Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 32 No. 4s (2025)

Comparative Analysis of SVM Variants for GST Fraud Detection

Vivek Vyas[1], Himanshu Thakkar[2], Siddharth Dabhade[3], Ramdas Gore[4], Hetal Thaker[5]
vivek.vyas@nfsu.ac.in[1], himanshu.thakkar@nfsu.ac.in[2], siddharth.dabhade@nfsu.ac.in[3], ramdas.gore@nfsu.ac.in[4], hetal.thaker@nfsu.ac.in[5]
[1,2,3,4,5]National Forensic Sciences University, Gandhinagar, Gujarat, India

Article History:

Received: *04-10-2024*

Revised: *30-11-2024*

Accepted: *09-12-2024*

Abstract:

This study explores various Support Vector Machine (SVM) variants and provides an analytical comparison to identify their effectiveness in detecting Goods and Services Tax (GST) fraud. GST fraud poses significant challenges to regulatory authorities and businesses, necessitating robust detection methods. SVM, a powerful machine learning algorithm, offers promise in this domain due to its ability to handle complex data and nonlinear relationships. Through a comprehensive examination of SVM variants, including linear SVM, polynomial SVM, and radial basis function SVM, this study assesses their performance in GST fraud detection. Additionally, computational efficiency and scalability are investigated to gauge the practical viability of each variant. The findings contribute to advancing the understanding of SVM's applicability in fraud detection contexts and offer insights into selecting the most suitable variant for GST fraud identification. Ultimately, this research aids stakeholders, including tax authorities and businesses, in implementing effective strategies to combat fraudulent activities and uphold fiscal integrity.

Keywords: Support Vector Machines (SVM), GST fraud, Input Tax Credit.

1. Introduction

GST is a comprehensive, multistage, destination based indirect tax implemented in India on 1st July 2017. In recent years, the application of Goods and Services Tax in India has brought about significant changes to the country's tax system. Goods and Services Tax is an indirect tax levied in India since July 1, 2017. The implementation of GST has been hailed as the biggest tax reform in India since 1947. With the introduction of GST, multiple indirect taxes have been replaced by a single tax structure. This tax reform has not only simplified the taxation process but also aimed to eliminate complications and double taxation. However, with any major change comes the potential for misuse and fraud. One of the challenges faced by the Indian government in implementing GST is dealing with fraudulent activities.

After implementation of GST the fraud in GST is increasing rapidly. As per answer given by Finance Minister in the Lok Sabha, from July 2017 to February 2023 there was close to Rs. 3.08 Lakh Crore GST evasion was reported, and 1402 offenders were detained, 1.03 lakh crore was recovered by the GST department.

Goods and Services Tax is a successful method for raising government revenue, but it is also susceptive to fraud. It is frequently exploited by taxpayers depleting the Government tax revenue. The volume of GST fraud is increasing day by day so that demand for forensic accountants is increasing rapidly.

According to GST enforcement probe (case under investigation) in 5000 crore GST Scam in Bhavnagar the fraudster made a gang in which they offered providing Government support and change phone number in Aadhaar card. Then based on Aadhaar card fraudster used to get PAN numbers and then get GST registrations for the creation of Shell Companies. According to Special Investigation Team fraudster have created Approx 1500 shell companies, transactions worth RS. 40,000 Crore. It is the largest GST fraud in the history of Gujarat. GST fraud is serious concern and it undermine the financial stability of the nation. For GST fraud various modes operandi has been observed in the GST fraud.

The objective of this paper is to explore the various types of GST fraud prevalent in India. The crucial role of forensic accountants in detecting, investigating, and preventing the GST fraud in India. This paper aims to contribute towards discussing comprehensive strategies to tackle GST fraud.

