EAI Endorsed Transactions
on Scalable Information Systems Research Article **EAI.EU**

EcomFraudEX: An Explainable Machine Learning Framework for Victim-Centric and Dual-Sided Fraud Incident Classification in E-Commerce

Salman Farsi[1,*] and Mahfuzulhoq Chowdhury[2]

[1, 2] Department of Computer Science and Engineering, Chittagong University of Engineering and Technology (CUET), Chattogram – 4349, Bangladesh.

Abstract

The popularity of e-commerce businesses and online shopping is experiencing rapid growth all around the world. Nowadays, people are more inclined to shop online than in the actual shops. Due to this advancement, fraudsters have set new traps to deceive consumers. Whether it is true that customers often become victims of fraud, it also happens that a fraud customer tries to deceive the seller and hassle the seller intentionally in several ways. To address these issues, an automated system is required so that fraud incidents can be classified. This will facilitate taking legal action and reporting to consumer rights authorities. Existing research on fraud detection and prevention didn't cover customer and seller-side fraud simultaneously. Besides, most of the work focused on fraud detection rather than post-fraud incident classification. To overcome these gaps, this research endeavor conducts a thorough online survey of customers and sellers to gather incident-specific victim data on fraud cases and it addresses the issue for both customer and seller. This paper proposes a machine learning (ML) based explainable fraud incident classification framework EcomFraudEX, that can efficiently classify these fraud incidents and analyze the reason behind each incident. This framework particularly focuses on proper feature selection techniques, hyper-parameter tuning of models, and exploring different ML and ensemble models. Ensemble majority voting schemes consisting of Random Forest (RF), XGBoost, and CatBoost achieved the highest F1-score of 96% with the Chi-Square feature selection technique in the customer complaint dataset and 98% with the RF feature selection technique in the seller complaint dataset. To explain the incident reasoning, Local Interpretable Model Agnostic Explanation (LIME) and Shapely Additive Explanation (SHAP) were further utilized. The proposed scheme achieved a 1.57% higher F1-score and 2.13% higher accuracy than previous works.

Keywords: Fraud Detection, Fraud Prevention, E-Commerce, Explainable AI, Ensemble Majority Voting, Survey Data, Feature Selection

Received on 28 07 2024, accepted on 12 01 2025, published on 20 01 2025

doi: 10.4108/eetsis.6789

1. Introduction

E-commerce sites have become a popular alternative to brick-and-mortar shops for purchasing goods from home. This trend has surged during the coronavirus pandemic, with more people around the world opting for online shopping. However, the growth in e-commerce has also led to an increase in fraudulent activities in this industry. Though buyers sometimes commit fraud, actually it is more common for sellers to engage in fraudulent practices. Thus, the nature of fraudulent practices in e-commerce is bidirectional. Customers who commit fraud do so in various ways. Some provide incorrect shipping addresses or refuse to accept products delivered due to a change of

[*] Corresponding author. Email: salman.cuet.cse@gmail.com

mind and so on. These actions create significant challenges for sellers to cope with. On the other hand, fraudulent activities by sellers are another major concern for e-commerce customers. Sellers have several ways to deceive customers. They may deliver products late, sell faulty or damaged goods, or intentionally refuse to deliver products even after receiving advance payment. Other common tactics include misrepresenting products to make them appear more appealing than they actually are. These fraudulent practices undermine consumer trust and can lead to financial losses for buyers.

The global e-commerce market is anticipated to grow in the coming years. It is projected to reach $5,136 billion by 2024, with an expected annual growth rate of 8.5% from 2024 to 2028 [1]. Amidst this booming growth, it is estimated that fraud costs e-commerce enterprises $48 billion annually. In the past year, global losses from e-commerce fraud have increased by 16%. Last year, e-commerce companies lost 2.9% of their global revenue to fraud. For every $100 in fraudulent orders, businesses lose $207 [2]. Due to these alarming statistics, it is evident that it is not only important to build robust fraud detection systems, but also equally important to classify the one that has occurred already. Such an automated AI-based system will potentially help recover the losses due to the fraud incident and will accelerate the post-fraud recovery.

Over the past few years, many research studies have been conducted on detecting several types of fraud, such as credit card fraud [3], cashback fraud [4], identity theft [5], etc. Apart from that, studies have also focused on sentiment analysis of customer reviews to find out the potential fraud scenarios [6]. Some research endeavors have even identified the aspect of the reviews to find out complaints from the review. These aspects included whether the review pertains to delivery issues, product problems, or poor servicing. However, these studies lack both the seller and customer-side fraud consideration, multiple types of fraud consideration, direct victim interaction, and comprehensive victim data. The need for post-fraud analysis thus remained under-explored and is still considered as a potential research area. In summary, after the analysis, this research concludes in addressing the following key Research Questions (RQs):

- **RQ1:** How can a fraud incident classification system be developed to incorporate post-incident-specific information?
- **RQ2:** How can the system be extended to address frauds involving both sellers and customers?
- **RQ3:** How can various types of fraud be incorporated into the system to create a more generalized solution?
- **RQ4:** What strategies can be employed to handle bias mitigation and class imbalance in fraud incident datasets?
- **RQ5:** What are the most influential features in predicting fraud incidents, and how can these insights contribute to fraud prevention?

To address these gaps, this paper considered multiple types of sellers and customer-side fraud by collecting fraud-specific data directly from victims. Here, the definitions of fraud for both customers and sellers are as follows:

(i) **Seller Fraud:** When a seller engages in fraudulent activities and the customer is a victim, it is called seller-related fraud. The proposed framework will identify seller-related fraud, such as no delivery of the product, delivering counterfeit products, misrepresentation of the product, etc.

(ii) **Customer Fraud:** On the other hand, when a customer engages in fraudulent activities and the seller is a victim, it is called customer-related fraud. For customer-related fraud, it will detect issues like providing the wrong shipping address and changing their minds about purchases.

This study was performed through important questionnaires regarding fraud cases. When victims provide data via questionnaires, they can also provide additional details about the incident. It helps accurately identify the type of fraud incident that occurred. This approach will not only improve the prediction of fraud types but will also assist law enforcement authorities in taking the necessary actions against fraudsters to mitigate potential losses. It will also facilitate the initial fraud case screening process and help investigate cases efficiently with more information available. Additionally, the historical data will be valuable for future studies as well.

Hence, this paper aims to develop an ML-based framework 'EcomFraudEX' that can classify the type of fraud incident that happened with either customers or sellers. Alongside with that, by employing the explainability of the classification models, necessary information can be achieved that will be helpful in analyzing each incident. This analysis will help people stay cautious in certain scenarios so that they don't become victims too in the future. The practical implementation this framework will greatly minimize the instances of fraudulent practices in the e-commerce industry, promoting safer and smoother growth for the business. The proposed scheme will boost confidence amongst customers and sellers by successfully recognizing and mitigating fraud incidents. This will encourage more people to participate in online transactions and raise trust in e-commerce platforms as a whole, leading to increased economic activity and growth. Therefore, the potential answers to the Research Questions (RQs) based on the contribution of this article are listed below:

- **ARQ1:** This article developed a machine learning-based interpretable fraud incident classification framework, EcomFraudEX. To find the top-performing model, it explored different ML models with optimal hyper-parameter tuning and performed ensembles of them by leveraging efficient feature selection techniques.

- **ARQ2 & ARQ3:** This study collected two start-of-the-art datasets, one from customers and another from sellers covering various types of fraud cases. This will be valuable for the country's law enforcement authority and understanding of the fraudulent pattern.

- **ARQ4:** This study also handled the issue of class imbalances in the dataset by allocating weights to each class. Besides, the dataset was annotated by three annotators and checked by an expert to ensure that the bias effect in the dataset is reduced. This indicates that the experimentation was able to successfully develop a more generalized system.

- **ARQ5:** This paper finally delineated the reasoning of the model's prediction through explainable AI methods SHAP and LIME. It uncovered important aspects that indicate why and how the fraud cases occurred.

- This article also compared the performance of the proposed framework with the existing works in terms of accuracy and F1-score.

The remaining portion of this research article is arranged as follows: the review of the existing literature is given in Section 2, the methodology is described in depth in Section 3, error analysis and findings are covered in Section 4, an in-depth explainability study is presented in Section 5, limitations and future works are covered in Section 6, Section 7 brings the findings to a conclusion.

2. Literature Review

Numerous research studies have been conducted focusing on various types of fraud detection in online marketplaces. These studies range from detecting specific frauds like credit card fraud to classifying customer sentiment from reviews. Each of these research efforts has been crucial for incremental progress in this field. Among these methods, using respondents' evident-based datasets and training machine learning models on these data has emerged as a promising approach [7, 8]. This section covers the literature review of all these types of research work in the consecutive subsections 2.1, 2.2, and 2.3. This analysis is crucial to identify and find the potential research trends, gaps and future research direction for fraud prevention.

2.1. E-commerce Fraud Detection

Machine learning techniques were employed to identify fake e-commerce cashback transactions in Indonesia in this study [4]. The authors used CNN, KNN, and LSTM algorithms for fraud detection and found that the KNN algorithm outperformed others with an accuracy of 83.82%. The limitation of this study is that it was conducted utilizing transaction data from only one e-commerce platform in Indonesia, which limits the generalizability of the findings to other platforms. Also, the relationship between cashback fraud and other similar types of fraud, like credit card fraud, was not addressed properly.

The author in another endeavor [9] tackled the critical issue of payment card fraud detection in online transactions by exploring 13 statistical and machine-learning models. They utilized feature aggregation and a genetic algorithm for feature selection to enhance model performance. Among the tested models, Gradient Boosted Trees (GBT) emerged as the best performer, achieving an AUC score of 0.937. Decision Trees (DT) and Deep Support Vector Machines (DSVM) achieved the highest Mathews Correlation Coefficient (MCC) score of 0.964, while GBT reached an MCC of 0.869. However, the study relied on a single model for developing the fraud detection framework. To improve the framework, hybrid models combining two or more algorithms could have been explored.

A system for credit card fraud detection was developed in a study [3] that used a genetic algorithm for selecting features to accelerate the performance of ML classifiers. Using a noble dataset of European cardholders, the study applied DT, RF, LR, ANN, and Naive Bayes models. The RF classifier, combined with GA-selected features, achieved a notable accuracy of 99.98%. In this study, the author didn't give proper reasoning on what prompted them to choose GA as the feature selection method, several other feature selection techniques could have been explored that were overlooked here, and no comparison with different feature selection techniques was provided.

The authors of [10] meticulously examined solutions developed over the past decade for the detection and prevention of click fraud exploited by machine learning and deep learning algorithms. The study emphasized several characteristics that are utilized in developing models that identify ad clicks as either "fraudulent" or "benign", with tree-based models like RF and gradient boosting models such as XGBoost standing out for their effectiveness. XGBoost achieved the best accuracy of 91%, adeptly handling missing data and preventing overfitting. However, most datasets used in click fraud detection have class imbalances, with a scarcity of fraudulent clicks compared to legitimate ones.

