of fraud, their fraud detection competencies are typically restrained for the reason that detection particularly is predicated on pre-defined easy guidelines special via domain professionals.

2. LITERATURE SURVEY

Herland et. al [4] employed an approach to predict a physician's expected specialty based on the type and number of procedures performed. From this approach, they generate a baseline model, comparing Logistic Regression and Multinomial Naive Bayes, in order to test and assess several new approaches to improve the detection of U.S. Medicare Part B provider fraud. These results indicate that this proposed improvement strategies (specialty grouping, class removal, and class isolation), applied to different medical specialties, have mixed results over the selected Logistic Regression baseline model's fraud detection performance. Through this work, they demonstrate that improvements to current detection methods can be effective in identifying potential fraud. Hancock et. al [5] conducted experiments with three Big Data Medicare Insurance Claims datasets. The experiments are exercises in Medicare fraud detection. They show that for each dataset, they obtain better performance from LightGBM and CatBoost classifiers with tuned hyperparameters. Since some features of the data, they are working with are high cardinality categorical features, they have an opportunity to try different encoding techniques in these experiments. They find that across the different encoding techniques, hyperparameter tuning Provides an improvement in the performance of both LightGBM and CatBoost. Bauder et. al [6] focused on the detection of Medicare Part B provider fraud which involves fraudulent activities, such as patient abuse or neglect and billing for services not rendered, perpetrated by providers and other entities who have been excluded from participating in Federal healthcare programs. Based on the performance and significance testing results, they posit that retaining more of the majority class information leads to better Medicare Part B fraud detection performance over the balanced datasets across the majority of learners. Herland et. al [7] focused on the detection of Medicare fraud using the following CMS datasets: (1) Medicare Provider Utilization and Payment Data: Physician and Other Supplier (Part B), (2) Medicare Provider Utilization and Payment Data: Part D Prescriber (Part D), and (3) Medicare Provider Utilization and Payment Data: Referring Durable Medical Equipment, Prosthetics, Orthotics and Supplies (DMEPOS). Additionally, they create a fourth dataset which is a combination of the three primary datasets.

Arunkumar et. al [8] provides an extensive study of detecting fraudulent claims in healthcare insurance by leveraging machine learning algorithms. By using the publicly available Medicare dataset, they are able to classify as fraud and non-fraud providers. Moreover, synthetically minority oversampling technique is used to avoid the class imbalance problem. Furthermore, a hybrid approach is used which is based on clustering and classification. Additionally, they have used other machine learning algorithms to check the efficiency of the best-suited algorithm. Chen et. al [9] proposed VAERM coupled with active learning strategy can assist healthcare industry experts to conduct cost-effective fraud investigation. Finally, they propose an online model updating method to reduce the computation and memory requirement while preserving the predictive performance. The proposed framework is tested in a real-world dataset, and it empirically outperforms the state-of-the-art methods in both automatic fraud detection and fraud investigation tasks. Yao et. al [10] proposed the Bagging algorithm based on the weighted threshold method named WTBagging and made ten model combinations using Bagging and WTBagging algorithms. The data is cleaned and sampled to construct three datasets with different class distributions. The 5-fold cross-validation process was applied to the model training and repeated ten times, and the F1 value was the performance metric to evaluate the model combination. The results show that the model combinations of the WTBagging achieved the highest F1 values under all datasets.

3. PROPOSED SYSTEM

Dataset description

57:Columns:PotentialFraud,BeneID,ClaimID,ClaimStartDt,ClaimEndDt,InscClaimAmtReimbursed, AttendingPhysician,OperatingPhysician,OtherPhysician,ClmDiagnosisCode_1,ClmDiagnosisCode_2 ,ClmDiagnosisCode_3,ClmDiagnosisCode_4,ClmDiagnosisCode_5,ClmDiagnosisCode_6,ClmDiag

nosisCode_7,ClmDiagnosisCode_8,ClmDiagnosisCode_9,ClmDiagnosisCode_10,ClmProcedureCode_1,ClmProcedureCode_2,ClmProcedureCode_3,ClmProcedureCode_4,ClmProcedureCode_5,ClmProcedureCode_6,DeductibleAmtPaid,ClmAdmitDiagnosisCode,AdmissionDt, DischargeDt,DiagnosisGroupCode,AdmitForDays,DOB,DOD,Gender,Race,RenalDiseaseIndicator,Satate,County,NoOfMonths_PartACov,NoOfMonths_PartBCov,ChronicCond_Alzheimer,ChronicCond_Heartfailure,ChronicCond_KidneyDisease,ChronicCond_Cancer,ChronicCond_ObstrPulmonary, ChronicCond_Depression,ChronicCond_Diabetes,ChronicCond_IschemicHeart,ChronicCond_Osteoporasis,ChronicCond_rheumatoidarthritis,ChronicCond_stroke,IPAnnualReimbursementAmt,IPAnnualDeductibleAmt,OPAnnualReimbursementAmt, OPAnnualDeductibleAmt, Age,WhetherDead
558212- Rows

