Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 8644 - 8654
Received 16 February 2021; Accepted 08 March 2021.

Machine Learning Based Approaches for Healthcare Fraud Detection: A Comparative Analysis

S. Lavanya[1]*,S.ManojKumar[2],P.Mohan Kumar[3]
[1]Assistant Professor, Department of Information Technology, Karpagam College of Engineering, Coimbatore, Tamilnadu, India
[2]Professor, ,Department of Information Technology, Karpagam College of Engineering, Coimbatore, Tamilnadu, India
3Assistant Professor, Department of Biomedical Engineering, PSNA College of Engineering and Technology, Dindigul, Tamilnadu, India
*lavanya.skar29@gmail.com

ABSTRACT

Fraudulent activities in healthcare are exponentially increasing and it is the big burden to society. Mitigation of healthcare fraud is one of the most desirable artificial intelligence (AI) services since many organizations experiencing huge benefits which is because of machine learning (ML) systems that spot and prevent fraud in real-time. Predictive analytics is a branch of data analytics designed at making predictions about future outcomes based on past data and analytics approaches such as machine learning. In this article, a comparative analysis on healthcare fraud detection methods is done by using various machine learning algorithms. It clearly shows that Multilayer Perceptron Algorithm provides significant performance when compared to the other approaches.

Keywords
Artificial Intelligence, Deep Learning, Machine Learning, Medical Fraud, Predictive Analytics

Introduction

The term fraud is referred as the crime of obtaining money of the organization or single person without necessarily in the lead to direct legal impact. In this competing world, fraud will become an overcritical problem if it is typical and if the prevention policies are not done. Fraud detection, if done manually using the process like screening/checking then it will become typical process that does not lead to the prevention, so the automation is the only tactics to do this in efficient way. So fraud detection can be effectually done using artificial intelligence to perform predictive analysis in all areas.

Health is the rudimentary human right and a universal social goal. It is very important for the accomplishment of primary human requirement and for the betterment of life. To ensure this right to each and every citizen, India has many health care schemes. Per year India spends more than 1.4% of GDP on healthcare. The fraud targets are captivated by proportion of the health department and the limitless proportion of finance involved in that. According to a recent survey 15% of total claims made is estimated as the false claim and India is losingRs 600- Rs 800 crores incurred on fraudulent claim annually [5-1 ref]. Hence in order to make Healthcare a feasible sector, it is essential to focus on elimination or minimization this fraud.

- Some of the healthcare frauds which may happen are,
- Billing for service not accomplished
- Misrepresentation of type of service provided
- Billing for the costlier medical equipment.
- Submitting the duplicate bills.
- Billing for unnecessary procedures.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 8644 - 8654
Received 16 February 2021; Accepted 08 March 2021.

The manual process to be followed for fraud detection is to investigate around fraud index. There are usually two methods for this, first - framing rules that define that the instance required to be sent for analysis, second - creating a checklist with outcome for various indicators of fraud.
Fraud detection in Medicare claims seems to be very tedious and unsuitable when it is done manual auditing and investigating when correlated to machine learning and data mining approaches. There are several challenges involved in Fraud detection using healthcare data. The situation is specified by the four Vs of big data: Volume, Variety, Velocity, and Veracity [3].
In order to make Healthcare a feasible sector, it is essential to focus on elimination or minimization this fraud. To mitigate the health care frauds one of the most desirable artificial intelligence (AI) services may be used, since many organizations experiencing huge benefits which is because of machine learning (ML) systems that spot and prevent fraud in real-time.
This health care fraud detection can be done using many automation methods like machine learning approach using available claim data[1][21][22], rule based system[2] [20], neural network[3][18], advanced analytics using machine learning[4], supervised and unsupervised learning[6], decision support system[7],data mining[8][23], computation intelligence model[9], deep learning[10][19] and big data analytics[11]. . So fraud detection can be effectually done using artificial intelligence to perform predictive analysis in all areas. In this paper we are going to compare the performance of predictive analysis for health care fraud detection with all other available methods.

