Fraud Detection And Analysis For Insurance Claim Using Extra Trees Classifier

C T Ancymol[1], B Priya[2], S Bebisha [3]
[1, 2, 3] Dept of CSE
[1, 2, 3] Vins Christian College Of Engineering

Abstract- *Insurance Company working as commercial enterprise from last few years has been experiencing fraud cases for all type of claims. Amount claimed by fraudulent is significantly huge that may causes serious problems, hence along with government, different organization also working to detect and reduce such activities. Such frauds occurred in all areas of insurance claim with high severity such as insurance claimed towards auto sector is fraud that widely claimed and prominent type, which can be done by fake accident claim. So, develop a project that work on insurance claim data set to detect fraud and fake claims amount. The project implements machine learning algorithms to build model to label and classify claim. Two types of predictive analytics can be distinguished depending on the measurement level of the target: regression and classification. In regression, the target variable is continuous and varies along a predefined interval. This interval can be limited (e.g., between 0 and 1) or unlimited (e.g., between 0 and infinity). A typical example in a fraud detection setting is predicting the amount of fraud. In classification, the target is categorical which means that it can only take on a limited set of predefined values. In proposed focusing on detecting the auto vehicle fraud by using, Extra trees machine learning technique. This can help to calculate accuracy, precision..*

Keywords- Fraud detection, Classification, Extra trees machine learning technique

I. INTRODUCTION

An insurance claim is a formal request by a policyholder to an insurance company for coverage or compensation for a covered loss or policy event. The insurance company validates the claim (or denies the claim). If it is approved, the insurance company will issue payment to the insured or an approved interested party on behalf of the insured.

Insurance claims cover everything from death benefits on life insurance policies to routine and comprehensive medical exams. In some cases, a third-party is able to file claims on behalf of the insured person. However, in the majority of cases, only the person(s) listed on the policy is entitled to claim payments.

A paid insurance claim serves to indemnify a policyholder against financial loss. An individual or group pays premiums as consideration for the completion of an insurance contract between the insured party and an insurance carrier. The most common insurance claims involve costs for medical goods and services, physical damage, loss of life, liability for the ownership of dwellings (homeowners, landlords, and renters), and liability resulting from the operation of automobiles.

For property and causality insurance policies, regardless of the scope of an accident or who was at fault, the number of insurance claims you file has a direct impact on the rate you pay to gain coverage (typically through installment payments called insurance premiums). The greater the number of claims that are filed by a policyholder, the greater the likelihood of a rate hike. In some cases, it's possible if you file too many claims that the insurance company may decide to deny you coverage.

If the claim is being filed based on the damage to property that you caused, your rates will almost surely rise. On the other hand, if you aren't at fault, your rates may or may not increase. For example, getting hit from behind when your car is parked or having siding blow off your house during a storm are both events that are clearly not the result of the policyholder.

However, mitigating circumstances, such as the number of previous claims you have filed, the number of speeding tickets you have received, the frequency of natural disasters in your area (earthquakes, hurricanes, floods), and even a low credit rating can all cause your rates to go up, even if the latest claim was made for damage you didn't cause.

When it comes to insurance rate increases, not all claims are created equal. Dog bites, slip-and-fall personal injury claims, water damage, and mold can all act as signals of future liability for an insurer. These items tend to have a negative impact on your rates and on your insurer's

willingness to continue providing coverage. Surprisingly, speeding tickets may not cause a rate hike at all. At least for your first speeding ticket, many companies will not increase your prices. The same goes for a minor automobile accident or a small claim against your homeowner's insurance policy.

Fraud is one of the largest and most well-known problems that insurers Focuses on claim data of a car insurance company. Fraudulent claims can be highly expensive for each insurer. Therefore, it is important to know which claims are correct and which are not. It is not doable for insurance companies to check all claims personally since this will cost simply too much time and money. Take advantage of the largest asset which insurers have in the fight against fraud: Data. Employ various attributes about the claims, insured people and other circumstances which are included in the data by the insurer. Separating different groups of claims and the corresponding rates of fraud within those groups provide new insights.