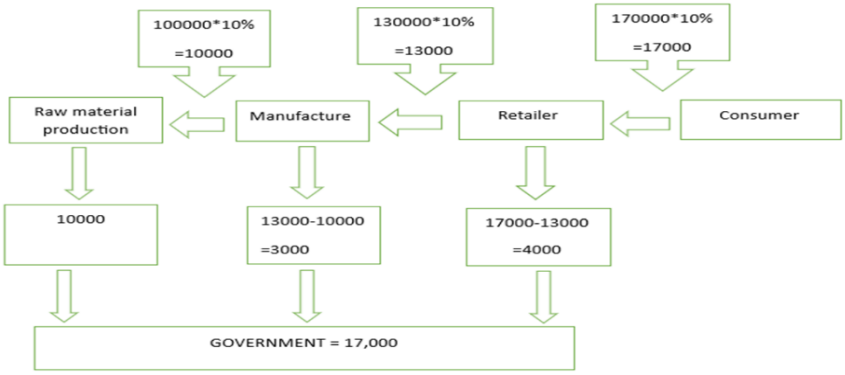

Fig. 1. Tax Flow Diagram of GST

This paper introduces various machine learning methods for effectively forecasting GST fraud, including SVM Quadratic, SVM-Coarse Gaussian, SVM-Cubic, SVM-Fine Gaussian, SVM-Linear, and SVM-Medium Gaussian. Consequently, this research aims to address the question of accurately predicting GST fraud by utilizing these models and determining which one performs better. Furthermore, these models will serve as initial models for predicting GST fraud.

2. Literature Review

In this Paper [1] found that with the help of transaction level data analysis 800 ghost firms has been identified by the tax authority of Ecuador. These ghost firms were not only involved in evasion of Value Added Tax but also in corporate income taxes. Author studied [2] some tax evasion techniques widespread in Goods and Services Tax in Telangana State. Various modus operandi of GST Input Tax Credit evasion scam has been identified with the help of sensitive parameters. It is observed that shell companies were opened in different states. For tax evasion fraudster shows purchase from other state and investigating such malicious dealers is difficult. Input tax credit has been availed with the help of fake billing.

Tax system should be simple, and it should be levied and collect the tax at a single point based on algorithms illegitimate transactions have been identified. The result suggests that circular trading evasion technique have been used by the perpetrators where malicious dealers do heavy fake sales and purchase transactions among themselves only that go around in a circular manner in short period

of time. It generates vouchers and avail fake input tax credit without any meaningful value addition [3]. It provides a detailed exploration of preventative strategies to minimize Goods and Services Tax (GST) fraud, which is essential for preserving government revenue. To prevent GST fraud researcher has suggested that further research should be carried out on identification of the conduct and profiling of impostors so that such GST fraud can be detect at the early stages.

Through the use of qualitative research methods, [4] looked at cases of GST fraud that occurred in Malaysia. Their research turned up a number of different types of GST fraud, including phony claims, manipulated sales data, neglected registration requirements, filing incomplete GST reports, using GST avoidance strategies, and taking part in carousel fraud schemes. The researchers also observed that those with average educational backgrounds and medium-sized firms appear to be more likely to commit GST fraud. [5]; These studies concluded that fraud could potentially be curbed through preventive measures. Organizations need to implement strong internal controls and systems, as well as educate their employees on ethical practices. This will help in creating a culture of integrity and compliance, making it less likely for individuals to engage in GST fraud.

In this study, various types of Support Vector Machines (SVM) are utilized, employing a range of kernel functions. SVM serves as a supervised machine learning technique applied in tasks involving classification and regression [6]. The kernel function within SVM plays a crucial role in defining the decision boundary that segregates distinct classes.

2.1. SVM Quadratic

The Support Vector Machine (SVM) algorithm utilized here is the standard one with a quadratic kernel function. By mapping data points to a higher dimensional space, it becomes capable of handling non-linearly separable data [7]. Although training the quadratic kernel can be computationally expensive, it proves to be effective for intricate datasets. In this case, the SVM with a quadratic kernel employs a quadratic polynomial as the kernel function, resulting in a decision boundary that takes the form of a quadratic curve in the input space. This particular type of SVM is well-suited for capturing more intricate relationships between features.

