The Eurasia Proceedings of Science, Technology, Engineering & Mathematics (EPSTEM)
ISSN: 2602-3199

The Eurasia Proceedings of Science, Technology, Engineering & Mathematics (EPSTEM), 2024

Volume 29, Pages 173-181

ICRETS 2024: International Conference on Research in Engineering, Technology and Science

Computer-Aided System for Customs Fraud Analytics Based on Artificial Intelligence Techniques

Veska Gancheva
Technical University of Sofia

George Popov
Technical University of Sofia

Kamelia Raynova
Technical University of Sofia

Antoaneta Popova
Technical University of Sofia

Ivaylo Georgiev
Bulgarian Academy of Science

Abstract: Globalization has stimulated the opening of the market and the accumulation of huge amounts of data, which consequently leads to an increase in the importance of the control of customs operations. However, customs data is highly imbalanced and this poses challenges in its integration and processing. Therefore, it is of prime importance to find automatic computationally intelligent solutions for customs management. The purpose of the research presented in this paper is to propose a computer-aided system for customs fraud analytics based on artificial intelligence techniques, ensuring the application and verification of methods and algorithms for integration, management, analysis and visualization of data on customs violations. The architecture of the customs violation data analysis system consists of the following components: data sources, data storage, data integration and preprocessing, real-time data flow, modeling, analysis and storage of analytical data, and visualization of the results. A machine learning approach for detecting customs fraud through unstructured data analysis is proposed. An artificial neural network designed for data analysis is designed, and the input data is divided into training data and testing data. A reduced set of statistical records related to the analysis of heterogeneous databases of different institutions, which is stored in a data warehouse, are used as experimental data. The first 80% of the data are used to train the neural network and the remaining 20% to test the trained network. Experimental results show that the calculated accuracy increases with increasing epochs and is higher for the training data and lower for the validation test data. Thus, the trained model can be saved and used to monitor for anomalies. The trained model is applied to the system to calculate new input parameters that were not used in either training or validation.

Keywords: Artificial intelligence, Customs fraud, Computer-aided system, Data analytics, Neural network.

Introduction

Financial fraud is a problem with far-reaching consequences in the financial industry, government, corporate sectors and for ordinary consumers. Increasing reliance on new technologies such as cloud and mobile computing in recent years has compounded the problem. Not surprisingly, financial institutions are turning to automated processes using statistical and computational methods.

Growing volumes of international trade put strain on regulatory oversight, which is faced by customs administration. Regulators thus use data mining to concentrate their little resources on the fraud cases that have the highest likelihood of occurring. Standard learning algorithms are most frequently used in tax studies' fraud detection applications. However, there are new difficulties because of the extreme imbalance in customs data.

Governments and customs administrations realize that the growing demand for free and secure commerce (including e-commerce) requires data standardization. This is the only approach by which governments can fulfill their missions. The World Customs Organization (WCO) Data Model provides an appropriate framework of standard and harmonized data sets and standard electronic messages that are transmitted by trade for customs and other regulatory purposes to complete arrival, departure, transit and release formalities of goods in international cross-border trade. Thus, standardized data sets and electronic messages using international customs code standards are a key mechanism for effective and efficient information exchange between businesses and governments. In practice, through the presented information model, unification of the requirements for data exchange is obtained, and thus it is possible to create a single electronic structure allowing effective exchange of information in a global aspect. The WCO data model also incorporates the data requirements of other government regulatory bodies, enabling a single window environment that allows traders to provide information only once to one official body, preferably customs, to fulfill all regulatory requirements related to import or export.

According to European and international requirements, customs codes are standardized. This standardization, as well as access to information, are regulated by law. Each register is a structured electronic database. The registers contain data provided by economic operators in accordance with European and national customs legislation, and data and circumstances entered by the Customs Service in connection with the implementation of its functions and tasks. Data from the registers are exchanged with the customs authorities of the Member States and the European Commission in cases where the customs legislation requires it, and are determined by the technical specifications of the relevant electronic systems.

Customs administrations lessen the grave risks that smuggling and tax evasion pose to the public through oversight. Customs establishes a set of rules to filter out high-risk goods based on the information provided in import declarations because it is challenging to thoroughly inspect every item given the volume of trade and the limited resources available (budget, officer count, etc.). Implementing an effective customs selection or fraud detection system is therefore essential to accelerating the customs clearance process (Kim S. et al., 2020, 2021, 2023). By predicting potentially fraudulent things, customs authorities can assess each item's level of examination; the most questionable items require a physical inspection by human inspectors. The difficulty is figuring out which set of very suspicious items should be the focus.

