

Machine Learning-based GST Fraud Detection using Support Vector Machine (SVM) for Enhanced Tax Governance

Darshan Rao M H¹, Soujanya B K², Kavya B S³, Janhavi Pramod Kulkarni⁴ and Gourav R⁵

¹⁻⁵School of CSE, REVA University, Bengaluru, India

Email: darshanrao126@gmail.com, sbk923@gmail.com, kavyabshivanand08@gmail.com, kulkarni.janhavi282@gmail.com, gouravr2210@gmail.com

Abstract— The Goods and Services Tax (GST) was implemented as a significant tax reform to increase revenue collection, simplify taxation, and promote transparency. Nonetheless, GST fraud is still a major problem that has an impact on both economic stability and government revenue. False invoicing, taking advantage of the Input Tax Credit (ITC), and circular trade are examples of fraudulent acts that cause significant financial losses and interfere with ethical company procedures. In addition to lowering tax collections, these fraudulent techniques make it more difficult for regulatory bodies to conduct thorough audits and investigations. This study suggests a machine learning-based fraud detection solution that employs the Support Vector Machine (SVM) classifier, which achieves an accuracy of 97.51%. Through the examination of important financial indicators, including invoice amounts, taxable values, ITC claims, and buyer-seller transaction patterns, the system can accurately spot suspicious activity. To enhance model performance, data preparation methods such as Recursive Feature Elimination (RFE) for feature selection and SMOTE for class imbalance handling are used. The re-search's conclusions help to strengthen tax governance and guarantee a stronger, more open GST system.

Index Terms— GST Fraud, Tax Compliance, Input Tax Credit, Regulatory enforcement, Revenue Integrity, Financial losses, Fraud Detection, Support Vector Machine, Feature Selection and Risk Assessment.

I. INTRODUCTION

Taxation plays a crucial role in a country's economic structure; they are categorized into two types: Direct and Indirect Taxes. Direct taxes, such as income tax and corporate tax, are levied directly on individuals and businesses. By contrast, indirect taxes are levied on goods and services and are collected from customers through middlemen. In order to improve their tax collecting methods, many developing countries have implemented the Goods and Services Tax (GST), one of the most important types of indirect tax.

Since a significant amount of their economies function informally, developing nations frequently use indirect taxes to broaden their tax bases. The development of parallel markets, where a sizable portion of the value chain is left unmonitored by regulatory bodies, is a key concern in these economies. In order to tackle this issue, multi-point taxation schemes like the Goods and Services revenue (GST) have been implemented, guaranteeing revenue collection across the supply chain.

On July 1, 2017, India's Value Added Tax (VAT), Excise Duty, and Service Tax were replaced by the Goods and Services Tax (GST) [1].

GST has increased revenue collection, but it has also made fraud more likely, which has resulted in a large amount of tax evasion. Approximately ₹3.08 lakh crore in GST evasion was recorded between July 2017 and February 2023, with ₹1.03 lakh crore recovered, according to public disclosures made in Parliament. Additionally, ₹1.51 lakh crore in GST evasion was found in the fiscal year 2023–2024 alone [2]. The GST system is susceptible to manipulation since it relies on self-assessment.

The government imposes taxes as obligatory financial duties in order to pay for public services. Laws control the administration and organization of the collection and compliance procedures. Since indirect taxes, like GST, are imposed at several points throughout a product's value chain, they make a substantial contribution to the national coffers. Every transaction creates comprehensive records that are recorded on official tax invoices, including tax classifications and associated rates. Due to frequent changes in tax legislation, exemptions, and variable tax rates for different sectors, tax compliance is still difficult even with the structured framework.

The exchequer suffers significant losses as a result of fraudulent activities like falsifying invoices, underreporting sales, and misreporting transactions, which eventually undermine public services.

