International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 05 | May - 2023 SJIF 2023: 8.176 ISSN: 2582-3930

Predictive Analytics for Fraud Detection of Realtime Financial Data by Using Machine Learning Techniques

Kunchala Vamshi Krishna[1], Mittapally Shyam Sunder[2], Dodle Sai Karthik Reddy[3], Nimmala Rohan Reddy[4]

[1,2,4] IT Department, Guru Nanak Intuitions Technical Campus, Hyderabad, India.
[3]ECE Department, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India.

Abstract: Financial fraud affects tremendously both the financial industry and everyday life. Fraud can reduce confidence in the industry, destabilize savings and affect the cost of living. Financial institutions use a variety of fraud prevention models to address this problem. This report presents an approach that utilizes Machine Learning techniques to build a model that detects fraudulent transactions and flags them. The approach utilizes a dataset that contains a collection of observation points on transactions and which can be useful in understanding the nature of transactions. The high accuracy results of the models are indicative of their readiness to use in a real-world setting. This performance means that the likelihood of a fraudulent case passing through is quite low. This paper, seeks to carry out comparative analysis of financial fraud detection techniques, like machine-learning techniques, who plays an important role in fraud detection, as it is often applied to extract and uncover the hidden truths behind very large quantities of data. This makes organizations adapt to high-level security and data handling technology solutions like machine learning, deep learning and predictive analytics which are efficient enough to deal with highly sensitive data, predict frauds and unwanted behavioural patterns in this data. This paper reviews the different advance technologies commonly used to deal with this type of data forms a comparison among them and suggests the most efficient and informative method to use in this sector. Through the end of the review, feature engineering and its selection of parameters for achieving better performance are discussed.

Keywords: Predictive Analytics, Artificial Intelligence, Machine learning, streaming data, real-time applications, deep learning.

I.INTRODUCTION

Financial transactions make the most of this world as people exchange goods, services and more. These transactions usually occur following an agreement made between two or more people upon performing the said services (Westermeier, 2020). Therefore, they range from a small amount to large business transactions involving huge sums. Examples of financial transactions include the reception of cash, deposits, purchases, invoices, charges of services, expenses and more. Financial fraud affects tremendously both the financial industry and everyday life. Fraud can reduce confidence in the industry, destabilize savings and affect the cost of living. Financial institutions use a variety of fraud prevention models to address this problem. However, fraudsters are adaptive, and over time, they conceive several ways of intruding such protective models. Despite the best effort of financial institutions, law enforcement and government, financial fraud continues to grow. Fraudsters today can be a very inventive, intelligent and fast fraternity. Due to the increase in online payments and transactions, fraudulent activities have increased leading to the stealing of personal information and misuse activities.
Machine learning has come up as a solution to these real-world problems allowing accurate predictive analytics, and data management with coverage to low false-positive rates. Machine learning has been proved contributory to solve problems containing sensitive data, such as email spam detection, accurate product recommendation,

accurate medical diagnosis, etc. like some technological solutions to the digital world. Fraudsters are able to access the financial records and customers' details internally and hence are not caught in the radar of authentication. This makes fraud detection an important challenge for the organization, its data science team to come up with technological strategies, with complete automation and analytics solutions to predict potential fraudulent activity before it bounds to occur.

II. FINANCIAL FRAUD

Fraud definition, according to the Association of Certified Fraud Examiners (ACFE) "ACFE Association of Fraud Examiners Certificates", fraud includes any intentional or deliberate act of depriving another of property or money by cunning, deception or other unfair acts [12].

