Fraud Detection in Shipping Industry using K-NN Algorithm

Ganesan Subramaniam[1]
College of Computing and Informatics
Universiti Tenaga Nasional
Malaysia

Moamin A. Mahmoud[2]
Institute of Informatics and Computing in Energy (IICE)
Universiti Tenaga Nasional
Malaysia

***Abstract*—The shipment industry is going through tremendous growth in volume thanks to technological innovation in e-commerce and global trade liberalization. Volume growth also means a rise in fraud cases involving smuggling and false declaration of shipments. Shipping companies and customs are mostly relying on routine random inspection thus finding fraud is often by chance. As the volume increases dramatically it would no longer be sustainable and effective for both shipment companies and customs to pursue traditional fraud detection strategies. Other related papers on this area have proven that intelligent data-driven fraud detection is proven to be far more effective than routine inspections. However, the challenge in data-driven detection is its effectiveness are often reliant on the availability of data and the various fraud mechanism used by fraudsters to commit shipment related fraud. As such in this paper, we review and subsequently identify the most optimized approaches and algorithms to detect fraud effectively within the shipping industry. We also identify factors that influence fraud activity, review existing fraud detection models, develop the detection framework and implement the framework using the Rapidminer tool.**

***Keywords*—Fraud detection; shipping industry; k-nn algorithm**

I. Introduction

World Customs Organization's (WCO) Illicit Trade Report 2016 states that the year 2016 was marked as the year of Digital Customs where the administrations were encouraged to actively showcase and promote the use of Information and Communication Technologies (ICT) by using a data-driven approach to collect and safeguard duties, control the flow of goods, people finally to secure cross-border trade [1]. CIO Magazine from IDG identified fraud detection as one of the IT projects primed for machine learning [2]. As such, there is a pressing need within the industry for a more intelligent fraud detection system that can considerably improve the detection of wrong declarations and smuggling compared to random checks.

Liberalization in trade and technological innovation such as advancement in e-commerce has accelerated international shipping volumes significantly over the last few years. The latest statistics from The International Air Transport Association (IATA) released full-year 2017 data for global air freight markets showing that demand, measured in freight ton-kilometers (FTKs) grew by 9.0%. This was more than double the 3.6% annual growth recorded in 2016 [3]. Such a rapid rise in volumes will lead to an increase in safety and compliance issues as tremendous volume creates a strain for both the shipping companies and customs authorities to perform safety and compliance audits on most shipments. Meanwhile, on the end of the spectrum, customers are demanding e-commerce providers and shipping companies to provide faster deliveries. By fulfilling urgent deliveries shipping companies are also able to expand their business to a new market segment such as urgent medical deliveries which provides a higher margin of profit to shipping companies. Thus, placing more manual checks on shipments by both shipping companies and customs will only cause more delays on shipments which directly impacts the profitability of shipping companies and their customer base. As such, there is an urgent need to automate and increase the effectiveness of the current random checking done by both the shipping companies and the customs. There are various violations committed by shippers such as illicit trade, smuggling that violates shipping restrictions between countries or miscoding of the shipment items shipped that saves custom duties payments. There could be also security and safety issues if shipments are not audited thoroughly. Imagine the impact of dangerous or flammable goods such as mobile batteries being declared as safe goods in an air freight shipment that can cause devastating impacts such as plane crashes. According to the UK P&I Club, 27% of incidents on cargo ships in 2013 and 2014 were attributable to mis-declared hazardous cargo, second only to poor packaging [3].

Issues and challenges highlighted in [4] [5] paper are generally concept drift or dynamic fraud patterns, overlapping data, capability to support real-time detection requirements, skewed distribution, integrating a vast amount of data, and data quality-related issues. Concept drift is the challenge to deal with sudden customer behavioral changes which could turn out to be a false positive outcome. The solution to overcome concept drift is to use an adaptive FDS algorithm that learns and improves over time by factoring in all the possible input variables that may influence the change in expected behaviors. Overlapping data is the issue where fraudulent transactions are made to look like genuine data which becomes true negative cases. The skewed distribution is the issue of having a very low ratio of fraudulent cases which may not be sufficient to train supervised classification-based FDS algorithms. Data quality issues also need to be reviewed as this factor directly impacts the efficiency of fraud detection.

Each issue and challenge impact the respective fraud domain area differently. There are mainly 5 business domain areas where FDS has applied namely banking,

telecommunication, insurance, online business, and shipping [6]. The specific area involved in banking is credit card-related fraud. Meanwhile, in insurance, the specific area will be medical insurance claims and vehicle insurance claim-related fraud [7]. In an online business, the typical fraud area is online auction-related fraud. Shipping-related fraud is usually related to smuggling and miscoding which is a false declaration of the goods being shipped. The critical issue for the shipment domain is getting efficient real-time results within a huge data set. The bigger the data set the better the efficiency thus we need to have a solution architecture that can process an optimal volume of the desired dataset that is efficient enough to be executed in a real-time mode. In the express shipping domain, accuracy and real-time performance is very critical as the life cycle of a shipment only varies between 3-5 days depending on the weight and location. Thus, identifying the fraud before the shipment gets delivered is very critical. The earlier the shipments are intercepted the bigger the cost benefits for both organizations and customers. Immediate detection avoids revenue leakage and improves customer's trust and confidence towards the organization's brand. It also ensures fraud culprits are identified effectively and handed over much earlier to authorities that may help to reduce future fraud cases.