The Authorities carry out audits and compliance inspections to combat tax evasion. However, manual auditing is ineffective due to the number and complexity of financial transactions. Systematic methods that examine transaction data for are necessary to spot fraud trends and inconsistencies. A variety of techniques are used to identify tax fraud, with an emphasis on irregularities, discrepancies, and taxpayer behavior patterns. This paper explores GST fraud, common fraudulent practices, and measures to improve tax compliance. It aims to provide a structured understanding of the challenges associated with tax evasion and potential strategies to enhance monitoring and enforcement mechanisms.

II. RELATED WORKS

A study on the economic effects, regulatory reactions, and patterns of GST tax evasion in India was carried out in 2024 by Helly Mukesh Maniya and Shobha V. [3]. In order to investigate several evasion strategies, including underreporting sales, creating false invoices, and abusing the Input Tax Credit (ITC), their research employed case study analysis. The impact of GST evasion on public trust, government revenue, and economic stability was also assessed in the study. In order to evaluate the efficacy of regulatory measures in lowering tax fraud, it also looked at enforcement actions and policy revisions. To reduce fraudulent activity, the authors underlined the necessity of better compliance monitoring, regulatory changes, and taxpayer education. Policymakers and tax authorities looking to improve GST enforcement and guarantee equitable taxation methods will find the study to be a useful resource.

A big data analytics strategy was proposed in 2019 by Priya Mehta, Jithin Mathews, Sandeep Kumar, K. Suryamukhi, Ch. Sobhan Babu, and S. V. Kasi Visweswara Rao to identify fraudulent transactions in the tax system, with a particular focus on Goods and Services Tax (GST) fraud in India [4]. A two-step process for identifying fraudulent groups involved in circular trading was proposed by their research. Initially, the problem was formulated as a social network analysis task, in which merchants and their transactions were arranged in a graph structure. To identify illegal activity, particular fraudulent transaction cycles were eliminated in the second stage. The report emphasized the dangers of circular trade, in which organizations set up shell corporations and fake transactions in order to avoid paying taxes. Their method's practical usefulness was demonstrated when it was used for the Commercial Taxes Department of the Government of Telangana, India. The study underlined the need for sophisticated monitoring systems to stop revenue losses brought on by tax evasion.

To determine how well different Support Vector Machine (SVM) variations detect GST fraud, Vivek Vyas, Himanshu Thakkar, Siddharth Dabhade, Ramdas Gore, and Hetal Thaker carried out a comparison study in 2025 [5]. By examining various SVM models, such as linear SVM, polynomial SVM, and radial basis function SVM, their study aimed to detect fraudulent tax transactions. To identify the best model for GST fraud detection, the study looked at scalability, classification accuracy, and computing efficiency. The results demonstrated how machine learning approaches may be used to detect fraudulent behaviors such the creation of shell companies, phony invoices, and misreported transactions. In order to help tax authorities reduce revenue losses and improve compliance, the authors underlined the significance of improving fraud detection techniques. Their research offers important new information on how to choose the best classification models for tax system fraud detection. Anju Sharma and Dr. Kapil Kumar investigated the identification of GST fraud in 2021 using a survey-based study and suggested incorporating sophisticated detection techniques [6]. Their study brought to light the rising

issues of tax evasion, false invoicing, and abuse of the GST system's Input Tax Credit (ITC). The study evaluated the shortcomings of conventional detection techniques and identified important fraud strategies by obtaining information from tax professionals and financial specialists. The authors suggested improved regulatory enforcement and systematic compliance monitoring, emphasizing the necessity of ongoing advancements in fraud detection techniques. In order to guarantee improved tax governance and revenue integrity, their findings emphasized the significance of creating proactive fraud detection procedures.

