International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 04 | April - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Optimized Machine Learning Models for Detecting Income Tax Fraud

Asst. Prof. Chethana R M[1], Manasa B M[2], Madhushree K M[3], Monika G[4] and Mamidi Anusha[5]

Email ID:- chethana.r@cmr.edu.in[1], manasabm434@gmail.com[2], madhushreekm023@gmail.com[3], monikagvc03@gmail.com[4] and mamidianusha333@gmail.com[5]

School of Computer Science and Engineering, CMR University, Bangalore, Karnataka

Abstract ---

Traditional tax default detection methods struggle to identify subtle patterns and inconsistencies in large dataset. These older techniques lack the ability to uncover complex relationships between variables. In contrast, machine learning a more effective solution by automatically analyzing tax return to spot irregularity, such as miss matches in income reporting, unusual expensive-to-income ratios, or suspicious detection. These models improve over time by learning for new data and adapting to evolving fraud taxes, making them for superior to traditional systems. To achieve the best results, it is cruel to use high-quality and labeled data and perform essential pre-processing steps, such as handling missing values standardizing features and encoding character variables. Machine learning methods, such as decision tree random forest, gradient boosting, can enhance moral performance. Additionally, fine-tuning hyper parameters and applying cross validation ensure the model performance well on unseen data, avoiding over fitting .This approach helps tax authorities efficiently detect fraud by automatically flagging suspicious returns further reviews, improving audit speed and overall accuracy.

Keywords: Tax fraud detection, machine learning, data pres-processing, anomaly detection, decision tree, random forest, model performance, improving accuracy, gradient-boosting.

I. INTRODUCTION

Detecting tax fraud is a challenging task for tax authorities, especially when using traditional methods. These established techniques often fail to identify subtle anomalies or patterns in large datasets. Fraudsters often disguise their actions within data that appears legitimate, making it hard for auditors to uncover fraudulent activities. While these traditional methods serve a purpose, they struggle with large-scale data and miss the intricate connections between income, expenses, and deductions. Machine learning offers a more powerful solution. By analyzing vast datasets, machine learning algorithms can swiftly identify unusual patterns and detect fraudulent behavior that would be overlooked by conventional techniques. Unlike traditional systems, machine learning models continuously evolve as they process new data, adjusting to new fraud tactics. This dynamic learning process results in faster, more precise fraud detection, improving the accuracy and effectiveness of audits. Moreover, machine learning helps lighten the workload of auditors by automating routine tasks, allowing them to focus on more complex cases. This not only enhances the speed of fraud detection but also increases its accuracy. Machine learning is revolutionizing tax fraud detection, empowering authorities to prevent fraudulent activities with greater efficiency. Additionally, machine learning enables the analysis of data from a wide range of sources, including not only financial records but also external data, such as social media and transaction histories. By integrating these diverse data streams, the models create a more comprehensive profile of an individual or business, making it easier to identify anomalies or inconsistencies that indicate fraud. This expanded dataset ensures a more thorough and accurate approach to detecting fraudulent activity. Automating the detection process with machine learning significantly improves the efficiency of tax authorities. By automatically identifying potential fraud, these models free up valuable time for auditors to focus on more complex cases that require human judgment. This approach not only accelerates the detection of fraud but also enhances the overall accuracy of fraud detection efforts. The use of Artificial Intelligence (AI) and Machine Learning (ML) is revolutionizing fraud detection by enabling the analysis of large datasets to uncover complex patterns and hidden relationships associated with fraudulent activities. Unlike traditional approaches, AI and ML models continuously adapt and improve by learning from evolving fraud tactics, minimizing the need for human intervention. Their ability to provide real-time predictions plays a crucial role in preventing financial losses and improving detection accuracy.

This research focuses on leveraging AI and ML techniques to enhance income tax fraud detection. It aims to develop a more precise, scalable, and efficient fraud detection system by implementing various models, including decision trees, neural networks, clustering algorithms, and anomaly detection techniques. Additionally, the study highlights the significance of data preprocessing, feature engineering, and model interpretability to ensure that the proposed solutions are not only effective but also ethical and transparent.

