A Proposed Fraud Detection Model based on e-Payments Attributes a Case Study in Egyptian e-Payment Gateway

Mohamed Hassan Nasr[1], Mona Mohamed Nasr[3]
Faculty of Computers and Information System
Helwan University, Cairo, Egypt

Mohamed Hassan Farrag[2]
Faculty of Computers and Information System
Fayoum University, Fayoum, Egypt

***Abstract*—As per Payfort's 2017 report, titled State of payments in the Arab world; Egypt had a 22% yearly increase in the overall volume of internet payments in 2016, which was assessed at $6.2 billion. e-Payments are the major point of life nowadays in Egypt and the whole world; with tens of e-payments companies in Egypt and more than 5 million transactions done every day and 60 billion EGP volume of payments in 2018. Online and mobile fraud was estimated at $10.7 billion in 2015, as per Juniper Research, and is expected to reach $25.6 billion by the end of the decade. As the whole e-payments business is affected by fraud, e-payments firms and their consumers lose a lot of money. On the other hand, one of the most powerful techniques that could be used for fraud predictive is data mining techniques such as the decision tree. This paper introduces a prediction model for managing the risk of fraud in the Egyptian e-payment market that helps to reduce the loss of money. This model is developed using a real dataset from one of Egypt's top e-payment gateways based on the e-payment transaction attributes importance like transaction time, transaction amount, transaction limit, and transaction customer No. repetition limit. The importance of these attributes was determined using IBM SPSS modeler's decision tree and its predictors' importance. The model significantly assisted in the reduction of fraud cases by a very high rate, with an accuracy of 88.45% and a precision of 93.5% resulting in a savings of 101970.52 EGP out of 131297.83 EGP.**

***Keywords*—Data mining; decision tree; e-payments; fraud detection; e-payment gateways; e-commerce**

I. Introduction

Egypt scored the highest growth of online shopping in the Arab world with a 32% increase in the volume of e-payments. According to the internet, world stats report (2017) internet users in Egypt By the end of March 2017, Egypt had officially hit 36.5 percent of the population, half of whom use e-commerce services for everything from purchasing goods and services to paying bills. There are many definitions for e-payments, of which an e-payment system is a form of financial commitment that involves the buyer and the seller facilitated via the use of electronic communications [6]. Another definition defines e-payment as any form of fund transfer via the internet [6].

Using e-commerce; business payments have taken the form of exchanging money electronically and are called electronic payments [2]. Nowadays, most organizations, companies, and government agencies have adopted electronic commerce to increase their productivity or efficiency in trading products or services in areas such as credit cards, telecommunication, healthcare insurance, automobile insurance, online auction, etc. [1]. The success of a particular electronic payment system is determined by how well it overcomes the practical and analytical hurdles that various online payment methods face.

These challenges include issues of laws and regulations (buyer and seller protection), technological capabilities of e-payment service providers, commercial relationships, and security considerations such as verification and authentication issues [2]. E-commerce systems are used by both legitimate users and fraudsters. Hence, they become more vulnerable to large-scale and systematic fraud. Internet Crime Complaint Centre (IC3) is a valuable resource for both victims of internet crime and law enforcement agencies in identifying, investigating, and prosecuting these crimes. In 2020, the IC3 gathered 15,421 Tech Support Fraud complaints from victims in 60 countries. The losses were over $146 million, an increase of 171 percent over the previous year (IC3, 2020).

The Egyptian e-payments market is also affected by fraud crimes. Many customers have been defrauded in the Egyptian e-payments market by someone calling them and pretending to be from the e-payments company sales team, attempting to convince them to give their account details in order to receive a bonus. The motivation of this model is to protect and minimize the losses of customers who have been deceived.

One of the most important and rapidly growing payment methods in Egypt is e-payments gateway companies with more than 5 million transactions done every day. Those companies have more than 789 services that customers can use and about 294 thousand outlets spreading in all places in Egypt with 60 billion EGP volume of payments in 2018 and expectation to reach 90 billion in 2019. With high efficiency, ease, and speed, e-payment has become a significant facilitating engine in e-commerce through e-business success.

Hence, the paper is proposing a model for fraud detection in e-payments, especially in Egyptian e-payments companies and the proposed model will be applied to one of those companies. A decision tree, which is one of the most effective data mining techniques, was used to create the model depending on the importance of the e-payment transaction's attributes. The paper is organized into seven sections. The

second section presents dentitions of e-payment gateway, decision tree, C5.0 algorithm, and related work. In Section 3 the methodology is presented. Section 4 shows the decision tree model and Section 5 presents results and discussion. Conclusion and future work are presented in Sections 6 and 7.