A study on GST Input Tax Credit (ITC) fraud was carried out in 2023 by Dr. Himanshu Thakkar, Ms. Vedanshi Joshi, and Ms. Rishita Bhatt. The study looked at the fraudulent schemes and how they operate [7]. The study demonstrated how scammers take advantage of the GST system by using shell corporations and false billing to avoid paying taxes, which causes the government to lose a significant amount of money. The Delhi shell company scam and the Bhavnagar false billing scandal were two significant GST fraud incidents in India that

A 2019 study on GST fraud in Malaysia by Zaleha Othman, F. F. Nordin, Zainol Bidin, and Muzainah Mansor examined different fraudulent tactics and how they affected tax compliance [8]. Falsifying claims, sales manipulation, non-submission of GST forms, failure to register, GST avoidance, and carousel fraud are the six main types of GST fraud that were found in their investigation. According to the survey, fraudsters use tax law loopholes to avoid paying taxes, which causes the government to lose a lot of money. The authors stressed the need for more robust regulatory frameworks and improved detection methods due to the growing sophistication of fraud in the GST system. Additionally, they suggested steps to reduce fraud, such as more stringent enforcement, improved compliance monitoring, and raised taxpayer awareness. Their findings provide insights into improving tax governance and aid in the study of GST fraud typologies.

A thorough mapping study on tax compliance and fraud detection in trading operations for goods and services was carried out in 2021 by Douglas Silva and Sergio Carvalho [9]. The difficulties in tax collection, such as inaccurate tax returns, false claims, and non-compliance, were examined in their study. The authors identified important issues in tax fraud, including underreported transactions and evasion strategies, by evaluating the body of existing literature. In order to improve tax governance and reduce revenue losses, their findings highlighted the need for organized compliance structures and effective fraud detection techniques. The study lays the groundwork for future investigations into enhancing fraud prevention techniques and fortifying tax systems.

In order to identify financial fraud, Hao Sun et al. [10] investigated the use of textual risk disclosures in financial reports. In order to assess the textual risk disclosures of 8999 firm-year observations from U.S. energy businesses (2006–2019), the study included linguistic features, specifically part-of-speech (POS) tags. Quantified POS characteristics, including nouns, verbs, and adjectives, were fed into machine learning models such as Random Forest, Support Vector Machine, Logistic Regression, and XGBoost. With the maximum AUC of 73.89% obtained using XGBoost when integrating POS features with conventional financial variables, empirical data showed that the POS features considerably improved fraud detection capabilities. The usefulness of forward-looking textual data in enhancing structured financial indicators for fraud detection was demonstrated by this method.

Numerous research has examined the use of machine learning and data mining techniques for risk management and fraud detection, each of which has offered valuable information about how to enhance financial and tax system monitoring. Self-Organizing Maps (SOMs) were suggested by Wilson T. Mongwe et al. [11] (2020) as a way to identify fraud in municipal financial accounts. Since SOMs are an unsupervised learning method, they don't need labels, which makes them perfect for datasets with a lack of fraud labels. Fraud detection is based on these clear patterns in financial behavior, and their research showed that financial ratios such as debt-to-revenue and current ratios may be used effectively to cluster towns into fraudulent and non-fraudulent categories.

Desy Jayanti et al. [12] (2021) carried out a systematic assessment of machine learning applications in the taxation domain, highlighting the increasing usage of supervised learning algorithms for fraud detection. Their research examined 29 studies that successfully used support vector machines (SVMs), decision trees, and neural networks to identify tax fraud and ensure compliance. They found that when it came to identifying taxpayers who were in compliance and those who weren't, these models—in particular, neural networks and SVMs—achieved accuracy rates between 80% and 95%. These models are crucial weapons in the battle against tax evasion because of their remarkable success in identifying trends that traditional methods could miss.

A systematic review of the literature on machine learning (ML) methods for financial fraud detection was carried out by Ludivia Hernandez Aros et al. [13]. With a focus on supervised and unsupervised methods used on actual datasets, they examined 104 papers released between 2012 and 2023. The review emphasized how common supervised learning techniques like Random Forest, Logistic Regression, and Support Vector Machines are for identifying business, insurance, and credit card fraud. Furthermore, they found that actual datasets were prioritized over synthetic ones, with the latter being adopted at a low rate (less than 7%). Significant geographic

contributions from Saudi Arabia, China, and India were also identified by the study. The findings, which include performance parameters including accuracy, precision, recall, and F1 score, illustrate the function of machine learning in identifying irregularities in financial systems.