Types of financial fraud

There are several types of financial fraud; we present here a brief description of some of the main types of fraud. Insurance fraud can occur at many points in the insurance process (e.g., application, eligibility, rating, billing, and claims), and can be committed by consumers, agents and brokers, insurance company employees, healthcare providers, and others [1, 2] Securities and commodities fraud, the FBI [3] provides brief descriptions of some of the most prevalent securities and commodities frauds encountered today, for example, "Market Manipulation, High Yield Investment Fraud, The Ponzi Scheme, The Pyramid Scheme, Prime Bank Scheme, Advance Fee Fraud, Hedge Fund Fraud, Commodities Fraud, Foreign Exchange Fraud, Broker Embezzlement and Late-Day Trading." According to another definition by CULS [4], securities frauds include theft from manipulation of the market, theft from securities accounts, and wire fraud.

Money Laundering is the process by which criminals conceal or disguise the proceeds of their crimes or convert those proceeds into goods and services. It allows criminals to inject their illegal money into the stream of commerce, thus corrupting financial institutions and the money supply and giving criminals unwarranted economic power [5].

Gao and Ye [6] similarly define money laundering as the process by which criminals "wash dirty money" to disguise its illicit origin and make it appear legitimate and "clean." Financial statement fraud (corporate fraud), financial statements are a company's basic documents to reflect its financial status [10]. It had an objective as.

- Fraud these statements to make the business more profitable
- Improvement of the performance of the actions
- Reduction of tax obligations
- Attempt to exaggerate performance due to managerial pressure

Credit card fraud is essentially of two types; application and behavioural fraud [26]. Application fraud is where fraudsters obtains new cards from issuing companies using false information or other people's information.