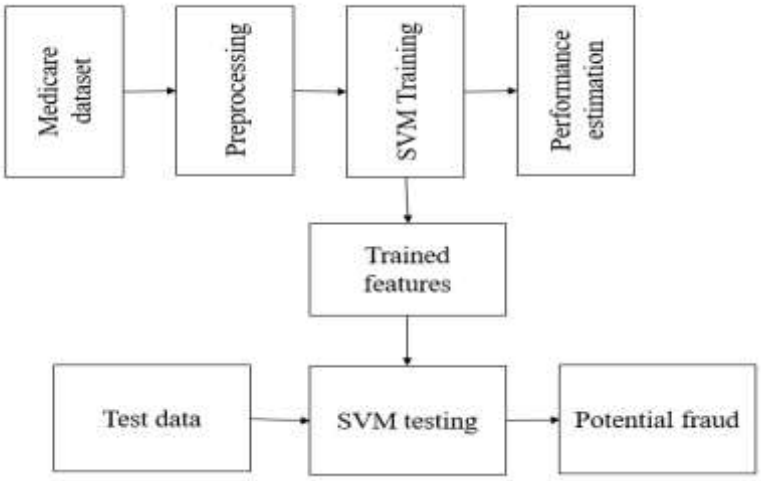


Fig. 1: Block diagram of proposed system.

Data Preprocessing in Machine learning

Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. When creating a machine learning project, it is not always a case that we come across clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put it in a formatted way. So, for this, we use data pre-processing tasks.

Why do we need Data Pre-processing?

Real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data pre-processing requires tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

- Getting the dataset
- Importing libraries
- Importing datasets
- Finding Missing Data
- Encoding Categorical Data
- Splitting dataset into training and test set
- Feature scaling

Splitting the Dataset into the Training set and Test set

In machine learning data pre-processing, we divide our dataset into a training set and test set. This is one of the crucial steps of data pre-processing as by doing this, we can enhance the performance of our machine learning model.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, No. 3, April : 2023

Suppose if we have given training to our machine learning model by a dataset and we test it by a completely different dataset. Then, it will create difficulties for our model to understand the correlations between the models.

If we train our model very well and its training accuracy is also very high, but we provide a new dataset to it, then it will decrease the performance. So we always try to make a machine learning model which performs well with the training set and also with the test dataset. Here, we can define these datasets as:

Training Set: A subset of dataset to train the machine learning model, and we already know the output.
Test set: A subset of dataset to test the machine learning model, and by using the test set, model predicts the output.

Support Vector Machine Algorithm

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

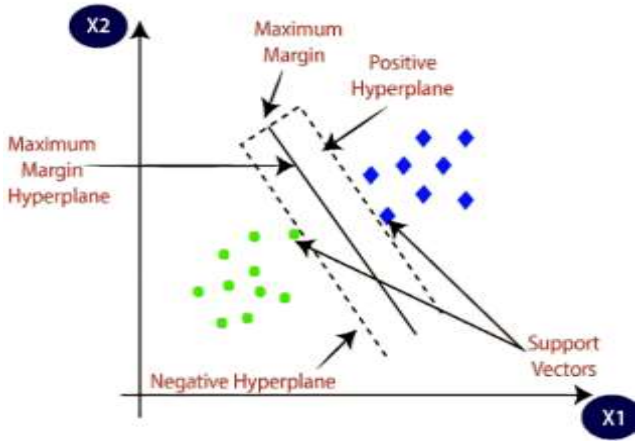


Fig. 2: An example of SVM model.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:

Advantages of proposed system

- SVM works relatively well when there is a clear margin of separation between classes.
- SVM is more effective in high dimensional spaces.
- SVM is effective in cases where the number of dimensions is greater than the number of samples.
- SVM is relatively memory efficient.

4. RESULTS AND DISCUSSION

Module implementation

- Medicare Dataset

- Preprocessing
- SVM Training
- Performance estimation
- Trained features
- Test data
- SVM Testing
- Potential Fraud

Sample Training Data

	Provider	PotentialFraud	BeneID	ClaimID	ClaimStartDt	ClaimEndDt	InscClaimAmtReimbursed	AttendingPhysician	OperatingPhysician	OtherPhysic
0	PRV51001	No	BENE11727	CLM733300	2009-12-17	2009-12-17	20	PHY383007	NaN	PHY383
1	PRV51001	No	BENE24646	CLM372475	2009-05-22	2009-05-23	700	PHY405682	NaN	PHY402
2	PRV51001	No	BENE31617	CLM748221	2009-12-28	2009-12-28	900	PHY345302	NaN	N
3	PRV51001	No	BENE32715	CLM272936	2009-03-29	2009-03-30	500	PHY318842	PHY392882	N
4	PRV51001	No	BENE36012	CLM58316	2009-07-04	2009-07-08	36000	PHY340163	NaN	N

5 rows × 58 columns

OtherPhysician	...	ChronicCond_IschemicHeart	ChronicCond_Osteoporasis	ChronicCond_rheumatoidarthritis	ChronicCond_stroke	IPAnnualReimbursementAmt	IF
PHY383007	...	1	0	0	0	0	
PHY402512	...	1	1	0	1	0	
NaN	...	1	0	1	0	0	
NaN	...	1	0	0	0	2020	
NaN	...	1	0	0	1	36000	

