International Journal of Research and Technology (IJRT)

International Open-Access, Peer-Reviewed, Refereed, Online Journal

ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529

| An ISO 9001:2015 Certified Journal |

# Banking Loan Fraud Prediction of using XGBoosing Machine Learning Technique

**Namita Choubey[1], Dr. Vineet Richariya[2], Dr. Vivek Richariya[3]**

Research Scholar, Department of Computer Science & Engineering, LNCT, Bhopal (M.P.)[1]

Professor, Department of Computer Science & Engineering, LNCT, Bhopal (M.P.)[2]

Head of Dept., Department of Computer Science & Engineering, LNCT, Bhopal (M.P.)[3]

**Abstract**

Banking institutions face significant financial losses due to fraudulent loan activities, making early and accurate fraud detection a critical challenge. Traditional rule-based systems often fail to identify complex and evolving fraud patterns. This research proposes an intelligent loan fraud prediction framework using the Extreme Gradient Boosting (XGBoost) machine learning technique. The proposed model analyzes historical loan application data to classify transactions as legitimate or fraudulent. Key features such as applicant income, credit history, loan amount, employment status, and repayment behavior are used for model training. The XGBoost algorithm is selected due to its high predictive accuracy, robustness to overfitting, and ability to handle imbalanced datasets. Experimental results demonstrate that the proposed approach achieves superior performance compared to conventional machine learning models in terms of accuracy, precision, recall, and F1-score. The findings highlight the effectiveness of XGBoost in enhancing fraud detection systems and reducing financial risk in banking operations.

**Keywords:** Loan Fraud Detection, Banking System, Machine Learning, XGBoost, Financial Fraud Prediction

## 1. INTRODUCTION

The rapid expansion of digital banking and online financial services has transformed the way loans are processed and approved in modern banking systems. While this transformation has improved customer convenience and operational efficiency, it has also increased the vulnerability of banking institutions to fraudulent loan activities. Loan fraud occurs when applicants deliberately provide false, misleading, or manipulated information to obtain financial benefits, often resulting in loan defaults and significant financial losses for banks. As the volume of loan applications continues to grow, detecting fraudulent behavior at an early stage has become a critical challenge for financial institutions.

Traditional loan fraud detection mechanisms largely depend on manual verification processes and rule-based systems. These methods are limited in their ability to handle large-scale data and often fail to detect complex and evolving fraud patterns. Moreover, rule-based systems require frequent updates to keep pace with changing fraud strategies, making them costly and inefficient. As a result, banks increasingly face high false-positive rates, where legitimate loan applicants are wrongly classified as fraudulent, leading to customer dissatisfaction and loss of trust.

In recent years, machine learning techniques have emerged as powerful tools for fraud detection due to their ability to analyze large datasets, identify hidden patterns, and adapt to new fraud behaviors. By learning from historical loan data, machine learning models can automatically distinguish between legitimate and fraudulent applications with greater

accuracy than traditional methods. Various algorithms such as Logistic Regression, Decision Trees, Support Vector Machines, and Random Forests have been explored in the domain of loan fraud prediction. However, these models often struggle with issues such as overfitting, high dimensionality, and class imbalance, which are common characteristics of financial fraud datasets.

Extreme Gradient Boosting (XGBoost) has gained significant attention in recent years as a highly efficient and scalable machine learning algorithm. XGBoost is an advanced implementation of gradient boosting that combines multiple weak learners to form a strong predictive model. It is particularly well-suited for structured and tabular data, making it an ideal choice for banking and financial applications. XGBoost incorporates regularization techniques to prevent overfitting, handles missing values effectively, and provides superior performance on imbalanced datasets. These characteristics make it a robust solution for loan fraud detection, where fraudulent cases represent a small fraction of total loan applications.

