WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Digital Fraud Analytics: Detection and Prevention of Online Payment Scams

Ms. Sreeja S[1], Karpaga Varshini V[2], Bhuvaneshwaran S J[3], Kavya R[4], Saran D[5]

[1] *Faculty of Artificial Intelligence and Data Science, United Institute of Technology, Coimbatore - 641020, Tamilnadu, India*
[2], [3], [4], [5] *Artificial Intelligence and Data Science, United Institute of Technology, Coimbatore – 641020, Tamilnadu, India*
[1]sreejaprasadcse@gmail.com, [2]vkarpagavarshini@gmail.com, [3]sjbhuvaneshwaran@gmail.com, [4]kavyarajanbala03@gmail.com, [5]sarantrade123@gmail.com

A B S T R A C T

Digital fraud has become a significant concern with the rise of online transactions. Cybercriminals exploit vulnerabilities in payment systems, leading to financial losses and reputational damage. This project aims to develop a machine learning model using the Random Forest algorithm to detect fraudulent transactions. The dataset, sourced from Kaggle, undergoes pre-processing, feature engineering, and data balancing using SMOTE to improve fraud detection accuracy. The model is evaluated based on performance metrics such as accuracy, precision, recall, and F1-score. A Streamlit-based dashboard is implemented for real-time fraud detection visualization, offering an interactive interface for financial institutions and users. The project contributes to enhancing cybersecurity measures by providing an efficient and scalable fraud detection system. The system also generates a Fraud Level score, indicating the potential risk associated with each transaction. This Fraud level score helps in categorizing transactions based on their risk levels, enabling quicker decision – making for fraud prevention.

Keywords: Fraud Detection, Online Payment Scams, Digital Fraud Analytics, Machine Learning, Ensemble Learning, Decision Tree, Random Forest, Logistic Regression, Synthetic Minority Over-sampling Technique (SMOTE).

1. INTRODUCTION

The rapid increase in digital transactions has led to a parallel rise in fraudulent activities, posing threats to financial security. Traditional rule-based fraud detection methods are often ineffective due to evolving fraud patterns. Machine learning offers a dynamic and data-driven approach to identifying anomalies and suspicious activities. This project employs a supervised learning model Random Forest to predict fraudulent transactions with high accuracy. By leveraging data preprocessing techniques and synthetic data generation (SMOTE) to address class imbalance, the model aims to minimize false positives and false negatives. The results are visualized through an interactive dashboard using Streamlit, making it accessible for fraud analysts and financial institutions.

Problem Statement

Online payment fraud poses a significant threat to the security and integrity of digital financial systems. Traditional fraud detection methods often struggle to adapt to evolving fraud patterns, resulting in delayed detection and increased financial losses. Moreover, highly imbalanced datasets in fraud detection lead to biased models that fail to accurately identify fraudulent transactions. The absence of a system that not only detects fraud but also assesses its severity and provides insights into the cause exacerbates the problem. Therefore, this project aims to develop a machine learning-based system that can effectively detect fraudulent transactions, quantify the fraud level, and offer a concise explanation of the cause to facilitate better decision-making.

Objective

The Main objective is to develop a fraud detection system utilizing ensemble learning techniques that combine K-Nearest Neighbors (KNN), Decision Tree, Random Forest, and Logistic Regression to enhance prediction accuracy.

1. To implement the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance and improve the performance of fraud detection models.

2. To establish a mechanism that assigns a fraud level score to identified fraudulent transactions, providing an indication of the severity of the threat.

3. To provide a brief yet informative explanation outlining the probable cause of a fraudulent transaction, assisting stakeholders in understanding and mitigating security risks.

4. To validate the model's effectiveness through rigorous testing on diverse datasets, ensuring reliability and accuracy in identifying and classifying fraudulent transactions.

2. LITERATURE SURVEY

Fraud detection has evolved from rule-based methods to machine learning for better accuracy and adaptability. Breiman (2001) introduced Random Forest, improving fraud classification with ensemble learning. Chawla et al. (2002) proposed SMOTE to handle class imbalance and enhance fraud detection. Dal Pozzolo et al. (2015) found that ensemble models outperform single classifiers in credit card fraud detection. Carcillo et al. (2021) emphasized real-time fraud detection using adaptive machine learning models. West and Bhattacharya (2016) highlighted the benefits of dashboard integration for real-time insights. This project leverages Random Forest with SMOTE and a Streamlit dashboard for effective fraud detection and monitoring.