To build a solution that detects fraud effectively we need to identify the parameters or the data elements that influence the most in actual fraud cases. In a study done in a leading global logistics company, it was identified that location is one of the key parameters that influence fraud cases. The location for shipment can be either the origin or the destination of the shipment. According to a report published by World Customs Organization (WCO), shipment origin and destination location can be a major factor as most frauds tend to originate from or being sent to a specific location. Based on data provided by a major global logistics organization the data that will be extracted for our simulation will be the origin and destination respective latitude and longitude values. Since these values are numerical it can precisely identify a location. With these precise numerical-based location values, a specific fraudulent shipment origin or destination and its surrounding area within a defined radius will be tagged as fraudulent by the algorithm. By having numerical data, the processing speed of the algorithm will also be much faster as opposed to using text or image data.

II. Literature Review

In subsequent sections, we will be reviewing various papers that are related to fraud detection.

A. Methods

The search strategy is the definition and selection process to find the most relevant papers are described in the following. The digital databases searched in this review include IEEE Xplore, Springer, and Science Direct. The reason for the selection of these four databases is due to the availability of highly cited and reliable papers in the fields of computer science and its related applications. The review objective is to find all primary research work associated with fraud detection systems within the shipping or logistics domain. The earlier phrase that was searched is "fraud detection system and shipping or cargo or freight or logistics" but since there were not many fraud detection system papers in the shipping domain thus most of the returns were only relevant to fraud detection. There were only 3 papers related to the shipping domain [8]. Finally, only the term "fraud detection system" was used. The initial query resulted in a total of 5866 papers: 598 from IEEE Explore, 964 from Science Direct, and 4304 from Springer. The filtered articles were published between 2000 and 2018. For Science Direct and Springer besides the year filter based on the topic is also applied to ensure non-computer science-related papers are excluded. The reason the year was narrowed down between 2000 and 2018 was due to the no of results which came up to thousands. After sifting through some of the papers we have divided into survey papers and specialized papers which specifically delve into specific techniques or business domain area such as financial which is a credit card or insurance, healthcare, telecommunication, and internet-related marketing fraud.

B. Review

The earliest survey paper since the year 2000 is the paper from [9]. This paper reviews fraud detection from a statistical perspective. Just like most fraud detection-related papers this paper also categorizes basic statistics models for fraud detection methods into supervised and unsupervised. Besides categorization by models, it also surveys papers based on application area or domain. Among the application covered are in the area of credit card fraud, money laundering, telecommunications fraud, computer intrusion which is also known as hacking these days, medical and scientific fraud which also includes plagiarism in the education sector. This paper concluded that the key issue in fraud detection is the effectiveness of fraud detection. Factors such as the speed of detection are directly related to its effectiveness. As such a strategy to use a graded system of investigation is suggested where areas with very high suspicion and high fraud value merit immediate and intensive investigation. This paper also concluded that fraud detection can be achieved even in difficult circumstances but there are also many challenges and opportunities waiting to be tapped in the future. In 2004 another fraud detection survey paper by [10] was published. This paper focuses more on fraud detection techniques. Domain areas covered are credit card fraud detection, telecommunication fraud detection, and computer intrusion detection. Common techniques applied in credit cards are outlier detection which is an unsupervised method that does not rely on historical data. Outliers are based on observation of deviation against the normal or average pattern. It's suitable to detect fraud that has not previously occurred. To detect fraud pattern which previously occurred then supervised method using historical labeled data are used. Neural network-related techniques which is a set of interconnected weighted nodes designed to function like a human brain are also applied widely for credit card fraud detection, but this technique requires an actual data set that is rarely made available to the public. For computer intrusion detection several techniques such as expert system, neural networks, model-based reasoning, data mining, and state transition analysis are applied. The challenge in the computer intrusion domain is to deal with heaps of the audit trail data, dealing with false alarms rate, difficulty in testing, simulating potential scenarios, and poor portability as the ruleset is very specific to a particular environment. Lastly in

telecommunication fraud detection among the techniques used are rule-based, a neural network that includes Bayesian network and also visualization methods. The challenge of managing the data load in supervised learning for the rule-based and neural network can be mitigated by using unsupervised learning to filter out normal behavior data. To create a more robust selection process for rule base technique a non-greedy rule-selection approach can be explored further. The telecommunication environment is very dynamic and always evolving thus it requires accurate definitions of thresholds and parameters that in tune with the changing landscape of this domain.

In 2010 data mining-based fraud detection research surveyed papers from the year 1998 till 2010 [11]. This paper also highlights types of fraudsters and affected industries. The type of fraudsters is divided broadly into managers, employees, or external parties. The most challenging fraudsters are the external parties as they are many of them and they can make use of various complex and new fraud mechanisms. This is the area where we need to apply data analytics or data mining techniques as it will be cost-effective compared to conventional manual methods to find the riskiest parties by using suspicion scores, rules, and visual anomalies that can be investigated and refined. This paper also identifies the fraud domain area as internal which is fraud committed by management and staff within the organization, insurance, credit card, and telecommunications. Credit transactional fraud detection has received the most attention from researchers. There are also other emerging fraud areas such as e-business and e-commerce related fraud in the online world. Two main challenges of data mining-based fraud detection research are the lack of publicly available real data to perform research on and also the lack of well-researched methods and techniques. To overcome the challenge of data availability a solution to use simulated data that closely matches the actual data which are often very sensitive to be shared in public domains. These were proposed in some papers such as [12] [13] [14] [15]. To overcome the issue well-researched methods and techniques some performance matrices and measures are critical to ensure fraud detection gets well-deserved attention from business stakeholders to invest and provide funding that flourishes research and development in these initiatives. Among the measures taken are such as placing a monetary value on predictions that can maximize cost savings/profits by having their own cost and benefit model customized according to their respective business needs. Other considerations to determine the methods of fraud detection are speed of fraud detection and also the styles/types of detection such as online/real-time or batch mode.