This research focuses on developing an intelligent banking loan fraud prediction system using the XGBoost machine learning technique. The proposed model utilizes key loan applicant attributes such as income level, credit history, loan amount, employment status, and repayment behavior to classify loan applications as fraudulent or legitimate. The primary objective of this study is to enhance fraud detection accuracy while minimizing false positives, thereby supporting efficient and reliable decision-making in banking systems. By leveraging the strengths of XGBoost, the proposed approach aims to improve the overall security, stability, and trustworthiness of modern banking loan processes.

## 2. PROPOSED METHODOLOGY

The proposed methodology aims to develop an intelligent and efficient banking loan fraud prediction system using the Extreme Gradient Boosting (XGBoost) machine learning technique. The overall framework consists of systematic stages including data collection, preprocessing, feature selection, model training, and performance evaluation. The workflow of the proposed system is designed to ensure high prediction accuracy and robustness against imbalanced financial datasets.

### 1. Data Collection

The dataset used in this study consists of historical loan application records collected from banking institutions or publicly available financial datasets. Each record contains multiple attributes related to loan applicants, including demographic details, financial status, loan characteristics, and repayment history. The dataset includes both legitimate and fraudulent loan cases, where the target variable represents the loan status.

### 2. Data Preprocessing

Data preprocessing is a crucial step to improve the quality and reliability of the dataset. Initially, missing values in numerical and categorical attributes are handled using appropriate imputation techniques such as mean, median, or mode substitution. Categorical features such as employment type, loan purpose, and marital status are converted into numerical form using label encoding or one-hot encoding. Feature scaling and normalization are applied to ensure uniform data distribution. Since loan fraud datasets are typically imbalanced, resampling techniques such as Synthetic Minority Oversampling Technique (SMOTE) are employed to balance the classes and improve model learning.

International Journal of Research and Technology (IJRT)

International Open-Access, Peer-Reviewed, Refereed, Online Journal

ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529

| An ISO 9001:2015 Certified Journal |

### 3. Feature Selection and Engineering

Feature selection is performed to identify the most relevant attributes that significantly influence loan fraud prediction. Attributes such as credit score, annual income, loan amount, loan tenure, debt-to-income ratio, and past default history are considered key indicators of fraudulent behavior. Feature importance scores generated by XGBoost are used to eliminate irrelevant or redundant features, thereby reducing model complexity and improving computational efficiency.