An examination of the scientific literature indicates that the application of AI-based methods significantly improves the detection of customs fraud. The Brazilian program, which aims to combat customs fraud, is a well-known example. The system integrates data from many administrative and customs sources to enable more precise and effective control over the import and export of commodities. The system uses a range of artificial intelligence and data analytics techniques to identify anomalies, disparities, and potential fraud (Digiampietri, 2008).

Three steps and three algorithms are used in a proposed approach to identify smugglers from unstructured social media data (Dangsawang & Nuchitprasitchai, 2024). The goal is to locate those who sell goods and services on social media without authorization in order to evade paying taxes levied by the government. The stability and economics of the nation may suffer from this practice. The model classifies imported products using techniques including logistic regression, gated recurrent unit, and long short-term memory by gathering labeling keywords and classifying them into three groups.

The World Customs Organization has implemented information technology in customs operations to identify security dangers resulting from the smuggling of high-risk items, acknowledging the difficulties involved. Using pictures, deep learning algorithms are utilized to automatically identify and discourage the smuggling of commodities into training containers (Jaccard et al., 2016).

The research presented in this paper is part of a scientific project aimed to create artificial intelligence-based solutions for an information system that prevents and detects financial and customs infractions. The system for their detection and prevention must notify users when a violation is being planned, created, and executed in order to put an end to financial and customs violations. If financial customs violations have already taken place,

they should be made public to extract revenge on the perpetrators and create a sense of punishment that will deter similar transgressions in the future.

Method

Computer-Aided System for Customs Fraud Analytics

The architecture of the SOA-based system intended to process different types of customs data is shown in Figure 1. In addition to facilitating the smooth integration and interaction of various applications and data stores, SOA supports the sharing and exchange of data through standardized interfaces, ensuring compatibility between various systems and data sources and enabling the distributed deployment of services in various media and computing environments. Covering different data and functions separately, these services work together to automate complicated tasks. Benefits include cross-compatibility, adaptability, simplicity of maintenance, support for various technologies, and flexibility in response to shifting company needs. It also makes distributed application creation and administration easier.