2.2. Customer's Review Based Study

The authors of this study [11] categorized customer complaints on food products. They used a dataset of 2217 customer complaints categorized into five labels: 'Hygiene', 'Texture', 'Package/Label', 'Foreign body', and 'Taste/Smell'. Machine learning classifiers such as LR, RF, KNN, XGBoost, etc. were employed utilizing two feature extraction techniques (TF-IDF and word2vec). XGBoost with TF-IDF representation outperformed other models with an F1-score of 84%, which improved to 88% after applying the Chi-Square feature selection method. The small dataset size is a primary limitation of this study.

A new dataset, categorized into complaints and non-complaints named UIT-ViOCD was introduced in the continuation of the complaint identification [6]. The best-performing model was the pre-trained language model PhoBERT, which obtained an F1-score of 92.16%. Special techniques such as the use of pre-trained word embedding models (PhoW2V and fastText) and the RDRSegmenter from VnCoreNLP [12] were employed to enhance the performance.

Another recent study on sentiment analysis incorporates user and product attributes to enhance classification accuracy [28]. The study proposed the Interactive Attributes Attention Network (IAAN), which captures interactive relationships (e.g., user–product, user–text) using bilinear interaction terms. A hierarchical architecture and a multiloss objective function were employed to integrate text and attribute features effectively. Evaluated on IMDB, Yelp, and Amazon datasets, IAAN achieved state-of-the-art accuracy (95.4% on IMDB), outperforming traditional models like HAN. This work demonstrates the effectiveness of leveraging interactive attributes for personalized sentiment analysis.

2.3. Questionnaire-based Fraud Detection

In a survey-based research effort [7], the authors presented a product-oriented fraud prediction method by taking into account several characteristics when a consumer buys a product from an online store. The study gathered important customer, product, and seller attributes and issues through survey questionnaires given to the customers. It found that CatBoost had the best accuracy 93.28% when no feature selection technique was used, while SVM provided 93.09% accuracy when Extra Tree Classifier was used to choose features. The author in this paper overlooked some very crucial elements that could have improved the model, such as seller badge or rating, seller's offline outlet availability, product guarantee or warranty, buyer awareness of online fraud schemes, buyer caution regarding security, etc. It could also explore ensemble models.

This social science study [13] primarily investigated online shopping fraud and its focus on customer behavior and preventive actions. It relied on secondary survey data to analyze how individuals protect themselves from online shopping fraud. The research underlined the significance of enhancing prevention advice for online shoppers to mitigate the risk of fraud. However, one drawback of this study is the paucity of insights from individuals who have experienced online shopping fraud. No particular data evaluation tools were used like the data science tools and machine learning models to further analyze the data and come to a conclusion regarding the fraud cases.

3. Methodology

This section includes a step-by-step explanation of the methodology utilized in this research endeavor which is depicted in Figure 3.

3.1. Dataset Preparation

Dataset preparation includes several key steps before model training as illustrated in Figure 1.

3.1.1. Data Collection

For data collection, questionnaires were created using ‘Google Forms’, which were then distributed to respondents through social media and emails. The questionnaires were made independently for customers and sellers, having a total of 28 questions for customers and 18 for sellers. The survey was conducted among the customers and sellers of the Bangladeshi e-commerce market. The collected dataset is made public and available on GitHub[1].

Customer Complaint Dataset (CCD): There are several social media groups and communities where e-commerce consumers report and share their recent fraud incidents. The questionnaires were provided to these groups and communities. For example, on Facebook, groups like ‘Bangladesh e-Commerce Consumer Society’, ‘Consumer Rights Organization of Bangladesh (CROB)’, ‘Fraud Alert BD’, and ‘Exposing Online Shopping Fraud in Bangladesh (EOSF-BD)’ was targeted to collect the dataset for seller related fraud.

Seller Complaint Dataset (SCD): For sellers, the questionnaires were distributed to Facebook-based small to medium-sized e-shops. These sellers were found in social media groups such as ‘Startup Community Bangladesh’, ‘Bangladesh E-commerce Business Forum (BEBF)’, and ‘E-Commerce Sellers Group’. Additionally, data was gathered from emerging online shops entering the market, as well as from the Consumer Right Authority of Bangladesh, which is responsible for handling complaints regarding the e-commerce market of the country.

3.1.2. Data Annotation

The data annotation for each response was conducted by three annotators. They all are undergraduate students. Each annotator independently reviewed and annotated the data. Majority voting was used to finalize the annotations.

[1] https://www.github.com/Salman1804102/EcomFraudQD

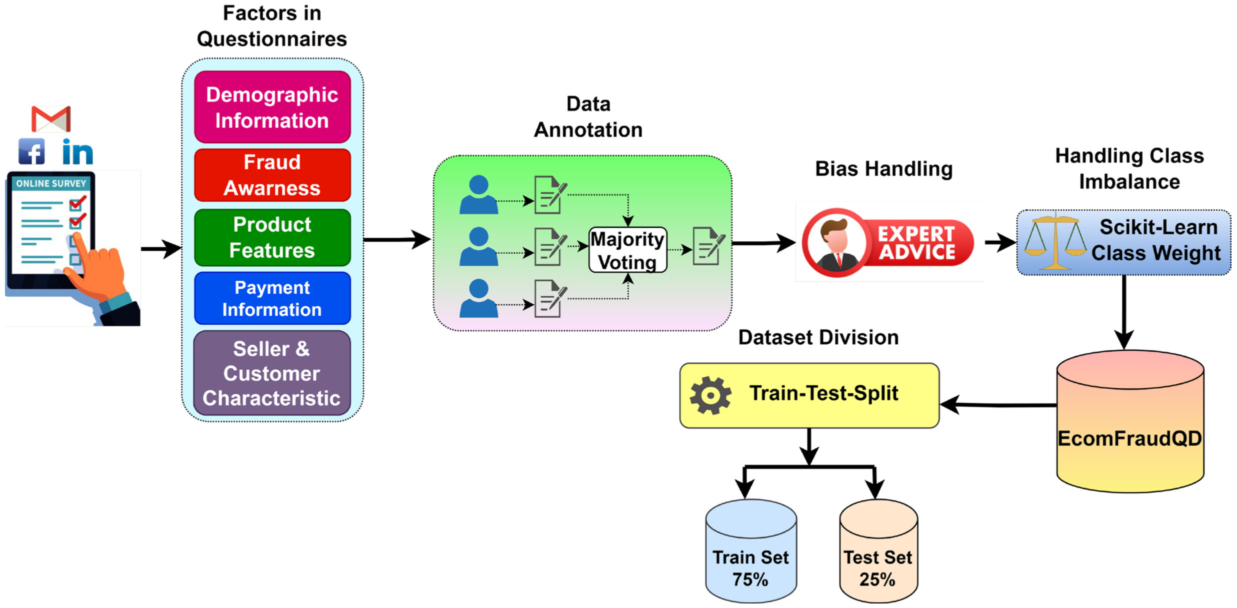
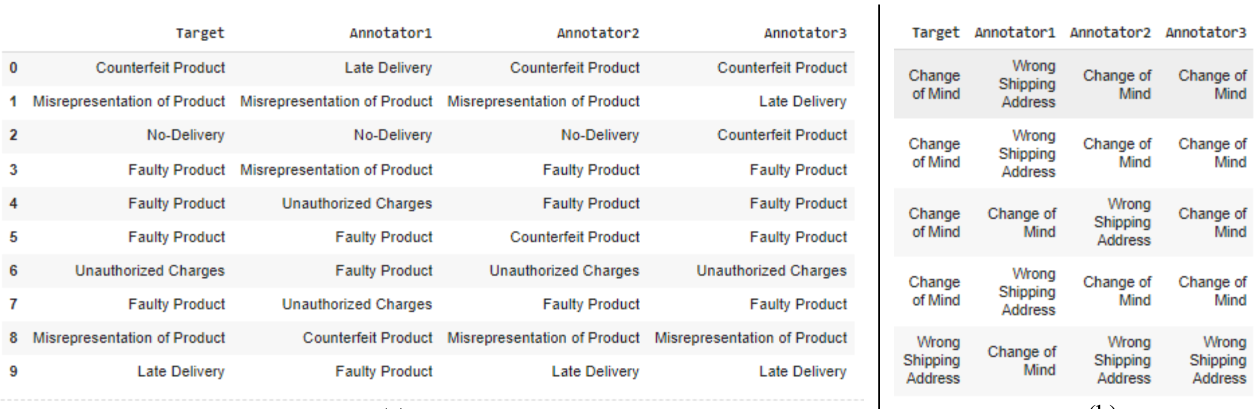


Figure 1. Visual representation of the dataset preparation



	Target	Annotator1	Annotator2	Annotator3
0	Counterfeit Product	Late Delivery	Counterfeit Product	Counterfeit Product
1	Misrepresentation of Product	Misrepresentation of Product	Misrepresentation of Product	Late Delivery
2	No-Delivery	No-Delivery	No-Delivery	Counterfeit Product
3	Faulty Product	Misrepresentation of Product	Faulty Product	Faulty Product
4	Faulty Product	Unauthorized Charges	Faulty Product	Faulty Product
5	Faulty Product	Faulty Product	Counterfeit Product	Faulty Product
6	Unauthorized Charges	Faulty Product	Unauthorized Charges	Unauthorized Charges
7	Faulty Product	Unauthorized Charges	Faulty Product	Faulty Product
8	Misrepresentation of Product	Counterfeit Product	Misrepresentation of Product	Misrepresentation of Product
9	Late Delivery	Faulty Product	Late Delivery	Late Delivery

(a)

Target	Annotator1	Annotator2	Annotator3
Change of Mind	Wrong Shipping Address	Change of Mind	Change of Mind
Change of Mind	Wrong Shipping Address	Change of Mind	Change of Mind
Change of Mind	Change of Mind	Wrong Shipping Address	Change of Mind
Change of Mind	Wrong Shipping Address	Change of Mind	Change of Mind
Wrong Shipping Address	Change of Mind	Wrong Shipping Address	Wrong Shipping Address

(b)

Figure 2. Annotation example for (a) Customer Complaint Dataset (b) Seller Complaint Dataset

In cases where there was ambiguity among the annotators, an expert was consulted to make the final decision. If any particular response seemed biased, such as a response filled out without any context, the annotator marked it as 'biased' instead of assigning it to the expected classes. Eventually, these types of responses were removed to ensure an unbiased dataset. The exclusion of biased responses helped us to identify and filter out untrustworthy data, thereby improving the overall quality of our dataset. To measure the effectiveness of data annotation, the Kappa score was calculated. Table 1 and Table 2 show the Kappa score [14] to illustrate the relevance of the inter-annotator's agreement. An average Kappa score of 0.91 for CCD and 0.95 for SCD was achieved.

However, Figure 2(a) and Figure 2(b) illustrate some examples of the annotations for the CCD and SCD respectively.