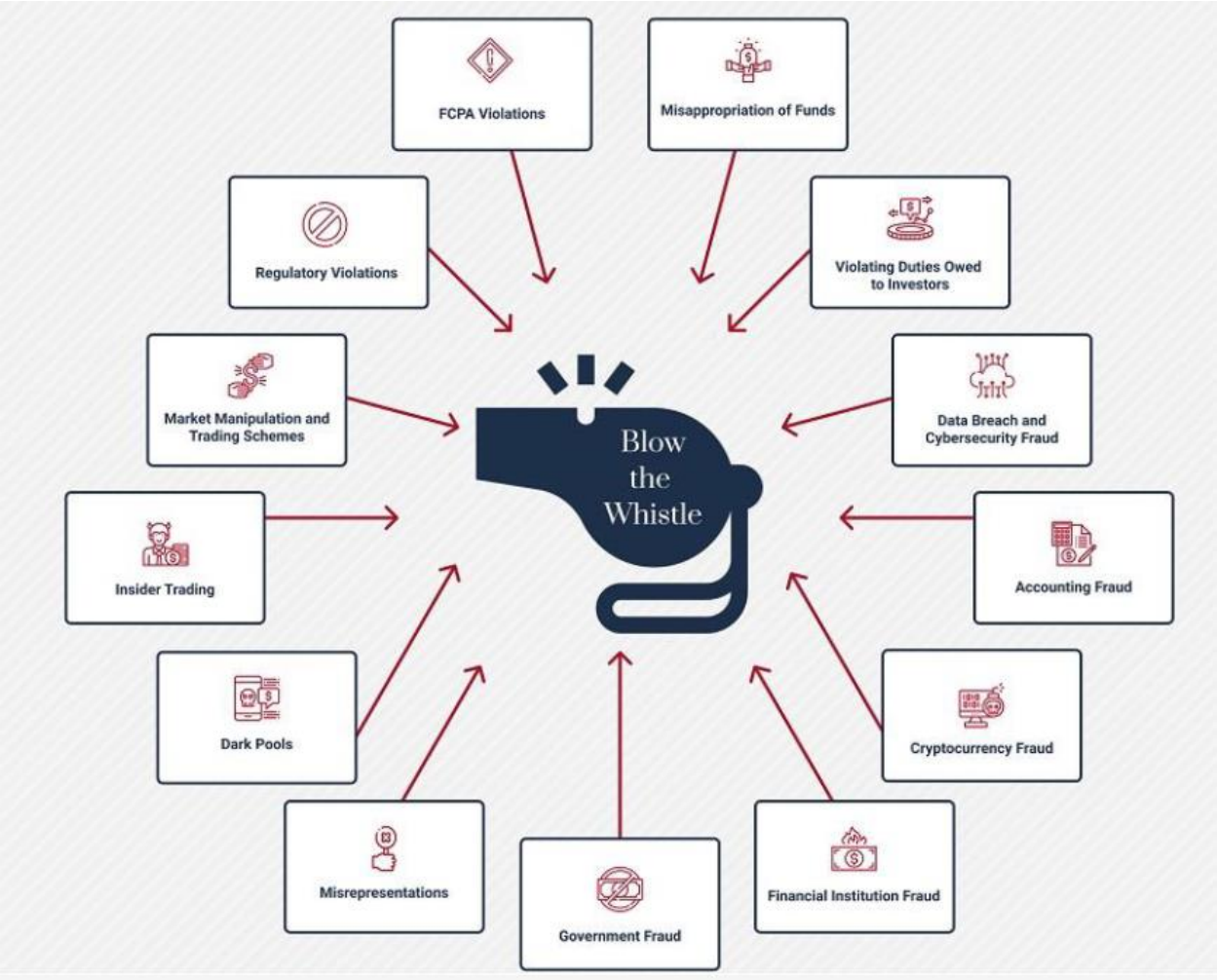


Fig 1. Types of financial frauds

Behavioural fraud can be of four types: mail theft, stolen/lost card, counterfeit card and 'card holder not present' fraud [25] Mortgage Fraud is a specific form of financial fraud that refers to the manipulation of a property or mortgage documents. It is often committed to distort the value of a property for the purpose of influencing a lender to finance a loan for it [5].

II. LITERATURE REVIEW

Earlier approaches and research studies have formulated solutions to fraud detection using core machine learning algorithms from a supervised approach, unsupervised, hybrid to neural networks [14]. These approaches have been successful to a rate of nearly some predictable result and this has been clearly observed in credit card fraud detection or finance industry business problems containing financial transaction record and customer personal information carried under human review [9]. The traditional approach of using a rule-based approach or core machine learning algorithm for fraud detection in real-time financial data, which tends to change with time and varying results, the security is still not under proper hold as the predictions focus on present data

and hence doesn't the problems that came across the review have analyzed for the later use of implementing an efficient machine learning model using predictive analytics algorithms. In review, few industries are analyzed in account with their applied advanced techniques in financial data. Organizations currently using machine learning and predictive analytics are sown in the table below [12].

The introduction of loaning for paybacks with interest also increased the possibilities of financing (Green, 2019). Banks soon caught up and introduced ways to obtain loans with the intention of their generation of profit from the interest. Further, since banks were designed to hold the representation of keeping funds for others, they soon got into the investment scene where they could utilize the funds for other cases. The concept of loaning also led to the introduction of seeking funds for business, personal cases, or in cases of countries running their economies.

III. PREDICTIVE ANALYTICS

Predictive analytics is an advanced technological solution above the core machine learning algorithms, a process of using historical trends in the data to make future predictions with better and appropriate results. In broad terms, it also refers to the field of data science that involves making predictions about future events. Analytics involves statistics, data mining, modelling and finally risk management and after the final process, the predicted value is used by the organization to incorporate fraud detection and risk management with the data [15]. By using different statistical
techniques like linear regression an algorithm analyzes past data to make predictions about future events. The concept uses complex techniques like data mining, machine learning, statistics, regression, classification techniques, and popular decision trees to help analysts make future business forecasts. The predictive models and analysis, based on these techniques are typically used to forecast future probabilities, indicating any unnatural pattern in the data or indicating risks occurring data points.

In predictive analysis, the purpose is to build an analytic model that predicts target objects of interest. Logistic Regression (LR) logistic regression is a type of generalized linear model. Using simple linear regression is inappropriate when the variable to be predicted is binary; due to normality assumptions. [7].