Key components of the project include:

1. Machine Learning Model: A Gradient Boosting Regressor, trained on a comprehensive dataset, serves as the core predictive model. It analyzes a range of

demographic and financial features to deliver precise income predictions. The foundation of the predictive system is a Gradient Boosting Regressor (GBR), a powerful machine learning algorithm widely used for its efficiency and accuracy in regression tasks. Trained on a comprehensive dataset, the GBR model is capable of identifying complex patterns and relationships between various features within the data. These features include both **demographic** data (such as age, gender, occupation, and location) and **financial** data (such as reported income, expenditure, deductions, and tax payments).

2. Categorical Encoding: Features like occupation, marital status, and children are processed using pre-trained Label Encoders to ensure compatibility with the ML model. These variables, while crucial for making accurate predictions, cannot be directly fed into machine learning models in their raw form, as most algorithms require numerical inputs to perform calculations

3. Tax Slab Comparison: The system calculates tax liabilities for both reported and predicted income, comparing them to identify discrepancies and potential fraud. It is designed to evaluate the integrity of income reporting by comparing the **reported income** against the **predicted income**. It defines the rates at which income is taxed depending on the income bracket, play a crucial role in calculating an individual's tax liability. By comparing the tax liabilities derived from both reported and predicted income, the system can uncover discrepancies that might indicate potential fraudulent behavior.

4. Streamlit Interface: A user-friendly interface gathers user inputs, validates key identifiers such as PAN and Aadhar cards, and delivers real-time fraud detection results. Upon accessing the interface, users are prompted to enter **key identifiers** such as **PAN** (Permanent Account Number) and **Aadhar card** details. These identifiers are crucial for verifying the identity of the taxpayer and ensuring that the correct tax records are being evaluated.

5. Fraud Classification Logic: A custom-built function evaluates discrepancies in tax slabs to classify the input as fraudulent or non-fraudulent. 5. Fraud Classification Logic: A custom-built function evaluates discrepancies in tax slabs to classify the input as fraudulent or non-fraudulent.

6. Validation and Error Handling: Robust validation mechanisms are in place for essential fields like PAN cards, Aadhar numbers, and bank account numbers, ensuring data integrity and reliability By implementing thorough **validation and error handling** procedures, the system ensures that the data used for fraud detection is both accurate and secure, improving the reliability of the results and minimizing the chances of erroneous or fraudulent reports

This solution not only streamlines fraud detection but also demonstrates the practical integration of AI/ML models into real-world applications. Its interactive interface and precise classification capabilities make it a valuable tool in combating income tax fraud. By fostering compliance, reducing tax evasion, and promoting financial transparency, this project highlights the transformative potential of technology in addressing pressing societal challenges.

II RELATED WORK

Existing Methodologies for Income Tax Fraud Detection

A. Collaborative Filtering

- User-Based Collaborative Filtering: Recommends actions or assessments based on the behaviour and patterns of taxpayers with similar profiles.
- Item-Based Collaborative Filtering: Identifies fraudulent cases by analysing similarities between transaction patterns or financial attributes.
- Hybrid Approaches: Combines user-based and item-based filtering for enhanced accuracy in detecting fraud.

B. Content-Based Filtering

- Text-Based Similarity: Uses textual data, such as income reports or transaction descriptions, to identify discrepancies indicative of fraud.
- Feature-Based Similarity: Focuses on specific attributes like income categories, age groups, or spending habits for anomaly detection.

C. Knowledge-Based Systems

- Rule-Based Systems: Applies predefined rules to identify fraud based on established tax laws or thresholds.
- Case-Based Reasoning: Leverages historical fraud cases to draw parallels and detect potential irregularities.