Asifuddin Nasiruddin Ahmed and Ravinder Saini's paper "A Survey on Detection of Fraudulent Credit Card Transactions Using Machine Learning Algorithms"[14] offers a thorough analysis of the different machine learning techniques applied to credit card fraud detection, highlighting the difficulty of highly imbalanced datasets, where fraudulent transactions make up only 0.173% of the total data. The study demonstrates the effectiveness of ensemble methods such as Random Forest and XGBoost, the latter of which in some circumstances can achieve accuracy of up to 99.96%. In order to balance datasets and enhance the identification of fraudulent cases, the authors additionally stress the significance of sampling techniques like NearMiss (undersampling) and SMOTE (oversampling). For example, the accuracy of 99.97% with 85.71% recall and 91.30% precision was achieved by combining SMOTE with Random Forest. In order to improve fraud detection, the study highlights the importance of preprocessing steps like data cleaning and feature engineering and makes recommendations for future research topics including neural networks and hybrid models.

III. METHODOLOGY

The proposed system utilizes machine learning classification models for GST fraud detection. To comprehend patterns and correlations in the data, these models must be trained on a well-chosen dataset. Datasets used for fraud detection frequently show class imbalance, with a disproportionately low number of fraudulent transactions compared to valid ones. Class imbalance can be addressed by over-sampling or undersampling approaches such as SMOTE (Synthetic Minority Over-sampling Technique), although oversampling increases dataset size and computational complexity while under-sampling can decrease accuracy. Consequently, it is crucial to choose an appropriate dataset that reduces these issues[15]. GSTR-1, GSTR-3B, Sellers Invoice Amount, Buyers Invoice Amount, Sellers Taxable Amount, Buyers Taxable Amount, ITC Claimed as per GSTR-3B, SEZ Supply, Buyer's Place of Registration, and Buyer's Place of Receipt are among the key financial and transactional features that are reflected in the labeled GST data used for this project. These characteristics assist in identifying irregularities and fraudulent activity through incorporating crucial information into the transactions.

The selection of the dataset followed by the selection of the best classification model. Support Vector Machine (SVM), Gradient Boosting (GB), Random Forest (RF), XGBoost (XGB), Multi-Layer Perceptron (MLP), Decision Tree (DT), and K-Nearest Neighbors (KNN) were among the models that were assessed. Accuracy, precision, recall, F1-score, and total train-test performance were used to evaluate each model. The SVM model was the best option for spotting fraudulent trends among them since it performed well across a variety of metrics. The model's outstanding performance was a result of its capacity to manage complicated distributions and successfully distinguish between cases that were fraudulent and those that appeared authentic. Additionally, XGBoost and Gradient Boosting showed dependable outcomes with strong accuracy and precision. They were more likely to overlook some false examples, though, as seen by their slightly poorer recall ratings when compared to SVM. Overall, Random Forest did well, although its relatively greater train score suggested that it tended to overfit. However, Multi-Layer Perceptron, Decision Tree, and KNN were less appropriate for the fraud detection task due to their lower accuracy and recall.

SVM was selected as the optimal classification model for this system because of its balanced performance in terms of accuracy and recall. A grid search technique was used to optimize the model's hyperparameters in order to increase its predicted accuracy. To get the best findings, parameters including the gamma value, regularization parameter, and kernel type were improved. The data preprocessing phase consisted of the use of LabelEncoder to translate categorical variables, such as GSTINs and HSN codes, into numerical representations and the application of SMOTE to control class imbalance. Recursive Feature Elimination (RFE) was also used for feature selection in order to decrease the quantity of features that are insignificant and improve the interpretability of the model.

The system followed a systematic strategy after the model was trained and optimized. A user-friendly Streamlit interface allows users to enter GST data, which eventually gets sent to the backend API[16]. The SVM model receives the input data from the API and uses it to make prediction. The prediction is sent back to the API and shown to the user based on the model's output. When fraudulent behavior is found, the information is entered into the database for additional examination and research. In order to spot irregularities, features like Sellers Invoice Amount, Buyers Invoice Amount, Sellers Taxable Amount, and Buyers Taxable Amount are essential. Fraud detection includes cross-checking GSTR-1 and GSTR-3B filings and looking for irregularities in ITC

Claimed according to GSTR-3B and SEZ Supply transactions. Finding discrepancies is further aided by the geographic data from the Place of Receipt of Buyer and Place of Registration of Buyer. The system's smooth operation ensures precise and effective fraud detection in real time, giving users insightful data. The suggested solution supports regulatory compliance and reduces financial loss by improving identification of GST fraud through the use of the SVM model.