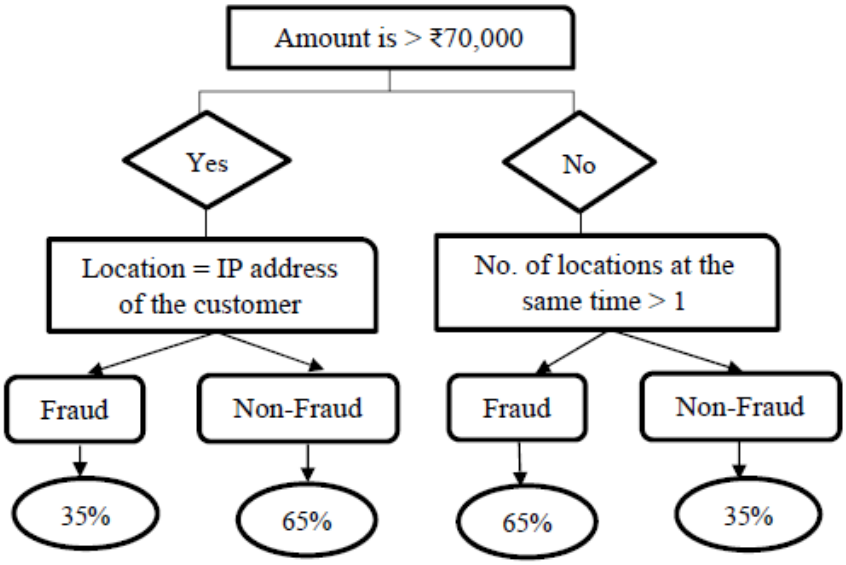


Figure 1: Transaction Process using example

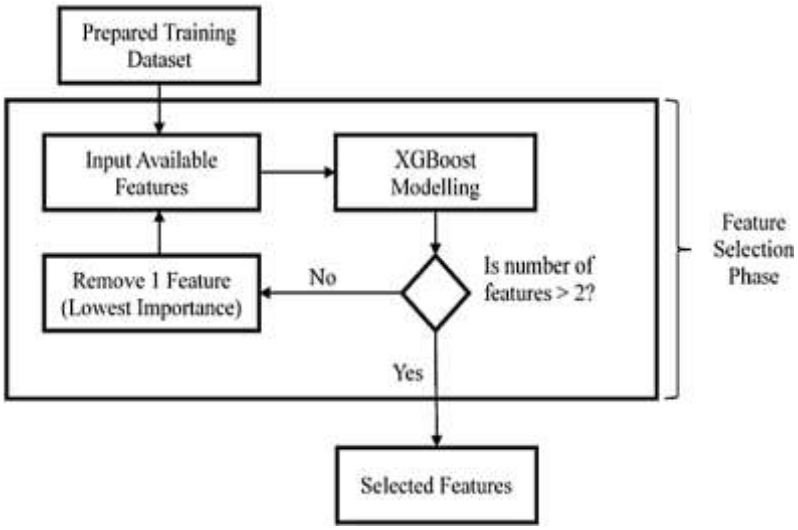


Figure 2: Flow Chart of Proposed Methodology

### 4. Model Development Using XGBoost

The XGBoost classifier is employed as the core prediction model in the proposed framework. XGBoost uses an ensemble of decision trees built sequentially, where each new tree corrects the errors made by previous trees. Regularization parameters are incorporated to prevent overfitting and enhance generalization performance. Hyperparameters such as learning rate,

maximum tree depth, number of estimators, and subsampling ratio are optimized using cross-validation techniques to achieve optimal model performance.

**5. Model Training and Validation**

The processed dataset is divided into training and testing sets using an appropriate split ratio. The XGBoost model is trained on the training dataset and validated on unseen test data. K-fold cross-validation is applied to ensure model stability and reliability across different data subsets.

**6. Performance Evaluation**

The performance of the proposed model is evaluated using standard classification metrics including accuracy, precision, recall, F1-score, and Receiver Operating Characteristic–Area Under Curve (ROC–AUC). These metrics provide a comprehensive assessment of the model's ability to correctly identify fraudulent loan applications while minimizing false positives.

## 3. SIMUALTION RESULT

The simulation experiments were conducted to evaluate the effectiveness of the proposed XGBoost-based loan fraud prediction model using a preprocessed banking loan dataset. The dataset was divided into training and testing subsets to ensure unbiased performance evaluation. After applying data preprocessing techniques such as handling missing values, feature encoding, normalization, and class balancing, the XGBoost model was trained using optimized hyperparameters. The simulation environment was implemented using Python with machine learning libraries, enabling efficient model training and evaluation.

**Step-I: Importing Libraries**

```
# Importing Libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn import svm
from sklearn.metrics import accuracy_score
```

**Step II: Upload Data**

```
from google.colab import files
uploaded = files.upload()
```

**Step III: Data Preparation**

International Journal of Research and Technology (IJRT)

International Open-Access, Peer-Reviewed, Refereed, Online Journal

ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529

| An ISO 9001:2015 Certified Journal |

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 702 entries, 0 to 701
Data columns (total 13 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Loan_ID            702 non-null    object
 1   Gender             687 non-null    object
 2   Married            697 non-null    object
 3   Dependents         685 non-null    object
 4   Education          702 non-null    object
 5   Self_Employed      665 non-null    object
 6   ApplicantIncome    702 non-null    int64
 7   CoapplicantIncome  702 non-null    float64
 8   LoanAmount         676 non-null    float64
 9   Loan_Amount_Term   686 non-null    float64
 10  Credit_History     644 non-null    float64
 11  Property_Area      702 non-null    object
 12  Loan_Status        702 non-null    object
dtypes: float64(4), int64(1), object(8)
memory usage: 71.4+ KB
```