II. Literature Review

A. Background

This section explains the main points of the area being researched and gives an overview of these points. It starts with explaining what is e-payment gateway then give a brief about decision tree technique, C5.0 algorithm, Splitting criteria, and information gain metric.

1) e-Payment gateway: An electronic payment gateway system is a software service that connects with retailer and service provider networks and enables consumers to make payments through these [8]. e-Payment companies offer financial services to consumers and businesses through various channels and a large network of agents. These financial services include paying bills, paying vouchers, reservations, donations, and other services. e-Payment helps businesses to reach more customers, increase productivity, and help consumers to save time and pay for services at any time in a cheaper, easier, faster, and real-time way.

2) Decision tree: Decision tree is the most important and widely used categorization and forecasting approach. A decision tree is a tree structure that looks like a flowchart, with each internal node representing an attribute test, each branch reflecting the test's outcome, and each leaf node (terminal node) storing a class label[12]. By learning simple decision rules derived from data properties, a decision tree is used to develop a model that forecasts required variable values. ID3, C4.5, C5.0, and CART are only a few of the algorithms for learning decision trees from a given data set that have been presented. The C5.0 algorithm will be used in our model because of its accuracy and ease of implementation.

3) C5.0 Algorithm: One of the most well-known algorithms is C5.0. The C5.0 technique has become the best choice for generating decision trees since it works successfully for most kinds of challenges straight out of the box. The decision trees of the C5.0 algorithm work nearly as well as more difficult and complex machine learning approaches (such as Neural Networks and Support Vector Machines), but are significantly easier to understand and use.

4) Splitting criteria: Information Gain (Entropy) is the splitting criterion used by C5.0. The C5.0 model breaks the sample into fields based on whatever field provides the maximum information gain. Each sub-sample specified by the first split is split a second time, generally on a different field, and the technique is repeated until the subsamples cannot be split anymore. Lastly, the lowest-level splits are inspected again, and those that do not add significantly to the model's value are trimmed or removed.

5) Information gain: Information gain is a metric for how much data a feature offers about a class. It helps to specify the order of attributes in the decision tree's nodes. The entropy of the dataset before and after a transformation is used to calculate information gain. The entropy of a sub split can be defined as a measure of its purity. Entropy is always between 0 and 1. Below is how the Information Entropy is calculated for a dataset with N classes.

$$\mathrm{E} = -\sum_{i=1}^{N} p\mathrm{i} \log_2 p\mathrm{i} \quad (1)$$

Where pi is the probability of randomly picking an element of class i (i.e. the proportion of the dataset made up of class i). Below is the formula of calculating Information Gain based on the calculated Entropy.

$$Gain = Eparent - Eparent \quad (2)$$

B. Related Work

Techniques for identifying fraud in e-payments are growing very quickly. Some of the popular techniques are rule-based systems, neural networks, Decision trees, machine learning business intelligence, hidden Markov Model, etc. As per lae Chouiekha and EL Hassane Ibn EL Haj [3], they have proposed a model to detect fraudsters in mobile communication using deep learning techniques. In order to forecast fraudulent events in a mobile environment, researchers compared the performance of convolution neural networks to that of classic machine learning algorithms.

Shirley Wong and Sitalakshmi Venkatraman [10] have proposed financial fraud detection and propose a forensic accounting framework using business intelligence. This framework presents a three-phase methodology for performing financial analysis, such as ratio analysis, for a business case scenario using unique knowledge discovery techniques. In contrast to traditional methods of vertical and horizontal analysis for the business case study, the framework's implementation practically demonstrates how the technologies and investigative methods of trend analysis could be used to investigate fraudulent financial reporting using their accounting data.

In another work, Roland Rieke, Maria Zhdanova, Jürgen Repp, and Romain Giot [9] developed a tool for runtime predictive security analysis that analyses process behavior in relation to transactions within a money transfer service and attempts to match it with the expected behavior provided by a process model. The tool analyzes deviations from the given behavior specification for anomalies that indicate a possible misuse of the service related to money laundering activities.