Related Works

There are enormous studies conducted relating to the Healthcare fraud detection. In this section we are going to discuss some of the methodologies performed for the healthcare fraud detection using automotive techniques.
Our study focuses only on Health care providers and does not include financial frauds like fraud observation in other fields such as credit card deceit, finance concealment, tele-communication deceit, system intrusion and scientific deceit. In this study we are going to compare the performance of predictive analysis for health care fraud detection with all other available methods.
A recent study by Richard et al.(2018) employs machine learning method with excluded provider label for the observation of human care deceit using available medicare data by random undersampling. This study concludes that decision tree and logistics regression learners have the best fraud detection performance [1] [17] [24].
In [2], Justin et al.(2019) using the publicly available healthcare data detected the fraud executed in that field using neural network. The researchers have used random over sampling (ROS), random under sampling (RUS), and amalgam of ROS–RUS. In this study Justin et al. used baseline neural network planning and its framework to discover into a casual look-in procedure that judge models on a validation groups.
A study by Vrinda et al.(2018) address approaches which make use of data except the claim data, for the fraud detection process.in this paper Self Organising Map(SOM) is used for unsupervised clustering and PRIDIT - Principal Component Analysis of RIDIT(Relative to an Identified Distribution Integral Transformative) is used to generate the score to indicate the severity of the fraud[3].
Fraud detection in health insurance sector by Vipula et al. (2015) hybrid model in data mining. In this study data sets used are dynamic and new data set, so for clustering of data Evolving Clustering Method (ECM) is applied and for classification Support Vector Machine(SVM) is used. In this work the data are first segregated based to the deceit type and then classified to detect the deceit assert [4].

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 8644 - 8654
Received 16 February 2021; Accepted 08 March 2021.

In a preliminary study, Jing li et al. (2007) made a statistical survey for medical care deceit study is done on summarizing, categorizing and comparing the statistical deceit observation. The basic target of this paper is to find and sort the type of frauds in healthcare sector according to the priority. The process used is first-Data cleaning, Second-Handling missing values, third- Data transformation, fourth- Feature selection and auditing.[5]

In relation to the last preliminary study which uses ECM and SVM for healthcare fraud detection, Robert A.Sowah et al.(2019) uses Decision Support system(DSS) and Genetic Support Vector Machine(GSVV). This work uses the National Health Insurance Scheme application dataset collected from hospitals in Ghana for finding the medical insurance deceit and other abnormalities in the claims made by the people of that hospital. The steps used by the researchers in this study are first-data preprocessing, second- classification engine development, and data postprocessing [6][28][29].

HosseinJoudaki et al.(2015) created an overview of fraud detection in healthcare industry. They have mentioned 21 primary researches that applied data mining methods in health care setting and health insurance. These studies have used many clustering algorithms like Bayesian's clustering, outlier detection, support vector machine, association rule, neural network etc..,[7][25].

AmiraKamil Ibrahim Hassan1 et al.(2013) also formulated a survey on insurance fraud detection in healthcare sector. The researchers have reviewed 566 papers and concluded that the data mining methods used for insurance deceit findings are logistic framework, Decision tree, the Naïve Bayes, and support vector machine [8][26].

In [9], Daniel Lasaga et al.(2017) formulated a procedure to detect the over utilization in health care sector. The problem handled in this paper is framed as categorical outlier problem which is solved using Logistic regression. In this study Restricted Boltzmann Machine (RBM)is uses testing and training to prove that using real world datasets with noise also the performance level can be improved.

Clifton Phua et al [10][27] presented survey on data mining-based fraud detection. This article classifies, relates and describes all the research works done on automated fraud detection. This works structure various methods and techniques in terms of supervised, unsupervised and hybrid approaches.

The above list of papers describe about the medical fraud and the methods to detect and correct them in details. In this paper we are going to compare the performance of the algorithms used in above papers to detect and correct the medical and the related healthcare frauds. Detailed investigations on these methods are presented in next topic.

Data Set Description

Healthcare Providers fraud detection dataset set is collected from Kaggle repository. This repository consists Beneficiary, inpatient and outpatient data. Table 1 shows the attributes available in each relation. These datasets contain raw information. Raw data are pre-processed in terms of removing noises (handling missing values, removing irrelevant attributes), data normalization and other techniques.