In early 2016, Abdallah et al.[4] released a survey paper that covers papers from 1997 to 2014. This paper provides a good summary of the matters surrounding the fraud detection system. Fraud is defined by the Association of Fraud Examiners (ACFE) as the use of one's occupation for personal enrichment through the deliberate misuse or misapplication of the employing organization's resource or assets. There are 2 main types of fraud systems namely fraud prevention systems and fraud detection systems. Fraud prevention is the first line of defense against fraud which blocks the entry of any fraudsters into the system. Meanwhile, fraud detection is the next layer of defense that detects fraudsters who have already committed the fraudulence act. Over the years many fraud detection approaches and techniques have been applied.

In [16], the paper studied the highest level grouping been categorized by area of study. The 2 main groups are statistical modeling and machine learning. Statistical modeling is an area of mathematics that deals with collecting and analyzing data with some assumptions. Machine learning is a technique using programming algorithm models that learn from data and solve complex problems. There are 2 main methods of machine learning namely supervised and unsupervised types. Some approaches combine these two techniques which are known as semi-supervised. The difference between supervised and unsupervised is in the use of labeled data in supervised as oppose to unsupervised which does not use any labeled data. Labeled data is the identification of fraud data in the data set that are used to train the algorithm or model. Unsupervised techniques rely on a grouping of similar attributes or finding outliers that can identify unusual behavior or patterns that can be further investigated. An overview of the various methods and techniques is illustrated in Figure 1.

Table 1 provides a summary of fraud detection data mining tasks with commonly used algorithmic techniques and example use cases [17]. Table 2 provides a comparison summary of fraud detection data mining algorithms that would help to identify the suitable algorithm that can be applied in this paper's use case [17].

Another potential mechanism that could be used in Fraud detection is using multi-agent systems [19] [26] [27]. MAS could be integrated with social norms [18] [20] [21] [22], to identify and learn different customs norms and accordingly predict anomaly behaviour [23] [24] [25] [28].

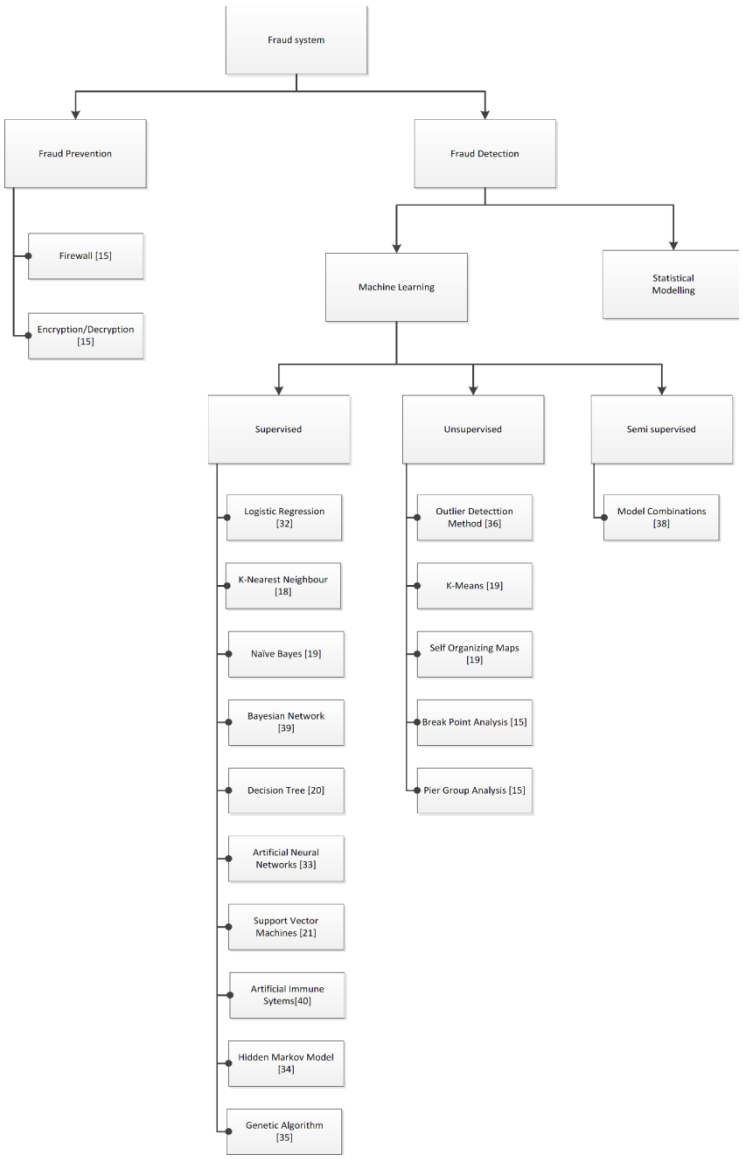

Fig. 1. Fraud Detection Algorithms.