Decision Trees (DT) is a tree structure, where each node represents a test on an attribute and each branch represents an outcome of the test. In this way, the tree attempts to divide observations into mutually exclusive subgroups [14]. Classification and Regression Tree (CART) is a computerized, non-parametric technique different from traditional statistical methods. CART applies the binary Recursive Partitioning Algorithm (RPA) to best classify samples into a number of non-overlapping regions, each of which corresponds to a terminal node in the tree [20].

Decision Trees C4.5 gives algorithm and solutions to a set of problems that have arisen over the years among decision tree researchers like handling various problems such as missing attribute values [19]. Cost-sensitive decision tree (CSDT) an induction algorithm developed to identify fraudulent credit card transactions are given. In the well-known decision tree algorithms, the splitting criteria are either insensitive to costs and class distributions or the cost is fixed to a constant ratio [29]. Neural Networks (NN) is a mature technology with an established theory and recognized application areas. A NN consists of a number of neurons, i.e., interconnected processing units. Associated with each connection is a numerical value, called "weight"[14]. Probabilistic neural network (PNN) is a feed-forward NN involving a one pass training algorithm used for classification and mapping

of data. It is a pattern classification network, based on the classical Bayes classifier, which is statistically an optimal classifier that seeks to minimize the risk of misclassification [11].

Support Vector Machines (SVM) use a linear model to implement nonlinear class boundaries by mapping input vectors nonlinearly into a high-dimensional feature space. In the new space, an optimal separating hyperplane is constructed [11]. Naïve Bayes (NB) a classification tool simply uses Bayes conditional probability rule. Each attribute and class label are considered random variable, and assuming that the attributes are independent, the naïve Bayes finds a class to the new observation that maximizes its probability given the values of the attributes.[9]. Bayesian belief network (BBN) allow for the representation of dependencies among subsets of attributes. A BBN is a directed acyclic graph, where each node represents an attribute and each arrow represents a probabilistic dependence [14].

Bayesian skewed logit model (BSL) this model incorporates the possibility of using asymmetric links in order to measure the probability of yi = 0 and yi = 1 in non-balanced samples [9] K-nearest neighbor (KNN) is used largely in detection systems. It is also proved that KNN works extremely well in credit card fraud detection systems using supervised learning techniques. [38]. Bivariate Probit Model (BP) is typically used where a dichotomous indicator is the outcome of interest and the determinants of the probable outcome includes qualitative information in the form of a dummy variable where, even after controlling for a set of covariates, the possibility that the dummy explanatory variable is endogenous cannot be ruled out a priori [40].This part, describes artificial and computational intelligence models, which is, a set of nature-inspired computational methodologies and approaches to address complex real-world problems to which mathematical or traditional modelling.

Genetic Algorithm (GA) in Genetic Algorithm i.e. inspired from natural evolution, randomly generated rules are considered as an initial population[15].

Genetic programming (GP) is an extension of genetic algorithms (GA). It is a search methodology belonging to the family of evolutionary computation. GP randomly generates an initial population of solutions. Then, the initial population is manipulated using various genetic operators to produce new populations [11].

Scatter Search (SS) is an evolutionary algorithm, which shares some common characteristics with the GA. It operates on a set of solutions, the reference set, by combining these solutions to create new ones [27].

Hidden Markov Model (HMM) it differs from the normal statistical Markov model by having invisible states, but each state randomly generates one of the visible states. A hidden Markov model can be presented as the simplest dynamic Bayesian network [36].

Iterative Dichotomiser 3 (ID3) for dealing with symbolic data by expressing the knowledge as a decision tree [39].

Artificial Immune System (AIS) the human biological immune system has a number of fundamental characteristics that can be adapted as design principles for AIS applications in various problem domains [28].