Insurance fraud detection is a challenging problem, given the variety of fraud patterns and relatively small ratio of known frauds in typical samples. While building detection models, the savings from loss prevention needs to be balanced with the cost of false alerts. Machine learning techniques allow for improving predictive accuracy, enabling loss control units to achieve higher coverage with low false positive rates. Insurance frauds cover the range of improper activities which an individual may commit in order to achieve a favourable outcome from the insurance company. This could range from staging the incident, misrepresenting the situation including the relevant actors and the cause of incident and finally the extent of damage caused.

Problem Definition The goal of this project is to build a model that can detect auto insurance fraud. The challenge behind fraud detection in machine learning is that frauds are far less common as compared to legit insurance claims.

Focusing on detecting the auto vehicle fraud by using machine learning technique. Also, the performance will be compared by calculation of confusion matrix. This can help to calculate accuracy, precision. Proposed system is using linear regression and Extra trees classifier. Advantages are good prediction and accuracy. Application using to find fraudulent transaction validation.

Machine learning is a subfield of artificial intelligence (AI). The goal of machine learning generally is to understand the structure of data and fit that data into models that can be understood and utilized by people. Although machine learning is a field within computer science, it differs from traditional computational approaches. In traditional computing, algorithms are sets of explicitly programmed instructions used by computers to calculate or problem solve.

Machine learning algorithms instead allow for computers to train on data inputs and use statistical analysis in order to output values that fall within a specific range. Because of this, machine learning facilitates computers in building models from sample data in order to automate decision-making processes based on data inputs.

Machine learning is a subfield of artificial intelligence (AI). The goal of machine learning generally is to understand the structure of data and fit that data into models that can be understood and utilized by people.

Although machine learning is a field within computer science, it differs from traditional computational approaches. In traditional computing, algorithms are sets of explicitly programmed instructions used by computers to calculate or problem solve. Machine learning algorithms instead allow for computers to train on data inputs and use statistical analysis in order to output values that fall within a specific range. Because of this, machine learning facilitates computers in building models from sample data in order to automate decision-making processes based on data inputs.

In supervised learning, the computer is provided with example inputs that are labeled with their desired outputs. The purpose of this method is for the algorithm to be able to "learn" by comparing its actual output with the "taught" outputs to find errors, and modify the model accordingly. Supervised learning therefore uses patterns to predict label values on additional unlabeled data.

For example, with supervised learning, an algorithm may be fed data with images of sharks labeled as fish and images of oceans labeled as water. By being trained on this data, the supervised learning algorithm should be able to later identify unlabeled shark images as fish and unlabeled ocean images as water.

A common use case of supervised learning is to use historical data to predict statistically likely future events. It may use historical stock market information to anticipate upcoming fluctuations, or be employed to filter out spam emails. In supervised learning, tagged photos of dogs can be used as input data to classify untagged photos of dogs.

In unsupervised learning, data is unlabeled, so the learning algorithm is left to find commonalities among its input data. As unlabeled data are more abundant than labeled

data, machine learning methods that facilitate unsupervised learning are particularly valuable.

The goal of unsupervised learning may be as straightforward as discovering hidden patterns within a dataset, but it may also have a goal of feature learning, which allows the computational machine to automatically discover the representations that are needed to classify raw data.

Unsupervised learning is commonly used for transactional data. You may have a large dataset of customers and their purchases, but as a human you will likely not be able to make sense of what similar attributes can be drawn from customer profiles and their types of purchases.

With this data fed into an unsupervised learning algorithm, it may be determined that women of a certain age range who buy unscented soaps are likely to be pregnant, and therefore a marketing campaign related to pregnancy and baby products can be targeted to this audience in order to increase their number of purchases.

Without being told a "correct" answer, unsupervised learning methods can look at complex data that is more expansive and seemingly unrelated in order to organize it in potentially meaningful ways. Unsupervised learning is often used for anomaly detection including for fraudulent credit card purchases, and recommender systems that recommend what products to buy next. In unsupervised learning, untagged photos of dogs can be used as input data for the algorithm to find likenesses and classify dog photos together.