TABLE I. FRAUD TASKS SUMMARY

Tasks	Description	Algorithms	Approach	General Usage
Classification	Datapoint prediction within predefined groups. Learning-based prediction from known data set.	Decision trees, neural networks, Bayesian models, induction rules, k-nearest neighbors	Supervised	Classifying customers into known groups. e.g., Profitable customers/fraudsters
Regression	Numeric target label prediction of a data point. Learning-based prediction from known data set.	Logistic regression	Supervised	Subsequent period fraud predicting fraud and estimating losses
Anomaly detection	Outlier prediction of datapoint against other data within the data set.	Distance-based, density based, local outlier factor (LOF)	Unsupervised	Detection of Credit Card Fraud and network intrusion
Clustering	Natural clusters identification based on inherent properties within data sets.	k-means, density-based clustering	Unsupervised	Fraud segments identification within organization using transaction, web, and customer call data

TABLE II. ALGORITHMS USED FOR FRAUD DETECTION COMPARISON SUMMARY TABLE

Classification						
Algorithm	**Model**	**Data**	**Outcome**	**Strength**	**Weakness**	**Operational Usage**
Decision Trees  Branching out data into subsets where each contains responses of one class	Data set partitioning based on different predictors values	Unrestricted variable type for predictors	The label must be categorical.	Simpler to present to business users. Predictors normalization is not required	Data overfit. Input data changes can cause significantly different trees. Challenging to choose the parameter	Marketing segmentation, fraud detection
k-Nearest Neighbors  Lazy learner where no model is defined. New unknown data point is compared with similar known training set data point	Model is the entire training data set.	Unrestricted but distance calculations work better with numeric data. Normalized data required.	Target variable prediction, which is categorical.	Faster to build model. Missing attributes handled well Works with nonlinear relationships. .	Operational runtime and storage requirement will be high. Value of k randomly selected. No model description	Image processing, application, fraud detections
Naïve Bayesian  Bayes theorem output class prediction by calculating class conditional probability and prior probability.	For each attribute with an output class need a lookup table of probabilities and conditional probabilities	Unrestricted but calculation of probability works better with categorical attributes	Probability prediction for all class values, together with the winning class.	Faster modeling and deployment. Suitable algorithm for benchmarking. Strong statistical foundation	Training data set needs represent population sample well and needs to have complete input/output combination. Independent attributes required	Detections of Spam and mining of text.
Artificial Neural Networks  Biological nervous a system inspired mathematical model. Actual /prediction tuning using network weights	Processing of data based on layers of network topology and weights.	All attributes should be numeric.	Prediction of target (label) variable, which is categorical	Suitable for modeling nonlinear relationships. Fast response time during runtime.	Complex inner working of the model. Requires preprocessing of data. Missing attributes cannot be handled.	Image recognition, fraud detection, quick response time applications.

Regression						
Algorithm	**Model**	**Data**	**Outcome**	**Strength**	**Weakness**	**Operational Usage**
Support Vector Machines  Boundary detection algorithm that illustrates multidimensional boundaries separating data points that belong to different classes.	Vector equation model that enables classification of new data points into different regions or groups.	All attributes should be numeric.	Prediction of target (label) variable, which can be categorical or numeric.	Underfit data and tolerates high variance. Small changes to input data does not influence boundary that results from inconsistency. Suitable for nonlinear relationships.	Training phase computational performance is slower. Additional effort is also needed to optimize parameter combinations.	Optical character recognition, fraud detection, modeling unpredictable events.

Clustering						
Algorithm	**Model**	**Data**	**Outcome**	**Strength**	**Weakness**	**Operational Usage**
k-means  Finding k centroids once data set is divided into k clusters or groups	Find k centroids using algorithm and data points are associated to the nearest centroids, that forms a cluster or group.	Data should be normalized. Works with all types of data but for distance calculations works better with numeric data.	Data set is appended by one of k cluster labels.	Simple implementation. Can be used to reduce number of dimension.	K specification may not be Precise and Cluster may Not be natural clusters. Sensitive to outliers.	Customer segmentation, anomaly detection, Applicable for natural globular clustering.
Anomaly Detection						
Algorithm	**Model**	**Data**	**Outcome**	**Strength**	**Weakness**	**Operational Usage**
Distance Based  Outlier Identification based from kth nearest neighbor	Distance score assigned for all data point based on nearest neighbor.	Data must be normalized due to distance calculation Numeric and categorical attributes accepted.	Distance score assigned to every data point. The further the distance, higher the probability of an outlier.	Easier implement mentation. Suitable for numeric attributes	Specification of k is arbitrary.	Fraud detection, pre-processing technique.
Density Based  Identification of outlier based on data points in low-density regions.	Neighborhood based density score for all data points	Data must be normalized due to density calculation. Both numeric and categorical attributes accepted.	Density score assigned to every data points. The lower the density, the probability of an outlier is higher.	Easier Implement ation. Higher precision with numeric attributes.	Distance parameter specification by the end user. Challenge in identifying varying density regions.	Fraud detection, preprocessing technique.
Local outlier factor  Outlier identification based on relative density calculation within neighborhood	Neighborhood based relative density score for all data points.	Data must be normalized due to density calculation. Both numeric and categorical attributes accepted.	Density the score assigned to every datapoints. The lower the relative density, the probability of an outlier is higher.	Handles varied density scenario.	Distance parameter specification by the end user.	Fraud detection, preprocessing technique.