IPAnnualReimbursementAmt	IPAnnualDeductibleAmt	OPAnnualReimbursementAmt	OPAnnualDeductibleAmt	Age	WhetherDead
0	0	300	110	80.0	0.0
0	0	720	10	67.0	0.0
0	0	1380	370	76.0	0.0
2020	1068	6700	2700	74.0	0.0
36000	1068	3520	140	69.0	0.0

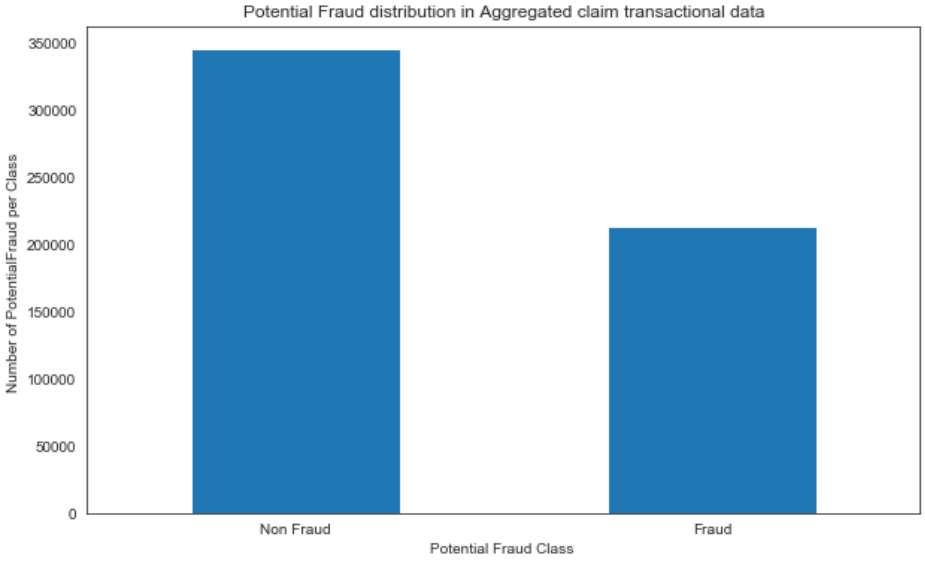
List of attributes

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, No. 3, April : 2023

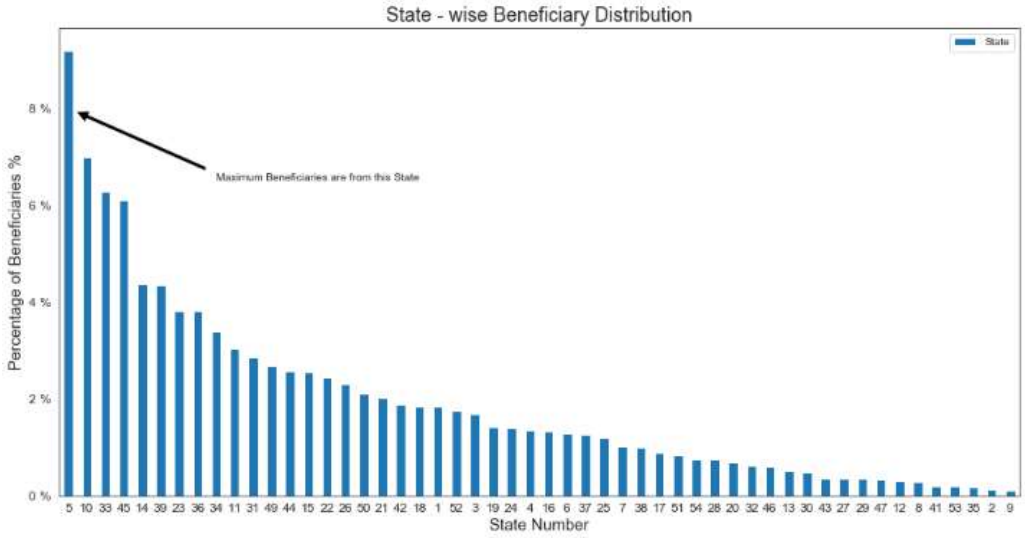
```
list(medicare_fraud.columns)
```

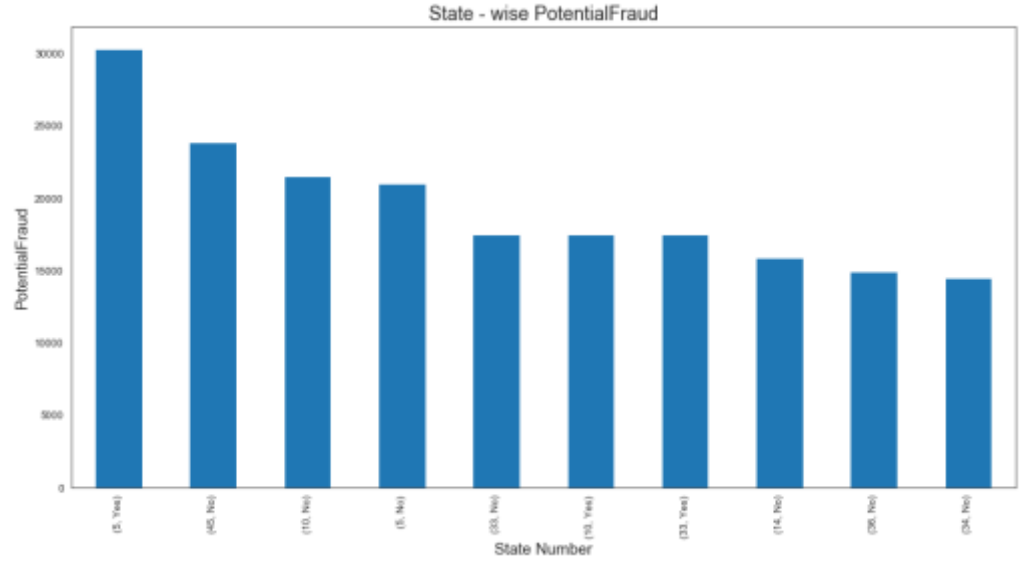
```
['Provider',
 'PotentialFraud',
 'BeneID',
 'ClaimID',
 'ClaimStartDt',
 'ClaimEndDt',
 'InscClaimAmtReimbursed',
 'AttendingPhysician',
 'OperatingPhysician',
 'OtherPhysician',
 'ClmDiagnosisCode_1',
 'ClmDiagnosisCode_2',
 'ClmDiagnosisCode_3',
 'ClmDiagnosisCode_4',
 'ClmDiagnosisCode_5',
 'ClmDiagnosisCode_6',
 'ClmDiagnosisCode_7',
 'ClmDiagnosisCode_8',
 'ClmDiagnosisCode_9',