II. **LITERATURE SURVEY**

Several state-of-the-art binary classification techniques are experimentally evaluated in the context of expert automobile insurance claim fraud detection. The predictive power of logistic regression, C4.5 decision tree, k-nearest neighbor, Bayesian learning multilayer perceptron neural network, least-squares support vector machine, naive Bayes, and tree-augmented naive Bayes classification is contrasted. For most of these algorithm types, we report on several operationalizations using alternative hyperparameter or design choices. We compare these in terms of mean percentage correctly classified (PCC) and mean area under the receiver operating characteristic (AUROC) curve using a stratified, blocked, ten-fold cross-validation experiment. We also contrast algorithm type performance visually by means of the convex hull of the receiver operating characteristic (ROC) curves associated with the alternative operationalizations per algorithm type. The study is based on a data set of 1,399 personal injury protection claims from 1993 accidents collected by the Automobile Insurers Bureau of Massachusetts. To stay as close to real-life operating conditions as possible, we consider only predictors that are known relatively early in the life of a claim.

The article proposes an expert system for detection, and subsequent investigation, of groups of collaborating automobile insurance fraudsters. The system is described and examined in great detail, several technical difficulties in detecting fraud are also considered, for it to be applicable in practice. Opposed to many other approaches, the system uses networks for representation of data. Networks are the most natural representation of such a relational domain, allowing formulation and analysis of complex relations between entities. Fraudulent entities are found by employing a novel assessment algorithm, Iterative Assessment Algorithm (IAA), also presented in the article. Besides intrinsic attributes of entities, the algorithm explores also the relations between entities. The prototype was evaluated and rigorously analyzed on real world data. Results show that automobile insurance fraud can be efficiently detected with the proposed system and that appropriate data representation is vital.

The aim of this article is to develop a model to aid insurance companies in their decision making and to ensure that they are better equipped to fight fraud. This tool is based on the systematic use of fraud indicators. We first propose a procedure to isolate the indicators which are most significant in predicting the probability that a claim may be fraudulent. We applied the procedure to data collected in the Dionne±Belhadji study (1996). The model allowed us to observe that 23 of the 54 indicators used were significant in predicting the probability of fraud. Our study also discusses the model's accuracy and detection capability. The detection rates obtained by the adjusters who participated in the study constitute the reference point of this discussion.

Internet of Things(IoT) is one of the rapidly growing fields and has a wide range of applications such as smart cities, smart homes, connected wearable, connected health-care, and connected automobiles, etc. These IoT applications generate tremendous amounts of data which needs to be analyzed to draw useful inferences required to optimize the performance of IoT applications. The artificial intelligence (AI) and machine learning (ML) play the significant role in building the smart IoT systems. The main objective of the paper is a comprehensive analysis of five well-known supervised machine learning algorithms on IoT datasets. The five classifiers are K-Nearest Neighbors (KNN), Naive Bayes (NB), Decision Tree (DT), Random Forest (RF) and Logistic Regression (LR). The feature reduction is performed using PCA algorithm. The performance of these five classifiers has

been compared on the basis of six characteristics of IoT dataset such as size, number of features, number of classes, class imbalance, missing values and execution time. The classifiers have also been compared on various performance metrics such as precision, recall, f1-score, kappa, and accuracy. As per our results, the DT classifier gives the best accuracy of 99% among all the algorithms for all datasets. The results also show the performance of RF and KNN as almost similar and the NB and LR perform the worst among all the classifiers.

Machine learning is predominantly an area of Artificial Intelligence which has been a key component of digitalization solutions that has caught major attention in the digital arena. In this paper author intends to do a brief review of various machine learning algorithms which are most frequently used and therefore are the most popular ones. The author intends to highlight the merits and demerits of the machine learning algorithms from their application perspective to aid in an informed decision making towards selecting the appropriate learning algorithm to meet the specific requirement of the application.