Memorie Mwanza [7] has proposed a model for detection of fraud on tax data for Zambia Revenue Authority using business intelligence model that implements data mining, outlier algorithms for fraud detection and is based on, Continuous Monitoring of Distance-Based and Distance-Based Outlier Queries was then designed and Hao ZHOU, Hong-feng CHAI and Maolin QIU [11] introduce machine learning algorithms to perform fraud detection of bankcard enrollment. They introduce several traditional machine-learning algorithms and finally choose the improved gradient boosting decision tree (GBDT) algorithm software library for use in a real system, namely, XGBoost also Shailesh S. Dhok, G. R.

Bamnote [4] model the sequence of operations in credit card transaction processing using a Hidden Markov Model (HMM) and show how it can be used for the detection of frauds. An HMM is initially trained on a cardholder's regular behavior. An incoming credit card transaction is considered fraudulent if the trained HMM does not accept it with an enough high probability. The hidden Markov Model aids in achieving high fraud coverage while minimizing false alarms.

With [5], two main techniques and using five classifiers, Sahu, Aanchal, G. M. Harshvardhan, and Mahendra Kumar created models to detect credit card fraud transactions. These two techniques are designed to address the problem of data imbalance which helps to detect fraud transaction.

III. Methodology

This section explains all the steps done to generate the model starting with describing the used data set, its preparation steps, and the used tools that helped build the model. The data set show available service categories and explains transaction details for the e-payment transaction done through the payments gateway.

A. Data Set

The data set contains real-life data of financial transactions of one of the top five e-payments companies in Egypt. Table I shows the services categories that e-payments companies in Egypt provide to their customers. There are governmental services such as electricity; gas and water there are mobile services for recharging and bill payment and many other services such as donation services, airline services DSL services, and many other services that any Egyptian consumer or corporates needs. Table II shows transaction details, some of those details are received by the system when the transaction is processed and some already exist related to the agent who made the transaction. The data set contains about 92718 records divided into two parts. Training data that contains 64903 transactions about (70%) and testing data set that contains 27814 transactions about (30%).

B. Data Preparation

As shown in Fig. 1 data preparation steps start with data gathering. The data was extracted directly and manually from the data source and exported to excel sheets then data cleaning and validation start by removing extraneous data and outliers, filling in missing values, conformed data into a specified pattern sensitive and private data was hidden, and removing error transactions. The next step was discovering and classifying the data. Data were classified and divided into months, weeks, and days for each agent's account based on three attributes (number of transactions, the amount, number of customer's number repeated).

Data analyzing was the final step. The data were divided into three groups based on the volume of transactions and the volume of their amounts in order to obtain better and more accurate results.

- Low volume rate group.
- Medium volume rate group.
- High volume rate group.

TABLE I. Service Category

#	Service category	Description
1	Mobile recharge	Used for recharging mobile
2	Mobile bill payment	Paying for mobile monthly bills
3	Mobile e-voucher	Generate vouchers for recharging mobile phone
4	Donations	Donation to charities
5	Airlines	Reserving airlines tickets
6	Gas	Gas bills
7	Water	Water bills
8	Electricity	Electricity bills
9	DSL	DSL bills
10	Cinema tickets	Reserving tickets online
11	Games	Paying for online games

TABLE II. Transaction Details

#	Attribute	Description
1	Create Date/Time	Date and time for creating the transaction
2	Update Date/Time	Date and time for receiving a response from the service provider
3	Transaction ID	Unique transaction ID in the system
4	Provider	Service provider name
5	Service	Service name
6	Customer number	Identifier used by customer EX: Phone number for recharging mobile
7	Amount	The amount that should be paid
8	Total amount	The amount that should be paid + Service charge
9	Status	Status of transaction success or error
10	Provider Response Code	Code sent by the service provider in the response
11	Provider Transaction ID	Service provider unique transaction Id in
12	Transaction Initiator	Name of agent that made the transaction
13	Transaction Deduction From	Account number for the agent who made the transaction
14	Interface	Type of interface he used (mobile, web
15	Outlet	Name of the outlet
16	Area	Area of the outlet

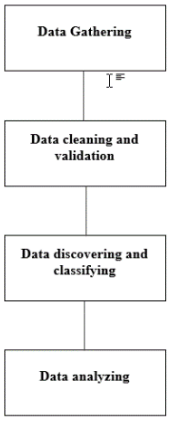

Fig. 1. Data Set Preparation Steps.