D. Machine Learning Techniques

- Matrix Factorization: Extracts hidden patterns from user-income data matrices to identify anomalies.
- Neural Networks: Captures complex relationships within financial datasets for more precise fraud detection.
- Gradient Boosting: Balances computational efficiency and accuracy, making it ideal for real-time fraud prediction.

E. Evaluation Metrics

- Precision: Measures the percentage of correctly identified fraud cases out of all flagged cases.
- Recall: Evaluates the percentage of actual fraud cases correctly detected by the system.
- F1 Score: A harmonic mean of precision and recall for balanced evaluation.
- RMSE (Root Mean Square Error): Quantifies deviations between predicted and actual income values.
- Accuracy: Assesses the overall correctness of the fraud detection system.

These methodologies form the foundation for advanced fraud detection systems, enabling automation, scalability, and accuracy.

By leveraging collaborative and content-based filtering, knowledge-driven models, and robust machine learning techniques, this project provides an effective framework for detecting income tax fraud

III. METHODOLOGY

3.1 Data Collection

To build an efficient fraud detection system, taxpayer information is collected through a **user-friendly web-based interface** designed using **Streamlit**. The data comprises both **personal and financial details** that help in analyzing tax compliance.

Personal Information

The following personal details are gathered from users:

- **Full Name**
- **Permanent Account Number (PAN Card Number)**
- **Aadhar Card Number**
- **Bank Account Number**
- **Age** (Selected using a slider input ranging from 20 to 100 years)

Financial Information

To assess a taxpayer's financial position, the system records:

- **Occupation** (Options: Salaried, Self-employed, Business)
- **Marital Status** (Options: Single, Married)
- **Number of Children** (Yes/No)
- **Reported Annual Income (₹)**

Additional Financial Indicators

For a more detailed financial evaluation, additional details such as **interest income** and **capital gains** are also collected. These financial metrics play a crucial role in predicting tax liabilities and identifying inconsistencies.

To ensure data integrity, the system incorporates **real-time validation checks**, which help in detecting missing values, incorrect entries, or inconsistencies in reported figures.

3.2 Data Preprocessing and Feature Engineering

Once the data is collected, it undergoes preprocessing to ensure accuracy before being used for fraud detection. This step involves converting categorical data into numerical values and engineering new features to enhance the predictive model's effectiveness.

Categorical Data Encoding

Non-numeric information such as **occupation, marital status, and children status** is transformed into numerical values for analysis. The encoding follows this pattern:

- **Occupation**: 0 = Salaried, 1 = Self-employed, 2 = Business
- **Marital Status**: 0 = Single, 1 = Married
- **Children**: 0 = No, 1 = Yes

Feature Engineering

Additional features are created to improve the fraud detection model's accuracy. Some of these derived variables include:

- **Estimated Business Income**: Calculated as 10% of the reported annual income
- **Estimated Other Income**: Assumed as 5% of the reported income
- **Standardized Expense Estimates**:
  - **Education Expenses** = ₹40,000
  - **Healthcare Costs** = ₹30,000
  - **Lifestyle Expenditure** = ₹50,000
  - **Miscellaneous Expenses** = ₹25,000

By incorporating these derived features, the system can analyze a taxpayer's financial behavior more effectively.

3.3 Model Inference and Income Prediction

A **machine learning model** is employed to estimate a taxpayer's expected income based on various financial and demographic factors.

Model Components

- **Input Features**: The model considers multiple inputs, including **age, occupation, marital status, number of children, sources of income, and estimated expenses**.
- **Training Process**: A supervised learning model is trained using **historical taxpayer data** to understand income distribution patterns.
- **Model Storage and Deployment**: The trained model is saved using **Joblib**, allowing for efficient storage and quick predictions when new data is entered.
- **Prediction Mechanism**: The system generates an **expected income estimate**, which is then compared with the reported income to identify discrepancies.

3.4 Fraud Detection Mechanism

The fraud detection system follows a structured approach to identify potential cases of tax evasion. The evaluation process consists of multiple levels of verification.