 'ClmDiagnosisCode_10',
 'ClmProcedureCode_1',
 'ClmProcedureCode_2',
 'ClmProcedureCode_3',
 'ClmProcedureCode_4',
 'ClmProcedureCode_5',
 'ClmProcedureCode_6',
 'DeductibleAmtPaid',
 'ClmAdmitDiagnosisCode',
 'AdmissionDt',
 'DischargeDt',
 'DiagnosisGroupCode',
 'AdmitForDays',
 'DOB',
 'DOD',
 'Gender',
 'Race',
 'RenalDiseaseIndicator',
 'State',
 'County',
 'NoOfMonths_PartACov',
 'NoOfMonths_PartBCov',
 'ChronicCond_Alzheimer',
 'ChronicCond_Heartfailure',
 'ChronicCond_KidneyDisease',
 'ChronicCond_Cancer',
 'ChronicCond_ObstrPulmonary',
 'ChronicCond_Depression',
 'ChronicCond_Diabetes',
 'ChronicCond_IschemicHeart',
 'ChronicCond_Osteoporasis',
 'ChronicCond_rheumatoidarthritis',
 'ChronicCond_stroke',
 'IPAnnualReimbursementAmt',
 'IPAnnualDeductibleAmt',
 'OPAnnualReimbursementAmt',
 'OPAnnualDeductibleAmt',
 'Age',
 'WhetherDead']
```