TABLE III. DATA CATEGORIES SPECIFICATION

		Low Volume Rate	Medium Volume Rate	High Volume Rate
Transaction	Daily	50	100	300
	Weekly	300	1000	3000
	Monthly	1200	5000	10000
Amount	Daily	1200	2500	8000
	Weekly	5000	20000	80000
	Monthly	20000	100000	500000
Customer No.	Daily	3	6	10
	Weekly	15	30	50
	Monthly	50	100	100

As shown in Table III, each group contains the maximum daily transactions number allowed for the agent, the maximum weekly number, the maximum monthly number also the maximum daily, weekly and monthly amount, and the same for customer No. repeated times.

The data set was divided into six data sets based on the three groups that were received from the data analysis phase. Each group has two data sets one for training and one for testing. The training data set contains 70% of transactions and the testing data set contains 30% of transactions for each group as shown in Table IV.

TABLE IV. DATA SET VOLUME

	Low Volume Rate	Medium Volume Rate	High Volume Rate	Total
Training	42426	28242	22050	64903
Testing	12727	8472	6615	27814

IV. DECISION TREE MODEL

A. Tools used

IBM SPSS modeler was used to create the decision tree model. It is a data-mining tool that allows you to create predictive models without programming. It helps you to specify groups, identify correlations between them and predicting events that will happen in the future. ASC5.0 is one of the decision tree algorithms included in IBM modeler, and it was used in the proposed model as shown in Fig. 2 due to its efficiency, as stated previously.

B. Splitting Criteria

As previously stated in the background section, C5.0 uses Information Gain (Entropy) as its splitting criteria. Fig. 3, Fig. 4, and Fig. 5 show the predictor importance for the e-payment transaction attributes for each group (low, medium, and high) as generated by IBM modeler. e-Payment transaction attributes are transaction allowed time, transaction (daily, weekly, monthly) allowed limit, transaction (daily, weekly, monthly) allowed amount, and transaction (daily, weekly, monthly) allowed customer number limit.

C. Resulting Decision Tree

The model will start by checking the transaction time if it is in the agent's allowed time or not and if it is in the allowed time the model will go to the next step to check if the transaction exceeds the maximum monthly transaction number if it did not exceed the model will check for the weekly and after that the daily. The model will check for Amount and customer, No. repeated in the same way as the transaction number and if the transaction passed all the conditions it will be accepted. Fig. 6 shows the whole procedure of the resulting decision tree.

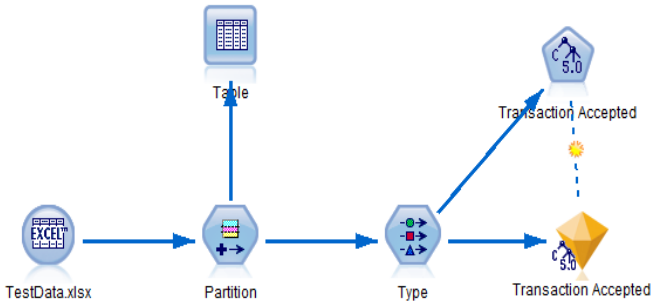

Fig. 2. IBM SPSS Decision Tree.

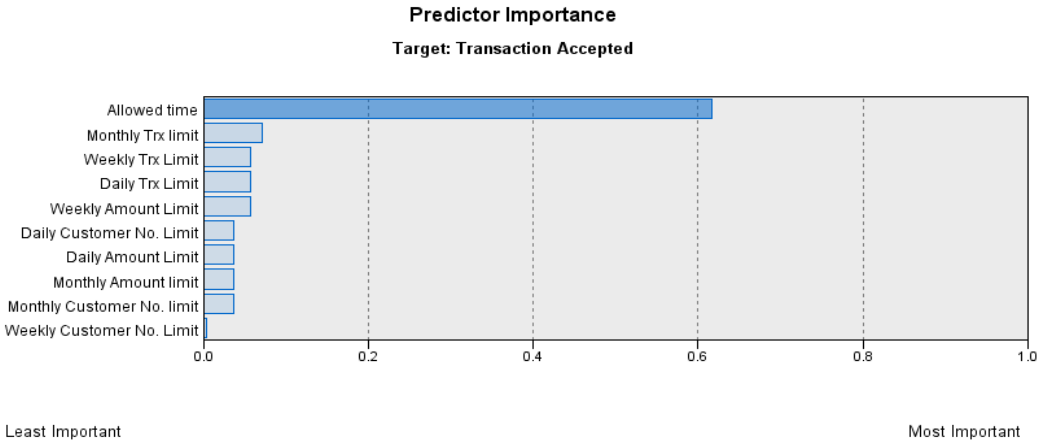

Fig. 3. Predictor Importance for Low Rate Group.