Table 1. Healthcare Providers Fraud Detection Repository

Beneficiary dataset	Inpatient dataset	Outpatient dataset
Beneficiary ID	Beneficiary ID	Beneficiary ID
DOB	ApplID	ApplID
DOD	ApplStartDt	ApplStartDt
Sex	ApplEndDt	ApplEndDt
Race	Provider	Provider
RenalDiseaseIndicator	InscClaimAmtReimbursed	InscClaimAmtReimbursed
State	AttendingPhysician	AttendingPhysician

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 8644 - 8654
Received 16 February 2021; Accepted 08 March 2021.

Country	OperatingPhysician	OperatingPhysician
IncurableCondi__Alzheimer	OtherPhysician	OtherPhysician
IncurableCondi__Heartfailure	AdmissionDt	ApplDiagnosisCode__1
IncurableCondi__KidneyDisease	ApplAdmitDiagnosisCode	ApplDiagnosisCode__2
IncurableCondi__Cancer	DeductibleAmtPaid	ApplDiagnosisCode__3
IncurableCondi__ObstrPulmonary	DischargeDt	ApplDiagnosisCode__4
IncurableCondi__Depression	DiagnosisGroupCode	ApplDiagnosisCode__5
IncurableCondi__Diabetes	ApplDiagnosisCode__1	ApplDiagnosisCode__6
IncurableCondi__IschemicHeart	ApplDiagnosisCode__2	ApplDiagnosisCode__7
IncurableCondi__Osteoporasis	ApplDiagnosisCode__3	ApplDiagnosisCode__8
IncurableCondi__rheumatoidarthritis	ApplDiagnosisCode__4	ApplDiagnosisCode__9
IncurableCondi__stroke	ApplDiagnosisCode__5	ApplDiagnosisCode__10
IPAnnualReimbursementAmount	ApplDiagnosisCode__6	ApplProcedureCode__1
IPAnnualDeductibleAmount	ApplDiagnosisCode__7	ApplProcedureCode__2
OPAnnualReimbursementAmount	ApplDiagnosisCode__8	ApplProcedureCode__3
OPAnnualDeductibleAmount	ApplDiagnosisCode__9	ApplProcedureCode__4
	ApplDiagnosisCode__10	ApplProcedureCode__5
	ApplProcedureCode__1	ApplProcedureCode__6
	ApplProcedureCode__2	DeductibleAmtPaid
	ApplProcedureCode__3	ClmAdmitDiagnosisCode
	ApplProcedureCode__4	
	ApplProcedureCode__5	
	ApplProcedureCode__6	

Comparative Analysis

The machine learning algorithms used for this comparative analysis and its description is mentioned in Table. 1. There are several works has been done for performing analysis of set of machine learning algorithms in various fields [12-14]. In this work, analysis of healthcare provider fraud detection is done.

Table 2. Machine Learning Algorithms used for Analysis

S. No	Algorithm [15,16]
1	Logistic Regression
2	Random Forest
3	Support Vector Machine
4	Adaboost
5	Multilayer Perceptron
6	Gradient Boosted Trees

Comparative analysis is performed by using python language. Table 3 shows the percentage of distribution of probable fraud type and Figure 1 shows its plot. Table 4 shows the percentage of
probable fraud distribution in individual claim providers Figure 2 shows the plot. Figure 3 presents state-wise beneficiary distribution. Healthcare fraud can be viewed in various aspects such as set of procedures involved in fraud, claim diagnosis / prognosis involved in fraud and doctors who involved in healthcare fraud. Figure 4 shows the Top-20 procedures intricated in Healthcare fraud. Figure 5 show Top-20 claim analysis involved in Healthcare Fraud. Figure 7 illustrated Top 20 doctors who convoluted in healthcare fraud. Age places vital role in reimbursing the insurance claims. This can be best illustrated with the help of Figure 7.