Table 1. Pairwise Kappa score for CCD

Pair	Kappa Score
P_1	0.92
P_2	0.87
P_3	0.95
Average	**0.91**

Table 2. Pairwise Kappa score for SCD

Pair	Kappa Score
P_1	0.95
P_2	0.96
P_3	0.94
Average	**0.95**

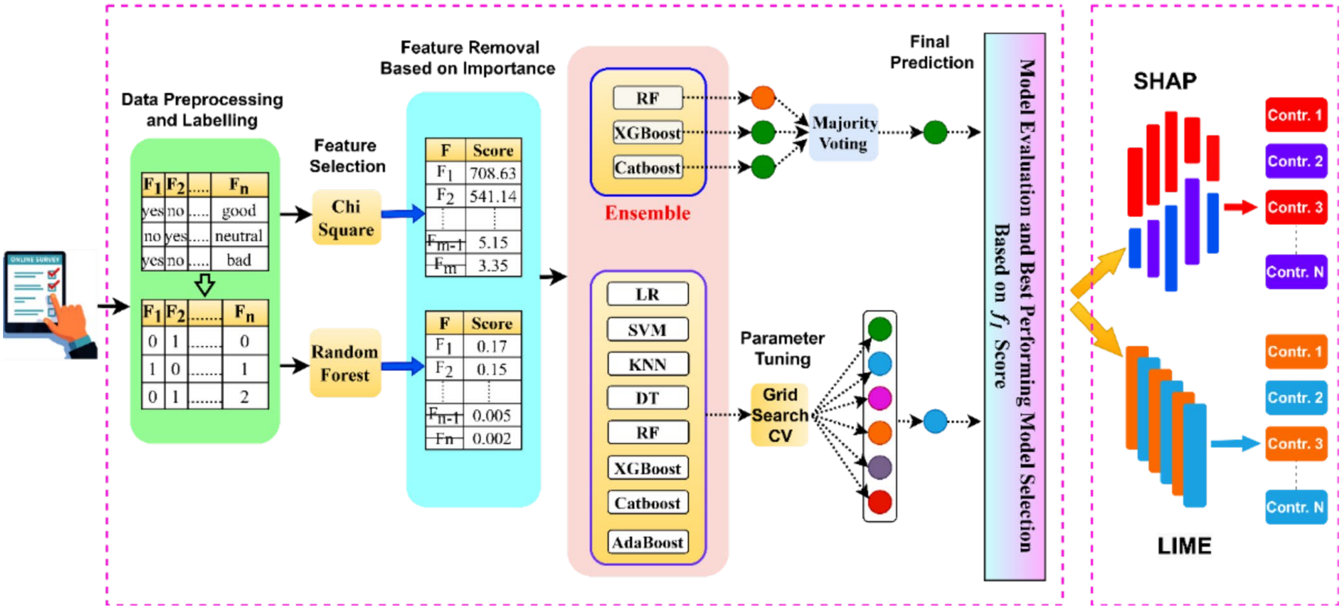


Figure 3. Visual representation of the proposed framework

The description of the classes in the Customer Complaint Dataset (CCD) is given below:

- **No-Delivery**: When the product is not delivered to the customer despite the seller taking payment in advance.
- **Faulty or Damaged Goods**: When the customer receives damaged or poor-quality products.
- **Misrepresentation of Products**: When the product has missing parts, discrepancies in packaging, or does not match the description.
- **Counterfeit Product**: When the customer receives a fake product or a different item than what was ordered.
- **Late Delivery**: When the product arrives significantly later than expected.
- **Unauthorized Charges**: When the customer notices unauthorized charges during the transaction.

The description of the classes in the Seller Complaint Dataset (SCD) is given below:

- **Change of Mind:** When the customer refuses to accept the product because they changed their mind.
- **Wrong Shipping Address:** When the customer intentionally provides an incorrect shipping address to harass the seller, especially in cash-on-delivery transactions.

3.1.3. Data Visualization

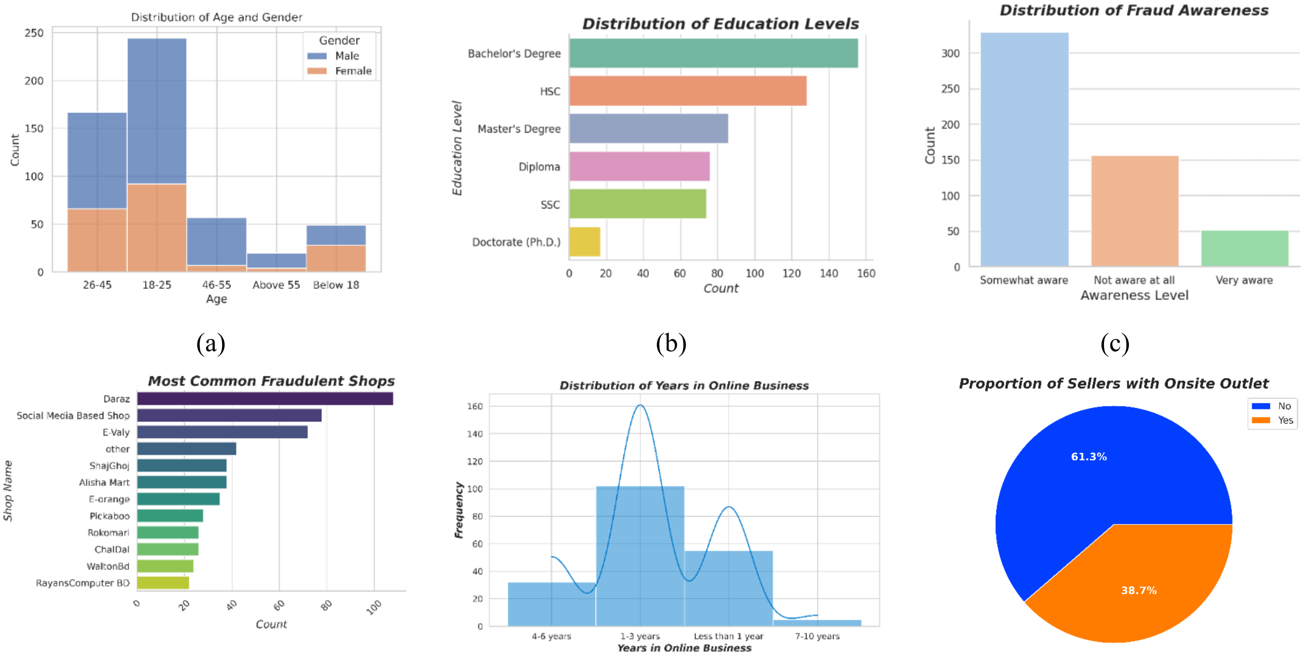
Data visualization played a crucial role in understanding the response patterns of the respondents.

Customer Complaint Dataset (CCD): Figure 4(a) illustrates the count of respondents by age and gender. It shows that the age range of '18-25' comprises the majority in the dataset, indicating that individuals in this age group are the primary buyers in the e-commerce sector and are therefore more likely to become victims of fraud. Additionally, a significant proportion of this age group is male. Figure 4(b) demonstrates that individuals with higher education levels are more inclined to engage in online shopping and subsequently fall victim to fraud. Figure 4(c) presents data on fraud awareness among respondents. It reveals that most fraud victims are 'somewhat aware' of fraudulent activities in online shopping, with more than 300 respondents indicating this level of awareness. 'Daraz', one of the country's popular online shops, is reported to have the highest number of fraud incidents, exceeding 100 cases. This indicates a significant transparency issue within the platform, leading to frequent victimization. The evidence is clear from the Figure 4(d).

Seller Complaint Dataset (SCD): Figure 4(e) indicates that most sellers have only 1-3 years of business experience when reporting fraud incidents. This lack of experience suggests that they might not have established proper fraud protection measures, unlike well-established shops that have been in business for many years. As the survey was distributed mainly to social media-based sellers, it is not surprising that most of them do not have a physical outlet. Consequently, 61.3% of respondents reported not having an onsite outlet. The statistics are shown in Figure 4(f).

3.1.4. Data Preprocessing

To preprocess the data, we first closely scrutinized the dataset to identify whether there were any null fields. Since there were no questions with optional answers, we found no fields with null values in the dataset. Next, we addressed the categorical features by mapping their answers to numerical values such as 0, 1, 2, etc., using the label encoder [15]. Label encoder is an encoding technique that maps the categorical feature values to numerical values.



(a) (b) (c)

(d) (e) (f)

Figure 4. (a) Distribution of age and gender in CCD (b) Distribution of education levels in CCD (c) Distribution of fraud awareness in CCD (d) Most common fraudulent shops reported by CCD respondents (e) Distribution of years in online business reported by SCD respondents (f) Proportion of sellers with onsite outlet in SCD

3.2. Feature Selection

Initially, some features related to the personal information provided by the respondents, such as email, transaction ID, and phone number were removed. However, to select the most relevant features for training, Chi-Square, and the RF feature selection method were utilized. Feature importance for both methods is shown in Table 4 and Table 5.

Chi-Square Feature Selection Method: This feature selection method [16] evaluates the inter-dependency between each feature and the target variable. It then calculates the Chi-Square value for each feature and selects those with the highest scores, indicating a stronger association with the target. The Chi-Square formula is:

$$\chi^2 = \sum\left(\frac{(OF_i - EF_i)^2}{EF_i}\right) \qquad (1)$$

where OF_i is the observed frequency and EF_i is the expected frequency of the feature.

Figure 5(a) and Figure 5(b) illustrate the F1-score versus the number of iterations as features are sequentially removed one by one from least to most significant using this technique. This was done to decide, how many features should be removed. As this couldn't be done randomly, so XGBoost and CatBoost were taken as base classifiers to check the saturation point of the F1 score for CCD and SCD respectively. Seven features from CCD and four features from SCD were removed eventually in this way.

Random Forest Feature Selection: This feature selection technique uses an ensemble of decision trees to assess the significance of each feature [17]. It measures the decrease in node impurity, averaged over all trees in the forest, attributed to each feature. This decrease is often quantified using metrics such as Gini impurity or entropy. In equation (2), $Impurity_b$ and $Impurity_a$ represent the impurity values before and after the split on the feature i in tree j.

$$FI_i = \frac{1}{T}\sum_{j=1}^{T}(Impurity_b - Impurity_a)_{i,j} \qquad (2)$$

Figure 5(c) and Figure 5(d) illustrate that features are sequentially removed one by one from bottom to top based on FI_i or importance. The same process is applied to the Chi-Square method as well.

3.3. Train-Test-Split

For testing the model after training, each of the customer and seller complaint datasets was partitioned into 75% training and 25% testing data. The distribution across the splits is shown in Table 3.