III. Solution Design

The processes defined for the proposed model as shown in Table 3. First, the fraud data will be prepared, normalized, and cleaned up by replacing missing values and removing duplicates. The data is simulated based on a study done in a global logistics organization based on their historical shipment origin and destination data for 5 years from 2012 till 2017. There are 5 columns namely Fraud label, Origin City Latitude, Origin City Longitude, Destination City Latitude, and Destination City Longitude. Origin or destination with high cases of fraud is tagged with fraud field as "Y". The number of rows simulated is 1500 records. A snapshot of the data is shown in Table 3.

Once the data preparation is completed data will be fed into the process model. Fraud attribute shall be labeled as target role. Once the label has been set up the validation process can be configured. Split type can be set as relative, and the ratio can be dynamically changed from a range of 0.6 to as 0.8 to increase the accuracy. A split ratio of 0.75 means 75% of the data will be used as training the algorithm and the remaining 25% data will be used to test the trained algorithm. Model design is illustrated in Figure 2. The flow in blue is using the labeled data are iterated with various combinations of the split ratio, various algorithms, and parameters related to the specific algorithm. Tuning of these variables will produce results that can be measured in terms of accuracy within an acceptable execution time.

Once an acceptable performance is achieved the algorithm chosen together with its known parameters can be applied to every new incoming shipment data. In this way, fraudulent shipments can be detected at the time the shipment is still in progress within the network before it gets delivered. The algorithm will be updating the prediction column to flag fraudulent shipments accordingly. Shipments flagged as fraud can be further investigated to check if it is genuine. If it's wrongly flagged further analysis needs to be done to improve the algorithm in the future. The analysis also needs to be done on shipments that were not flagged by the algorithm, but it was later found to be fraudulent.

TABLE III. Data Set

Fraud	Origin City Lat	Origin City Lng	Dest City Lat	Dest City Lng
Y	-28.5495	29.78	7.8704	9.78
N	38.0517	58.21	40.306	36.563
N	13.979	45.574	30.5333	105.5333
N	43.8436	-88.8386	-25.5096	-57.36
N	48.5095	-122.2344	35.2495	-81.1856
N	42.7666	-78.6172	38.758	-89.9839
N	10.9587	123.3086	23.1904	75.79
Y	-13.6396	-72.89	44.5372	135.5172
N	-33.5995	150.74	72.685	-78.0001
Y	40.8128	44.4883	-0.215	-78.5001

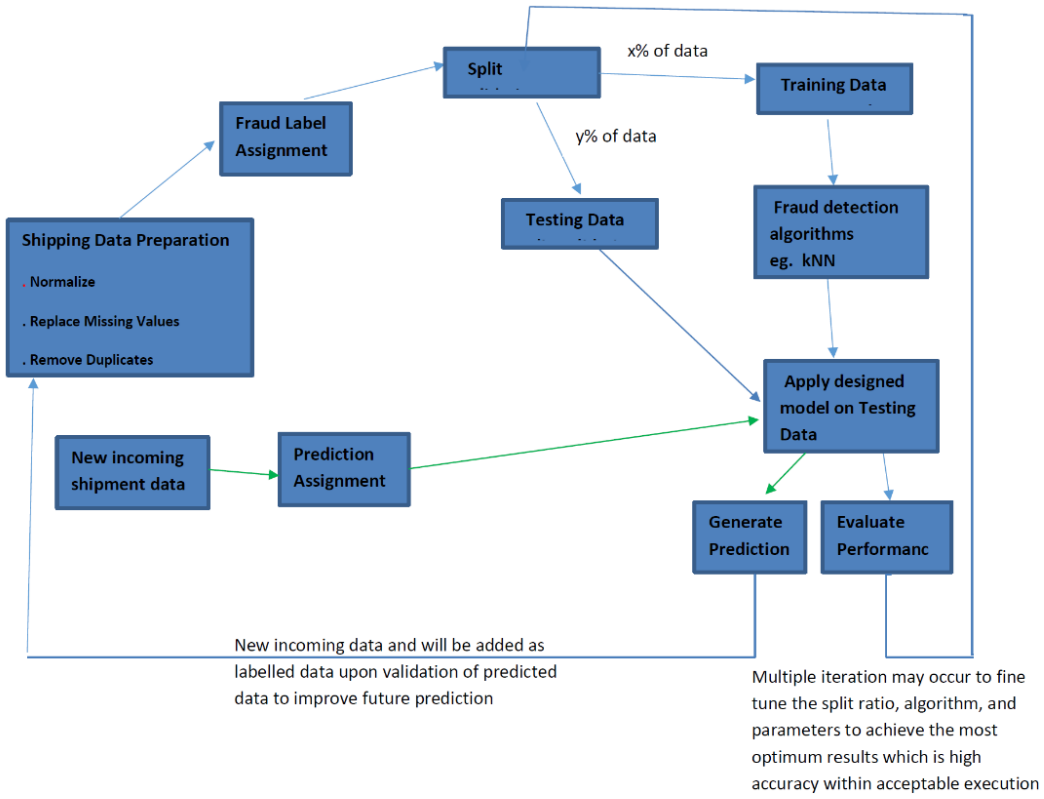

Fig. 2. Solution Design.