3. PROPOSED METHEDOLOGY

The research follows an experimental design to develop and evaluate a machine learning-based fraud detection system. The system is built to analyze historical transaction data, identify patterns indicative of fraudulent behavior, and assess the severity of detected fraud through a scoring mechanism. The approach combines supervised learning techniques with data balancing methods to ensure improved model performance. The research is conducted in multiple phases, starting with data preprocessing and feature selection, followed by model training, evaluation, and validation. The experimental design allows for iterative testing and fine-tuning of the models to achieve optimal accuracy and robustness in identifying fraudulent transactions.

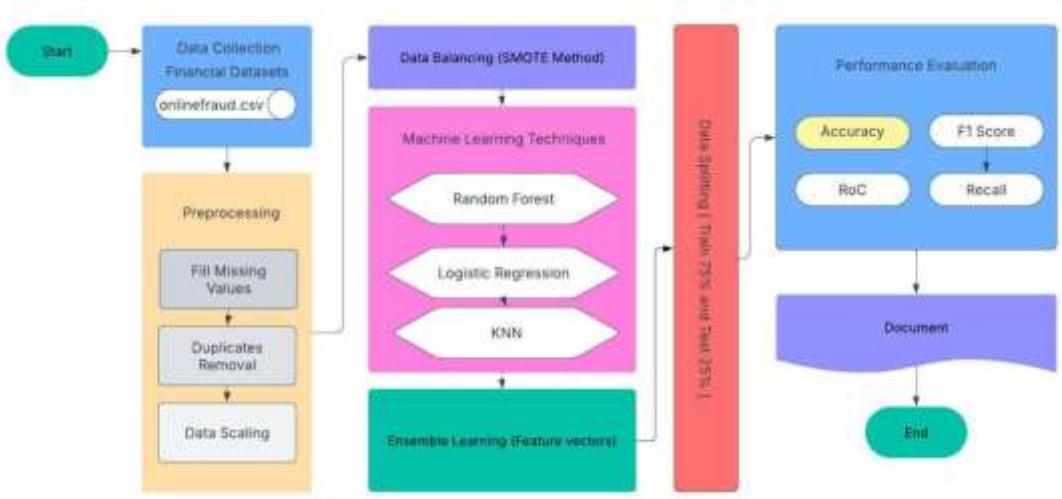

Fig. 1 – Proposed & Research Methodology

Data Collection

Source of Data: The dataset used in this project is obtained from Kaggle, which contains real-world online transaction records labeled as fraudulent or non-fraudulent. The dataset consists of multiple features, including transaction amount, transaction type, time, sender and receiver details, and fraud labels.

Preprocessing and Cleaning: To ensure high-quality input for the models, raw data undergoes rigorous preprocessing. Missing values, duplicate entries, and irrelevant features are handled through appropriate imputation and filtering techniques. Data normalization and encoding techniques are applied to transform categorical variables into machine-readable formats.

Balancing the Dataset: Fraud detection datasets are often highly imbalanced, with a very small percentage of fraudulent transactions compared to legitimate ones. To address this, Synthetic Minority Over-sampling Technique (SMOTE) is applied to oversample the minority (fraudulent) class, ensuring balanced representation and preventing model bias. This step significantly improves the model's ability to detect fraudulent transactions accurately.

Feature Selection: Key features contributing to fraud detection are selected based on their correlation with fraudulent behavior. Redundant or less significant attributes are eliminated using feature importance analysis and correlation techniques to enhance model efficiency.

Algorithm / Model Description

Logistic Regression

Logistic Regression is used to baseline model due to its simplicity, speed and effectiveness in binary classification problems, it provides interpretable results and is widely applied in fraud detection to understand the influence of input features.

K – Nearest Neighbors

KNN is chosen for its simplicity and effectiveness in small datasets, it doesn't require training and make predictions based on similarity, which can be helpful in detecting unusual patterns in fraud cases.

Random Forest

The Random Forest algorithm is selected for fraud detection due to its high accuracy, robustness, and ability to handle imbalanced data. It is an ensemble learning method that combines multiple decision trees to improve classification performance.