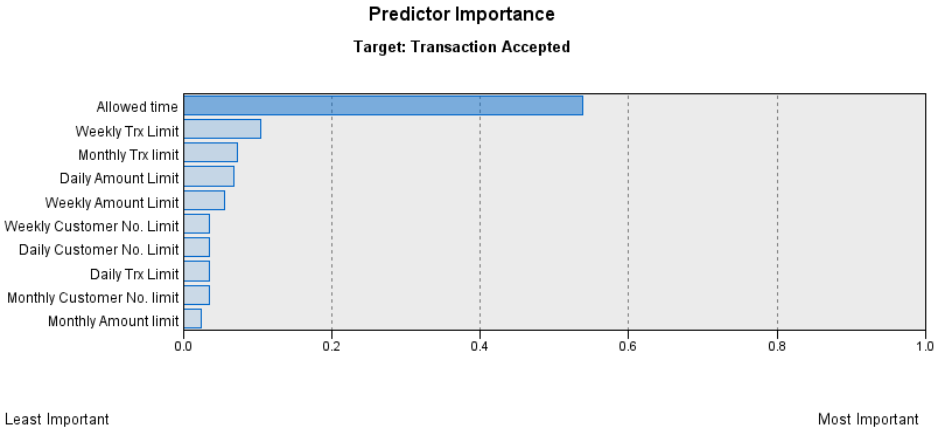

Fig. 4. Predictor Importance for Medium Rate Group.

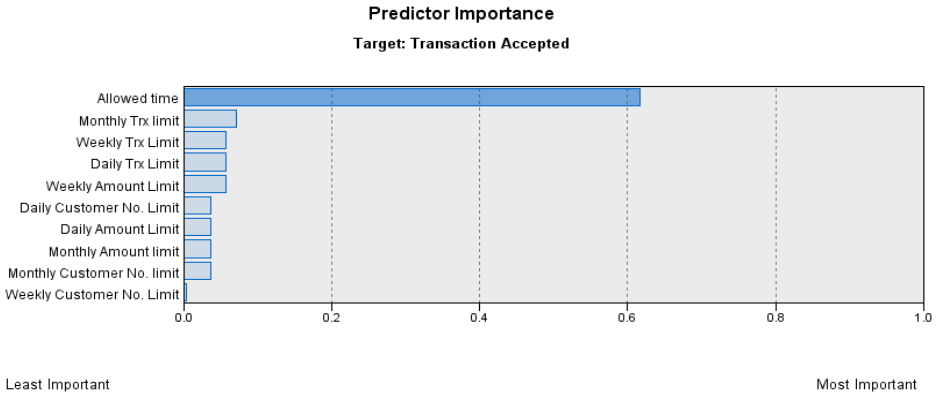

Fig. 5. Predictor Importance for High Rate Group.

D. Decision Tree Derived Rules

R = Result for condition statement.

1 =True, 0 =False.

F(Mtrx) = Monthly transaction.

F(Mamut) = Monthly Amount.

F(Mcus.no) = Monthly Customer No.

F(Wtrx) = Weekly transaction.

F(Wamut) = Weekly Amount.

F(Wcus.no) = Weekly Customer No.

F(Dtrx) = Daily transaction.

F(Damut) = Daily Amount.

F(Dcus.no) = Daily Customer No.

1) Low volume rate group

$$R = f(Mtrx) \int_{0,>\,1200}^{1,\leq 1200}, R = f(Wtrx) \int_{0,>\,300}^{1,\leq 300}, R = f(Dtrx) \int_{0,>\,50}^{1,\leq 50} \quad (3)$$

$$R = f(Mamount) \int_{0,>\,20000}^{1,\leq 20000}, R = f(Wamount) \int_{0,>\,5000}^{1,\leq 5000}, R = f(Damount) \int_{0,>\,1200}^{1,\leq 1200} \quad (4)$$

$$R = f(Mcus.no) \int_{0,>\,50}^{1,\leq 50}, R = f(Wcus.no) \int_{0,>\,15}^{1,\leq 15}, R = f(Dcus.no) \int_{0,>\,3}^{1,\leq 3} \quad (5)$$