Receiver_Operating_Characteristic (ROC) curve is applied to exemplify the analytic strength of a two-class classifier method as its discriminant entry method is varied. The ROC is curve is constructed by framing the True Positive Rate (TPR) versus the False Positive Rate (FPR) at various threshold sceneries. ROC curve for the method mentioned in the Table 1 is depicted from Figure 8-Figure 13 respectively. From those figures, it is clearly understood

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 8644 - 8654
Received 16 February 2021; Accepted 08 March 2021.

that Multilayer perceptron (MLP) neural network approach outperforms well. Table 3 has the record of six state-of-the-art machine learning algorithms accomplishment in phase of accuracy, sensitivity, specificity, Area_Under_Curve (AUC), Kappa statistic and F1-Measure. The table has record for both training and validation phase. From Table 4, it is clearly understood that MLP has higher performance than the other machine learning approaches.

Table 3.Percentage of Distribution of Potential Fraud Type is mentioned below:

Classification Type: Potential Fraud	**Percentage**
Yes	38.12
No	61.88

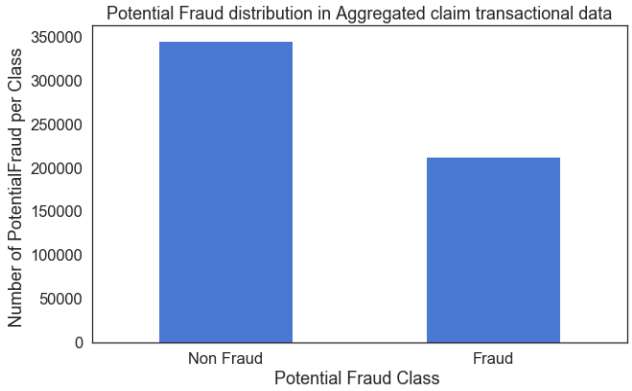

Figure 1. Plot of Potential Fraud distribution in Aggregated claim transactional data

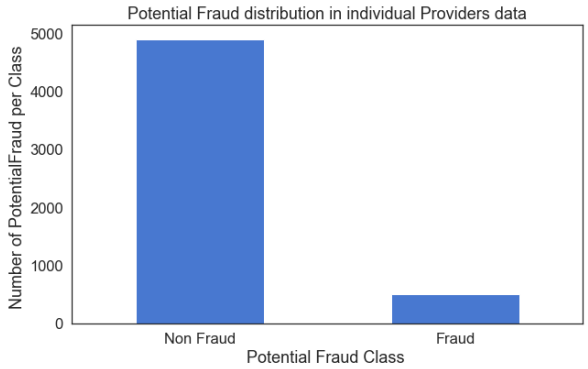

Figure 2. Plot of Potential Fraud distribution in Individual claim providers data

Table 4. Percentage of Distribution of Provide Class

Providers Class	**Percentage**
NonFraud	90.65
Fraud	9.35

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 8644 - 8654
Received 16 February 2021; Accepted 08 March 2021.