IV. Solution Modeling

Various tools can be used to perform fraud detections in the market like R, Rapid-Miner, SAS Enterprise Miner, IBM SPSS, etc. In this paper, we have selected Rapidminer as its simple to use and it also has many built-in ready-to-use algorithms. This allows various techniques can be tested against the available data. Furthermore, it's also provided as a freeware version for students. Once the data is ready then the Rapidminer tool can be used to set up the model. Data used for this modeling will be as per below.

Among the algorithm tested are Naive Bayes, Neural Net, Deep Learning, Decision Tree, Logistic Regression, SVM and finally k-Nearest Neighbors or k-NN as shown in Figure 3.

After several executions with various algorithms and split ratio combination, it was found that the optimal best result with the highest accuracy of about 98.4% was achieved using the k-NN algorithm using default parameters as shown in Figure 4.

As shown in Figure 5, in terms of execution speed it's found that most of the algorithm immediately returned the result except for Neural Net that took almost 2 seconds, and Deep Learning that took 6 seconds. As such these 2 algorithms are not suitable for fraud detection within the shipping domain as speed is one of the key criteria.

The above Figure represents the relationship between the k-NN key parameter which is the k nearest neighbor number of classes against the accuracy of the prediction. It's was found that the highest accuracy was recorded when k is either 1 or 2. Accuracy starts to drop once k is increased beyond 2.

Figure 6 represents the relationship between the k-NN key parameter which is the k nearest neighbor number of classes against the accuracy of the prediction. It's was found that the highest accuracy was recorded when k is either 1 or 2. Accuracy starts to drop once k is increased beyond 2. in this study, the k-NN algorithm has been identified as the best optimum results in terms of accuracy and speed criteria that were required in fraud detection within the shipping domain. Results in Figure 6 illustrates that genuine fraud was detected correctly for 88% of the total cases. Nonfraud cases were predicted correctly at 99.14% of the total cases.

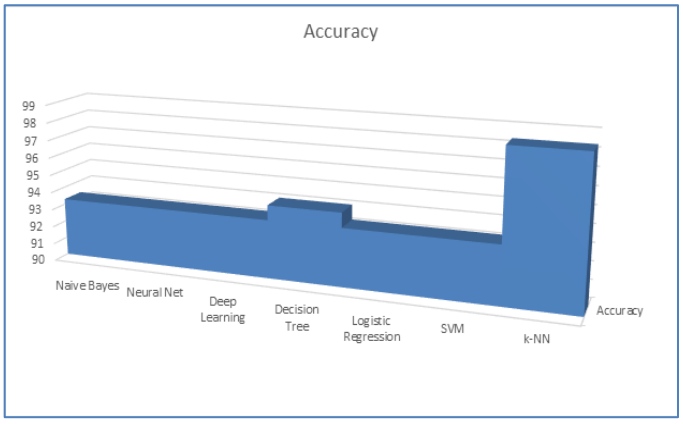

Fig. 3. Accuracy Results According to Algorithm Type.

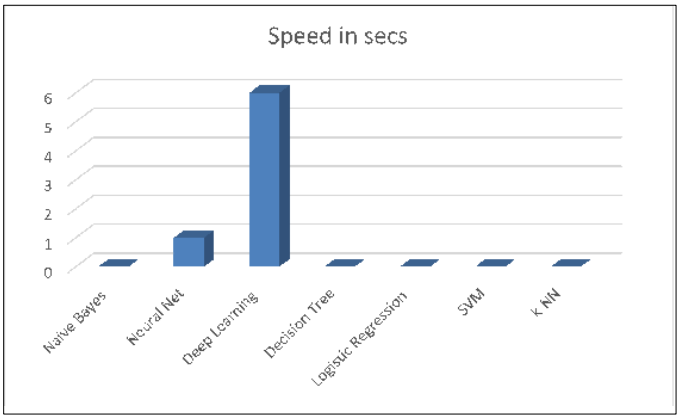

Fig. 4. Execution Speed According to Algorithm Type.

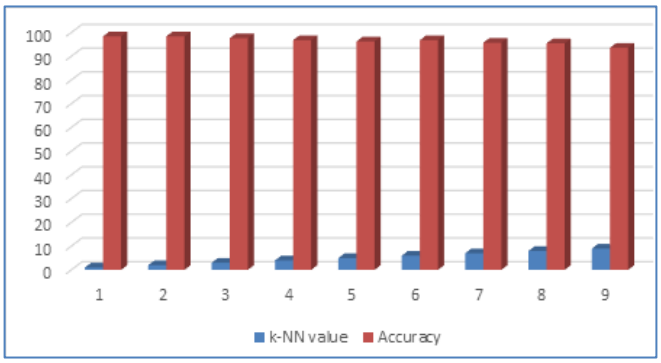

Fig. 5. Results According to different Values of k within the k-NN Algorithm.