2.2. SVM-Coarse Gaussian

This version employs a Gaussian kernel function with a wide width. The kernel exhibits a smooth decision boundary and is effective for data with smooth, continuous relationships between features [8]. Nevertheless, it may not be appropriate for capturing intricate, localized patterns. In the context of SVM, the Gaussian kernel (also referred to as Radial Basis Function or RBF) is a commonly favored option. The term "coarse" typically denotes a larger bandwidth or spread parameter in the Gaussian kernel. A wider bandwidth results in a smoother decision boundary and encompasses broader patterns in the data.

2.3. SVM-Cubic

This version utilizes a cubic kernel function, which exhibits a more distinct decision boundary in contrast to the Gaussian kernel [9]. It proves to be advantageous for datasets with well-defined class boundaries, but it may not yield satisfactory results for datasets with overlapping classes or noise. Similar to the SVM-Quadratic, the SVM-Cubic employs a cubic polynomial as its kernel function.

Consequently, the decision boundary takes the form of a cubic surface within the input space. This enables the algorithm to capture even more intricate relationships between features when compared to linear or quadratic kernels.

2.4. SVM-Fine Gaussian

This version utilizes a Gaussian kernel function with a narrow width. This kernel is highly responsive to small-scale fluctuations in the data and can effectively detect intricate patterns [10]. Nevertheless, it is prone to overfitting and noise sensitivity.

In contrast to the "Coarse Gaussian," the term "Fine Gaussian" probably denotes a reduced bandwidth or spread parameter in the Gaussian kernel. A smaller bandwidth leads to a more intricate decision boundary, thereby increasing the SVM's sensitivity to local patterns in the data.

2.5. SVM-Linear

The SVM-Linear employs a linear kernel function, allowing for a hyperplane decision boundary in the input space [11]. This is ideal for scenarios where the features exhibit a linear relationship. This approach directly maps data points to the feature space without any alteration, making it computationally efficient and effective for datasets with linearly separable classes. Nonetheless, it is not suitable for handling non-linearly separable data.

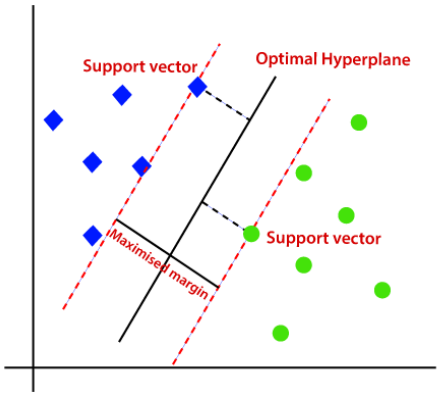

Fig. 2. SVM Linear hyperplane

2.6. SVM-Medium Gaussian

This version utilizes a Gaussian kernel function with a moderate width, striking a balance between the smoothness of the Coarse Gaussian and the sensitivity of the Fine Gaussian. It proves to be effective across a broader spectrum of datasets compared to other Gaussian variations. The "Moderate Gaussian" implies a middle ground for the spread parameter in the Gaussian kernel [12]. The decision boundary will exhibit a moderate level of smoothness, capturing patterns that are neither overly broad nor excessively detailed.

Various SVM variants with diverse kernels and parameter configurations may exhibit varying performances on different dataset types. The selection of the SVM variant hinges on the data characteristics and the complexity of the relationships between features. It is customary to experiment with different kernels and parameters to determine the setup that best fits the specific problem at hand.

3. Research Methodology

The study report details the use of a unique dataset with five thousand records. These files were loaded into machine learning models could be built. The authors described a set of procedures used to identify and categorize instances of GST fraud.

3.1. Data Preprocessing

The dataset is cleaned and transformed during the data preprocessing stage in order to make it ready for use as an input for machine learning models. This procedure entails choosing the seven relevant columns as shown in Table 1. The existence of null values is then checked in order to verify the accuracy of the results, since the algorithm's conclusions can change if there are null values in the data. Thereafter the cross tabulation of the data based on several attributes may be examined to obtain descriptive visualization of the data.