Artificial Immune Recognition System (AIRS) both self/non-self cells and detector cells are represented as feature vectors. In order to reduce redundancy, ARB (Artificial Recognition Ball) is used which is representative of similar memory cells [31].

Artificial neural network (ANN) artificial neural networks were first created with the purpose to imitate the behavior of the human brain. A neural network is the connection of elementary objects called the simple neuron [32].

Multilayer Perception Algorithm (MPL) is an artificial neural network and is a nonparametric estimator that can be using for classifying and detecting intrusions [32].

Parenclitic Network (PN) a network reconstruction technique that allows highlighting the differences between one instance and a set of standard [33].

Multi-layer feed forward neural network (MLFF-NN) is one of the most common NN structures, as they are simple and effective, and have found home in a wide assortment of machine learning applications [11].

All predictive analytics applications involve three fundamental components:

• Data: A main functional component for analytics, is responsible for the quality and effectiveness of every predictive model

• Statistical modelling: From basic to complex, various some statistical modelling techniques, in which regression being common, are involved for insights and derivations of results.

• Assumptions: This being another important component helps to integrate the collected and analyzed data and design the pattern based on historical data and events.

IV. SYNTHESIS AND DISCUSSION

It could be observed that almost, all implemented algorithms, do not work in real time. As can be seen, the detection of credit card fraud uses several ML techniques, especially those of artificial intelligences and combines them with optimization techniques such as aggregation, for the detection of frauds of financial statements it is based mainly on text processing techniques. For fraud insurance, the non-necessity of the real-time processing, makes the detection of the fraud easier, nevertheless the difficulty resides in the fact that these deceptions are human and can be well masked. Comparing the logistic regression and the Bayesian results, we see that the Bayes logistic model gives posterior estimations for the true positive rate.

In financial statement fraud, results based on the accuracy indicated that the PNN was the best performing (98.09%) following by Genetic algorithm (95%) who gave marginally lower accuracies in most cases. Naives bays and SVM gives good results with NSL-KDD dataset (99,02%, 98,8%) for credit card fraud.

Also we found that: − The aggregation period has a major impact on the performance for fraud detection. Aggregating a product improves the prediction rate for all techniques except for CART.

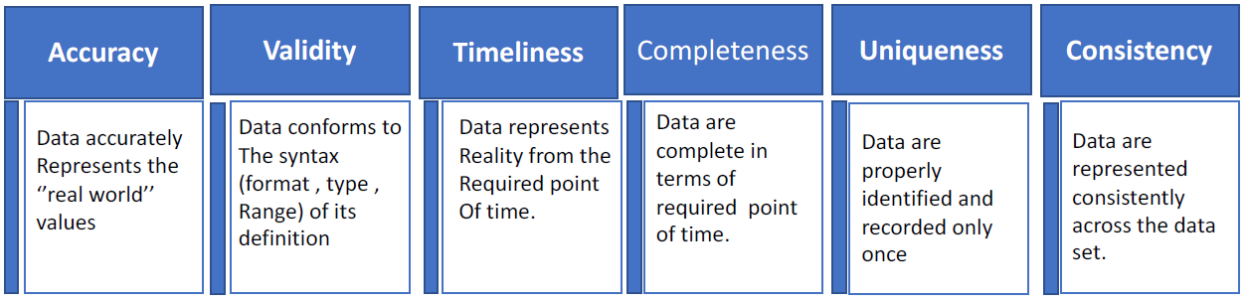
− SOM Clustering helps to identify new patterns hidden in the input data, which otherwise cannot be identified by traditional statistical methods, transaction filtering for further examination reduces overall cost as well as processing time.