Table 3. Dataset distribution after split

Splits	Customer	Seller
Train	402	144
Test	135	49
Total	537	193

Table 4. Feature importance comparison between RF and Chi-Square methods in CCD

Features (Sorted)	Random Forest	Features (Sorted)	Chi-Square
Counterfeit_Item_Received	0.174221	Counterfeit_Item_Received	708.631695
Product_Authenticity_Verified	0.155112	Product_Authenticity_Verified	541.141486
Delivery_Delay_Days	0.111713	Delivery_Delay_Days	463.491656
Product_Returned	0.103769	Product_Returned	410.907370
Delayed_Delivery	0.098315	Delayed_Delivery	372.086795
Unauthorized_Charges_Noticed	0.085526	Unauthorized_Charges_Noticed	307.027271
Unusual_Pricing	0.058751	Item_Received	157.815435
Seller_Response	0.036849	Unusual_Pricing	152.335529
Item_Received	0.026454	Evidence	130.889533
Product_Description_Read	0.023454	Fraudulent_Shop	42.222782
Payment_Method	0.021820	Tracking_Information_Provided	42.191581
Fraudulent_Shop	0.018152	Age	43.266368
Evidence	0.015994	Education_Level	37.895878
Education_Level	0.015856	Payment_Method	35.843810
Tracking_Information_Provided	0.015155	Product_Description_Read	30.468181
Age	0.014027	Return_Policy_Read	24.829232
Seller_Contact	0.010277	Onsite_Outlet	24.293223
Fraud_Awareness	0.008075	Payment_Info_Shared	17.876356
Return_Policy_Read	0.008002	Seller_Response	12.046979
Onsite_Outlet	0.007966	Gender	8.758525
Location	0.005259	Fraud_Awareness	5.533153
Gender	0.005050	Seller_Contact	5.144082
Payment_Info_Shared	0.002946	Location	3.347699

Table 5. Feature importance comparison between RF and Chi-Square methods in SCD

Features (Sorted)	Random Forest	Features (Sorted)	Chi-Square
refused_product	0.351642	damaged_or_used_returned_product	68.993612
damaged_or_used_returned_product	0.219705	incorrect_address	64.544206
incorrect_address	0.215217	refused_product	57.700173
response_to_refusal	0.051342	advance_payment	11.650509
advance_payment	0.045496	registered	6.472640
payment_terms_violation	0.030175	payment_terms_violation	5.866328
years_online	0.025462	years_online	4.421705
registered	0.019165	evidence_for_complaint	0.880636
refund_return_policy	0.018655	response_to_refusal	0.808379
onsite_outlet	0.008388	refund_return_policy	0.573761
evidence_for_complaint	0.007854	onsite_outlet	0.541540
communication_with_customer	0.006899	communication_with_customer	0.140408

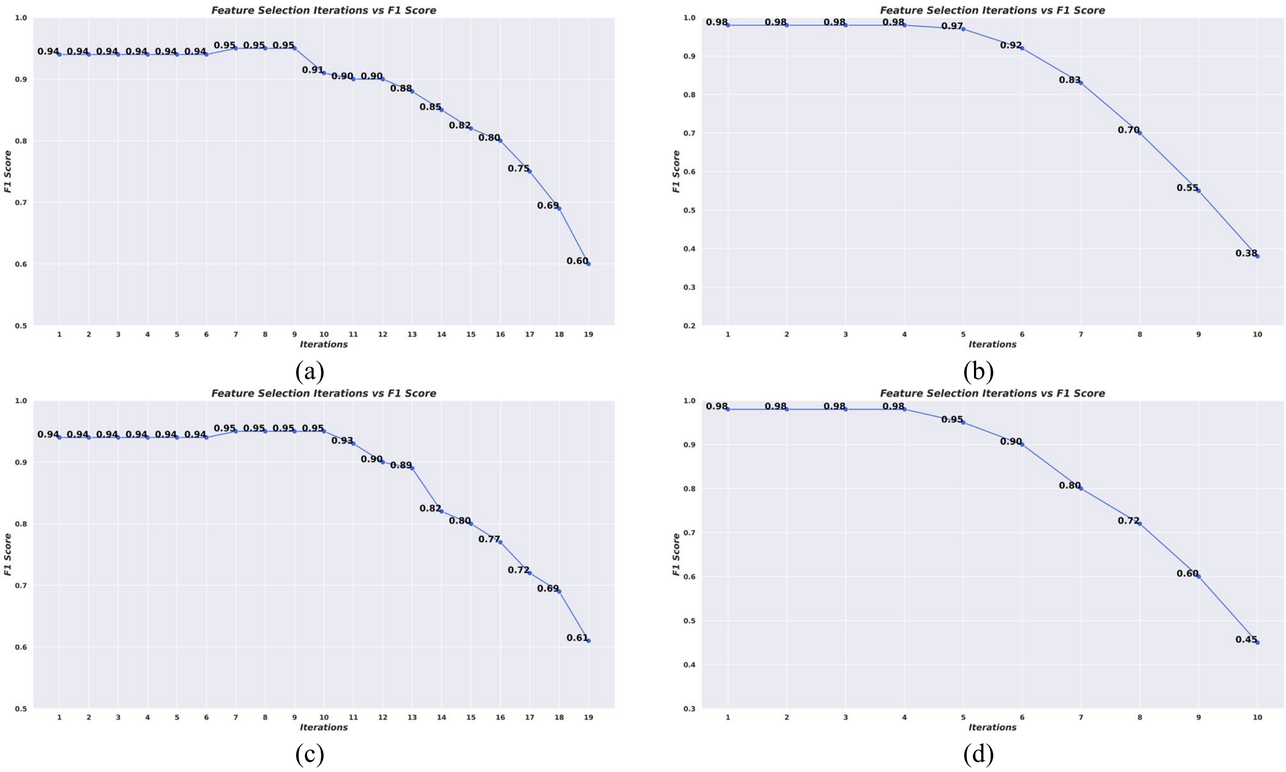


Figure 5. (a) F1-score vs feature removing iterations on CCD for Chi-Square (b) F1-score vs feature removing iterations on SCD for Chi-Square (c) F1-score vs feature removing iterations on CCD for RF (d) F1-score vs feature removing iterations on SCD for RF

3.4. Class Imbalance Handling

In order to handle the class imbalances in the dataset the Algorithm 1 was utilized. The assignment of class weight allowed it to effectively face the class imbalance. Table 6 and Table 7 reflects the class distribution for both datasets.

Algorithm 1: Class weight calculation procedure

```
Compute Class Weights Algorithm
$SET \rightarrow N, M \,\&\, list$
$weights \leftarrow [\,]$
$for\ i \rightarrow 1\ to\ M$:
    # Compute the weight for class $i$
    $currWeight \leftarrow \frac{N}{M * list[i]}$
    # Append the computed weight
    $weights[i] \leftarrow currWeight$
$endfor$
return $weights$
```

Where,

- $weights_i$ denotes assigned weight for the class i
- M denotes the number of fraud classes in the dataset
- The total number of samples is denoted by N
- The total number of samples for i^{th} class is $list_i$

The below tables show the class weights and their count.

Table 6. Class distribution and weights for CCD

Class	Count	Weight
No-Delivery	82	1.091
Unauthorized Charges	75	0.806
Late Delivery	111	0.942
Faulty Product	95	1.193
Counterfeit Product	103	0.869
Misrepresentation of Product	71	1.261

Table 7. Class distribution and weights for SCD

Class	Count	Weight
Change of Mind	103	0.942
Wrong Shipping Address	91	1.066

3.5. Model Training

In order to train both datasets, a total of eight state-of-the-art ML models and four ensemble models consisting of different combinations of RF, XGBoost, CatBoost, and AdaBoost were explored. The ensemble combinations were made by taking three classifiers each time out of four.

Logistic Regression (LR): LR [18] is a very popular ML model primarily utilized for problems of binary classification. This model measures the likelihood that an input given is a member of a specific class. The model uses the logistic (sigmoid) function to convert predicted values into probabilities. The probability that an outcome y is 1 given an input X is expressed as follows:

$$P(y = 1|X) = \frac{1}{1 + e^{-(\alpha+\beta X)}} \tag{3}$$

where α and β are coefficients estimated from the data.

Support Vector Machine (SVM): SVM [19] is a kind of classifier that was designed to determine the hyperplane that best divides data into distinct classes. The aim of SVM is to figure out the best hyperplane that tries to maximize the ‘margin’ among the classes. For tasks involving nonlinear classification, SVM utilizes a function called kernel that maps the inputs into a space of higher dimension, facilitating more efficient division.

K-Nearest Neighbour (KNN): This algorithm [20] is a straightforward, instance-based learning method that assigns a possible class to a certain data point based on the classifications of its closest neighbors. The algorithm determines the proximity between data points using a distance metric, such as ‘Euclidean Distance’. Here, $(p1, q1)$ and $(p2, q2)$ are two neighbor data points.

$$D \;= \sqrt{(p1 - p2)^2 + (q1 - q2)^2} \tag{4}$$

Decision Trees (DT): DT [21] is a non-parametric model that separates data points into subsets focusing on the input feature’s values. It generates a tree structure where each feature is denoted by internal nodes, a decision rule is signified by each branch, and the final outcome or class is indicated by each leaf node.

Random Forest (RF): RF [22] is a particular kind of ensemble learning that generates several decision trees and combines them to get a prediction that is more reliable and accurate. To make sure the trees are uncorrelated, it builds each tree using feature randomness and bootstrap sampling.

Xtreme Gradient Boosting (XGBoost): XGBoost [23] is a distributed gradient-boosting algorithm that has been developed to achieve maximum efficiency and versatility. It constructs the model in a stage-wise approach by optimizing the loss function using gradient descent. XGBoost starts with an initial prediction, which is often a constant value. For regression, this could be the mean of the target variable. For classification, it might be the log-odds. The sum of the primary prediction along with the predictions generated from every tree constructed in the subsequent steps is the final prediction. Formally, the prediction for an instance x at stage t can be expressed as:

$$\acute{y}^{(t)} = \acute{y}^{(t-1)} + f_t(x) \tag{5}$$

where $\acute{y}^{(t)}$ is the prediction at the stage t, $\acute{y}^{(t-1)}$ is the prediction at the stage $t - 1$, and $f_t(x)$ is the prediction of the $t - th$ tree. After t trees are added, the final prediction is given by:

$$\acute{y} = \; \sum_{t=0}^{T} f_t(x) \tag{6}$$

This is the sum of the initial prediction and all the corrections made by the T trees.

Categorical Boosting (CatBoost): CatBoost [24] leverages a gradient-boosting algorithm that successfully manages categorical features while preventing overfitting. It employs a technique called ordered boosting to manage categorical features. It employs target encoding or target statistics to translate categorical variables into numerical values. For each categorical feature, this model computes the average label (target variable) value for each category, using only a subset of the data to prevent overfitting. The transformation can be represented as:

$$Encoding\left(X_{i,catB}\right) = \frac{\sum_{j\in Q(i)} y_j}{|Q(i)|} \tag{7}$$

where $X_{i,catB}$ is the categorical feature of the instance i, y_j is the target value of an instance j, and $Q(i)$ is a subset of the data used to calculate the encoding.