Table 1. Attributes Associated with GST Fraud

Attribute	**Values**
GST Number	**15 digit alphanumeric number**
Sales	**Sales in Indian Rupees**
Purchases	**Purchases in Indian Rupees**
Total Tax Liability	**Total Tax Liability in Indian Rupees**
Total Input Tax Credit	**Total Input Tax Credit in Indian Rupees**
Percentage of Input Tax Credit	**Percentage of Input Tax Credit (0-100)**

3.2. Feature selection

The next step is to investigate the correlation between the 7 attributes. If the attributes are not correlated one to another, all 7 attributes can be included in the model. However, if there is a strong correlation between two attributes, one of them should be dropped. The correlation method utilizes the Pearson correlation. The results of the correlation study point to the seven attributes' poor relationships with one another, indicating their independence. Consequently, all seven characteristics are kept and included to the model for additional examination.

3.3. Data Splitting

The 5000-record dataset is split into training and testing subsets at this stage. The split is done at random, with the train-test data division followed by an 80:20 ratio. The ideal fraction for the train-test data split is found to be this ratio.

3.4. Model Training

The training procedure begins with the training data as input and builds the models. For this, different variants of the Support Vector Machine (SVM) algorithms are used separately. These algorithms' parameters are adjusted and calibrated to improve the models' accuracy.

3.5. Model Evaluation

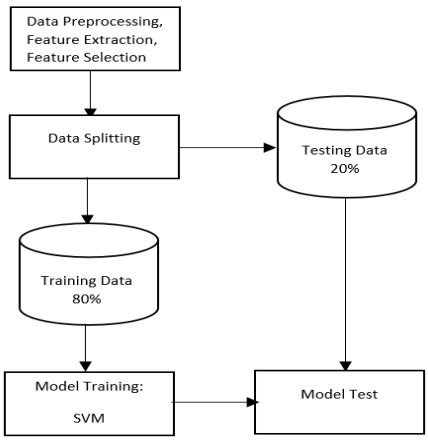

Fig. 3. Machine Learning Model

The created models are finally applied to the testing data. The results are displayed as a confusion matrix, comparing the real positive and negative examples with the ones that were predicted. In this case, "negative" indicates that the GST fraud class does not exist, whereas "positive" indicates that the GST fraud class does exist. TP (True Positive: actual and predicted values are positive), FP (False Positive: actual value is negative but predicted value is positive), TN (True Negative: actual and predicted values are negative), and FN (False Negative: actual value is positive but predicted value is negative) are the four categories that make up the confusion matrix. Metrics like accuracy, precision, F-measure, and recall for different SVM versions can be computed using the confusion matrix.

Accuracy = TP + IN/ (TP+IN+FP+ FN)

Precision = TP/(TP + FP)

Recall = TP/(TP + FN)

4. Result and Discussion

In this paper, the results are the most significant aspects associated with GST Fraud and models developed using different variant of SVM algorithms to predict the GST Fraud.

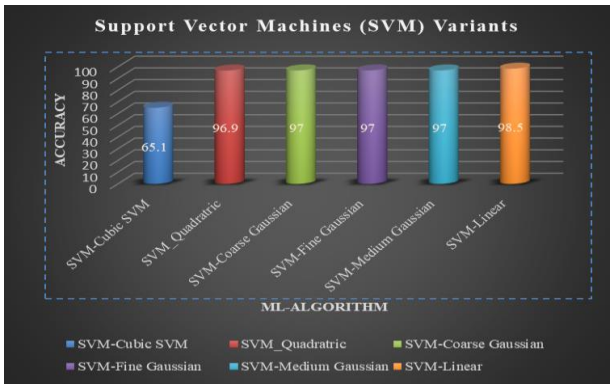

Fig. 4. Accuracy of Various SVM Variants

As per methodology used different types of Support Vector Machines (SVM) with various kernel functions. SVM is a supervised machine learning algorithm used for classification and regression tasks. The kernel function in SVM determines the type of decision boundary that the algorithm creates to separate different classes. In this work GST data has been trained and tested 06 different regression models, including Quadratic Support Vector Machine (SVM Quadratic), Coarse Gaussian Process Regression (GPR Coarse), Fine Gaussian Support Vector Machine (SVM Fine Gaussian), Linear Support Vector Machine (SVM Linear), and Quadratic Support Vector Machine with Gaussian Process Regression (SVM Quadratic GPR). GST data comprehensive analysis using various machine learning algorithms.