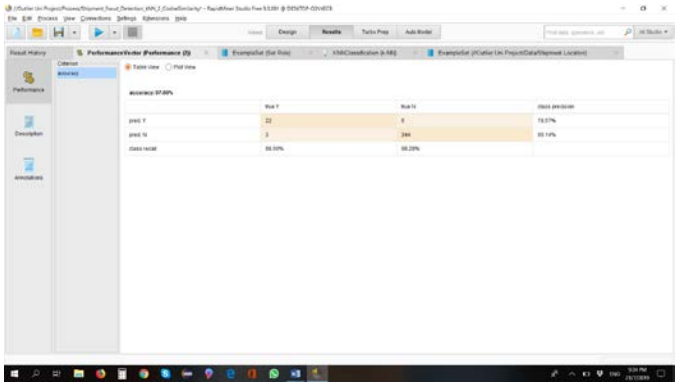

Fig. 6. Results.

As such to achieve high accuracy with optimum performance k-Nearest Neighbour algorithm technique is proposed to detect fraud in the shortest possible time within the shipping domain. This technique is proposed based on the modeling simulation done in Rapidminer. This algorithm indicates that this technique usually provides an acceptable response time during execution which is within a second. Thus, the model for this solution will be as shown in Figure 7 below where the shipping data will be first pre-processed to ensure there are no missing values. The pre-processed data will be then split between training and test data. The data set available can be split between training and test data with 75% for training the algorithm and 25% for testing the algorithm which is also close to the split recommended by [17].

Performance Evaluation with new incoming data and subsequent execution with more historical data. New unlabeled data can be routed to the model to predict if it could be fraudulent as shown below in 7e 7. The set of new data will be analyzed by the trained algorithm and will predict each row of data with origin longitude latitude and destination longitude-latitude with a prediction flag to each row. Suspected fraudulent locations will be tagged as Y and non-fraudulent will be identified as "N". The identification is primarily based on how close the locations to the labeled fraudulent cases are provided in the learning stage. Thus, if new locations are present in the data these data will not be identified as a fraud as there is no historical data linked to it. To resolve this challenge new location data can be identified in outlier techniques and analyzed distinctly before it's can be used as part of the main dataset.

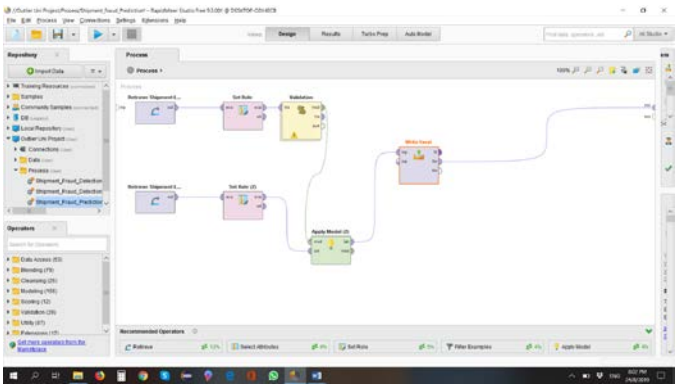

Fig. 7. New Data Test Model.

Alternatively, if the customer data profile is available k-NN distance-based outlier techniques can be applied by looking for any outlier locations within a customer's shipment data set. If there are no new locations expected from a particular customer and if there are any new locations detected within the customer's data set then this data can be classified as potential fraud using outlier techniques. As such combining machine learning, k-NN and outlier techniques can be a complementary strategy to increase the effectiveness of fraud detection in shipping domains.

V. Conclusion

As a conclusion, we are recommending k-NN algorithm machine learning to address fraud detection within the shipping domain. It's proven that fraud detection using machine learning is much more efficient compared to manually identifying fraud while a shipment is in progress. Identifying fraud before the shipment arrives at the destination is very crucial to ensure that fraud items do not get delivered to the consignee. This is only possible by automating the fraud detection process as some international shipments can be delivered within the same day depending on the location. Using the k-NN algorithm ensures fraud can be detected within the duration of shipment which is an effective way to stop the current fraud and to reduce future fraud cases. To overcome some challenges in identifying new cases of fraud k-NN machine learning technique can be combined with distance-based outlier techniques on data set that are grouped by customer profiles. Getting hold of actual shipping data from logistic companies was quite challenging due to the sensitive nature of shipping data. As such exploration of the various detection approaches, analyzing the strength and weakness of each before choosing the most optimum approach was done with simulated data which was based on parameters identified in a study done in a shipping company. Future papers may use these approaches and algorithms from this paper to simulate and perform further testing with actual data if they have access to it. Besides location parameters which were used in this paper, other parameters influence fraud in the shipping domain such as shipment weight, payment method, and the profile of the customer which was not evaluated in this paper due to lack of actual production data. As such in future papers these parameters can be considered to get results with higher accuracy.

Acknowledgment

This work is sponsored by Universiti Tenaga Nasional (UNITEN) under the Bold Research Grant Scheme No. J510050002.

References

[1] Shelley, L. I. (2018). Dark commerce: How a new illicit economy is threatening our future. Princeton University Press.

[2] Horsey, L. L. (2017). Data Analytics for Fraud Prevention and Detection in State Government (Doctoral dissertation, Utica College).

[3] International Air transport Association(IATA), 2018. http://www.iata.org/pressroom/pr/Pages/2018-01-31-01.aspx.

[4] Abdallah A., Maarof M. A., Zainal A., Fraud detection system: A survey, Journal of Network and Computer Applications, Volume 68, 2016, Pages 90-113,.