2) Medium volume rate group

$$R = f(Mtrx) \int_{0,>\,5000}^{1,\leq 5000}, R = f(Wtrx) \int_{0,>\,1000}^{1,\leq 1000}, R = f(Dtrx) \int_{0,>\,100}^{1,\leq 100} \quad (6)$$

$$R = f(Mamount) \int_{0,>\,100000}^{1,\leq 100000}, R = f(Wamount) \int_{0,>\,20000}^{1,\leq 20000}, R = f(Damount) \int_{0,>\,2500}^{1,\leq 2500} \quad (7)$$

$$R = f(Mcus.no) \int_{0,>\,100}^{1,\leq 100}, R = f(Wcus.no) \int_{0,>\,30}^{1,\leq 30}, R = f(Dcus.no) \int_{0,>\,6}^{1,\leq 6} \quad (8)$$

3) High volume rate group

$$R = f(Mtrx) \int_{0,>\,10000}^{1,\leq 10000}, R = f(Wtrx) \int_{0,>\,3000}^{1,\leq 3000}, R = f(Dtrx) \int_{0,>\,300}^{1,\leq 300} \quad (9)$$

$$R = f(Mtrx) \int_{0,>\,500000}^{1,\leq 500000}, R = f(Wtrx) \int_{0,>\,80000}^{1,\leq 80000}, R = f(Dtrx) \int_{0,>\,8000}^{1,\leq 8000} \quad (10)$$

$$R = f(Mtrx) \int_{0,>\,100}^{1,\leq 100}, R = f(Wtrx) \int_{0,>\,50}^{1,\leq 50}, R = f(Dtrx) \int_{0,>\,10}^{1,\leq 10} \quad (11)$$