Adaptive Boosting (AdaBoost): AdaBoost [25] is an ensemble technique that is designed to incorporate the outcomes of various weak learners in order to form a strong learner. It operates by sequentially adjusting the amount of weight assigned to incorrectly classified samples so that the following learners prioritize correcting these mistakes. AdaBoost progressively trains a set of weak learners G_t which are usually decision trees with shallow depth or stumps sequentially. At each iteration t,

$$\acute{y} = sign\,(\sum_{t=1}^{t=T} \alpha_t * G_t(x)) \tag{8}$$

where T is the number of weak learners. The weight α_t indicates how much to trust the predictions of G_t in the final ensemble. It depends on the error rate ϵ_t, with lower error rates resulting in higher weights.

$$\alpha_t = \frac{\ln\,(\frac{1-\epsilon_t}{\epsilon_t})}{2} \tag{9}$$

Ensemble Model: This paper experiments with four ensemble models. The ensemble combinations are shown in Table 8 with their assigned acronym.

Table 8. Ensemble combinations

Acronym	Classifiers
Ensemble1	RF, CatBoost, and AdaBoost
Ensemble2	RF, XGBoost, and AdaBoost
Ensemble3	XGBoost, CatBoost, and AdaBoost
Ensemble4	RF, XGBoost, and CatBoost

Table 9. Machine learning model’s hyper-parameters configuration

Model	Hyper-Parameters	
	Customer Complaint Dataset	Seller Complaint Dataset
LR	‘C’: 1, ‘max_iter’: 50, ‘penalty’: l2, ‘solver’: newton-cg	‘C’: 0.1, ‘max_iter’: 100, ‘penalty’: l2, ‘solver’: newton-cg
SVM	‘C’: 10, ‘gamma’: 0.001, ‘Kernel’: sigmoid, ‘class_weight’: custom_weight	‘C’: 1, ‘gamma’: 0.01, ‘Kernel’: sigmoid, ‘class_weight’: None
DT	‘criterion’: entropy, ‘max_depth’: None, ‘max_features’: None, ‘min_samples_leaf’: 1, ‘min_samples_split’: 5, ‘splitter’: random	‘criterion’: gini, ‘max_depth’: None, ‘max_features’: None, ‘min_samples_leaf’: 4, ‘min_samples_split’: 2, ‘splitter’: random
KNN	‘metric’: manhattan, ‘n_neighbors’: 7, ‘weights’: distance	‘metric’: manhattan, ‘n_neighbors’: 9, ‘weights’: distance
RF	‘bootstrap’: False, ‘min_samples_leaf’: 1, ‘min_samples_split’: 2, ‘n_estimators’: 200	‘bootstrap’: True, ‘min_samples_leaf’: 2, ‘min_samples_split’: 5, ‘n_estimators’: 300
XGBoost	‘learning_rate’: 0.01, ‘n_estimators’: 200, ‘max_depth’: 10	‘learning_rate’: 0.1, ‘n_estimators’: 300, ‘max_depth’: None
CatBoost	‘depth’: 4, ‘iterations’: 100, ‘l2_leaf_reg’: 3, ‘learning_rate’: 0.5	‘depth’: 8, ‘iterations’: 100, ‘l2_leaf_reg’: 1, ‘learning_rate’: 0.5
AdaBoost	‘learning_rate’: 0.1, ‘n_estimators’: 200	‘learning_rate’: 0.01, ‘n_estimators’: 300

4. Results and Error Analysis

4.1. Experimental Setup

Each model's hyperparameters were chosen using GridSearchCV [26], which systematically searches for the optimal hyperparameters by evaluating a predefined set of hyperparameter combinations. This approach ensures that the best configuration for each model is found by maximizing their performance on the given task. Table 9 shows the optimal hyper-parameters for each model.

4.2. Evaluation Matrices

This paper considers four evaluation matrices in order to evaluate the performance of models. Their definitions are:

Precision: It is the percentage of accurately predicted positive fraud samples to the total predicted positives. It calculates the accuracy of positive predictions.

$$Precision = \frac{True\_Positives}{True\_Positives\ +\ False\_Positives} \quad (10)$$

Recall: It is the percentage of accurately predicted positive fraud samples to all fraud samples for that class.

$$Recall\ = \frac{True\_Positives}{True\_Positives\ +\ False\_Negative} \quad (11)$$

F1-Score: It is the harmonic mean between precision and recall that offers harmony between these two measures.

$$F1 - Score = \frac{2\ \times (Precision\ \times\ Recall)}{(Precision\ +\ Recall)} \quad (12)$$

Accuracy: It is the percentage of accurately predicted fraud samples to the number of total fraud samples.

$$Acc\ = \frac{No.of\ Correct\ Predictions}{No.of\ Total\ Samples\ in\ Dataset} \quad (13)$$

4.3. Performance Analysis

Customer Complaint Dataset (CCD): From Table 10, it is seen that the CatBoost outperformed all other ML models with no feature selection. It achieved a 94% F1-score and 95% accuracy. But when the ensemble classifiers were explored (shown in Table 11), the Ensemble4 model outperformed all the employed models including CatBoost as well. This model achieved a 95% F1-score and 96% accuracy. So, a clear 1% increase is seen when compared to CatBoost. However, LR achieved the lowest F1-score of 89%, and accuracy of 91%. When a feature selection technique like Chi-Square was employed, the performance of most of the models improved. Only KNN’s performance degraded by 1%, 93% to 92% in terms of both F1-score and accuracy.

With the Chi-Square technique, XGBoost achieved a 95% F1-score, though the Ensemble4 model outperformed all others with an F1-score of 96%. Therefore, this obtained F1-score by Ensemble4 is the highest for the CCD because the highest F1-score with RF feature selection was achieved by the Ensemble1 model, which was 95%. So, 1% less than the Ensemble4 with Chi-Square. Thus the Ensemble4 model with the Chi-Square technique is superior among all models employed in CCD. The detailed results of CCD are shown in Table 10 & Table 11.

Table 10. Performance comparison of traditional ML classifiers on customer complaint dataset. Here F1, Acc, P, and R indicate macro-average F1-score, accuracy, precision, and recall respectively

Classifier	**No Feature Selection**				**Chi-Square Feature Selection**				**RF Feature Selection**			
	P	**R**	**F1**	**Acc**	**P**	**R**	**F1**	**Acc**	**P**	**R**	**F1**	**Acc**
LR	0.88	0.89	0.89	0.91	0.91	0.92	0.91	0.92	0.89	0.90	0.90	0.90
SVM	0.91	0.90	0.90	0.92	0.93	0.92	0.93	0.93	0.93	0.93	0.94	0.95
KNN	0.93	0.93	0.93	0.94	0.92	0.91	0.92	0.93	0.91	0.91	0.91	0.92
DT	0.89	0.90	0.90	0.92	0.93	0.92	0.92	0.94	0.93	0.93	0.93	0.93
RF	0.93	0.94	0.93	0.93	0.94	0.94	0.94	0.95	**0.95**	**0.96**	**0.95**	**0.96**
XGBoost	0.92	0.92	0.92	0.94	**0.95**	**0.95**	**0.95**	**0.95**	0.94	0.93	0.94	0.95
CatBoost	**0.94**	**0.94**	**0.94**	**0.95**	0.95	0.94	0.95	0.95	0.95	0.94	0.95	0.96
AdaBoost	0.92	0.92	0.92	0.94	0.93	0.93	0.93	0.93	0.94	0.95	0.94	0.95

Table 11. Performance comparison of ensemble models on customer complaint dataset. Here F1, Acc, P, and R indicate macro-average F1-score, accuracy, precision, and recall respectively

Classifier	**No Feature Selection**				**Chi-Square Feature Selection**				**RF Feature Selection**			
	P	**R**	**F1**	**Acc**	**P**	**R**	**F1**	**Acc**	**P**	**R**	**F1**	**Acc**
Ensemble1	0.94	0.93	0.93	0.93	0.95	0.95	0.95	0.95	**0.94**	**0.95**	**0.95**	**0.95**
Ensemble2	0.92	0.92	0.92	0.92	0.93	0.92	0.93	0.93	0.94	0.94	0.94	0.94
Ensemble3	0.94	0.94	0.94	0.94	0.92	0.92	0.92	0.93	0.93	0.94	0.94	0.94
Ensemble4	**0.95**	**0.96**	**0.95**	**0.96**	**0.96**	**0.96**	**0.96**	**0.96**	0.95	0.95	0.95	0.95

It is observed that tree ensemble models performed way better than models like LR, SVM, and KNN. The reason behind their superior performance is that the dataset is heavily categorical and those models can efficiently generate trees to find the desired classes. Unfortunately, it didn't happen in the same way for LR, SVM, and KNN.

Seller Complaint Dataset (SCD): From Table 12, it can be noticed that with no feature selection technique, AdaBoost achieved the highest F1-score and accuracy of 97%. The ensemble models couldn't beat AdaBoost with no feature selection technique in terms of F1-score. Only the Ensemble4 model could achieve the same F1-score as AdaBoost. Surprisingly, it surpassed AdaBoost in accuracy and achieved the highest accuracy of 98%, which is basically 1% greater than AdaBoost.

With the Chi-Square technique, the performance didn't improve significantly, rather in most cases it remained almost equal. But with the RF feature selection technique, the Ensemble4 model obtained the best F1-score of 98% and accuracy of 98% (see Table 13). This result is the highest for the SCD considering all the employed models. Unfortunately, due to feature selection and removal, a few models like DT and KNN performed poorly. This happened because of losing the necessary information due to feature removal. Models like KNN lost some nearest neighbors for this reason which made difficulty in accurate predictions. The number of features in SCD is less than in CCD. Hence, a clear impact of feature removal causes issues for these models. However, the performance is reduced by 1% when Chi-Square is employed. The tree ensemble models performed better in this dataset like the CCD because of the categorical characteristics of features.

4.4. Error Analysis

Customer Complaint Dataset (CCD): From the first confusion matrix in Figure 6, it is seen that exactly one sample from each of the 'No Delivery', and 'Faulty Product' classes were misclassified. The predicted classes were 'Late Delivery', and 'Counterfeit Product' respectively. Here, two samples from the 'Counterfeit Product' class were misclassified as 'Faulty Product' and 'Misrepresentation of Product'. Also, two samples from 'Misrepresentation of Product' were misclassified as faulty products. In case of the other confusion matrices such as in Figure 7 and Figure 8, the misclassifications were noticed mostly for the similar classes explained in the case of the first confusion matrix.