[5] S. Makki et al., "Fraud Analysis Approaches in the Age of Big Data - A Review of State of the Art," 2017 IEEE 2nd International Workshops on Foundations and Applications of Self* Systems (FAS*W), Tucson, AZ, USA, 2017, pp. 243-250, doi: 10.1109/FAS-W.2017.154.

[6] Panigrahi, S., Kundu, A., Sural, S., & Majumdar, A. K. (2009). Credit card fraud detection: A fusion approach using Dempster–Shafer theory and Bayesian learning. Information Fusion, 10(4), 354-363.

[7] Wang, Y., & Xu, W. (2018). Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. Decision Support Systems, 105, 87-95.

[8] Triepels, R., Daniels, H., & Feelders, A. (2018). Data-driven fraud detection in international shipping. Expert Systems with Applications, 99, 193-202.

[9] Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. Statistical science, 17(3), 235-255.

[10] Kou, Y., Lu, C. T., Sirwongwattana, S., & Huang, Y. P. (2004, March). Survey of fraud detection techniques. In IEEE International Conference on Networking, Sensing and Control, 2004 (Vol. 2, pp. 749-754). IEEE.

[11] Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. arXiv preprint arXiv:1009.6119.

[12] Barse, E., Kvarnstrom, H. & Jonsson, E., "Synthesizing Test Data for Fraud Detection Systems", Proc. of the 19th Annual Computer Security Applications Conference, 384-395 , 2003.

[13] Chen, R., Chiu, M., Huang, Y. & Chen, L. , "Detecting Credit Card Fraud by Using Questionaire-Responded Transaction Model Based on Support Vector Machines", Proc. of IDEAL2004,800-806 , 2004.

[14] Aleskerov, E., Freisleben, B. & Rao, B. , "CARDWATCH:A Neural Network-Based Database Mining System for Credit11 Card Fraud Detection" , Proc. of the IEEE/IAFE on Computational Intelligence for Financial Engineering, 220-226 , 1997.

[15] Pathak, J., Vidyarthi, N. & Summers, S. , "A Fuzzy-based Algorithm for Auditors to Detect Element of Fraud in Settled Insurance Claims", Odette School of Business Administration , 2003.

[16] Richard J. Bolton and David J. Hand , "Statistical Fraud Detection: A Review" , Imperial College, 2002.

[17] Vijay K , Bala D , "Predictive Analytics and Data Mining" , Elsevier Inc , 2015.

[18] Mahmoud, M. A., Ahmad, M. S., Ahmad, A., Yusoff, M. Z. M., & Mustapha, A. (2011, July). Norms detection and assimilation in multi-agent systems: a conceptual approach. In Knowledge Technology Week (pp. 226-233). Springer, Berlin, Heidelberg.

[19] Jassim, O. A., Mahmoud, M. A., & Ahmad, M. S. (2015). A multi-agent framework for research supervision management. In Distributed Computing and Artificial Intelligence, 12th International Conference (pp. 129-136). Springer, Cham.

[20] Mahmoud, M. A., Ahmad, M. S., Ahmad, A., Yusoff, M. Z. M., Mustapha, A., & Hamid, N. H. A. (2013, May). Obligation and Prohibition Norms Mining Algorithm for Normative Multi-agent Systems. In KES-AMSTA (pp. 115-124).

[21] Mahmoud, M. A., Ahmad, M. S., Ahmad, A., Mustapha, A., Yusoff, M. Z. M., & Hamid, N. H. A. (2013). Building norms-adaptable agents from potential norms detection technique (PNDT). International Journal of Intelligent Information Technologies (IJIIT), 9(3), 38-60.

[22] Mahmoud, M. A., Mustapha, A., Ahmad, M. S., Ahmad, A., Yusoff, M. Z. M., & Hamid, N. H. A. (2013). Potential norms detection in social agent societies. In Distributed Computing and Artificial Intelligence (pp. 419-428). Springer, Cham.

[23] Mahmoud, M. A., Ahmad, M. S., Yusoff, M. Z. M., & Mostafa, S. A. (2018, February). A regulative norms mining algorithm for complex adaptive system. In International Conference on Soft Computing and Data Mining (pp. 213-224). Springer, Cham.

[24] Mahmoud, M., Ahmad, M. S., Mostafa, S., & Subramainan, L. (2020). How Norm Assimilation Using Agent-Based Systems. Journal of Systems Science and Complexity, 33(4), 849-881.

[25] Mahmoud, M. A., Ahmad, M. S., & Mostafa, S. A. (2019). Norm-based behavior regulating technique for multi-agent in complex adaptive systems. IEEE Access, 7, 126662-126678.

[26] Mahmoud, M. A., & Ahmad, M. S. (2016, August). A prototype for context identification of scientific papers via agent-based text mining. In 2016 2nd International Symposium on Agent, Multi-Agent Systems and Robotics (ISAMSR) (pp. 40-44). IEEE.

[27] Mahmoud, M. A., & Ahmad, M. S. (2015, August). A self-adaptive customer-oriented framework for intelligent strategic marketing: A multi-agent system approach to website development for learning institutions. In 2015 International Symposium on Agents, Multi-Agent Systems and Robotics (ISAMSR) (pp. 1-5). IEEE.

[28] Mahmoud, M. A., Ahmad, M. S., Yusoff, M. Z. M., & Mustapha, A. (2014, December). Norms assimilation in heterogeneous agent community. In International Conference on Principles and Practice of Multi-Agent Systems (pp. 311-318). Springer, Cham.