− Cost-sensitive decision tree approaches is used in credit card fraud detection and show that it outperforms the models built using the traditional data mining methods such as decision trees, ANN and SVM

− Logistic Regression works well with linear data for credit card fraud detection.
− Support vector machine method is capable of detecting the fraudulent activity at the time of transaction.
− Complex networks can be used as a way to improve data mining models; they may be integrated as complementary tools in a synergistic manner in order to improve the classification rates obtained by classical data mining algorithms.
− KNN method can suit for detecting fraud with the limitation of memory. By the meantime, outlier detection mechanism helps to detect the credit card fraud using less memory and computation requirements. − Outlier detection works fast and well on online large datasets.

V. DATA PRE-PROCESSING

In the data pre-processing stage, we have to check for the quality of our data based on the 6 key dimensions. Our dataset can be assessed through these dimensions to decide whether the data can be used or not. First, the data needs to be complete and no missing values in it. Second, the dataset should follow the same format when representing the values in the attributes. Third, there should be no conflicting information between the values which could mislead our analysis. Fourth, the dataset should be accurate and up to date. Fifth, search for any duplicated values in the dataset. Finally, checking for missing data or not referenced.




```
> names(fraud)
 [1] "step"          "type"          "amount"
 [4] "nameOrig"      "oldbalanceOrg" "newbalanceOrig"
 [7] "nameDest"      "oldbalanceDest" "newbalanceDest"
[10] "isFraud"       "isFlaggedFraud"
```

Figure 2: Attributes in The Dataset

Figure 2 shows the names of the attributes in the dataset. We have a total of 11 names and each name represents an attribute in our dataset.

- 'step' = maps to a unit of time in the real world
- 'type' = is the type of transaction made
- 'amount' = amount of money transferred
- 'nameOrig' = person who initiated the transaction
- 'oldbalanceOrg' = the amount before the transaction
- 'newbalanceOrg' = the amount after the transaction
- 'nameDest' = the person who is the recipient of the transaction
- 'oldbalanceDest' = initial balance of the recipient before the transaction
- 'newbalanceDest' = new balance of the recipient after the transaction
- 'isFraud' = the transaction is fraud
- 'isFlaggedFraud' = the transaction is flagged fraud

```
> str(fraud)
spec_tbl_df [1,048,575 × 11] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ step          : num [1:1048575] 1 1 1 1 1 1 1 1 1 1 ...
 $ type          : chr [1:1048575] "PAYMENT" "PAYMENT" "TRANSFER" "CASH_OUT" ...
 $ amount        : num [1:1048575] 9840 1864 181 181 11668 ...
 $ nameOrig      : chr [1:1048575] "C1231006815" "C1666544295" "C1305486145" "C840083671"
 ...
 $ oldbalanceOrg : num [1:1048575] 170136 21249 181 181 41554 ...
 $ newbalanceOrig: num [1:1048575] 160296 19385 0 0 29886 ...
 $ nameDest      : chr [1:1048575] "M1979787155" "M2044282225" "C553264065" "C38997010" .
 $ oldbalanceDest: num [1:1048575] 0 0 0 21182 0 ...
 $ newbalanceDest: num [1:1048575] 0 0 0 0 0 ...
 $ isFraud       : num [1:1048575] 0 0 1 1 0 0 0 0 0 0 ...
 $ isFlaggedFraud: num [1:1048575] 0 0 0 0 0 0 0 0 0 0 ...
```

Model	Train Accuracy	Test Accuracy	Train F1-Score	Test F1-Score	Train AURROC	Test AURROC
SVM	0.9998	0.9993	0.9997	0.7286	0.9997	0.9279
Decision Tree	1.0	0.9997	1.0	0.8257	1.0	0.8878
Random Forest	0.9826	0.9895	.9799	00.1557	0.9814	0.9603

Table 1: Results of Models

The result indicates an accuracy of 99.9%, 72.8% F1-score and an Area Under Curve ROC of 92.8% on the validation dataset. The precision shows that for the dataset, the model has a chance of 1 in predicting legitimate transactions, and a 0.634 chance of predicting fraud transactions. The recall results indicate that the CatBoost model has a chance of 1 in going through the non-fraud cases but a 0.856 chance in going through the fraud transactions.