It is clear that most of the misclassifications happened in pairwise. For example, between 'No Delivery' and 'Late Delivery', 'Faulty Product' and 'Counterfeit Product', 'Misrepresentation of Product' and 'Faulty Product', 'Misrepresentation of Product' and 'Counterfeit Product'. The reason for these pairwise misclassifications has occurred because these classes are very close in the case of

Table 12. Performance comparison of traditional ML classifiers on seller complaint dataset. Here F1, Acc, P, and R indicate macro-average F1-score, accuracy, precision, and recall respectively

Classifier	**No Feature Selection**				**Chi-Square Feature Selection**				**RF Feature Selection**			
	P	**R**	**F1**	**Acc**	**P**	**R**	**F1**	**Acc**	**P**	**R**	**F1**	**Acc**
LR	0.95	0.94	0.95	0.96	0.96	0.96	0.96	0.96	0.93	0.93	0.93	0.93
SVM	0.95	0.94	0.95	0.95	0.94	0.94	0.94	0.94	0.96	0.96	0.96	0.96
KNN	0.96	0.97	0.96	0.97	0.96	0.96	0.96	0.96	0.97	0.96	0.97	0.97
DT	0.94	0.94	0.94	0.94	0.95	0.94	0.94	0.95	0.95	0.96	0.95	0.96
RF	0.96	0.96	0.96	0.96	0.95	0.94	0.94	0.95	0.96	0.97	0.97	0.97
XGBoost	0.96	0.97	0.96	0.97	0.96	0.96	0.96	0.96	0.97	0.96	0.97	0.97
Catboost	0.96	0.96	0.96	0.96	**0.97**	**0.96**	**0.97**	**0.97**	0.95	0.96	0.96	0.96
AdaBoost	**0.97**	**0.96**	**0.97**	**0.97**	0.96	0.96	0.96	0.96	**0.97**	**0.97**	**0.97**	**0.97**

Table 13. Performance comparison of ensemble models on seller complaint dataset. Here F1, Acc, P, and R indicate macro-average F1-score, accuracy, precision, and recall respectively

Classifier	**No Feature Selection**				**Chi-Square Feature Selection**				**RF Feature Selection**			
	P	**R**	**F1**	**Acc**	**P**	**R**	**F1**	**Acc**	**P**	**R**	**F1**	**Acc**
Ensemble1	0.94	0.94	0.94	0.95	0.96	0.95	0.95	0.96	0.95	0.96	0.96	0.97
Ensemble1	0.94	0.94	0.94	0.95	0.96	0.95	0.95	0.96	0.95	0.96	0.96	0.97
Ensemble3	0.95	0.95	0.95	0.96	0.97	0.97	0.97	0.97	0.96	0.97	0.97	0.98
Ensemble4	**0.97**	**0.96**	**0.97**	**0.98**	**0.98**	**0.97**	**0.97**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**

their definition and the nature of reported incidents. They have a similar form of incident data reported by victims, which made the model to get confused sometimes. Though the misclassifications are not that much higher, still in some case, this similarity couldn't be resolved accurately.

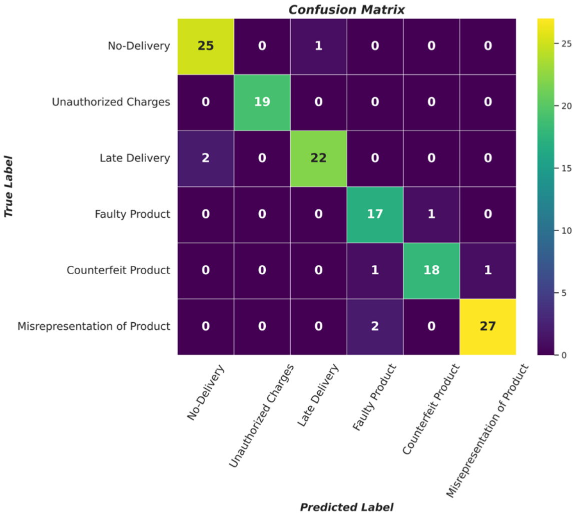


Figure 6. Confusion matrix – 1 by Ensemble4 model with Chi-Square feature selection technique

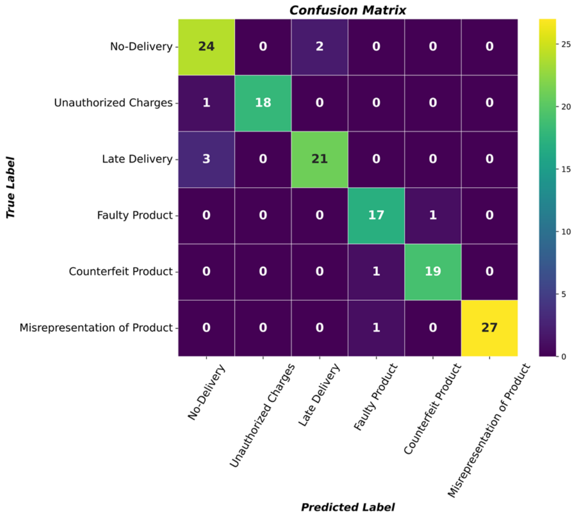


Figure 7. Confusion matrix - 2 by Ensemble4 model with RF feature selection technique

A total of 6, 9, and 10 misclassifications occurred for the confusion matrices 1, 2, and 3 respectively. This demonstrates the excellent performance of the proposed model. We hope that by adding more incident-specific information will remove ambiguity for closely related classes. In the near future, the text-based responses can be considered from the victims to make a more robust model.

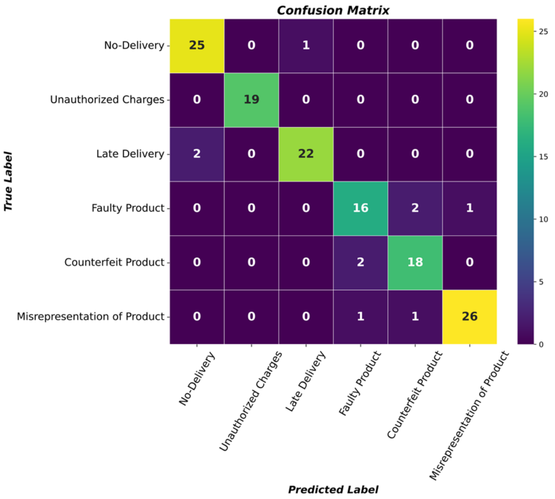


Figure 8. Confusion matrix - 3 by Ensemble4 model with no feature selection technique

Seller Complaint Dataset (SCD): As the seller complaint dataset contains binary classes and the model performed outstandingly well, hence the misclassification wasn't noticed that much. From the first confusion matrix in Figure 9, it can be observed that only one misclassification happened, while confusion matrices 2 and 3 both have the same number of misclassifications, which is two (see Figure 10 and Figure 11 for further details).

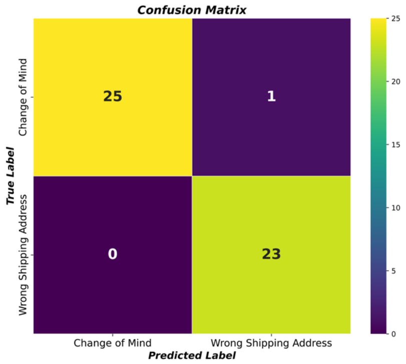


Figure 9. Confusion matrix - 1 by Ensemble4 model with RF feature selection technique

Here the misclassification occurred probably due to the annotation mistake because the annotation accuracy was not 100% due to some unavoidable human error. As the seller dataset consists of fewer training samples and classes, as a future work a comprehensive seller complaint dataset needs to be collected for wider coverage of the fraud incidents that are usually faced by the sellers. More classes need to be incorporated into the datasets for widespread coverage of potential fraud scenarios.

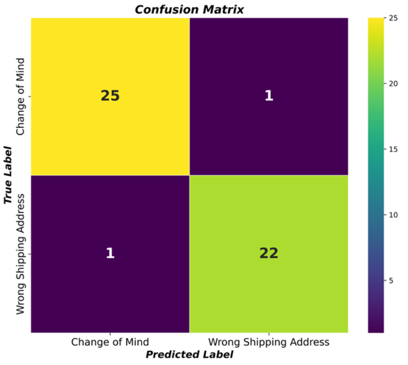


Figure 10. Confusion matrix - 2 by Ensemble4 model with Chi-Square feature selection technique

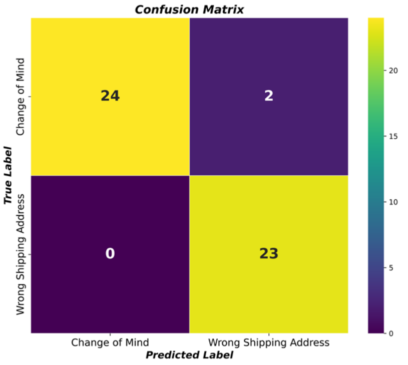


Figure 11. Confusion matrix - 3 by Ensemble4 model with no feature selection technique

4.5. Comparison with Existing Works

As far as we know, no particular study in this regard and way has yet been undertaken for the e-commerce industry. As a result, this study collected a new dataset containing victim information through a questionnaire and the proposed framework is compared by applying it to similar questionnaire-based datasets. In study [7], the author utilized a questionnaire dataset for predicting fraud riskiness on online purchases. The proposed framework outperformed this work in both F1-score and accuracy by 1.57% and 2.13% as shown in Table 14.

Table 14. Performance comparison of our proposed framework

Ref	Accuracy	**Accuracy (Proposed)**	F1-score	**F1-score (Proposed)**
[7]	93.3% Cat-Boost	95.41%	92% Cat-Boost	93.57%
[8]	96% XG-Boost	97.12%	95% XG-Boost	95.38%

In study [8], the author utilized a questionnaire dataset for airline customer satisfaction prediction. The proposed framework outperformed this work by 0.38% in F1-score and 1.12% in accuracy. The superior performance is due to our effective feature selection methods, exploration of ensemble models, optimal hyper-parameter tuning, and assigning class weights to handle the class imbalance.

5. Explainability Analysis of Models

5.1. Shapley Additive Explanations (SHAP)

By giving each feature, a contribution score for each prediction, the SHAP technique [27] adds meaning to the outcome of machine learning models. To divide the model output across the features fairly, it makes use of game theory and Shapley values. The formula for a feature i is:

$$\emptyset_i = \sum\nolimits_{X \subseteq N} \frac{|X|!\,(|N| - |X| - 1)!}{|N|!}\, y\big(X \cup \{i\} - y(X)\big) \quad (14)$$

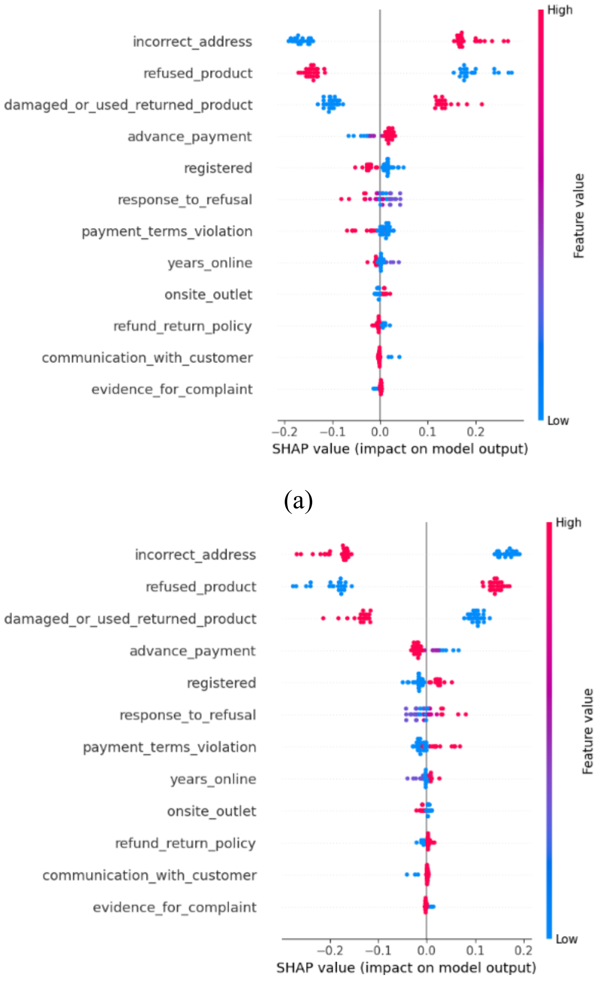
i^{th} feature's SHAP value is $\emptyset_i$, X indicates the subset of features that do not include i, y is the output of the model, and N represents the set that includes all features.