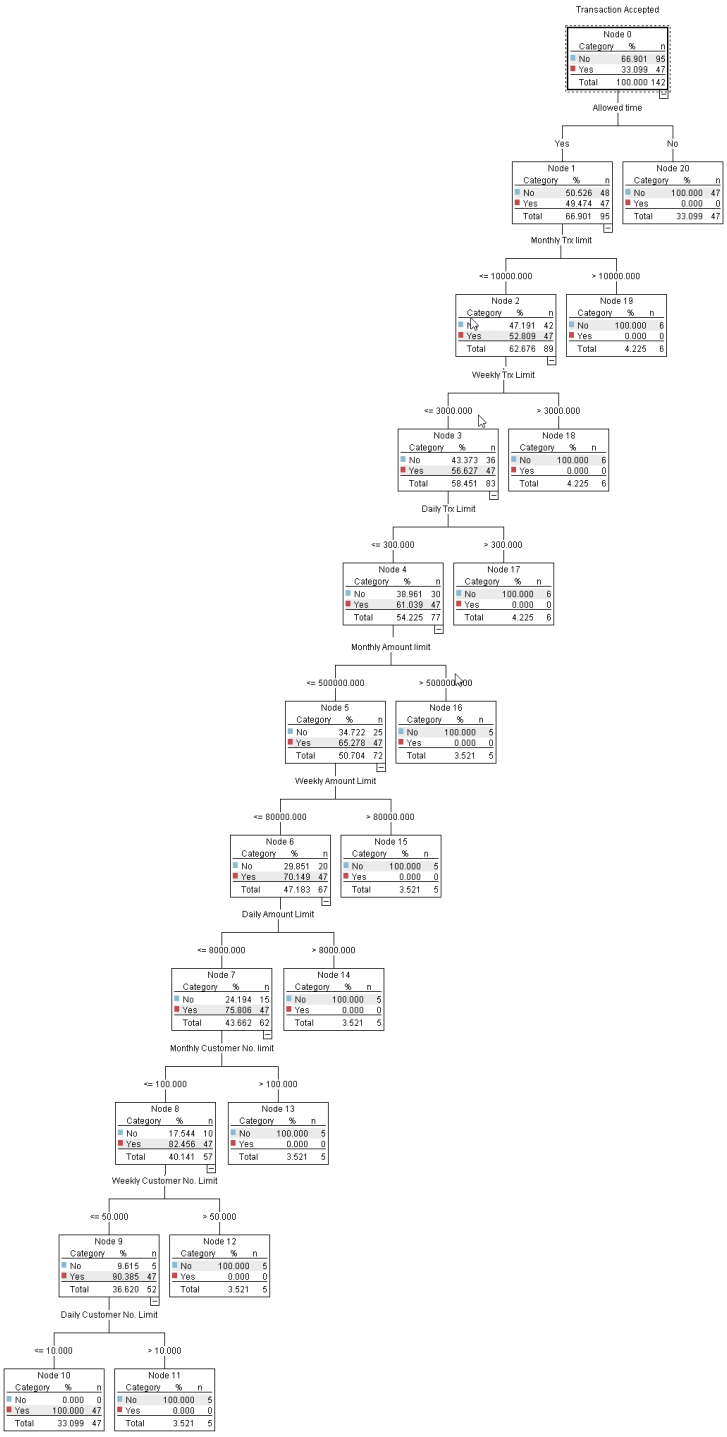

Fig. 6. Decision Tree Model for Low, Medium and High Volume Rate Groups.

V. Result and Discussion

A. Results and Findings

The model was implemented and a real data set that containing fraud transactions were used to test the implemented model. The used data set is a real dataset of a total of 1590 transactions with 590 fraud transactions totaling 131297.83 EGP in losses. With a total of 433 out of 590 fraud transactions detected, the model was able to identify more than 73%. Out of a total of 131297.83 EGP, the model saved 77% of the total amount lost due to fraud, equating to 101970.52 EGP. The model significantly reduced fraud transactions, demonstrating the need to analyze and verify that all customer transactions are carried out by them, not only that they come from their account. The confusion matrix was used to calculate the accuracy, precision, and false alarm rate for the model as shown in Table V.

TABLE V. CONFUSION MATRIX DEFINITIONS

TP	True positive (number of transactions that were fraudulent and were also classified as fraudulent by the model)
TN	True negative (number of transactions that were legitimate and were also classified as legitimate)
FP	False positive (number of transactions that were legitimate but were wrongly classified as fraudulent transactions)
FN	F False negative (number of transactions that were fraudulent but were wrongly classified as legitimate transactions by the model)

Accuracy is the fraction of transactions that were correctly classified.

Accuracy (ACC)/Detection rate = (TN + TP) / (TP + FP + FN + TN)

Precision (also known as the detection rate), the number of transactions either genuine or fraudulent were correctly classified.

Precision/Detection rate/Hit rate = TP / TP + FP

False Alarm rate measures out of total instances classified as fraudulent or how many were wrongly classified.

False Alarm Rate = FP/FP+TN.

TP= 433 TN= 1000 FP = 30 FN= 157.

Accuracy (ACC)/Detection rate = 1000 + 433 / (433 + 30 + 157 + 1000) = 88.45 %.

Precision/Detection rate/Hit rate = 433/433 + 30 = 93.5 %F.

False Alarm Rate = 30/30+1000 = 2.9%VI. Accuracy & Precision & False Alarm Rate.

Accuracy	Precision	False Alarm Rate
88.45 %	93.5 %	2.9 %

B. *Limitation of Work*

One of the limitations that we faced is the lack of previous research studies on the topic of the Egyptian e-payment market and the absence of official statistics that help estimate the extent of the problem. As well as the difficulties we faced as a result of the payment companies' lack of cooperation and refusal to provide any numbers of fraud losses suffered by their customers.

VI. CONCLUSION

The paper introduces a fraud detection model using data mining. The model used the decision tree technique. The model is based on real data from one of the Egyptian e-payment gateways. With an accuracy of 88.45 percent, the suggested model helps in the detection of any up normal transactions that differ from the typical behavior of users' transactions. Fig. 7 shows that the model detected 433 fraud transactions out of 590, resulting in a savings of 101970.52 EGP out of 131297.83 EGP as shown in Fig. 8. Using this model, a secure payment environment will enhance user confidence and reduce money loss.

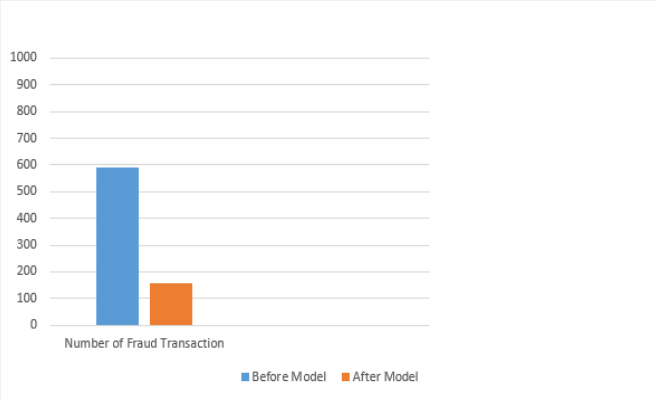

Fig. 7. Number of Fraud Transactions before and after using the Model.