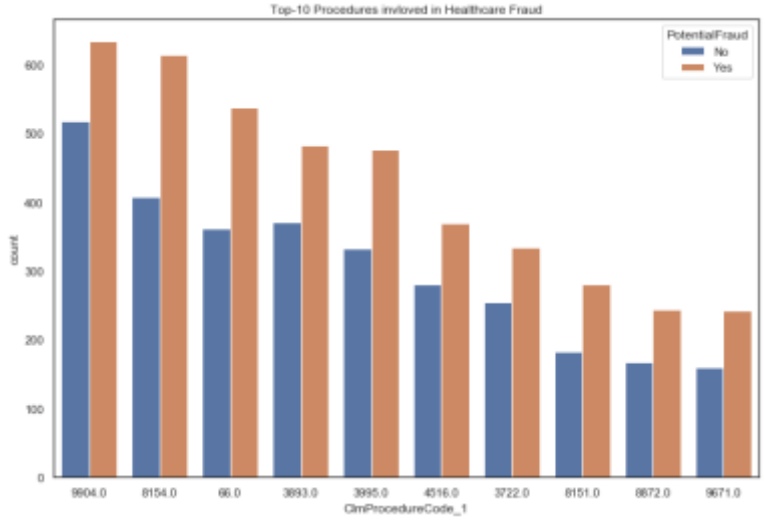
Exploratory Data Analytics



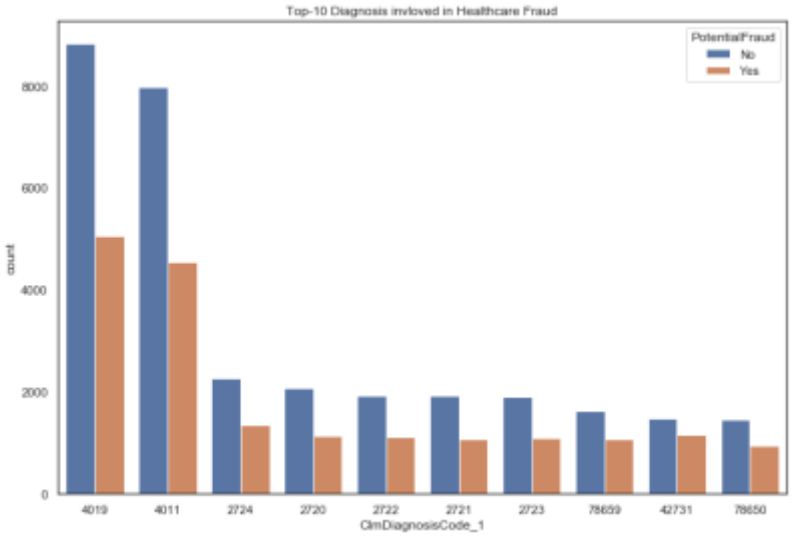
Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, No. 3, April : 2023

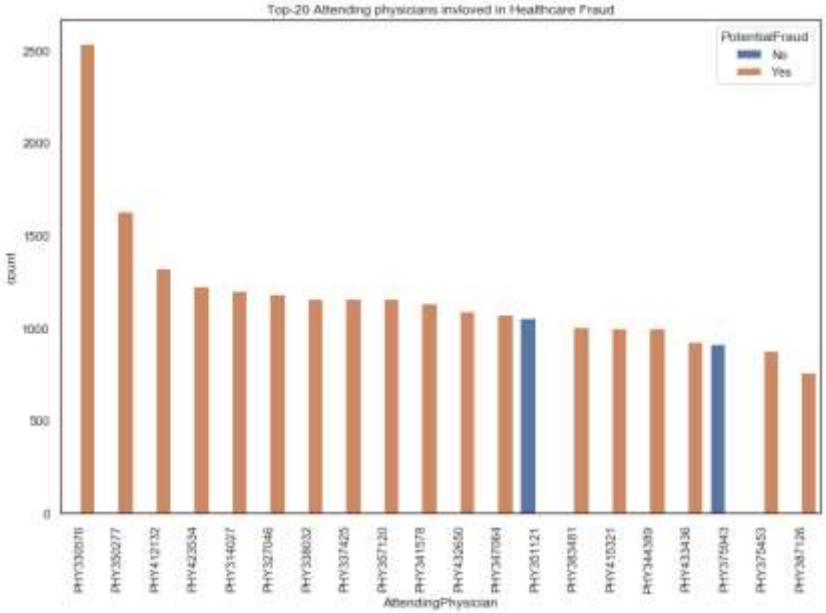


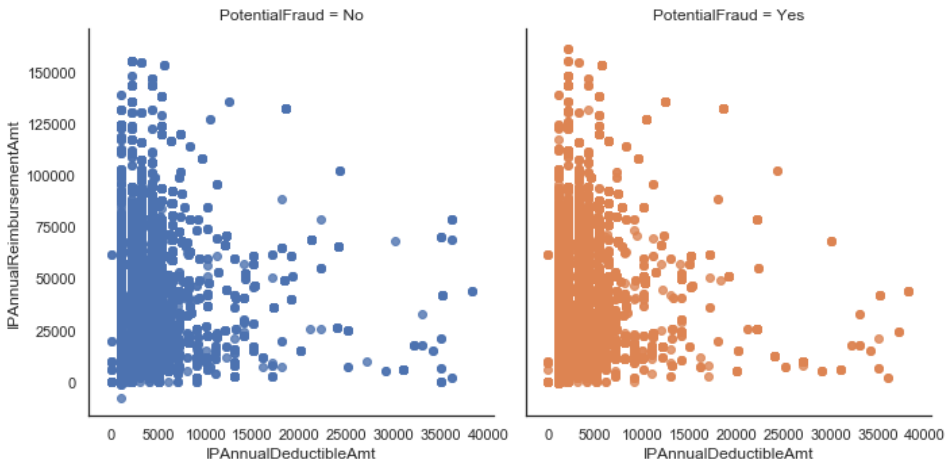




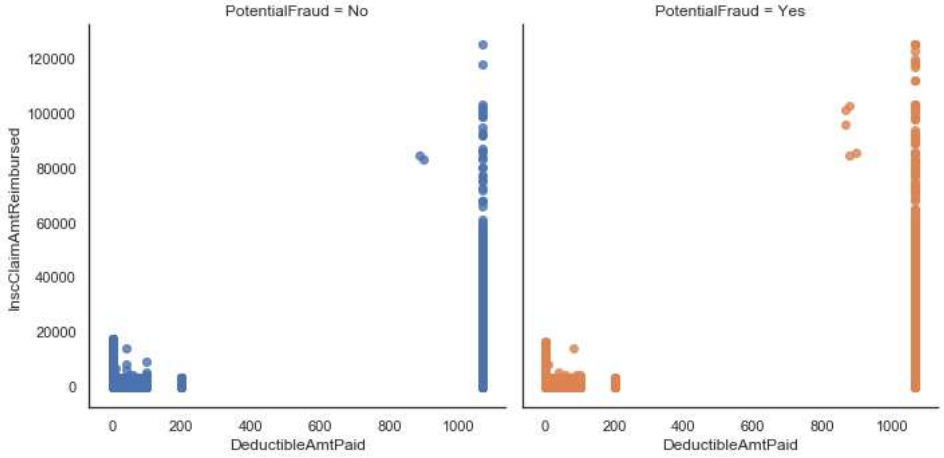
Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, No. 3, April : 2023