5. Conclusion

This article focuses on the use of Support Vactor Machine for the detection of Goods and Services Tax (GST) Fraud at the early stages. The study highlights machine learning tools are very helpful for the GST fraud detection. By employing Support Vector Machine performance of various GST fraud metrics like Sales, Purchases, Input Tax Credit and Tax Liability have been analyzed and we got the 98.5% accuracy for SVM-Linear. Further research can be carried out on the use of different machine learning tools for anomaly detection in the GST fraud. In addition, integration of machine learning tools can be explored to ensure real time fraud detection by the GST Investigation department.

References

[1] Carrillo, P., Donaldson, D., Pomeranz, D., & Singhal, M. Ghosting the Tax Authority: Fake Firms and Tax Fraud in Ecuador (2022).
[2] Mehta, P., Mathews, J., Rao, S. K. V., Kumar, K. S., Suryamukhi, K., & Babu, C. S. Identifying malicious dealers in goods and services tax. In 2019 IEEE 4th International Conference on Big Data Analytics (ICBDA) (pp. 312-316). IEEE. (2019, March).
[3] Mehta, P., Mathews, J., Kumar, S., Suryamukhi, K., Sobhan Babu, Ch., & Kasi Visweswara Rao, S. V. Big data analytics for nabbing fraudulent transactions in taxation system. Lecture Notes in Computer Science, 95–109. (2019).
[4] Othman, Z., Nordin, F.F., Bidin, Z. and Mansor, M. "GST fraud: unveiling the truth", International Journal of Supply Chain Management, Vol. 8 No. 1, pp. 941-950. (2019),
[5] Halbouni, S.S., Obeid, N. and Garbou, A. "Corporate Governance and information technology in fraud prevention and detection evidence from the UAE", Managerial Auditing Journal, Vol. 31 No. 6-7, pp. 589-628 (2016).
[6] Sen, P.C., Hajra, M., Ghosh, M. Supervised Classification Algorithms in Machine Learning: A Survey and Review. In: Mandal, J., Bhattacharya, D. (eds) Emerging Technology in Modelling and Graphics. Advances in Intelligent Systems and Computing, vol 937. Springer, Singapore (2020).
[7] Zhou, J., Tian, Y., Luo, J. et al. Novel non-Kernel quadratic surface support vector machines based on optimal margin distribution. Soft Comput 26, 9215–9227 (2022).
[8] Bhoopesh Singh Bhati and C.S. Rai, Intrusion detection technique using Coarse Gaussian SVM, International Journal of Grid and Utility Computing Vol. 12, Issue. 1, 2021
[9] U. Jain, K. Nathani, N. Ruban, A. N. Joseph Raj, Z. Zhuang and V. G.V. Mahesh, "Cubic SVM Classifier Based Feature Extraction and Emotion Detection from Speech Signals," 2018 International Conference on Sensor Networks and Signal Processing (SNSP), Xi'an, China, pp. 386-391, 2018.
[10] Omar Chamorro-Atalaya, Dora Arce-Santillan, Guillermo Morales-Romero, César León-Velarde, Primitiva Ramos-Salaza, Elizabeth Auqui-Ramos, Miguel Levano-Stella, Sentiment analysis through twitter as a mechanism for assessing university satisfaction, Vol. 28, No. 1, pp. 516~524, ISSN: 2502-4752, October 2022
[11] S. Ghosh, A. Dasgupta and A. Swetapadma, "A Study on Support Vector Machine based Linear and Non-Linear Pattern Classification," 2019 International Conference on Intelligent Sustainable Systems (ICISS), Palladam, India, pp. 24-28, 2019.
[12] Tao Wang, Chia-Hung Su, Medium Gaussian SVM, Wide Neural Network and stepwise linear method in estimation of Lornoxicam pharmaceutical solubility in supercritical solvent,Journal of Molecular Liquids, Volume 349