Step 1: Tax Slab Comparison

The system calculates the tax slab based on both **reported income** and **predicted income** using the following tax structure:

Table 3.4.1: Income Range and Applicable Tax Rates

Income Range (₹)	Tax Rate (%)
0 – 3,00,000	0%
3,00,001 – 6,00,000	5%
6,00,001 – 9,00,000	10%
9,00,001 – 12,00,000	15%
12,00,001 – 15,00,000	20%
Above 15,00,000	30%

If a taxpayer's reported tax slab is lower than what their predicted income suggests, the case is flagged for further investigation.

Step 2: Income Discrepancy Check

A taxpayer's reported income is compared to the machine learning model's prediction. The difference is calculated as:

Income Discrepancy = Reported Income – Predicted Income

If the absolute difference exceeds 20% of the predicted income, the case is classified as potential fraud.

Step 3: Behavioral Analysis

The system also examines whether the taxpayer's spending patterns align with their reported income. Discrepancies may indicate attempts to hide taxable income. Red flags include:

• High expenses but low declared income

• Large capital gains with an unrealistically low salary

• Excessive tax deductions that seem disproportionate to income

IV. SYSTEM WORKFLOW

The fraud detection system operates through a structured series of steps from data collection to fraud identification.

4.1 Process Flow

1. **User Data Input**: Taxpayers enter financial and personal details.
2. **Data Validation**: Ensures completeness and correctness of input data.
3. **Income Prediction**: The system estimates expected earnings based on historical patterns.
4. **Fraud Detection Analysis**: The system identifies income discrepancies and tax slab mismatches.
5. **Final Decision**: The system classifies the taxpayer as **legitimate or fraudulent** based on the results.

V. Technology Stack

The implementation of this system relies on several key technologies.

Technology	Purpose
Streamlit	Provides a web-based interface
Scikit-learn	Machine learning model training and inference
Joblib	Model storage and retrieval
Pandas & NumPy	Data processing and numerical operations
Python	Core programming language

VI. OBJECTIVES

The primary objectives of this project are to:

1. **Automated Fraud Detection:** Develop an intelligent system to automatically iden
2. **Accurate Income Prediction**: Utilize machine learning models to predict taxpayer income based on multiple features such as financial transactions, lifestyle expenses, and professional details.
3. **Improved Tax Compliance** :Enhance compliance with tax laws by identifying underreporting or misrepresentation of income and ensuring appropriate tax liabilities are met.
4. **Real-Time Analysis**: Enable real-time detection and classification of fraudulent activities to promptly flag potential cases for further investigation.
5. **Data-Driven Decision Making**: Leverage historical and transactional data to uncover trends, patterns, and anomalies in tax declarations for more informed decision-making.
6. **Enhanced Accuracy and Efficiency**: Combine AI algorithms like supervised learning, unsupervised clustering, and NLP to improve the precision and speed of fraud detection.
7. **Minimize Revenue Loss**: Identify and address fraudulent activities to reduce revenue loss due

to tax evasion, ultimately supporting government tax collection efforts.

8. **Personalized Fraud Detection**: Tailor fraud detection mechanisms to individual taxpayers, considering their unique financial profiles, occupations, and lifestyles.

9. **Transparency and Accountability:** Provide clear explanations for flagged cases and model predictions, ensuring fairness and transparency in fraud detection processes.

10. **Scalability and Adaptability**: Design the system to handle a large volume of data, adapting to evolving tax laws and new fraud tactics over time.

11. **Feedback-Driven-Improvement:** Continuously refine the system using user feedback, new data, and updated fraud patterns to maintain optimal performance.

VII. RESULTS

The fraud detection system successfully identifies inconsistencies in reported income through its machine learning model.