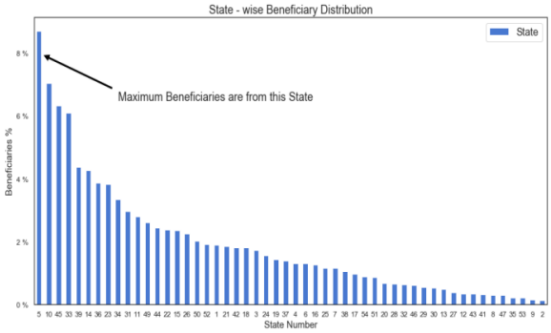

Figure 3. Plot of State with Beneficiary Distribution

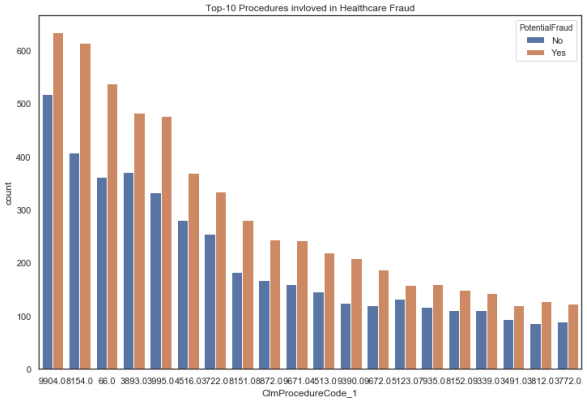

Figure 4. Plot of Top-20 procedures involved in Healthcare Fraud

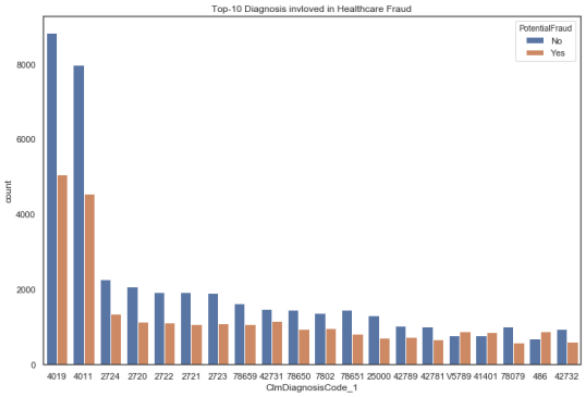

Figure 5. Plot of Top-20 claim diagnosis involved in Healthcare Fraud

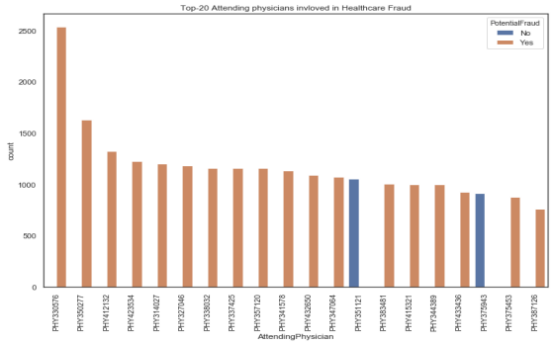

Figure 6. Plot of Top-20 doctors involved in Healthcare Fraud

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 8644 - 8654
Received 16 February 2021; Accepted 08 March 2021.