		Predicted	
		Positive (0)	Negative (1)
Actual	Positive (0)	167512	86
	Negative (1)	25	149

Table 2. Confusion matrix

The confusion matrix shows that 167512 transactions were correctly classified as legitimate cases and 149 ones were correctly labelled as fraud cases. These are the true positives and true negative predictions of the dataset from the t model. The misclassifications were 25 and 86 for false positive and false negatives respectively.

VI. COMPARATIVE RESULTS

With the review of different machine learning techniques and predictive analytics for analysis, it has been viewed of using predictive analytics over a financial dataset for fraud prediction is much more efficient with appropriate and better results rather than using core machine learning algorithms over it. A brief comparative summary of predictive analytics and machine learning over financial data is explained. Predictive analytics helps to build an answer on the basis of historical patterns and analytics used in the financial data statistics while machine learning core is able to provide personalized services to the customers on the basis of customer profile and preferences, making them access their financial transactions accordingly. Personalized graphs and charts illustrations help give a better perception to the management for decisions, with the help of predictive analytics while machine learning will help to automate the finance management. Machine learning delivers results or the next recommendation on the basis of insights in user profile whereas predictive analytics takes old of the entire data and its history and on that basis give the next result or decision count in business.

VII. CONCLUSION AND FUTURE WORK

The ability of machine learning models to analyze large amounts of data - both structured and unstructured – can improve analytical capabilities in risk management and compliance, is contributing to the financial organizations in an extremely effective and time-saving manner. This is helping companies make better and informed decisions, reducing loss factors. A lot of applications and development is empowering AI in the financial services and helping organizations serve better to their clients. Future research in this area and development olds to use automated machine learning and predictive analytics techniques for much better results and cost-saving architecture.

REFERENCES

[1] S. Maniraj, A. Saini, S. Ahmed and S. Deep Sarkar, "Credit Card Fraud Detection using Machine Learning and Data Science", International Journal of Engineering Research and, vol. 08, no. 09,2019. doi: 10.17577/ijertv8is09003.

[2] Ravindra Changala, “Secured Activity Based Authentication System” in " in Journal of innovations in computer science and engineering (JICSE), Volume 6, Issue 1,Pages 1-4, September 2016.ISSN: 2455-3506.

[3] A. Abdallah, M. Maarof and A. Zainal, "Fraud detection system: Asurvey", Journal of Network and Computer Applications, vol. 68, pp.90-113, 2016. doi: 10.1016/j.jnca.2016.04.007.

[4] R. Guha, S. Manjunath and K. Palepu, "Predictive Analytics For Insurance Fraud Detection - Wipro", Wipro.com.

[5] Ravisankar P, Ravi V, Raghava Rao G, and Bose, Detection of financial statement fraud and feature selection using data mining techniques, Elsevier, Decision Support Systems Volume 50, Issue 2, p491-500 (2011).

[6] Ravindra Changala, "Retrieval of Valid Information from Clustered and Distributed Databases" in Journal of innovations in computer science and engineering (JICSE), Volume 6, Issue 1,Pages 21-25, September 2016.ISSN: 2455-3506.

[7] K. Nagaraj and A. Sridhar, "A Predictive System for Detection of Bankruptcy Using Machine Learning Techniques", International Journal of Data Mining & Knowledge Management Process, vol. 5,no. 1, pp. 29-40, 2015. doi: 10.5121/ijdkp.2015.5103.

[8] Ravindra Changala, “Challenges and Solutions for the Semantic Web and Future of Document Management in Enterprises " in Journal of innovations in computer science and engineering (JICSE), Volume 6, Issue 1,Pages 10-13, September 2016.ISSN: 2455-3506.

[9] D.Zhang, L.Zhou, Discovering Golden Nuggets: Data Mining in Financial Application, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) (Volume: 34, Issue: 4), p513-522 (2004)