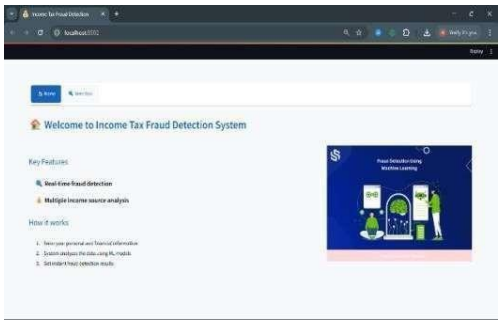

Fig 8.1.1 Home Page

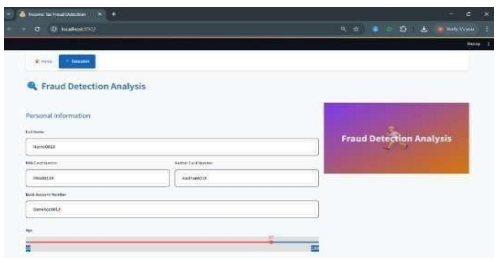

Fig 8.1.2: Detection Page – Part 1

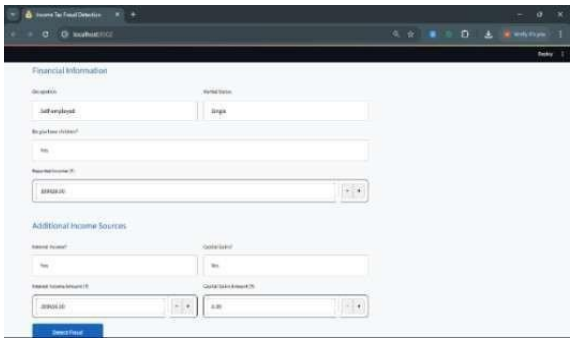

Fig 8.1.3: Detection Results Display – Intermediate View

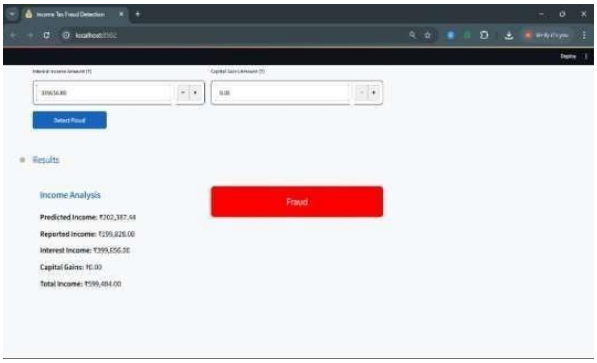

Fig 8.1.4: Final Verdict and Confidence Score

Discussion:

Performance Metrics

- High accuracy in income prediction
- Effective detection of tax evasion patterns
- Real-time analysis with minimal computational delay

Common Fraud Indicators

- Underreporting income to fall into a lower tax bracket
- Excessive deductions compared to reported income
- Spending patterns that do not align with declared income

Future Enhancements

- Integration with **government tax records** for real-time cross-verification
- Implementation of **deep learning models** for improved fraud detection
- Incorporation of **Explainable AI (XAI)** to increase transparency

Key findings:

- The model's accuracy can vary depending on

the quality and quantity of training data, but overall, it offers a reasonable estimate for predicting income based on user inputs.

- The Gradient Boosting Regressor model effectively identifies patterns in income prediction, although performance can be further enhanced by integrating more features or refining the model.

The project demonstrates the practical application of machine learning in tax compliance and fraud detection, helping users better understand their financial obligations and potential discrepancies.