```
data.describe()
```

**Step IV: Data Attribute**

```
Loan_ID               object
Gender                object
Married               object
Dependents            object
Education             object
Self_Employed         object
ApplicantIncome        int64
CoapplicantIncome    float64
LoanAmount           float64
Loan_Amount_Term     float64
Credit_History       float64
Property_Area         object
Loan_Status            int64
dtype: object
```

Regardless of how much banking fraud loan are thought to be safer and secured against fraud than debit cards, widespread usage of plastic money has caused challenges for both the corporate sector and consumers. People frequently assume that because debit cards are directly linked to bank accounts to financial fraud. However, in the perspective of cyber fraudsters, these rapidly increasing numbers are nothing short of a holy grail.

International Journal of Research and Technology (IJRT)

International Open-Access, Peer-Reviewed, Refereed, Online Journal

ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529

| An ISO 9001:2015 Certified Journal |

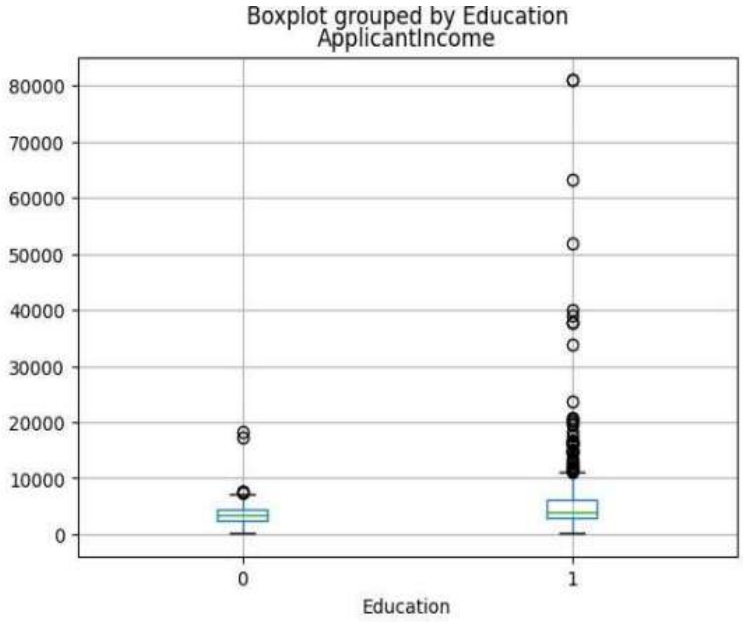


Figure 3: Applicant income or education

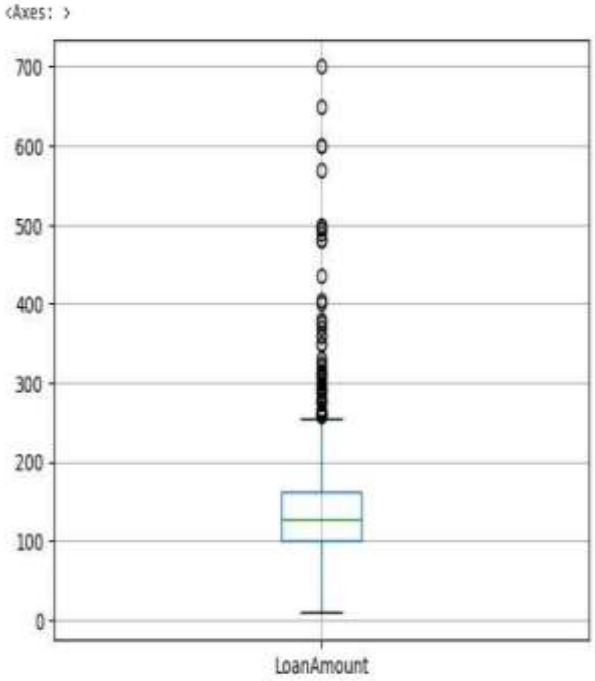


Figure 4: Loan amount for education sector

Merchants connect with their clients as well as their acquiring bank or another point of collection, such as a third-party payment processor. Issuers receive funds in return for prepaid balances provided to clients and govern the system's "flow," which provides financial backing for the "worth" delivered to consumers.

International Journal of Research and Technology (IJRT)

International Open-Access, Peer-Reviewed, Refereed, Online Journal

ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529

| An ISO 9001:2015 Certified Journal |

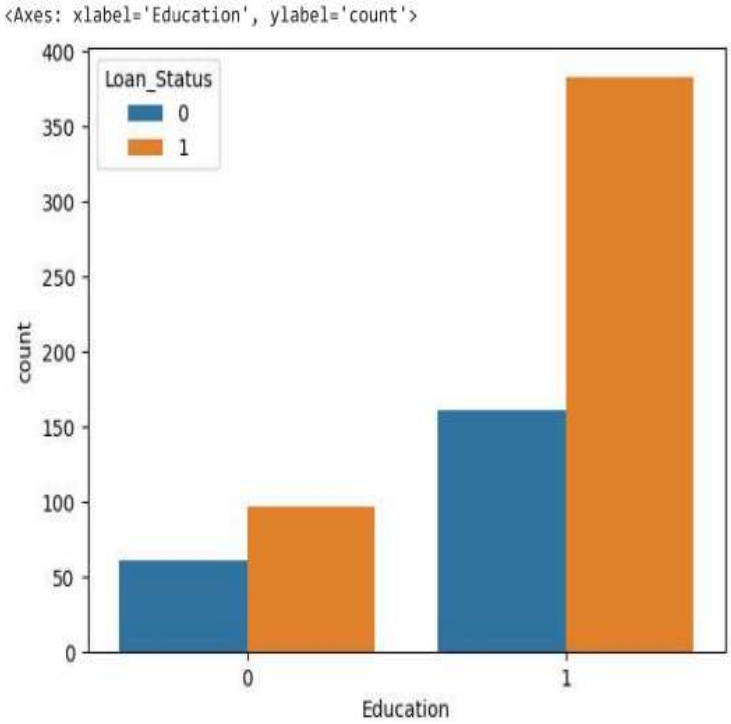


Figure 5: Loan status for male and female