In the last decade, the ease of online payment has opened up many new opportunities for e-commerce, lowering the geographical boundaries for retail. While e-commerce is still gaining popularity, it is also the playground of fraudsters who try to misuse the transparency of online purchases and the transfer of credit card records. This paper proposes APATE, a novel approach to detect fraudulent credit card transactions conducted in online stores. Our approach combines intrinsic features derived from the characteristics of incoming transactions and the customer spending history using the fundamentals of RFM (Recency - Frequency- Monetary); and network-based features by exploiting the network of credit card holders and merchants and deriving a time-dependent suspiciousness score for each network object. Our results show that both intrinsic and network-based features are two strongly intertwined sides of the same picture.

III. PROPOSED SYSTEM

Focusing on detecting the auto vehicle fraud is by using, machine learning technique. Also, the performance will be compared by calculation of confusion matrix. This can help to calculate accuracy, precision, and recall. In proposed using linear regression and Extra trees classifier. Advantages are good prediction and accuracy,Application is using to find fraudulent transaction validation, and to detect if an insurance claim is fraudulent or not.

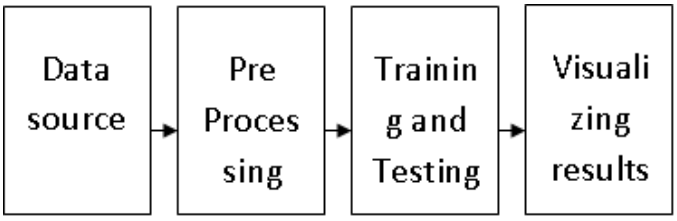


Fig 1 System Architecture

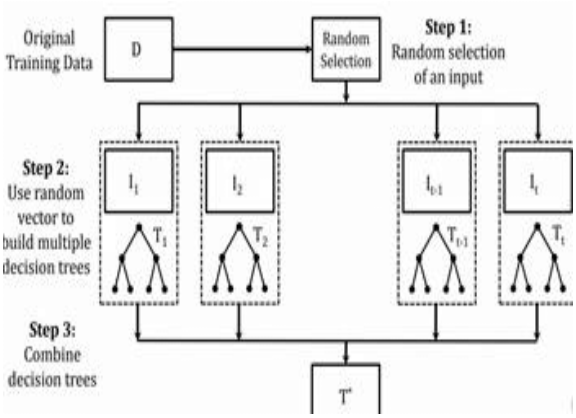
Machine learning model is built with different algorithms that are trained by information and data set provided which predict new classification as fraud or not These algorithms implemented for building model that is trained using historical data and that predict unseen data with most matching features. And then model is tested and validated to evaluate its performance. After the calculations comparison is made. For automobile insurance fraud detection shows the higher accuracy.

Linear Regression Algorithm

Linear Regression is an algorithm that belongs to supervised Machine Learning. It tries to apply relations that will predict the outcome of an event based on the independent variable data points. The relation is usually a straight line that best fits the different data points as close as possible. The output is of a continuous form, i.e., numerical value. For example, the output could be revenue or sales in currency, the number of products sold, etc. In the above example, the independent variable can be single or multiple.

Extra trees classifier

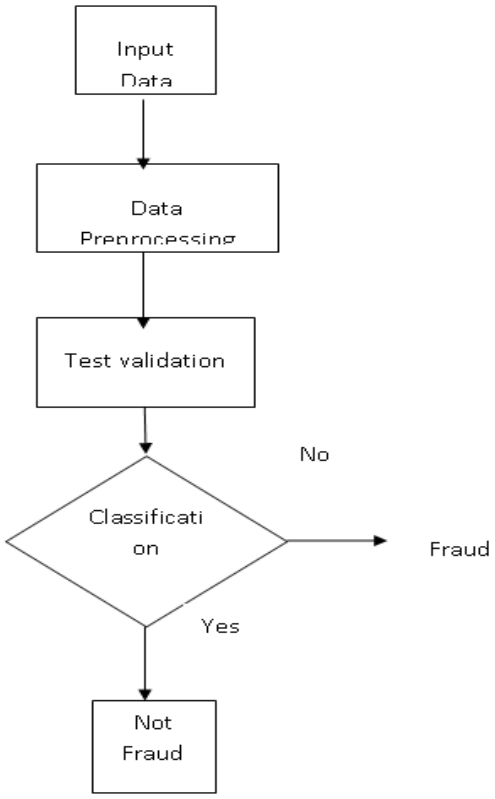
Extra trees short for extremely randomized trees is an ensemble supervised machine learning method that uses decision trees and is used by the Train Using AutoML tool. See Decision trees classification and regression algorithm for information about how decision trees work. This method is similar to random forests but can be faster.