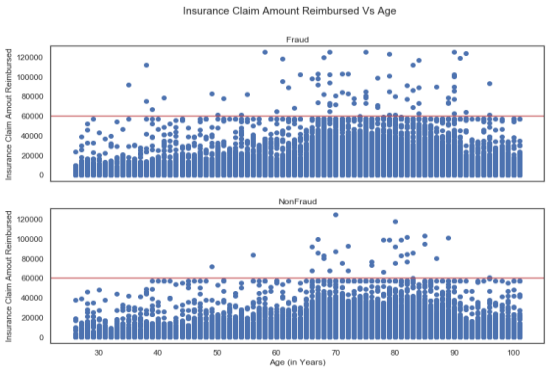

Figure 7. Plot of Reimbursement of Insurance Claim Vs Age

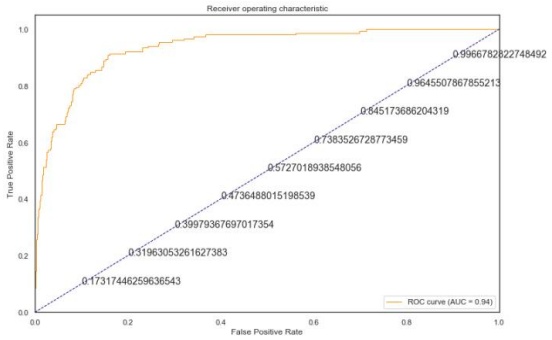

Figure 8. Logistic Regression ROC Curve

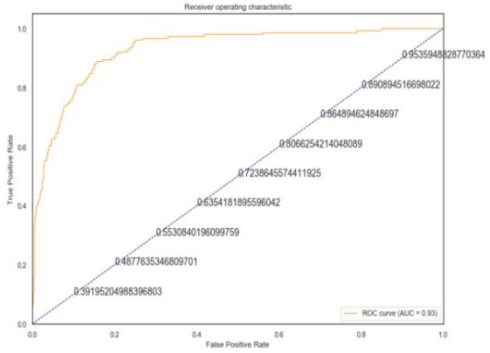

Figure 9. Random Forest ROC Curve

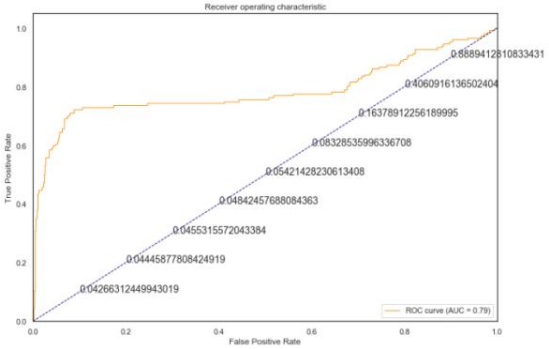

Figure 10. Support Vector Machine ROC Curve

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 8644 - 8654
Received 16 February 2021; Accepted 08 March 2021.

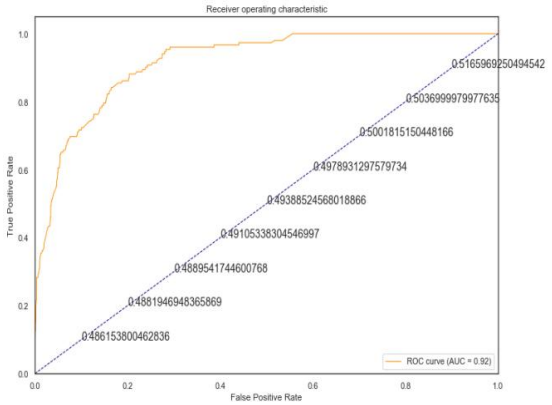

Figure 11. AdaBoost Classifier ROC Curve

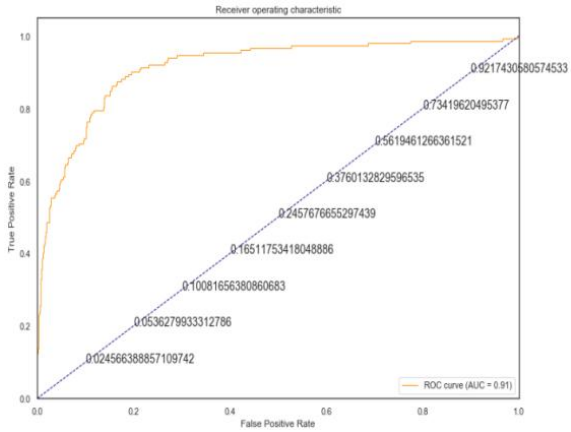

Figure 12. Multilayer Perceptron Classifier ROC Curve

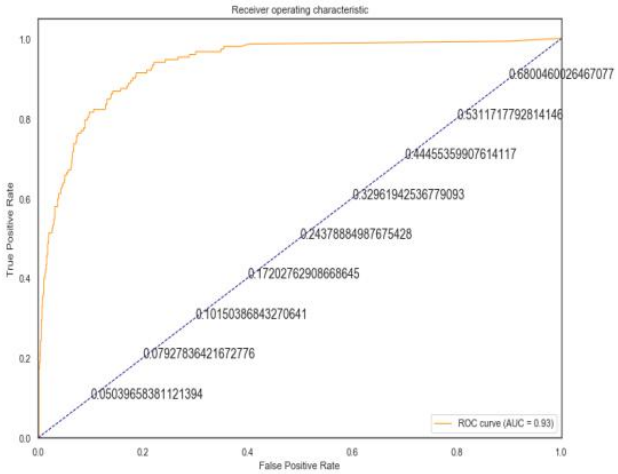

Figure 13. Gradient Boosted Tree Classifier ROC Curve

Table 5. Performance comparison of various machine learning algorithms

S. No	Algorithm	Accuracy		Sensitivity		Specificity		AUC	Kappa	F-Score	
		Training	Validation	Training	Validation	Training	Validation			Training	Validation
1	**Logistic Regression**	92.23	91.25	76.27	67.76	93.88	93.67	0.543	0.8072	64.74	59.19

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 8644 - 8654
Received 16 February 2021; Accepted 08 March 2021.