Application for Ensemble Learning

Ensemble learning techniques such as bagging and boosting are employed to combine predictions from multiple models, thereby enhancing the overall accuracy and minimizing variance. The use of Random Forest, which aggregates multiple decision trees, adds diversity to the model's decision-making process, while Logistic Regression and KNN contribute to fine-tuning decision boundaries.

Fraud Level Scoring and Cause Identification:

Upon identifying a fraudulent transaction, the system assigns a fraud level score to quantify the severity of the detected fraud. This score is calculated based on multiple factors such as transaction amount, frequency, and deviation from normal behavior. In addition, the system generates a concise explanation highlighting the probable cause behind the detected fraud, enabling stakeholders to gain deeper insights into suspicious activities.

4. IMPLEMENTATION

The implementation of the fraud detection system involves multiple steps, including data preprocessing, model training, testing, performance evaluation, and real-time dashboard deployment. The system is built using Python, with key libraries such as Pandas, Scikit-learn, Imbalanced-learn, and Streamlit. Below is a detailed breakdown of the implementation process.

Training and Testing Phase in Fraud Detection system

The Training and Testing phase plays a vital role in building a reliable fraud detection system. In this project, we employed three different machine learning algorithms Random Forest, K-Nearest Neighbors (KNN) and Logistic Regression to evaluate and compare their effectiveness in identifying fraudulent transactions. Each model was trained using historical transaction data and then tested on unseen data to assess its accuracy and performance. This comparative approach enabled the selection of the most accurate and efficient algorithm for detecting frauds.

Dataset Partitioning

Splitting the Dataset

The dataset is divided into two parts:

- **Training Set (75%)** – Train the model using Ensemble learning techniques.
- **Testing Set (25%)** – Used to evaluate the model's performance

Handling Class Imbalance with SMOTE

Fraudulent transactions are rare, causing class imbalance, where the majority class (non-fraud) dominates the dataset. To address this, SMOTE (Synthetic Minority Over-sampling Technique) is applied to generate synthetic fraud cases and balance the dataset. To avoid the Class Imbalance Issue, this will make model more reliable.

Training the Model

Fraudulent transactions are rare, causing class imbalance, where the majority class (non-fraud) dominates the dataset. To address this, SMOTE (Synthetic Minority Over-sampling Technique) is applied to generate synthetic fraud cases and balance the dataset. To avoid the Class Imbalance Issue, this will make model more reliable.

To develop a robust fraud detection system, we utilized three machine learning algorithms: Random Forest, K-Nearest Neighbors (KNN), Logistic Regression, each model was trained on a balanced dataset created using SMOTE to ensure fair learning.

1. Random Forest: Hyperparameters such as the number of trees (n_estimators), maximum depth (max_depth), and minimum samples per split (min_samples_split) were carefully defined. The ensemble nature of Random Forest helps in improving prediction accuracy and reducing overfitting by aggregating the results of multiple decision trees.
2. K-Nearest Neighbors: The KNN Model was trained by selecting an optimal value for k, which determines how many neighboring points influence the classification, the model relies on the distance metric to classify transactions based on their proximity to labeled instances.
3. Logistic Regression: This model was trained to establish a linear relationship between the input features and the probability of a transaction being fraudulent. Regularization Techniques were also considered to avoid overfitting and enhance generalization.

Testing the Model – Making Predictions

Once trained, the model is tested on the unseen testing dataset (25%) to assess its ability to detect fraud. The model takes the features (X_test) as input and predicts whether transactions are fraudulent (1) or non-fraudulent (0).

5. RESULT AND ANALYSIS

The fraud detection model using Random Forest achieved high accuracy and reliability in identifying fraudulent transactions. The model's performance metrics were:

- Accuracy: The base model had an accuracy of 96%, while our optimized Random Forest model achieved 99%, ensuring precise fraud detection.
- Precision: Reducing false positives.
- Recall: Capturing most fraudulent cases.
- F1-Score: Balancing precision and recall.
- AUC-ROC: Demonstrating strong classification capability.

The fraud risk scoring system categorized transactions into low, medium, and high-risk levels, helping financial institutions prioritize fraud investigations effectively. Comparative analysis showed Random Forest outperformed Logistic Regression and Decision Tree models, making it an optimal choice for fraud detection.