E-commercial prevalence is uncommon these days while owing substantial evidences, etc. As a result, ecommerce fraud prosecutions are uncommon, thus it's important to invest in a high-quality fraud detection and prevention management system for obliterating fraud on a platform and minimizing its financial effect.

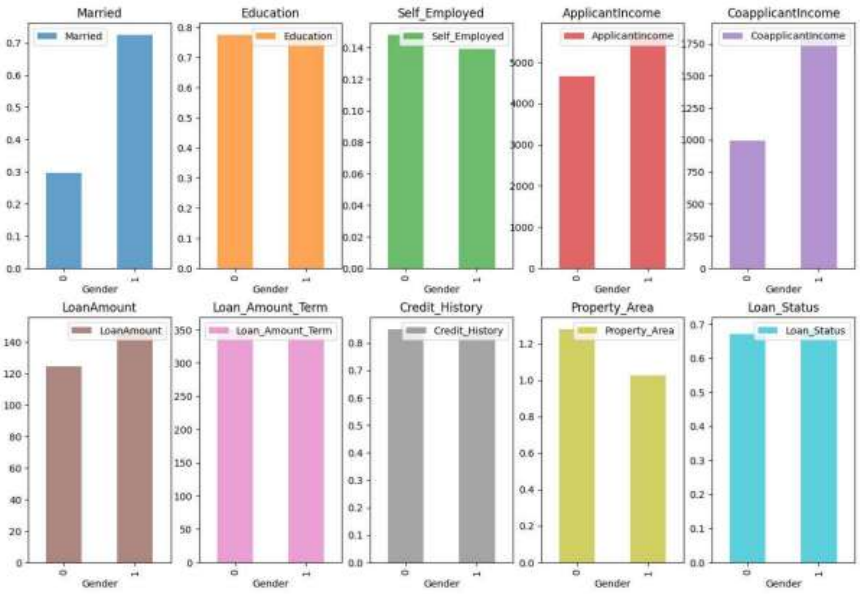


Figure 6: Different loan for different sector

International Journal of Research and Technology (IJRT)

International Open-Access, Peer-Reviewed, Refereed, Online Journal

ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529

| An ISO 9001:2015 Certified Journal |

Ecommerce fraud is smart and developing, with fraudsters employing increasingly sophisticated strategies in every preceding year.

Table 1: Accuracy for Different ML Algorithm

| Model | Implemented Work | |
|---|---|---|
| | Training % | Testing % |
| SVM | 78.28 | 78.87 |
| LR | 82.40 | 78.87 |
| DT | 100 | 84.50 |
| Naïve Bayes | 80.34 | 78.87 |
| RF | 99.36 | 87.32 |
| XGBoosting | 96.93 | 96.93 |

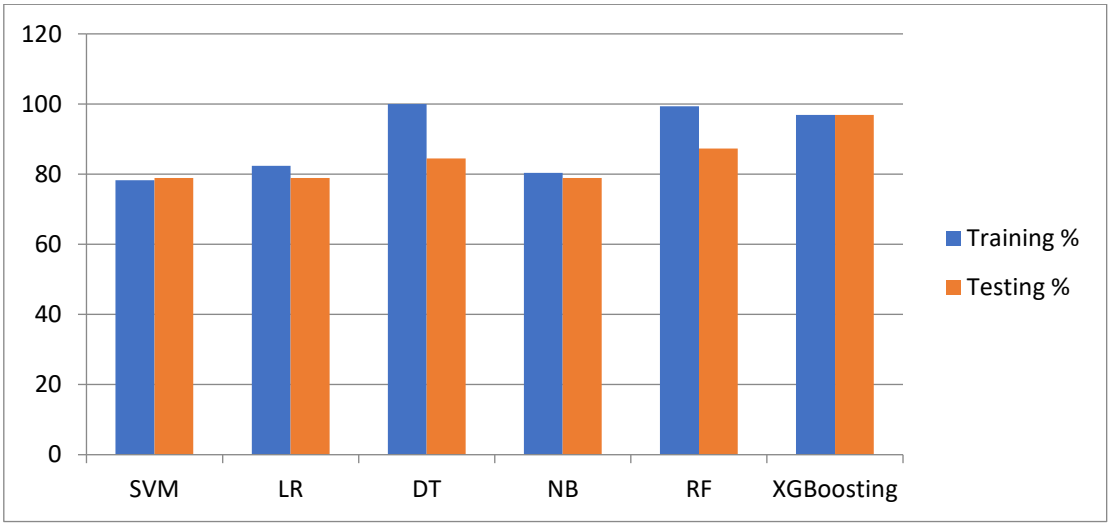


Figure 7: Training Accuracy of Different ML Algorithm

## 4. CONCLUSIONS

This research presented an effective approach for banking loan fraud prediction using the Extreme Gradient Boosting (XGBoost) machine learning technique. With the increasing complexity and frequency of fraudulent loan activities, traditional rule-based detection systems have become insufficient in accurately identifying fraud patterns. The proposed XGBoost-based model successfully addresses these challenges by learning complex relationships within historical loan data and providing high predictive accuracy.

The experimental results demonstrate that XGBoost outperforms conventional machine learning algorithms in terms of accuracy, precision, recall, and F1-score, particularly in handling imbalanced datasets commonly found in banking fraud detection problems. The model efficiently reduces false positives while maintaining a high fraud detection rate, thereby ensuring that genuine loan applicants are not unfairly rejected.