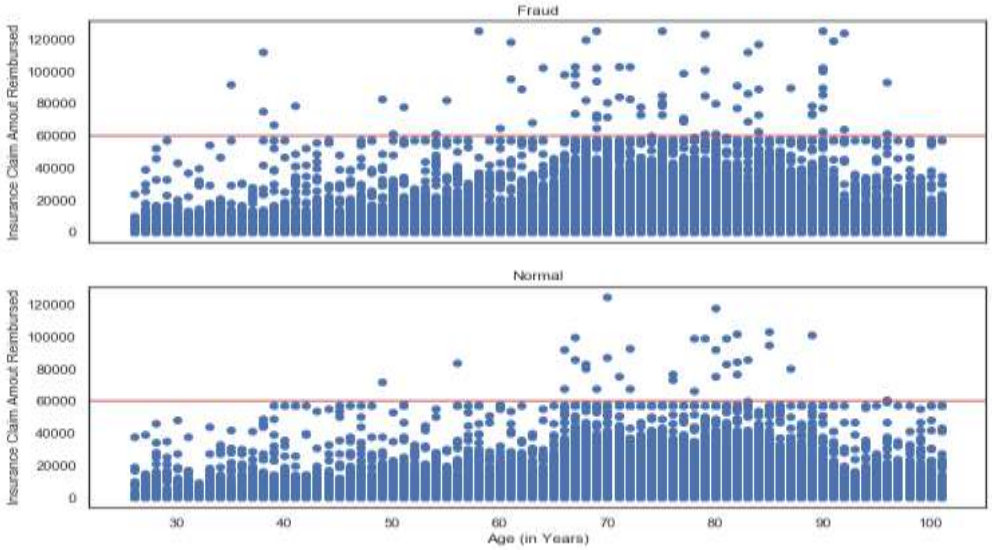




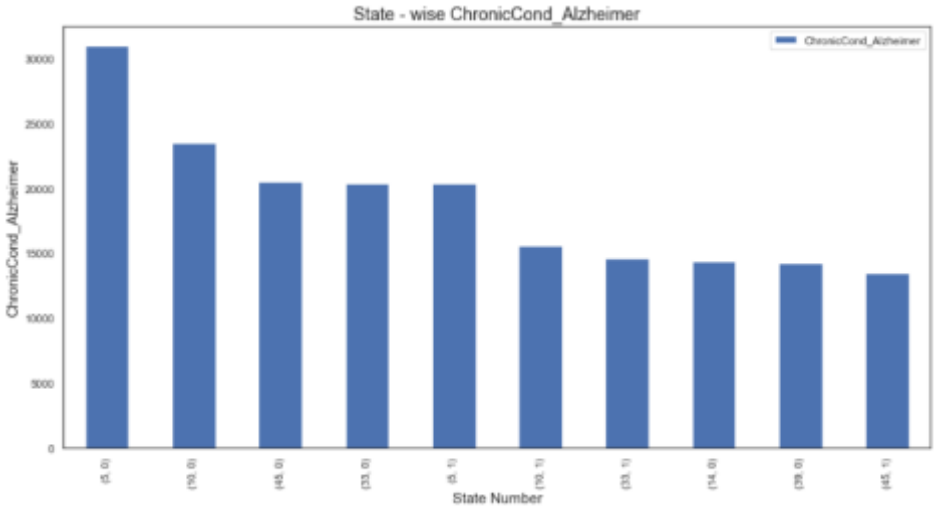
Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, No. 3, April : 2023



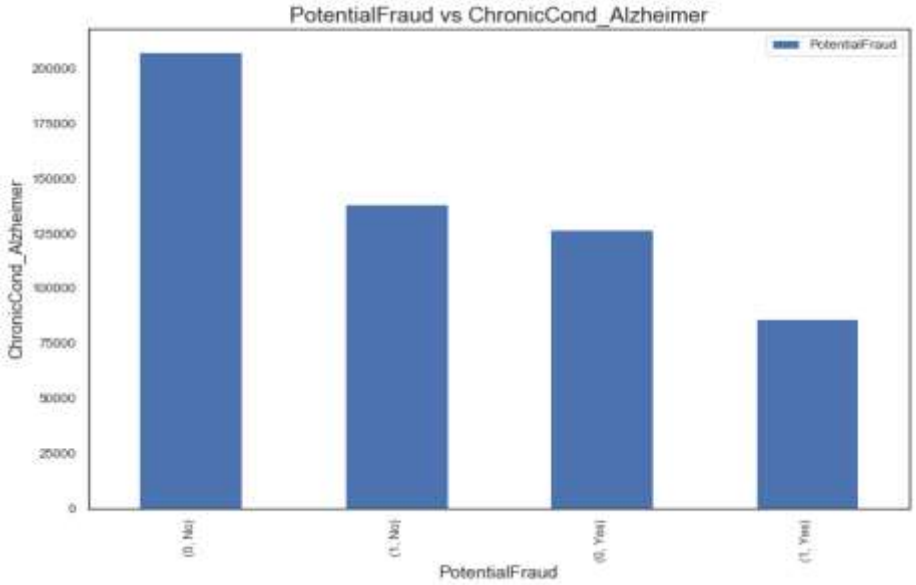
Insurance Claim Amount Reimbursed Vs Age