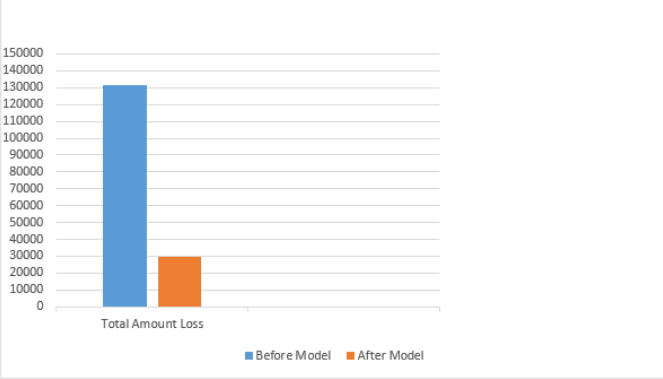

Fig. 8. Total Amount Loss before and after using the Model.

VII. FUTURE WORK

This paper introduced a fraud detection model using the decision tree technique for a specific type of e-payment, which are e-payment companies that allow users to pay for services through their system. Through this research, several points have arisen that must be discussed in the future. One of these points is to implement the proposed model to other different e-payment types and analyze the results in order to improve the Another point is to apply this model using other data mining techniques and compare it to decision tree technique also applying this model in another e-payments fields for example credit card payments and mobile banking payments.

REFERENCES

[1] Abdallah, Aisha, Mohd Aizaini Maarof, and Anazida Zainal. "Fraud detection system: A survey." Journal of Network and Computer Applications 68 (2016): 90-113.

[2] Bezovski, Zlatko. "The future of the mobile payment as electronic payment system." European Journal of Business and Management 8, no. 8 (2016): 127-132.

[3] Chouiekh, Alae, and EL Hassane Ibn EL Haj. "Convnets for fraud detection analysis." Procedia Computer Science 127 (2018): 133-138.

[4] Dhok, Shailesh S., and G. R. Bamnote. "Credit card fraud detection using hidden Markov model." International Journal of Soft Computing and Engineering (IJSCE) 2, no. 1 (2012): 231-237.

[5] Sahu, Aanchal, G. M. Harshvardhan, and Mahendra Kumar Gourisaria. "A dual approach for credit card fraud detection using neural network and data mining techniques." In 2020 IEEE 17th India Council International Conference (INDICON), pp. 1-7. IEEE, 2020.

[6] Kabir, Mohammad Auwal, Siti Zabedah Saidin, and Aidi Ahmi. "Adoption of e-payment systems: a review of literature." In International Conference on E-Commerce, pp. 112-120. 2015.

[7] Mwanza, Memorie. "Fraud detection on big tax data using business intelligence, data mining tool: A case of Zambia revenue authority." PhD diss., University of Zambia, 2017.

[8] Nasr, Mohamed Hassan, Mohamed Hassan Farrag, and Mona Nasr. "e-payment Systems Risks, Opportunities, and Challenges for Improved Results in e-business." International Journal of Intelligent Computing and Information Sciences 20, no. 1 (2020): 16-27.

[9] Rieke, Roland, Maria Zhdanova, Jürgen Repp, Romain Giot, and Chrystel Gaber. "Fraud detection in mobile payments utilizing process behavior analysis." In 2013 International Conference on Availability, Reliability and Security, pp. 662-669. IEEE, 2013.

[10] Wong, Shirley, and Sitalakshmi Venkatraman. "Financial accounting fraud detection using business intelligence." Asian Economic and Financial Review 5, no. 11 (2015): 1187-1207.

[11] Zhou, Hao, Hong-feng Chai, and Mao-lin Qiu. "Fraud detection within bankcard enrollment on mobile device based payment using machine learning." Frontiers of Information Technology & Electronic Engineering 19, no. 12 (2018): 1537-1545.

[12] Nuruzzaman, Md, Md Shahadat Hossain, Md Mostafijur Rahman, Ahete Shamul Haque Chowdhury Shoumik, Md Abbas Ali Khan, and Md Tarek Habib. "Machine Vision Based Potato Species Recognition." In 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 1-8. IEEE, 2021.