6. PERFORMANCE EVALUATION

To ensure model reliability and robustness, cross-validation techniques are applied during the training phase. Performance is evaluated using standard metrics such as accuracy, precision, recall, F1-score, and Area Under the Curve (AUC-ROC). The dashboard provides key insights, displaying whether a transaction is fraudulent or not, the associated risk score, transaction type, and risk level. These elements help in assessing fraud risk effectively, aiding in decision-making and fraud prevention.

6. 1. Model Evaluation Key metrics:

Accuracy: Measures the proportion of correctly classified transactions. However, accuracy alone is not sufficient due to data imbalance in fraud detection.

- Accuracy =TP+TN / TP+TN+FP+FN

Precision (Positive Predictive Value): Indicates how many transactions predicted as fraud are truly fraudulent.

- Precision = TP / TP+FP

Fraud Risk Score Calculation

The fraud risk score is a numeric value that represents the likelihood of a transaction being fraudulent. The model assigns a probability score (between 0 and 1) to each transaction, based on which a risk level is determined.

- The risk score is computed as: Risk Score = Probability (Fraud) × 100

6. 2. Fraud Risk level Classification:

Based on the risk score, transactions are categorized into different risk levels:

LOW, MEDIUM, HIGH

6. 3. Comparative Analysis with other models:

To validate the effectiveness of the Random Forest model, we compare its performance with other classifiers:

- Logistic Regression – Fast but less accurate for complex fraud patterns.
- Decision Tree Classifier – Works well but prone to overfitting.
- Support Vector Machine (SVM) – Effective but computationally expensive.
- XGBoost Classifier – Performs well but requires extensive hyperparameter tuning.

The results confirm that Random Forest achieves the best balance of accuracy, recall, and fraud detection efficiency.

7. REAL TIME FRAUD MONITORING DASHBOARD – UI

A Streamlit-based dashboard is implemented to visualize fraud detection insights in real time.

Key features include:

- Transaction Upload & Analysis: Users can upload transaction datasets for fraud prediction.
- Fraud Prediction Results: Displays fraud scores and levels for each transaction.

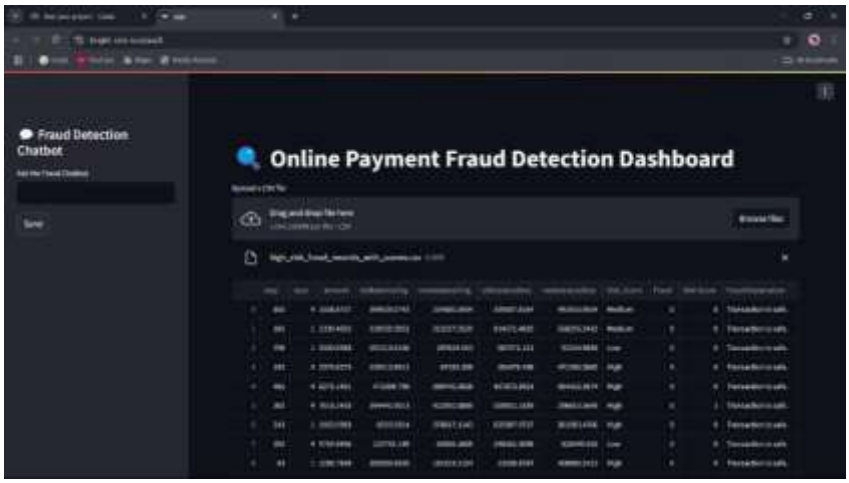

Fig. 2 – Fraud Online Payment Fraud Detection Dashboard

Fig.2 represents the home page of the Online Payment Fraud Detection System. This page allows users to upload individual transaction datasets in formats like CSV or Excel. It acts as the first step in the fraud detection workflow, enabling the system to begin analyzing transaction data for any suspicious or abnormal patterns. Once the dataset is uploaded, the backend processes it, and the results are displayed through various interactive dashboards and visualizations. This home page is essential for initializing the fraud detection process and sets the foundation for further analysis and decision-making.