2	**Random Forest**	88.85	87.12	90.39	81.57	88.69	87.90	0.4773	0.8463	60.26	54.26
3	**Support Vector Machine**	93.10	93.28	44.35	44.73	89.13	98.30	0.5210	0.7151	54.50	55.51
4	**Adaboost**	94.61	92.05	62.42	48.02	97.93	96.60	0.488	0.7231	68.42	53.09
5	**Multilayer Perceptron**	98.04	92.97	83.33	48.68	99.56	97.55	0.527	0.7311	88.95	56.48
6	**Gradient Boosted Classifier**	97.80	93.46	76.83	48.68	99.97	98.09	.5487	0.7339	86.76	58.26

Conclusion

In this research article, a comprehensive performance analysis of machine learning algorithms for medical fraud detection is presented. A detailed experimental evaluation show that Multilayer perceptron algorithm outperforms well in terms of accuracy, sensitivity, specificity and F-Measure. We have compared the performance of the algorithms used in above papers to detect and correct the medical and the related healthcare frauds that can be uses to higher accuracy in medical deceit detection. It clearly shows that Multilayer Perceptron Algorithm provides significant performance when compared to the other approaches. In future, the comparative analysis is performed by using deep learning algorithms.

References

[1] Richard A. Bauder&Taghi M. Khoshgoftaar, “The Detection of Medicare Fraud Using Machine Learning Methods with Excluded Provider Labels”, The Thirty-First Internationa Florida Artificial Intelligence Research Society Conference (FLAIRS-31), pp. 404-409, 2018.

[2] Justin M. Johnson &Taghi M. Khoshgoftaar, “Medicare fraud detection using neural Networks”, Journal of Big Data, Vol. 6, No. 63, pp. 1-35, 2019.

[3] VrindaGarg&PreeteshShukla, “Combating medical provider fraud with advanced analytics using machine learning”, EXL white paper, 2018.

[4] VipulaRawte& G Anuradha, “Fraud detection in health insurance using data mining techniques”, International Conference on Communication, Information & Computing Technology (ICCICT), Jan. 16-17, Mumbai, India, pp. 1-6, 2015.

[5] Jing Li &Kuei-Ying Huang &Jionghua Jin &Jianjun Shi, “A survey on statistical methods for health care fraud detection”, Health Care Manage Sci, Springer, Vol. 11, No. 3, pp. 275-287, 2008.

[6] Robert A. Sowah ,MarcellinusKuuboore, Abdul Ofoli, Samuel Kwofie, Louis Asiedu , Koudjo M. Koumadi, &Kwaku O. Apeadu, “Decision Support System (DSS) for Fraud Detection in Health Insurance Claims Using Genetic Support Vector Machines (GSVMs)”, Journal of Engineering, Hindawi Publications, Vol. 2019, pp. 1-19, 2019.

[7] HosseinJoudaki, ArashRashidian, BehrouzMinaei-Bidgoli, MahmoodMahmoodi, BijanGeraili, Mahdi Nasiri& Mohammad Arab, “Using Data Mining to Detect Health Care Fraud and Abuse: A Review of Literature”, Global Journal of Health Science; Vol. 7, No. 1, pp. 194-202, 2015.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 8644 - 8654
Received 16 February 2021; Accepted 08 March 2021.

[8] AmiraKamil Ibrahim Hassan &Ajith Abraham, Computational Intelligence Models for Insurance Fraud Detection: A Review of a Decade of Research, Journal of Network and Innovative Computing, Vol. 1, pp. 341-347, 2013.

[9] Daniel Lasaga&PrakashSanthana, "Deep Learning to Detect Medical Treatment Fraud", Proceedings of Machine Learning Research, KDD 2017: Workshop on Anomaly Detection in Finance, pp. 114-120, 2017.

[10] Clifton Phua1, Vincent Lee1, Kate Smith & Ross Gayler, "A Comprehensive Survey of Data Mining-based Fraud Detection Research", Arxiv Journal, pp. 1-14, 2010.

[11] https://www.kaggle.com/rohitrox/healthcare-provider-fraud-detection-analysis

[12] PM Arunkumar& S Kannimuthu, "Machine Learning Based Automated Driver-Behavior Prediction For Automotive Control Systems", Journal of Mechanics of Continua and Mathematical Sciences, Vol. 7, pp. 1-12, 2020.