Data flow diagram first input the datasets after preprocessing will occur. There are many stages involved in data preprocessing. Data cleaning attempts to impute missing values, removing outliers from the dataset. Data integration integrates data from a multitude of sources into a single data warehouse. Data transformation such as normalization, may be applied. For example, normalization may improve the accuracy and efficiency of mining algorithms involving distance measurement.

Data reduction can reduce the data size by dropping out redundant features. Feature selection and feature extraction techniques can be used.Different models are tested on the dataset once it is obtained and cleaned. On the basis of the initial model performance, different features of the model are engineered and tested again. Once all the options area unit designed, the model is made and run victimisation completely different completely different values and victimisation different iteration procedures. A predictive model is created that predicts if an insurance claim is fraudulent or not.

Binary Classification task takes place which gives answer between YES or NO. This report deals with classification algorithm to detect fraudulent transaction.



Flow Diagram

MODULES DESCRIPTION

Data Source

There are four data sources, describing beneficiaries, outpatient claims, inpatient claims, and provider fraud. This dataset collected from vehicle insurance csv file.

Pre-processing

For the successful application pre-processing is required. The data which is acquired from different resources are sometime in raw form. It may contain some incomplete, redundant, inconsistent data. Therefore in this step such redundant data should be filtered. Data should be normalized.

Data Training and Modelling

There are many stages involved in data preprocessing: Data cleaning attempts to impute missing values, removing outliers from the dataset. Data integration integrates data from a multitude of sources into a single data warehouse. Data transformation such as normalization may be applied. For example, normalization may improve the accuracy and efficiency of mining algorithms involving distance measurement. Data reduction can reduce the data size

by dropping out redundant features. Feature selection and feature extraction techniques can be used.

Training and Testing

The ExtraTreesClassifier is to fit a number of randomized decision trees to the data, and in this regard is from of ensemble learning. Particularly, random splits of all observations are carried out to ensure that the model does not over fit the data. The extra trees algorithm, like the random forests algorithm, creates many decision trees, but the sampling for each tree is random, without replacement. This creates a dataset for each tree with unique samples. A specific number of features, from the total set of features, are also selected randomly for each tree.

Visualizing the Results

Linear Regression is an algorithm that belongs to supervised Machine Learning. It tries to apply relations that will predict the outcome of an event based on the independent variable data points

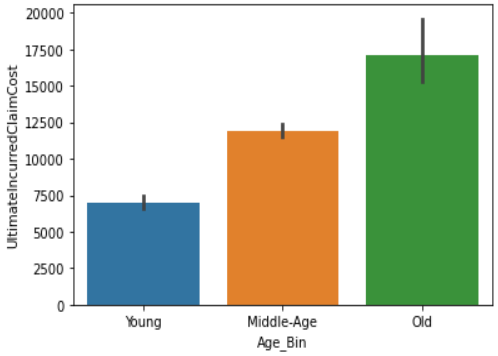
IV. SCREEN SHOTS

	Unnamed: 0	Unnamed: 1	DateReported	Unnamed: 3	Unnamed: 4	Unnamed: 5	Unnamed: 6	DependentsOther	Unnamed: 8	Unnamed: 9	Unnamed:
0	ClaimNumber	DateTimeOfAccident	NaN	Age	Gender	MaritalStatus	DependentChildren		NaN	WeeklyWages	PartTimeFullTime	HoursWorkedPerWe
1	WC8205482	2002-04-09T07:00:00Z	2002-07-05T00:00:00Z	48	M	M	0		0.0	500	F	
2	WC6922469	1999-01-07T11:00:00Z	1999-01-20T00:00:00Z	43	F	M	0		0.0	509.34	F	37
3	WC5442654	1996-03-25T00:00:00Z	1996-04-14T00:00:00Z	30	M	U	0		0.0	709.1	F	
4	WC9796897	2005-06-22T13:00:00Z	2005-07-22T00:00:00Z	41	M	S	0		0.0	555.46	F	