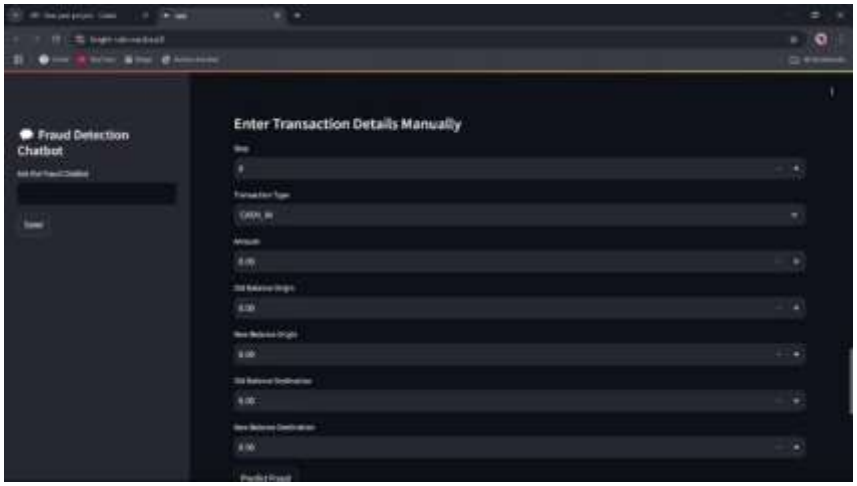

Fig. 3 – Fraud Manual Transaction Entry Page

Fig.3 displays the Manual Transaction Entry page, which allows users to manually input transaction details for real-time fraud analysis. Through this feature, users can enter customized transaction information to simulate and test the system's fraud detection capabilities. It provides a flexible way to validate individual transactions securely and observe how the model assesses the risk level associated with each entry. This page enhances the overall usability by supporting personalized testing and verification of transaction safety.

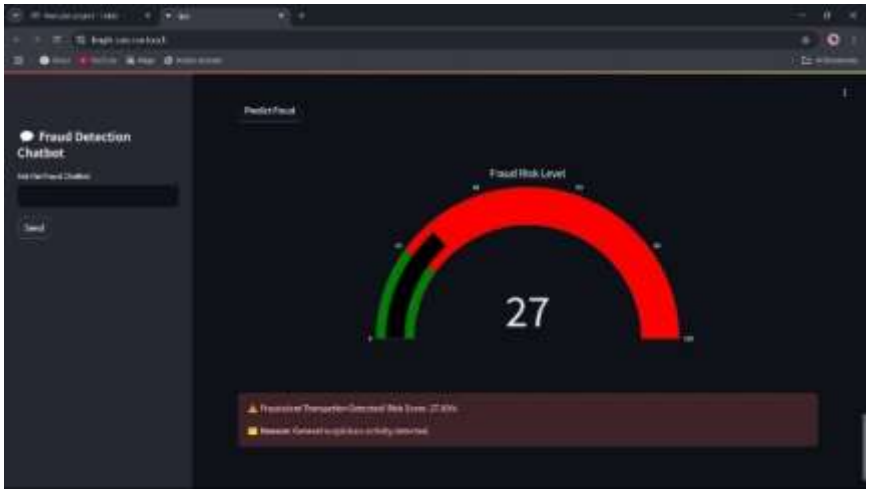

Fig. 4 – Fraud detection and Risk level score analysis

Fig.4 presents the Fraud Detection and Risk Score Level Analysis, which highlights the potential risk associated with each individual transaction. The system generates a fraud probability score that helps in identifying transactions that may be suspicious or high-risk. By analyzing these scores, users can easily differentiate between safe and potentially fraudulent activities. This feature enables quicker and more informed decision-making, allowing organizations or individuals to take preventive measures and strengthen their fraud management process.

Fig. 5 – Amount vs Risk Score Distribution

Fig.5 illustrates the Amount vs Risk Score Distribution, which visually represents the relationship between the transaction amounts and their corresponding risk levels. By analyzing this distribution, users can observe how the transaction value impacts the likelihood of fraud. It helps in identifying unusual patterns, such as instances where higher transaction amounts are associated with increased fraud risks. This deeper insight into amount-risk behavior assists in developing more targeted and effective fraud detection strategies, enabling better monitoring, prevention, and decision-making processes.

7. CHALLENGES AND LIMITATIONS

- Class Imbalance: Fraud cases are rare, requiring SMOTE and careful model tuning to improve fraud detection.
- Feature Selection Complexity: Some transaction features are less relevant, requiring rigorous feature importance analysis.
- Computational Cost: Training Random Forest with many trees can be resource-intensive, requiring optimization for efficiency.
- False Positives: Some legitimate transactions may still be misclassified as fraud, despite model optimizations.