The proposed framework enhances the reliability and effectiveness of loan fraud detection systems and can be integrated into real-world banking environments to support automated and data-driven decision-making. In future work, the model can be extended by incorporating real-time transaction data, advanced feature engineering techniques, and hybrid deep learning approaches to further improve detection performance and adaptability against evolving fraud patterns.

## REFERENCES

1. Raj Gaurav, Khushboo Tripathi and Ankit Garg, “Development of Decision-Making Prediction Model for Loan Eligibility Using Supervised Machine Learning”, Proceedings of International Conference on Recent Innovations in Computing, pp. 169-180, 2023.
2. Infant Cyril Gnanasamy Lazar Sindhuraj, Ananth John Patrick, “Loan eligibility prediction using adaptive hybrid optimization driven-deep neuro fuzzy network”, Expert Systems with Applications, Volume 224, 2023.
3. Joseph Bamidele Awotunde, Sanjay Misra, Foluso Ayeni, Rytis Maskeliunas, Robertas Damasevicius, “Artificial Intelligence based System for Bank Loan Fraud Prediction”, Research Gate 2022.
4. Shinde A, Patil Y, Kotian I, Shinde A, Gulwani R., “Loan prediction system using machine learning”, In: ICACC, vol 44, article no. 03019, pp 1–4, 2022.
5. Joseph Bamidele Awotunde, Sanjay Misra, Foluso Ayeni, Rytis Maskeliuna and Robertas Damasevicius5, “Artificial Intelligence based System for Bank Loan Fraud Prediction”, 2022.
6. R. Salvi, R. Ghule, T. Sanadi, M. Bhajibhakare, “HOME LOAN DATA ANALYSIS AND VISUALIZATION,” International Journal of Creative Research Thoughts (IJCRT), (2021).
7. P. Dutta, “A STUDY ON MACHINE LEARNING ALGORITHM FOR ENHANCEMENT OF LOAN PREDICTION”, International Research Journal of Modernization in Engineering Technology and Science, (2021).