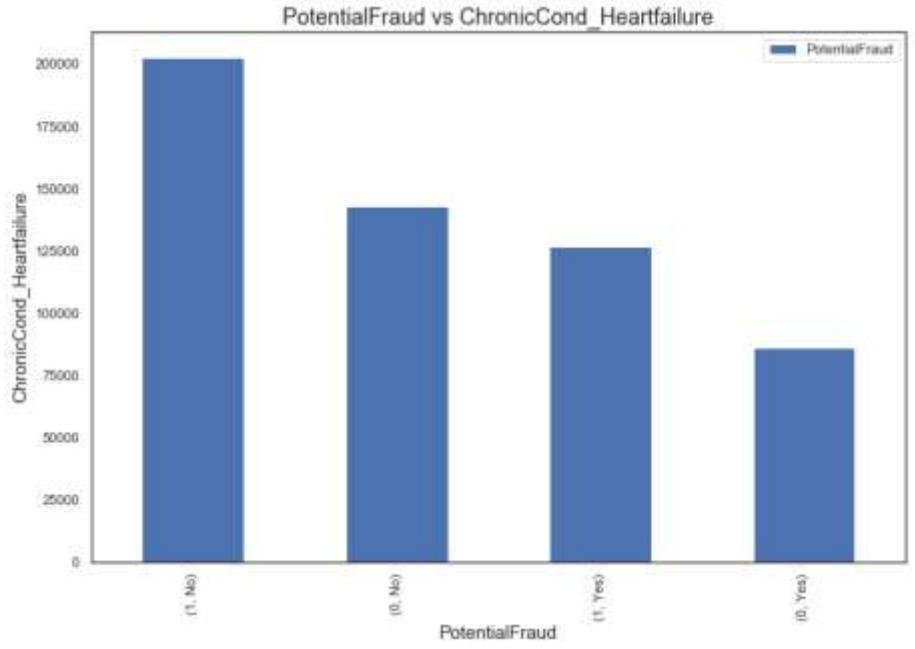


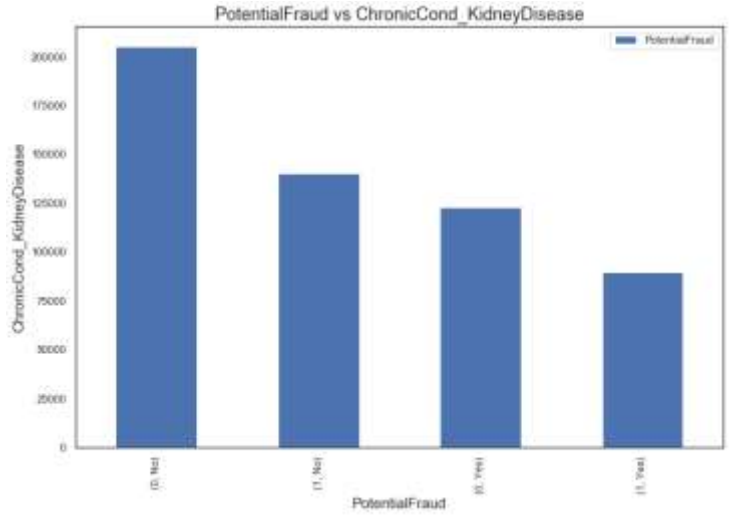
State - wise ChronicCond_Alzheimer



Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, No. 3, April : 2023







Training and testing data

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, No. 3, April : 2023

```
X_train : (3787, 156)
y_train : (3787,)
X_test : (1623, 156)
y_test : (1623,)
```

Logistic regression report

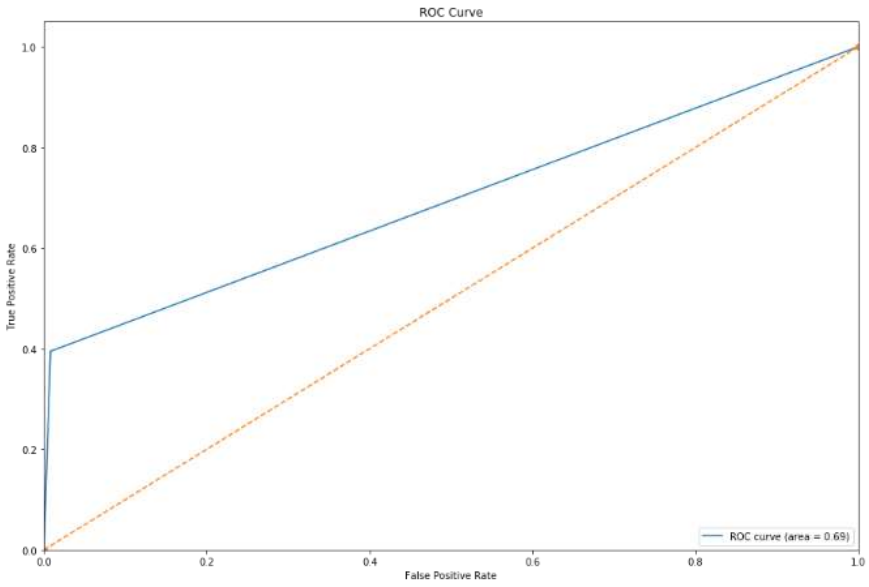
```
print(classification_report(y_test, y_predict))
```