[13] S Kannimuthu, KS Bhuvaneshwari, D Bhanu, A Vaishnavi& S Ahalya, "Performance Evaluation of Machine Learning Algorithms for Dengue Disease Prediction", Journal of Computational and Theoretical Nanoscience, Vol. 16, No. 12,, pp. 5105–5110, 2019.

[14] G Chellamuthu, S Kannimuthu& K Premalatha, "Data Mining and Machine Learning Approaches in Breast Cancer Biomedical Research", Handbook of Sentiment Analysis and Knowledge Discovery in Contemporary Business, pp. 175-204, 2019.

[15] Stephan Marsland, "Machine Learning: An Algorithmic Perspective", Second Edition (Chapman & Hall/Crc Machine Learning & Pattern Recognition), 2014.

[16] Tom M. Mitchell, "Machine Learning: A Guide to Current Research", McGraw Hill publishers, 1997.

[17] Madhumitha Ramamurthy, "Fraudster Mobile Apps Detector in Google Playstore", Journal of Computational and Theoretical Nanoscience, Vol 17, pp.1752-1757, 2020.

[18] Madhumitha Ramamurthy , Y. Harold Robinson , S. Vimal , A. Suresh, Auto encoder based dimensionality reduction and classification using convolutional neural networks for hyperspectral images, Microprocessors and Microsystems , Volume No. 79, 103280, 2020.

[19] Madhumitha Ramamurthy, Ilango Krishnamurthi, S.Vimal, Y.Harold Robinson, "Deep Learning based genome analysis and NGS – RNA LL identification with a novel hybrid model", Biosystems, Volume No. 197, 104211, 2020.

[20] Madhumitha Ramamurthy, Ilango Krishnamurthi, Sudhakar Ilango, Shanthi Palaniappan, 'Discrete Model based answer script evaluation using decision tree rule classifier', Cluster Computing, 22, 13499–13510 Nov(2019)

[21] G.Selva Marry, S. Manoj kumar "Self-verifiable Computational Visual Crptographic Protocol for Secure 2D Image Communication" International Journal of Measurement Science and Technology, Vol. 30, No. 12, 2019.

[22] G. Selva Marry, S. Manoj kumar "Secure grayscale image communication using significant visual cryptography scheme in real time applications" SPRINGER International Journal of Multimedia Tools and Applications, Pages 1-20, 2019

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 8644 - 8654
Received 16 February 2021; Accepted 08 March 2021.

[23] S. Manoj Kumar N.Rajkumar "SCT Based Adaptive Data Aggregation for Wireless Sensor Networks" International Journal SPRINGER - Journal of Wireless Personal Communication Volume 75, Issue 4 April 2014, Page 2121-2133

[24] S. Manoj Kumar N.Rajkumar W.Catherine "Dropping False Packet to Increase the Network Lifetime of Wireless Sensor Network using EFDD Protocol" International Journal SPRINGER - Journal of Wireless Personal Communication Volume 70, Issue 4 June 2013, Page 1697-1709

[25]Ponmagal, R.S., Karthick, S., Dhiyanesh, B. et al. Optimized virtual network function provisioning technique for mobile edge cloud computing. J Ambient Intell Human Comput (2020).

[26]Ramamoorthy, S., Ravikumar, G., Saravana Balaji, B. et al. MCAMO: multi constraint aware multi-objective resource scheduling optimization technique for cloud infrastructure services. J Ambient Intell Human Comput (2020).

[27]Basha, A.J., Balaji, B.S., Poornima, S. et al. Support vector machine and simple recurrent network based automatic sleep stage classification of fuzzy kernel. J Ambient Intell Human Comput (2020)

[28]Balaji, B.S., Balakrishnan, S., Venkatachalam, K. et al. Automated query classification-based web service similarity technique using machine learning. J Ambient Intell Human Comput (2020)

[29]Viji, C., Rajkumar, N., Suganthi, S.T. et al. An improved approach for automatic spine canal segmentation using probabilistic boosting tree (PBT) with fuzzy support vector machine. J Ambient Intell Human Comput (2020). IF 4.5