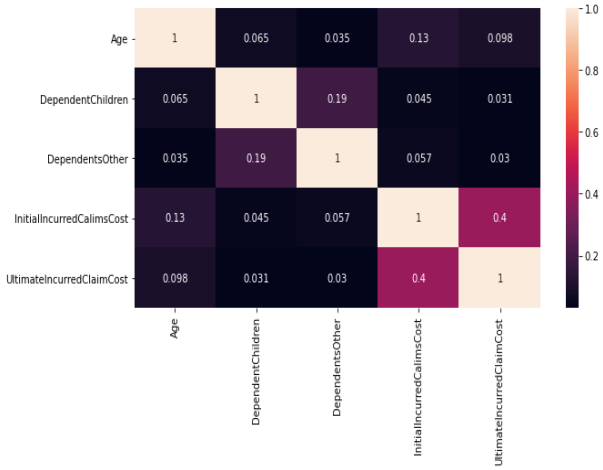
Loading data set

	ClaimNumber	DateTimeOfAccident	DateReported	Age	Gender	MaritalStatus	DependentChildren	DependentsOther	WeeklyWages	PartTimeFullTime	HoursWorkedPerWee
1	WC8205482	2002-04-09T07:00:00Z	2002-07-05T00:00:00Z	48	M	M	0	0.0	500	F	3
2	WC6922469	1999-01-07T11:00:00Z	1999-01-20T00:00:00Z	43	F	M	0	0.0	509.34	F	37
3	WC5442654	1996-03-25T00:00:00Z	1996-04-14T00:00:00Z	30	M	U	0	0.0	709.1	F	3
4	WC9796897	2005-06-22T13:00:00Z	2005-07-22T00:00:00Z	41	M	S	0	0.0	555.46	F	3
5	WC2603726	1990-08-29T08:00:00Z	1990-09-27T00:00:00Z	36	M	M	0	0.0	377.1	F	3

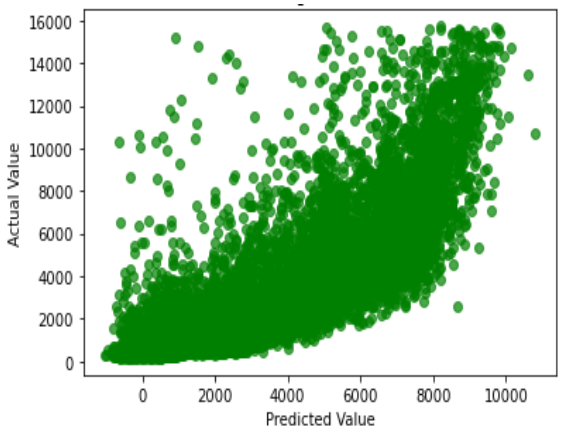
Named Datasets



Bivariate Analysis Based on Age Group



Correlation Plot



Prediction

V. CONCLUSION

As the different countries around the world evolve into a more economical-based one, stimulating their economy is the goal. To fight these fraudsters and money launderers was quite a complex task before the era of machine learning but thanks to machine learning and AI we are able to fight these kinds of attacks. The proposed solution can be used in insurance companies to find out if a certain insurance claim made is a fraud or not. The model was designed after testing multiple algorithms to come up with the best model that will detect if a claim is fraudulent or not. This is aimed at the insurance companies as a pitch to come up with a more tailored model for their liking to their own systems. The model should be simple enough to calculate big datasets, yet complex enough to have a decent successful percentile.