	precision	recall	f1-score	support
0	0.97	0.92	0.95	1471
1	0.49	0.74	0.59	152
accuracy			0.90	1623
macro avg	0.73	0.83	0.77	1623
weighted avg	0.93	0.90	0.91	1623

SVM report

```
print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
0	0.94	0.99	0.97	1471
1	0.83	0.39	0.54	152
accuracy			0.94	1623
macro avg	0.89	0.69	0.75	1623
weighted avg	0.93	0.94	0.93	1623



Prediction on test data

	Provider	PotentialFraud
0	PRV51002	Y
1	PRV51006	Y
2	PRV51009	Y
3	PRV51010	Y
4	PRV51018	Y

5. CONCLUSION AND FUTURE SCOPE

This work aimed at developing a novel fraud detection model for insurance claims processing based on genetic support vector machines, which hybridizes and draws on the strengths support vector machines. SVMs have been considered preferable to other classification techniques due to several

advantages. With other notable advantages, it has a nonlinear dividing hyper plane, which prevails over the discrimination within the dataset. The generalization ability of any newly arrived data for classification was considered over other classification techniques.

5.1 Future Scope

The proposed methodology provides the information that Random Forest performs better than Sequential CNN. The drawback of this methodology is that anyone would expect Sequential CNN can outperform any of the conventional ML methodologies, but it is not happening here. It may happen because the dataset is not enough to train and identify the hidden patterns to predict the future or upcoming data and the initialization of weights was very random that might affect the training process. It can be further improved in two ways. The first way is to tune the hyperparameters through optimization, and the second method is to apply the transfer learning methodology so that the performance of the proposed methodology is improved to detect the fraud transaction through Medicare in the healthcare sector.

REFERENCES

[1] Lakshman Narayana Vejendla and A Peda Gopi, (2019)," Avoiding Interoperability and Delay in Healthcare Monitoring System Using Block Chain Technology", Revue d'Intelligence Artificielle, Vol. 33, No. 1, 2019, pp.45-48.

[2] Gopi, A.P., Jyothi, R.N.S., Narayana, V.L. et al. (2020), "Classification of tweets data based on polarity using improved RBF kernel of www.jespublication.com PageNo:482 SVM". Int. j. inf. tecnol. (2020)

[3] Lakshman Narayana Vejendla and A Peda Gopi, (2017)," Visual cryptography for gray scale images with enhanced security mechanisms", Traitement du Signal, Vol.35, No.3-4, pp.197-208. DOI: 10.3166/ts.34.197-208

[4] Herland, M., Bauder, R.A. & Khoshgoftaar, T.M. Approaches for identifying U.S. medicare fraud in provider claims data. Health Care Manag Sci 23, 2–19 (2020). https://doi.org/10.1007/s10729-018-9460-8

[5] Hancock, J.T., Khoshgoftaar, T.M. Hyperparameter Tuning for Medicare Fraud Detection in Big Data. SN COMPUT. SCI. 3, 440 (2022). https://doi.org/10.1007/s42979-022-01348-x

[6] Bauder, R.A., Khoshgoftaar, T.M. The effects of varying class distribution on learner behavior for medicare fraud detection with imbalanced big data. Health Inf Sci Syst 6, 9 (2018). https://doi.org/10.1007/s13755-018-0051-3

[7] Herland, M., Khoshgoftaar, T.M. & Bauder, R.A. Big Data fraud detection using multiple medicare data sources. J Big Data 5, 29 (2018). https://doi.org/10.1186/s40537-018-0138-3

[8] Arunkumar, C., Kalyan, S., Ravishankar, H. (2021). Fraudulent Detection in Healthcare Insurance. In: Sengodan, T., Murugappan, M., Misra, S. (eds) Advances in Electrical and Computer Technologies. ICAECT 2020. Lecture Notes in Electrical Engineering, vol 711. Springer, Singapore. https://doi.org/10.1007/978-981-15-9019-1_1

[9] J. Chen, X. Hu, D. Yi, J. Li and M. Alazab, "A Variational AutoEncoder-Based Relational Model for Cost-Effective Automatic Medical Fraud Detection," in IEEE Transactions on Dependable and Secure Computing, 2022, doi: 10.1109/TDSC.2022.3187973.

[10] J. Yao, S. Yu, C. Wang, T. Ke and H. Zheng, "Medicare Fraud Detection Using WTBagging Algorithm," 2021 7th International Conference on Computer and Communications (ICCC), 2021, pp. 1515-1519, doi: 10.1109/ICCC54389.2021.9674545.