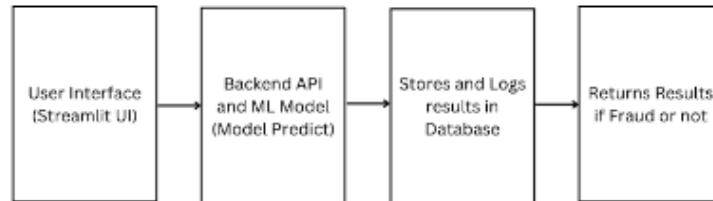


Fig.1 Block Diagram of the Entire System Flow

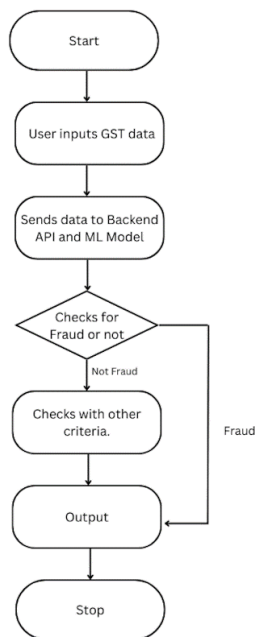


Fig.2 Flowchart of the Entire Process

IV. RESULTS

In the process of selecting the most suitable model for GST fraud detection, multiple machine learning algorithms were evaluated based on their performance metrics. The training and testing dataset included essential financial variables for efficient fraud detection. Important evaluation parameters like accuracy, precision, recall, and F1-score were taken into account to provide a fair comparison. A thorough performance comparison of the several models investigated for GST fraud detection is given in Table 1. The Gradient Boosting (GB) and Support Vector Machine (SVM) classifiers were the most accurate of these, with GB reaching 97.44% and SVM obtaining 97.51%. These models demonstrated high recall and precision, which made them trustworthy for detecting fraud.

Despite having a 99.68% training score, the Random Forest (RF) classifier demonstrated its efficacy in identifying fraud patterns by achieving 97.30% accuracy on the test set. Likewise, Multi-Layer Perceptron (MLP) and XGBoost (XGB) produced competitive results while retaining good recall and precision. With a 96.86% accuracy rate, the Decision Tree (DT) classifier performed admirably, but its precision and recall were marginally worse than those of RF and SVM. Last but not least, of the models examined, K-Nearest Neighbors (KNN) had the lowest accuracy (95.69%), which made it less appropriate for fraud detection in this situation. These findings showed that SVM, Random Forest, and Gradient Boosting were the best models for detecting GST fraud because of their balanced precision, recall, and accuracy scores.

TABLE I. PERFORMANCE OF VARIOUS CLASSIFICATION MODELS

Algorithm	Train Score	Test Score	Accuracy Score	Precision Score	Recall Score	F1 Score
SVM	98.12	97.51	97.51	99.66	95.00	99.62
GB	98.54	97.44	97.44	99.85	95.28	97.51
RF	99.68	97.30	97.30	99.00	95.83	97.39
XGB	98.12	97.22	97.22	99.71	95.00	97.30
MLP	97.86	96.93	96.93	98.85	95.28	97.03
DT	99.98	96.86	96.86	97.48	96.53	97.00
KNN	99.98	95.69	95.69	95.84	95.97	95.91