SHAP analysis for CCD:

- **No-Delivery**: Figure 13(a) shows that respondents who bought a product at an unusual price (marked in red with higher SHAP value) often did not authenticate the product before purchase. They were lured into making advance payments, and the fraud seller never delivered the product.
- **Faulty or Damaged Goods**: Figure 13(b) illustrates that people who do not authenticate the product, do not read the product description, or check the return policy are more likely to fall into this trap.
- **Misrepresentation of Products**: Figure 13(c) indicates that those who do not read the product description carefully and buy products based on package discounts are often deceived.
- **Counterfeit Product**: Figure 13(d) shows that customers who buy brand-name products at low prices frequently fall victim to this trap.
- **Late Delivery**: Figure 13(e) doesn't indicate any particular reason, the respondent's direct reporting on the late delivery contributes to this class's prediction.
- **Unauthorized Charges**: Figure 13(f) shows that the products with hidden charges, which customers overlooked because they made payment advances and didn't compare the product price with other online shops are the victims of this type.

SHAP analysis for SCD:

- **Wrong Shipping Address:** Figure 12(a) shows reasons specifically if the seller is not registered and doesn't have an onsite outlet then the fraud buyer took that chance of deception.
- **Change of Mind:** Figure 12(b) indicates that the customer who refused to take the product took the chances of the seller's weakness, like the seller not having a specific return policy.



(a)

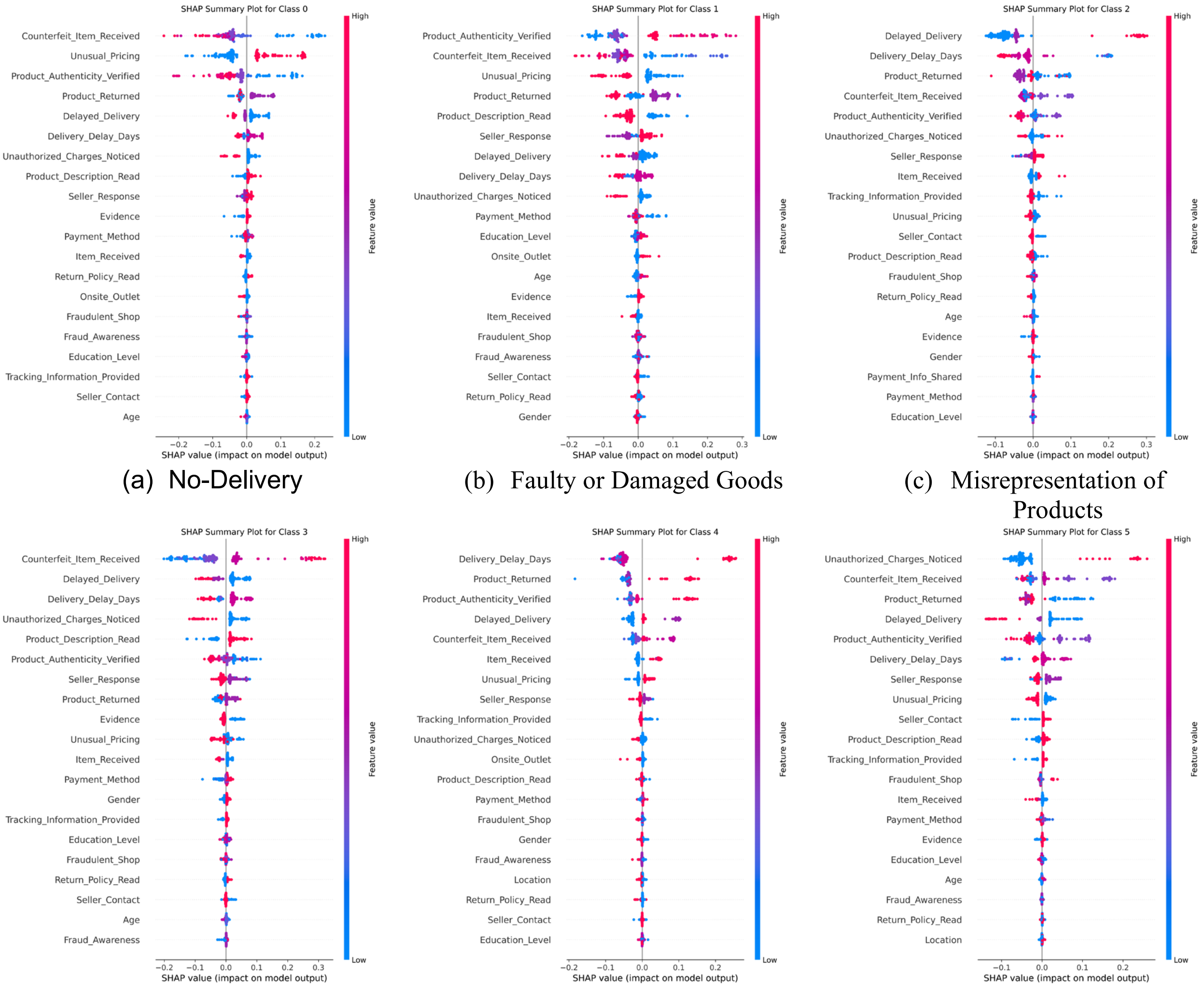
(b)

Figure 12. SHAP plot for seller complaint dataset, (a) Wrong Shipping Address (b) Change of Mind

5.2. Local Interpretable Model Agnostic Explanations (LIME)

It uses an explainable model, like linear regression, to approximate the specific local action of a model to explain individual predictions [28]. It minimizes the following function:

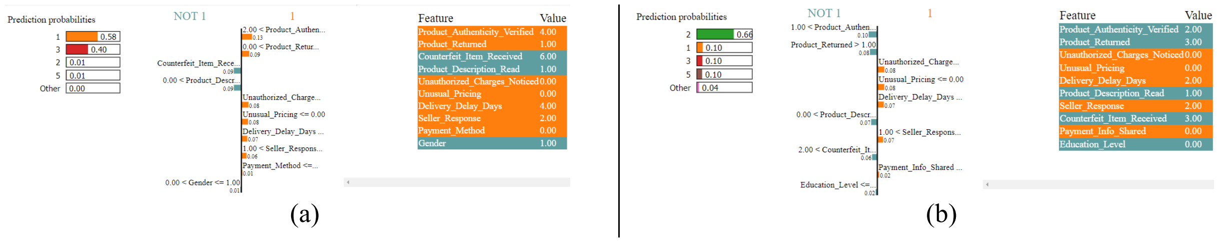
$$\xi(z) = argmin_{m \in M}\, L(F, m, \pi_z) + \Omega(m) \quad (15)$$



(a) No-Delivery (b) Faulty or Damaged Goods (c) Misrepresentation of Products

(d) Counterfeit Product (e) Late Delivery (f) Unauthorized Charges

Figure 13. SHAP for the customer complaint dataset (CCD)


Prediction probabilities
1 0.58
3 0.40
2 0.01
5 0.01
Other 0.00
NOT 1
1
2.00 < Product_Authen... 0.13
0.00 < Product_Retur... 0.09
Counterfeit_Item_Rece... 0.09
0.00 < Product_Descr... 0.09
Unauthorized_Charge... 0.08
Unusual_Pricing <= 0.00 0.08
Delivery_Delay_Days ... 0.07
1.00 < Seller_Respons... 0.06
Payment_Method <=... 0.01
0.00 < Gender <= 1.00 0.01

Feature	Value
Product_Authenticity_Verified	4.00
Product_Returned	1.00
Counterfeit_Item_Received	6.00
Product_Description_Read	1.00
Unauthorized_Charges_Noticed	0.00
Unusual_Pricing	0.00
Delivery_Delay_Days	4.00
Seller_Response	2.00
Payment_Method	0.00
Gender	1.00

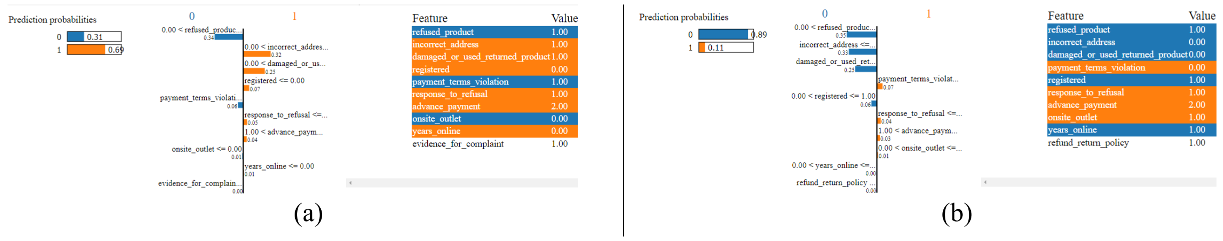
(a)

Prediction probabilities
2 0.66
1 0.10
3 0.10
5 0.10
Other 0.04
NOT 1
1
1.00 < Product_Authen... 0.10
Product_Returned > 1.00 0.08
Unauthorized_Charge... 0.08
Unusual_Pricing <= 0.00 0.08
Delivery_Delay_Days ... 0.07
0.00 < Product_Descr... 0.07
1.00 < Seller_Respons... 0.07
2.00 < Counterfeit_It... 0.06
Payment_Info_Shared ... 0.02
Education_Level <=... 0.02

Feature	Value
Product_Authenticity_Verified	2.00
Product_Returned	3.00
Unauthorized_Charges_Noticed	0.00
Unusual_Pricing	0.00
Delivery_Delay_Days	2.00
Product_Description_Read	1.00
Seller_Response	2.00
Counterfeit_Item_Received	3.00
Payment_Info_Shared	0.00
Education_Level	0.00

(b)

Figure 14. LIME for customer complaint dataset, (a) prediction for the 3rd sample, and (b) prediction for the 5th sample


Prediction probabilities
0 0.31
1 0.69
0
1
0.00 < refused_produc... 0.34
0.00 < incorrect_addres... 0.32
0.00 < damaged_or_us... 0.25
registered <= 0.00 0.07
payment_terms_violati... 0.06
response_to_refusal <=... 0.05
1.00 < advance_paym... 0.04
onsite_outlet <= 0.00 0.01
years_online <= 0.00 0.01
evidence_for_complain... 0.00

Feature	Value
refused_product	1.00
incorrect_address	1.00
damaged_or_used_returned_product	1.00
registered	0.00
payment_terms_violation	1.00
response_to_refusal	1.00
advance_payment	2.00
onsite_outlet	0.00
years_online	0.00
evidence_for_complaint	1.00

(a)

Prediction probabilities
0 0.89
1 0.11
0
1
0.00 < refused_produc... 0.35
incorrect_address <=... 0.33
damaged_or_used_ret... 0.25
payment_terms_violat... 0.07
0.00 < registered <= 1.00 0.06
response_to_refusal <=... 0.04
1.00 < advance_paym... 0.03
0.00 < onsite_outlet <=... 0.01
0.00 < years_online <=... 0.00
refund_return_policy ... 0.00

Feature	Value
refused_product	1.00
incorrect_address	0.00
damaged_or_used_returned_product	0.00
payment_terms_violation	0.00
registered	1.00
response_to_refusal	1.00
advance_payment	2.00
onsite_outlet	1.00
years_online	1.00
refund_return_policy	1.00

(b)

Figure 15. LIME for seller complaint dataset, (a) prediction for the 1st sample, and (b) prediction for the 9th sample

Where L is the fidelity measure between the complex model F and the interpretable model m, π_z is a proximity measure to the instance z, and $\Omega(m)$ is the complexity of the interpretable model.