- Scalability Issues: The model may need further tuning for real-time deployment in high-transaction environments.
- Adaptability: Fraud techniques constantly evolve, requiring regular model updates and retraining to stay effective.

8. CONCLUSION

This project successfully develops a machine learning-based fraud detection system using the Random Forest algorithm, addressing key challenges such as data imbalance through SMOTE and real-time detection via a Streamlit dashboard. The system significantly improves the security and accuracy of online payment transactions by dynamically adapting to evolving fraud patterns. Compared to traditional rule-based methods, it minimizes financial losses, reduces manual review efforts, and enhances operational efficiency. The model is trained and evaluated on relevant datasets, ensuring its robustness and generalization capability across diverse transaction scenarios. Additionally, the dashboard offers intuitive visualizations for quicker decision-making and proactive monitoring. Continuous model retraining strategies are also incorporated to maintain high detection performance against emerging fraud techniques. Overall, the project demonstrates the effectiveness of AI in combating digital fraud, offering a scalable, automated, and robust approach to modern fraud detection.

9. FUTURE SCOPE

The model can be enhanced for real-time fraud detection, allowing instant identification of fraudulent transactions. It can be improved in the future by adding more advanced models like deep learning to catch tricky fraud patterns, Deep learning techniques like LSTMs and neural networks can improve accuracy and adaptability. Adaptive learning will enable the system to update continuously with new fraud patterns, ensuring long-term effectiveness. Better methods can be use in deal with uneven data, and the project can be moved to the cloud to make it faster and easier to use on a larger scale. Scalability improvements will make it suitable for large-scale financial systems with high transaction volumes. Additionally, integrating multi-source data and AI-driven fraud response mechanisms will strengthen detection and automate high-risk transaction handling.

ACKNOWLEDGEMENTS

The completion of this project would not have been possible without the guidance and support of many people. We would like to express our sincere gratitude to Dr.H. Abdul Rauf, the Principal, United Institute of Technology for Providing us the necessary facilities for the project work.

We express our heartfelt gratitude to Dr.A. Kousalya, the Head of Department, Department of Artificial Intelligence and Data Science, United Institute of Technology for providing us the opportunity to carry out this project work and for his constant support and encouragement throughout the project.

Our Sincere thanks to our Project guide, Mrs. Sreeja, ME. Artificial Intelligence and Data Science, United Institute of Technology for her Constant support and Guidance.

References

1.Xia Y Zhang,Y & Chen, X, (2019). Fraud Detection in Financial Transactions Using Machine Learning Techniques. *International Journal of Financial Engineering*

2. Abdulwahab Ali Almazroi & nasir uib, (2023), Online Payment Fraud Detection Model using Machine Learning Techniques, *IEEE publication 4*

3. Abbasi. A, & liu H (2008). Fraud Detection in Online Auctions using Classification Algorithms. *International Conference on Data Mining 23 – 30.*

4. Raza. K, & Iqbal, W. (2014). A Comparative Study of Machine Learning Techniques in financial fraud detection. *International Journal of Computer Science 11(6)*

5. T. Brown, "Security in Online Payment Systems," Cybersecurity Journal, vol. 12, no. 2, pp. 15-20, 2022.

6. Ahmed, M,, & Mahmood, A, (2013). An Efficient Fraud detection for online Transactions, *International Journal of Information Technology, 6(2), 101 – 110.*

7. Kou, G,, Lu, L & Peng, Y, (2014), A Survey of Fraud Detection Techniques in Financial Transactions, *Expert Systems with Applications, 41(8), 4234 – 4260.*

8. Zhao, Y,, & he, L, (2018), An Overview of Machine Learning Methods for Fraud Detection in financial services, *Journal of Financial Services Technology, 3(2).*

9. Bojanowski, J,, & Indyk, P, (2015), Efficient Algorithms in Fraud Detection in Online Payments, *ACM Transactions on Knowledge Discovery from Data (TKDD)*

10. M. H. U. Sharif & M. A. Mohammed, "A Literature review of financial losses statistics for cyber security and future trend", *IEEE Access, vol. 15. pp – 138 – 156.*

11. S.Thudumu, P, Branch, J, Jin, and J. Singh, (2020). "A Comprehensive survey of Anomaly detection techniques for high dimensional big data". *J BIG DATA, Vol. 7*.