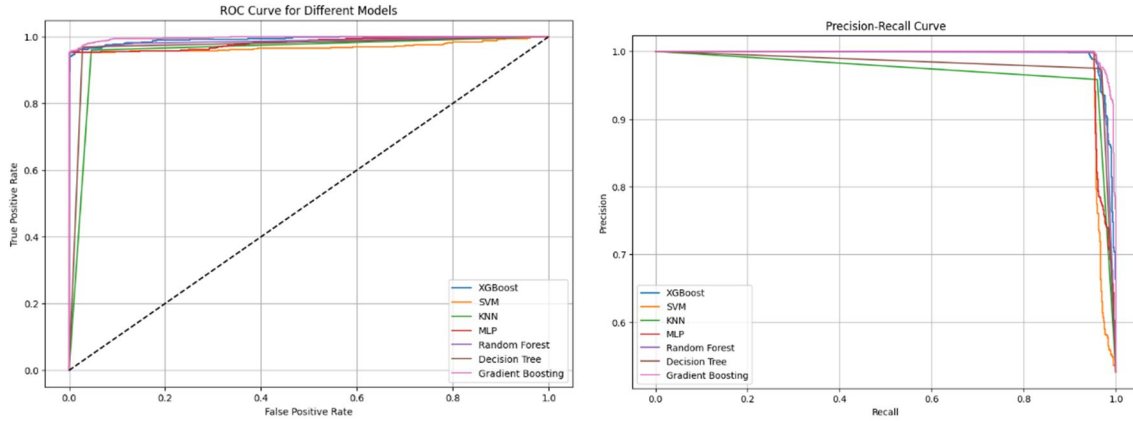


Fig. 3 Receiver Operator Characteristics (ROC) Curve Fig.4 Precision-Recall (PR) Curve for Various Models

After the rigorous model selection process, the selected model was evaluated to understand the level of performance of the model (as well as the whole system) across various aspects crucial to the project's objectives. The following metrics were used to analyze the performance of our system. The system yielded favorable results in the experiments. The performance evaluation metrics considered are as follows:

A. Accuracy

This metric calculates the ratio of correctly identified samples to the total number of samples in the dataset.

$$Accuracy = \frac{\text{Samples correctly classifies in the test data}}{\text{Number of samples in test data}}$$

B. Precision

It is the measure of how many predicted positive instances are actually correct. It is calculated as,

$$Precision = \frac{TP}{TP + FP}$$

C. Recall

Recall represents the model's ability to correctly identify all actual positive instances. It is given by,

$$Recall = \frac{TP}{TP + FN}$$

D. F1 score

It is the harmonic mean of Precision and Recall, providing a balanced measure between the two. It is computed as,

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

TP → True positive, FP → False positive, FN → False negative

V. CONCLUSION

This study presents the design and development of a machine learning-based GST fraud detection system. After a thorough examination of several classification models, the Support Vector Machine (SVM) classifier was selected due to its high accuracy of 97.51%, as well as its great precision and recall. To efficiently identify fraudulent activity, the system uses real GST datasets and analyses important financial components, such as GSTR-1, GSTR-3B, invoice amounts, taxable values, ITC claims, and buyer-seller transaction patterns. Model performance is improved by data preparation methods like Recursive Feature Elimination (RFE) for feature selection and SMOTE for handling class imbalance. For simple interaction and viewing of fraud detection data, the system incorporates the user-friendly Streamlit interface.

Future developments might concentrate on identifying particular fraud types, implementing automatic alerts for questionable GSTINs, enhancing tax compliance, reducing financial risks, and strengthening regulatory enforcement.