Figure 14(a) and Figure 14(b) indicate prediction overview for the two specific samples, 3rd and 5th in CCD. From Figure 14(a), it is seen that the 'Counterfeit Product' fraud was predicted by the model for the 3rd sample. The explanation for the predictions in these two samples will remain the same as the explanation using SHAP. Because similar features in both explainable techniques have higher values and contribute to the prediction in the exact same manner. Likewise, CCD, Figure 15(a) and Figure 15(b) delineate the two sample predictions in SCD. Figure 15(a) and Figure 15(b) illustrate the prediction for the 1st and 9th samples. The prediction for the 1st sample was 'Change of Mind', and for the 9th sample, the prediction was 'Wrong Shipping Address'.

6. Limitation and Future Work

The limitations of the findings of this paper are outlined below:

- Dataset is collected only from the Bangladeshi e-commerce market; this may limit generalizability in the global context
- Few training samples and fraud classes in the seller complaint dataset
- The dataset is heavily categorical and should have contained more continuous values
- No deep learning model like MLP is explored

Suggested future works that may guide the potential research scopes to improve this paper's findings:

- Expand the questionnaire and conduct funded surveys for comprehensive and unbiased data
- Include continuous value and open-ended text-based responses from the respondents to build a multi-modal classification model
- Conduct time series analysis on order dates to identify peak fraud times
- Develop a real-time IoT-based fraud prediction system for instant police support
- Explore additional feature selection methods
- Apply the methodology to other countries' e-commerce markets and consider incorporating data from more than one country
- Explore the deep learning model for better prediction
- Build a mobile application to implement the whole system in one place to facilitate user interaction with authority and report to the authority regarding their fraud cases

7. Conclusion

This paper presented a machine learning-based explainable fraud incident classification framework, EcomFraudEX. Two survey datasets were collected, one for customers and another for sellers. This framework achieved an impressive F1-score of 96% and an accuracy of 97% on the customer complaint dataset utilizing the Chi-Square feature selection method. Similarly, it attained an F1-score of 97% and an accuracy of 98% on the seller complaint dataset utilizing the RF feature selection method. The best performance was observed with the Ensemble4 model, which integrates RF XGBoost, and CatBoost algorithms.

By employing this framework in existing works, this paper was able to improve accuracy by 2.13% and F1-score by 1.57%. For the model's prediction analysis, two Explainable AI (XAI) techniques SHAP and LIME were utilized. The explanation suggests that customers should read the product description carefully before purchasing any product from online marketplaces, they shouldn't feel tempted to buy a product at an unusual discount or offer, they should read the return policy before purchase, they should analyze the seller's reputation, etc. On the other hand, sellers should register their shops with the relevant authority, make the return policy clear to the customer, implement payment-related fraud detection methods, and follow the guidelines provided by the government authorities.

Acknowledgments

Ethics Approval:
Not Applicable.

Conflict of Interest:
The author declares no conflict of interest.

Funding:
The author didn't receive any kind of funding for this work.

Availability of Supporting Data:
The collected and used dataset in this study is made public and it is available at the following URL provided below: https://www.github.com/Salman1804102/EcomFraudQD

Authors consent to publish:
All the authors have given their consent to publish this manuscript.

Author Contribution:
The contributions of the authors are provided below:

[**Salman Farsi**]: Ideation, Data Collection, Implementation, Methodology, Writing – Original Draft, Adjusting Reviews to Prepare the Final Manuscript.

[**Mahfuzulhoq Chowdhury**]: Ideation, Supervision, Manuscript Revision & Review, Resource Provision.

Acknowledgement:
A special thanks to Mr. Tarek Hossain, a practicing lawyer for consumer rights, for his involvement in this work as an expert. We also convey our heartfelt gratitude to all the respondents who took the time to fill up the Google form to give us valuable research data. Finally, thanks to Mahfuja Mannan Rimu, who helped us in data collection.

References

[1] E-Commerce Fraud Statistics: $48 Billion Lost Annually | wisernotify.com [Internet]. [Accessed: 2024 Jun 13]. Available from: https://wisernotify.com/blog/ecommerce-fraud-stats/

[2] 23+ eCommerce Fraud Statistics (2024) [Internet]. [Accessed: 2024 Jun 13]. Available from: https://explodingtopics.com/blog/ecommerce-fraud-stats

[3] Ileberi E, Sun Y, Wang Z. A machine learning based credit card fraud detection using the GA algorithm for feature selection. J Big Data. 2022 Dec 25;9(1):24.

[4] Karunachandra B, Putera N, Wijaya SR, Suryani D, Wesley J, Purnama Y. On the benefits of machine learning classification in cashback fraud detection. Procedia Comput Sci. 2023;216:364–9.

[5] Hu X, Zhang X, Lovrich NP. Forecasting Identity Theft Victims: Analyzing Characteristics and Preventive Actions through Machine Learning Approaches. Vict Offender. 2021 May 19;16(4):465–94.

[6] Nguyen NT, Ha PP, Nguyen LT, Nguyen KV, Nguyen NL. Vietnamese complaint detection on e-commerce websites. In New Trends in Intelligent Software Methodologies, Tools and Techniques 2021 (pp. 618-629). IOS Press.

[7] Sabih M, Ejaz M, Quershi KK, Asad MU, Gu J, Balas VE, et al. Fraud Prediction in Pakistani E-commerce Market. In: 2021 4th International Symposium on Advanced Electrical and Communication Technologies (ISAECT). IEEE; 2021. p. 01–6.

[8] Ramadhan, Ghaniaviyanto N, Putrada, Gautama A. XGBoost for Predicting Airline Customer Satisfaction Based on Computational Efficient Questionnaire. International Journal on Information and Communication Technology (IJoICT). 2023;9(2):120–36.

[9] Seera M, Lim CP, Kumar A, Dhamotharan L, Tan KH. An intelligent payment card fraud detection system. Ann Oper Res. 2024 Mar 8;334(1–3):445–67.

[10] Alzahrani RA, Aljabri M. AI-Based Techniques for Ad Click Fraud Detection and Prevention: Review and Research Directions. Journal of Sensor and Actuator Networks. 2022 Dec 31;12(1):4.

[11] BOZYİĞİT F, DOĞAN O, KILINÇ D. Categorization of Customer Complaints in Food Industry Using Machine Learning Approaches. Journal of Intelligent Systems: Theory and Applications. 2022 Mar 1;5(1):85–91.

[12] Vu T, Nguyen DQ, Nguyen DQ, Dras M, Johnson M. VnCoreNLP: A Vietnamese natural language processing toolkit. arXiv preprint arXiv:1801.01331. 2018 Jan 4.

[13] Whittaker JM, Edwards M, Cross C, Button M. “I Have Only Checked after the Event”: Consumer Approaches to Safe Online Shopping. Vict Offender. 2023 Oct 3;18(7):1259–81.

[14] Cohen J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychol Bull. 1968;70(4):213–20.

[15] LabelEncoder - scikit-learn 1.5.0 documentation [Internet]. [Accessed: 2024 Jun 13]. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html

[16] Alshaer HN, Otair MA, Abualigah L, Alshinwan M, Khasawneh AM. Feature selection method using improved CHI Square on Arabic text classifiers: analysis and application. Multimed Tools Appl. 2021 Mar 21;80(7):10373–90.

[17] Menze BH, Kelm BM, Masuch R, Himmelreich U, Bachert P, Petrich W, et al. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics. 2009 Dec 10;10(1):213.

[18] Hosmer Jr DW, Lemeshow S, Sturdivant RX. Applied logistic regression. John Wiley & Sons; 2013.

[19] Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. Support vector machines. IEEE Intelligent Systems and their Applications. 1998 Jul;13(4):18–28.

[20] Peterson LE. K-nearest neighbor. Scholarpedia. 2009;4(2):1883.

[21] Quinlan JR. Induction of decision trees. Machine learning. 1986 Mar;1:81-106.Breiman L. Random forests. Mach Learn. 2001;45:5–32.

[22] Chen T, Guestrin C. XGBoost. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM; 2016. p. 785–94.

[23] Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. Adv Neural Inf Process Syst. 2018;31.

[24] Freund Y, Schapire RE. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. J Comput Syst Sci. 1997 Aug;55(1):119–39.

[25] Eusha A, Farsi S, Islam A, Hossain J, Ahsan S, Hoque MM. CUET_Binary_Hackers@DravidianLangTech-EACL 2024: Sentiment Analysis using Transformer-Based Models in Code-Mixed and Transliterated Tamil and Tulu. In: Chakravarthi BR, Priyadharshini R, Madasamy AK, Thavareesan S, Sherly E, Nadarajan R, et al., editors. Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages [Internet]. St. Julian’s, Malta: Association for Computational Linguistics; 2024. p. 205–11. Available from: https://aclanthology.org/2024.dravidianlangtech-1.34

[26] Van den Broeck G, Lykov A, Schleich M, Suciu D. On the Tractability of SHAP Explanations. Journal of Artificial Intelligence Research. 2022 Jun 23;74:851–86.

[27] Vishwarupe V, Joshi PM, Mathias N, Maheshwari S, Mhaisalkar S, Pawar V. Explainable AI and Interpretable Machine Learning: A Case Study in Perspective. Procedia Comput Sci. 2022;204:869–76.

[28] Zhang Y, Wang J, Zhang X. Personalized sentiment classification of customer reviews via an interactive attributes attention model. Knowledge-Based Systems. 2021 Aug 17;226:107135.