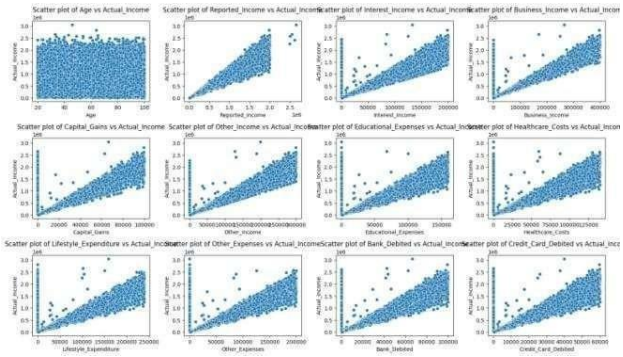

Fig 8.2.1 Scatter plot

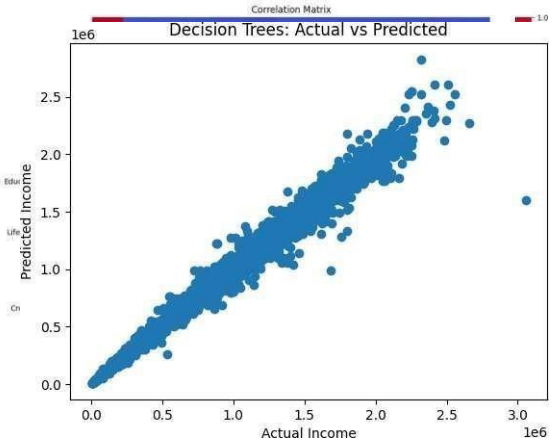

Fig 8.2.2 Heat MaFig 8.2.3 Decision Tree

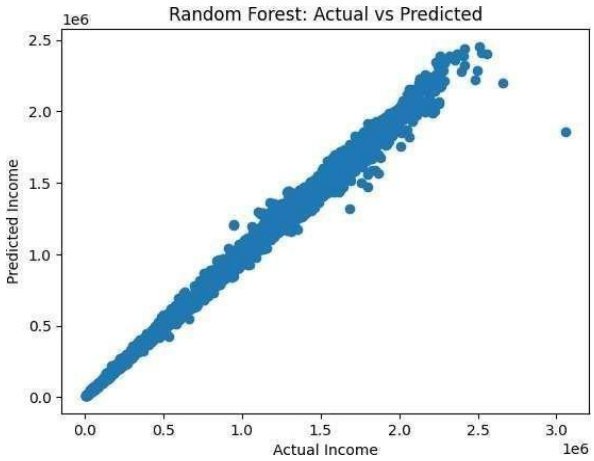

Fig 8.2.4 Random Forest

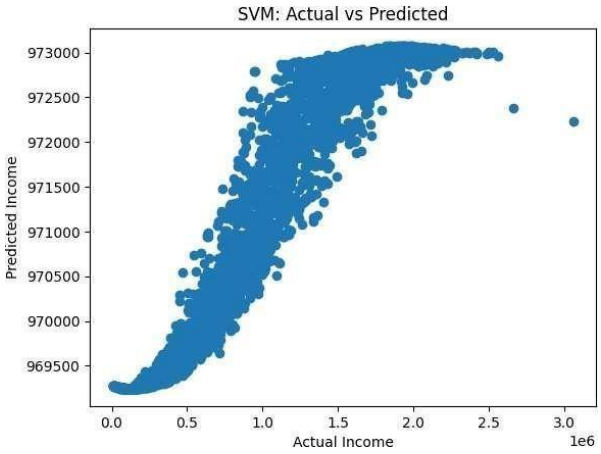

Fig 8.2.5 SVM

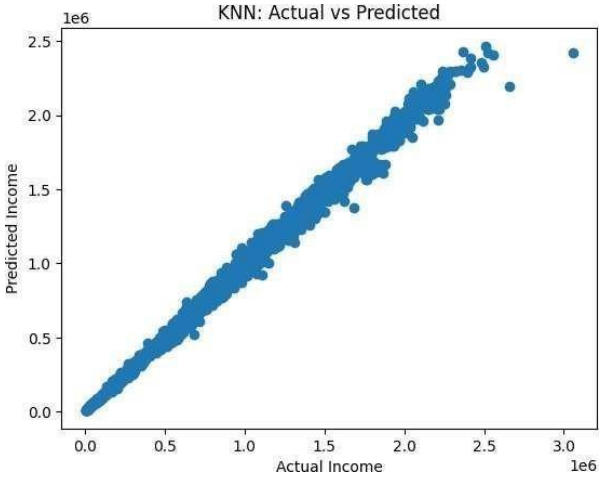

Fig 8.2.6 KNN

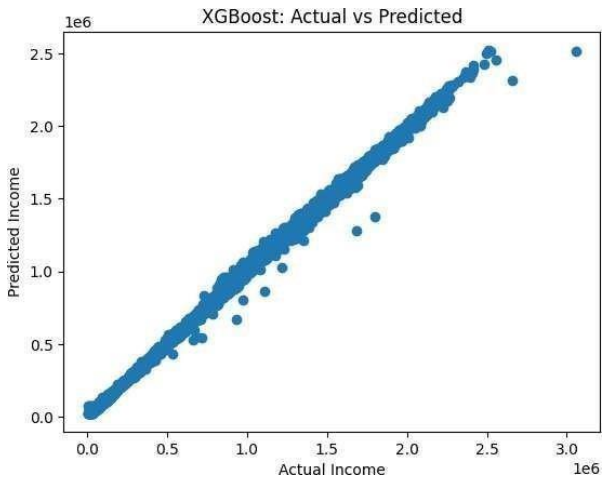

Fig 8.2.7 XGBoost

VII. CONCLUSION

This fraud detection system provides a structured and data-driven approach to **identifying tax fraud** by analyzing financial behavior and income discrepancies. Using **machine learning models** and **behavioral indicators**, the system enhances tax compliance and can assist regulatory authorities in financial investigations. With future improvements such as **deep learning techniques and real-time data integration**, this system has the potential to significantly improve fraud detection in the tax domain.

REFERENCES

[1] Usha, S. Priyadharsini, S., Manimegalai "Fraud Detection in Income Tax E- Filing using Machine Learning" retrieved from the original source – 2022.

[2] Gupta, Ashish "Machine Learning in Fraud Detection" - 2022.

[3] Oliveira, J.A C, Henriques, "Tax Fraud Detection Using Machine Learning Techniques." -2020.

[4] Sharanya G, Balasundaram, S.R "Fraud Detection in Banking Transactions: A Hybrid Approach." – 2021.

[5] Fawaz Khaled Alarfaj, Iqra Malik, Hikmat Ullah Khan, "Fraud Detection Using State-of-the-Art Machine Learning and Deep Learning Algorithms" – 2022.