REFERENCES

- [1] Press Information Bureau, Ministry of Finance. (2024, July 1). "Celebrating GST Day: A milestone in economic reform. Government of India." <https://www.pib.gov.in/PressNoteDetails.aspx?NoteId=151915>
- [2] CNBCTV18.com. (2023, April 23). "GST evasion: Taxmen to use data analytics in a bid to curb tax evasion in supply chain." CNBC TV18. <https://www.cnbctv18.com/finance/gst-evasion-taxmen-to-use-data-analytics-in-a-bid-to-curb-tax-evasion-in-supply-chain-16475711.htm>.
- [3] Helly Mukesh Maniya, Shobha V. (2024). "Studies on GST Tax Evasion: Patterns, Impacts, and Regulatory Responses in India," Vol. 20 No. 11s.
- [4] Verma, Priya & Mathews, Jithin & Kumar, K Sandeep & K., Suryamukhi & Babu, Ch & Rao (2019). "Big Data Analytics for Nabbing Fraudulent Transactions in Taxation System." 10.1007/978-3-030-23551-2_7.
- [5] Vyas, Vivek, Himanshu Thakkar, Siddharth Baburao Dabhade, Ramdas D. Gore and Hetal Thaker. "Comparative Analysis of SVM Variants for GST Fraud Detection." Communications on Applied Nonlinear Analysis (2024): n. pag. <https://api.semanticscholar.org/CorpusID:276069366>.
- [6] "DETECTING GST FRAUD THROUGH MACHINE LEARNING TECHNIQUES", International Journal of Emerging Technologies and Innovative Research (www.jetir.org), ISSN:2349-5162, Vol.8, Issue 3, page no.320-337, March-2021, Available :<http://www.jetir.org/papers/JETIR2103433.pdf>.
- [7] Thakkar, Himanshu & Joshi, Vedanshi & Bhatt, Rishita. (2024). "GST FRAUD: UNVEILING THE MODUS OPERANDI OF INPUT TAX CREDIT FRAUD SCHEMES." VIDYA - A JOURNAL OF GUJARAT UNIVERSITY. 3. 35-39. 10.47413/vidya.v3i1.357.
- [8] Othman, Zaleha & Bidin, Zainol & Mansor, Muzainah & Abdullah, Yeop. (2019). "GST fraud: Unveiling the truth. International Journal of Supply Chain Management." .
- [9] Silva, D., & Carvalho, S. (2021). "Data Science Methods and Techniques for Goods and Services Trading Taxation: a Systematic Mapping Study." EGOV-CeDEM-ePart-*. <https://api.semanticscholar.org/CorpusID:246081928>.
- [10] Hao Sun, Jianping Li, Xiaoqian Zhu, "Financial fraud detection based on the part-of-speech features of textual risk disclosures in financial reports", Procedia, Computer Science, Volume 221, 2023, Pages 57-64, ISSN 1877-0509, <https://doi.org/10.1016/j.procs.2023.07.009>.
- [11] W. T. Mongwe and K. M. Malan, "The Efficacy of Financial Ratios for Fraud Detection Using Self Organising Maps," 2020 IEEE Symposium Series on Computational Intelligence (SSCI), Canberra, ACT, Australia, 2020, pp. 1100-1106, doi: 10.1109/SSCI47803.2020.9308602.
- [12] D. Jayanti, S. Sulisty and P. I. Santosa, "Application of Machine Learning in Taxation: A Systematic Literature Review," 2024 International Seminar on Intelligent Technology and Its Applications (ISITIA), Mataram, Indonesia, 2024, pp. 687-691, doi: 10.1109/ISITIA63062.2024.10668268.
- [13] Hernandez Aros, L., Bustamante Molano, L.X., Gutierrez-Portela, F. et al. "Financial fraud detection through the application of machine learning techniques: a literature review." Humanit Soc Sci Commun 11, 1130 (2024). <https://doi.org/10.1057/s41599-024-03606-0>.
- [14] A. N. Ahmed and R. Saini, "A Survey on Detection of Fraudulent Credit Card Transactions Using Machine Learning Algorithms," 2023 3rd International Conference on Intelligent Communication and Computational Techniques (ICCT), Jaipur, India, 2023, pp. 1-5, doi: 10.1109/ICCT56969.2023.10076122.
- [15] Sopiyan, Muhamad & Kasyfi, Fauziah & Wijaya, Yunan. (2022). "Fraud Detection Using Random Forest Classifier, Logistic Regression, and Gradient Boosting Classifier Algorithms on Credit Cards." JUITA: Jurnal Informatika. 10. 77. 10.30595/juita.v10i1.12050.
- [16] "Streamlit Documentation". API reference. Streamlit. <https://docs.streamlit.io/develop/api-reference>