Future work, use various machine learning to predict which claims are likely to be fraudulent. This information can narrow down the list of claims that need a further check. It enables an insurer to detect more fraudulent claims

REFERENCES

[1] K. Ulaga Priya and S. Pushpa, "A Survey on FraudAnalytics Using Predictive Model in Insurance Claims," Int. J. Pure Appl. Math., vol. 114, no. 7, pp.755–767, 2017.

[2] E. B. Belhadji, G. Dionne, and F. Tarkhani, "A Model for the Detection of Insurance Fraud," Geneva Pap. Risk Insur. Issues Pract., vol. 25, no. 4, pp. 517– 538, 2000, doi: 10.1111/1468-0440.00080.

[3] "Predictive Analysis for Fraud Detection." https://www.wipro.com/analytics/comparative analysis-of-machine-learning-techniques-for-%0Adetectin/.

[4] F. C. Li, P. K. Wang, and G. E. Wang, "Comparison of the primitive classifiers with extreme learning machine in credit scoring," IEEM 2009 - IEEE Int. Conf. Ind. Eng. Eng. Manag., vol. 2, no. 4, pp. 685– 688, 2009, doi: 10.1109/IEEM.2009.5373241.

[5] V. Khadse, P. N. Mahalle, and S. V. Biraris, "An Empirical Comparison of Supervised Machine Learning Algorithms for Internet of Things Data," Proc. - 2018 4th Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2018, pp. 1–6, 2018, doi:10.1109/ICCUBEA.2018.8697476.

[6] S. Ray, "A Quick Review of Machine Learning Algorithms," Proc. Int. Conf. Mach. Learn. Big Data, Cloud Parallel Comput. Trends, Prespectives Prospect. Com. 2019, pp. 35–39, 2019, doi: 10.1109/COMITCon.2019.8862451.

[7] Rushika Ghadge, Juilee Kulkarni, Pooja More, Sachee Nene, Priya R , "Prediction of Crop Yield using Machine Learning", International Research Journal of Engineering & Technology, Vol 5, Issue 2, Feb 2018.

[8] C. Phua, V. Lee, K. Smith, R. Gayler (2010); "Comprehensive Survey of Data Mining-based Fraud Detection Research", ICICTA '10 Proceedings of the 2010 International Conference on Intelligent Computation Technology and Automation Volume 1, pp. 50-53.

[9] S. Cheng, J. Liu, X. Tang (2014); "Using unlabeled Data to Improve Inductive Models by Incorporating Transductive Models"; International Journal of Advanced Research in Artificial Intelligence, Volume 3 Number 2, pp. 33-38.

[10] Sonal S. Ambalkar, S. S. Thorat2, "Bone Tumor Detection from MRI Images using Machine Learning: A Review", International Research Journal of Engineering & Technology", Vol. 5, Issue 1, Jan -2018.

[11] Rajat Raina, Alexis Battele, Honglak Lee,Benjamin Packer, Andrew Y. Ng , "Self-taught Learning : Transfer of Learning from Unlabeled Data", Computer Science Department, Stanford University, CA, USA, Proceedings of 24th International Conference on Machine Learning Corvallis, OR, 2007.

[12] Jimmy Lin, Alek Kolcz, "Large-Scale Machine Learning at Twitter", Proceedings of SIGMOD '12, May 20–24, 2012, Scottsdale, Arizona, USA.

[13] Dr. Rama Kishore, Taranjit Kaur, "Backpropagation Algorithm: An Artificial Neural Network Approach for Pattern Recognition", International Journal of Scientific & Engineering Research, Volume 3, Issue 6, June-2012.

[14] Kedar Potdar, Rishab Kinnerkar, "A Comparative Study of Machine Algorithms applied to Predictive Breast Cancer Data", International Journal of Science & Research, Vol. 5, Issue 9, pp. 1550-1553, September 2016.

[15] D. Pelleg, A. Moore (2000): "X-means: Extending K-means with Efficient Estimation of the Number of Clusters"; ICML '00 Proceedings of the Seventeenth International Conference on Machine Learning, pp. 727-734.