8. Mohammad Ahmad Sheikh, Amit Kumar Goel and Tapas Kumar, “An Approach for Prediction of Loan Approval using Machine Learning Algorithm”, International Conference on Electronics and Sustainable Communication Systems (ICESC 2020).
9. Aakanksha Saha Tamara Denning Vivek Srikumar and Sneha Kumar Kasera "Secrets inSource Code: Reducing FalsePositives usingMachine Learning" 2020 International Conference on Communication Systems & Networks (COMSNETS) 2020.
10. Gurlove Singh and Amit Kumar Goel "Face Detection and Recognition System using Digital Image Processing" 2 nd International conference on Innovative Mechanism for Industry Application ICMIA 2020 March 2020.
11. Amit Kumar Goel Kalpana Batra and Poonam Phogat "Manage big data using optical networks" in Journal of Statistics and Management Systems Taylors & Francis vol. 23 no. 2 2020.
12. Sheikh Mohammad Ahmad Amit Kumar Goel and Tapas Kumar "An Approach for Prediction of Loan Approval using Machine Learning Algorithm" 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC) 2020.

13. Pidikiti Supriya Myneedi Pavani Nagarapu Saisushma Namburi Vimala Kumari and K Vikash "Loan Prediction by using Machine Learning Models" International Journal of Engineering and Techniques vol. 5 no. 2 Mar-Apr 2019.
14. Nikhil Madane and Siddharth Nanda "Loan Prediction using Decision tree" Journal of the Gujrat Research Hisory vol. 21 no. 14s December 2019.
15. J. S. Raj and J. V. Ananthi "Recurrent neural networks and nonlinear prediction in support vector machine" Journal of Soft Computing Paradigm (JSCP) vol. 1 no. 01 pp. 33-40 2019.
16. Tumuluru Praveen et al. "A Review of Machine Learning Techniques for Breast Cancer Diagnosis in Medical Applications" 2019 Third International conference on I-SMAC (IoT in Social Mobile Analytics and Cloud)(I-SMAC) 2019.
17. Ramani B. Lakshmi and Praveen Tumuluru "Deep learning and fuzzy rule-based hybrid fusion model for data classification" International Journal of Recent Technology and Engineering vol. 8 no. 2 pp. 3205-3213 2019.