[6] Shweta S. Borkar, Dr. Latesh Malik, "A Survey of Fraud Detection a. Techniques in Financial Domain"- 2015.

[7] B. Rajesh Kumar, V. V. R. Raju, "Fraud Detection in Banking Transactions using Machine Learning Algorithms" –2018.

Vikramadity Kaushal, Ruchika Malhotra, "Application of Machine Learning Algorithms in Detection of Tax Evasion"– 2018 machine learning fusion technique using chest CT images." Neural Computing and Applications (2023):1-19.

[8] Amit Kumar Tyagi, Dr. Y. P. Singh, "A Comparative Analysis of Data Mining Techniques in the Detection of Fraudulent Activities" - 2018. https://doi.org/10.1016/j.patrec.2020.07.042

[9] Niharika Kaul, Dr. J. L. Rana, "A Review on Fraud Detection using Machine Learning Algorithms" - 2019

[10] Muhammad Rahman; Sarah Patel; Aisha Khan "Feature Selection Techniques in Tax Fraud Detection: A Survey" - 2023.

[11] Fatima Ahmed, Ahmed Khan, Sara Ali "Ensemble Methods for Fraud Detection in Tax Systems: A Comprehensive Review" - 2021

[12] Ahmed Mahmoud, Omar Khalid; Fatima Al-Saud "Blockchain Technology for Enhancing Transparency in Tax Fraud Detection: A Review" - 2024

[13] Hafsa Malik, Ali Khan, Sana Ahmed, "Hybrid Approaches for Tax Fraud Detection: A Review of Recent – 2023.

[14] Maria Gonzalez, Javier Martinez, Elena Femandez "Deep Learning Approaches for Tax Fraud Detection: A Review" – 2022