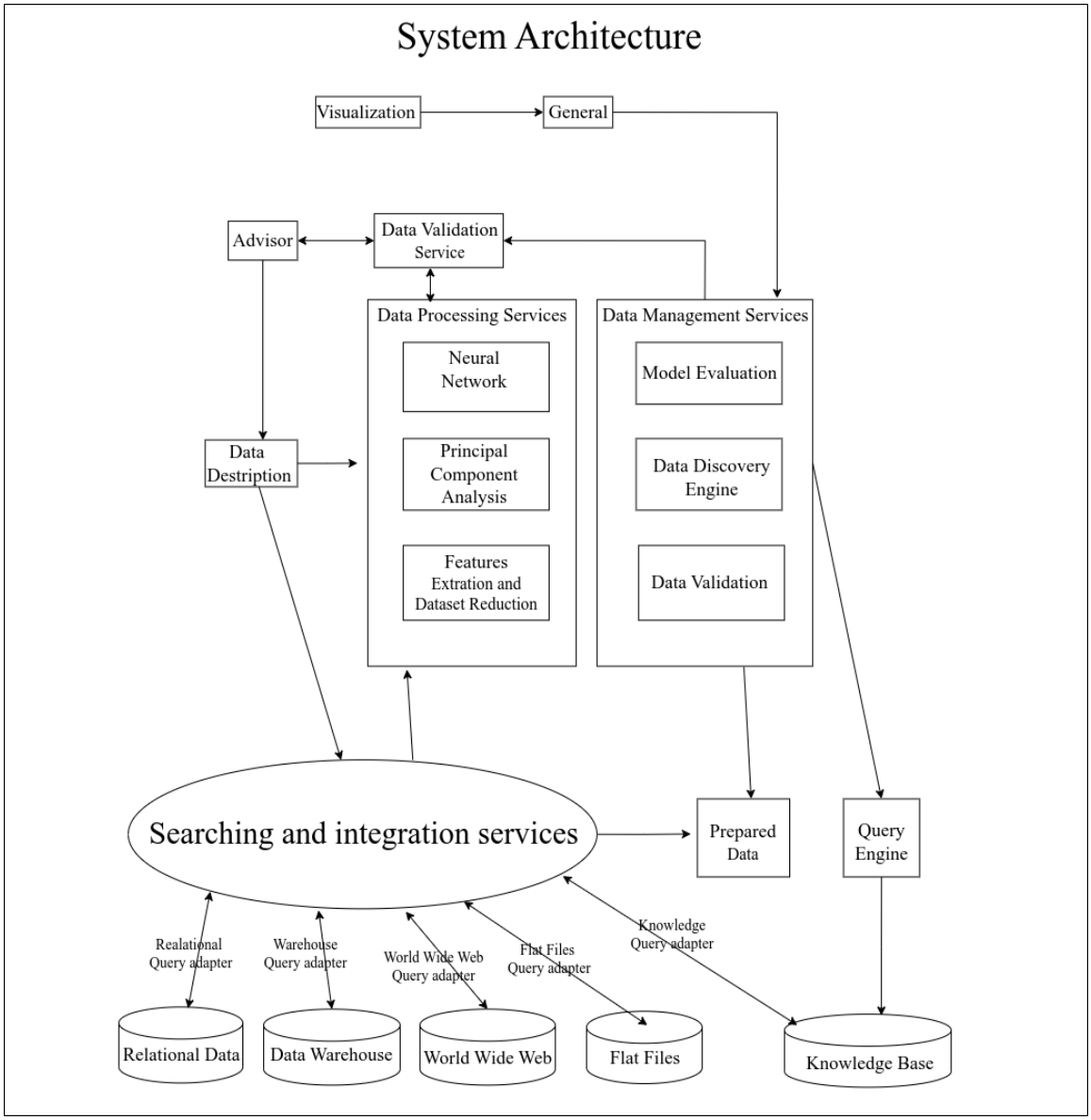


Figure 1. Computer-aided system architecture

In order to facilitate knowledge extraction and decision-making, the system provides intelligent solutions as well as automation of practical techniques, tools, and algorithms for integrating, storing, analyzing, and

visualizing customs data. The system offers an integrated and flexible framework for creating workflows to automate the computational process through a set of software tools. The capabilities encompass the following: 1) a user-friendly interface; 2) interactive tools for executing workflows for custom data analysis; 3) support for various data formats; 4) sharing and reusing workflows; 5) monitoring the outcomes of workflow execution and the steps involved in its creation; 6) the ability to swiftly add, remove, and scale functionalities as services; 7) enhanced speed and efficiency; and 8) the capacity to scale and adaptably increase resources

Essential data management activities are handled by the databases server and data warehouse, guaranteeing that actual data sets are prepared for processing and stored according to their nature. Databases, data warehouses, the Internet, text files, and different papers are examples of data sources. They are all rich in historical information that can be used to effectively extract knowledge. In order to solve the problem of insufficient or unreliable data, data preparation, integration, cleansing, and selection are essential first steps in the knowledge discovery process. It is imperative to apply diverse strategies for cleaning and selection in order to narrow down the extra data generated by multiple sources and select only the pertinent data for the study. The next step in the server's job is to retrieve the relevant data in response to the user's processing request. The system for managing data is created.

One important component is Service Advisor, that helps users choose services and data, helps them upload datasets, and leads them through the inquiry process. Efficient communication with the system is made possible by this module. It enables the wizard to communicate with the system in order to handle, interpret, and analyze different types of data, and it displays the outcomes in an easy-to-use interface.

Rapid handling of massive and diverse data sets in many formats—such as relational databases, NoSQL, and flat files—is made easier by integration services. These services combine information from multiple sources and convert general queries into targeted database queries. The system explores ongoing access to research updates for comparing findings with existing knowledge, and it tackles the difficulties in managing large, diverse, and complicated datasets that can be found in public databases.

By collaborating with the data discovery engine to pinpoint important patterns within the knowledge base, the pattern assessment module evaluates the significance of patterns that are found. A key component of the system, the data discovery engine uses a variety of algorithms to perform tasks such as classification, genetic algorithms, prediction, and clustering, and it consults the knowledge base to produce more dependable and accurate results. While database and data warehouse servers store and manage datasets that are ready for processing, query engines streamline interactions with intricate underlying data sources. By keeping a significant amount of historical data, these servers enable efficient knowledge extraction by retrieving pertinent information from a variety of factual sources in response to user requests.

The knowledge repository stores all identified patterns, models, and rules, supporting the entire knowledge discovery process, particularly in evaluating the significance of outcome models or focusing on specific demands. Preprocessing involves search and optimization techniques for data integration, cleansing, and selection based on relevance and accuracy. Iterative machine learning optimizes feature sets, and post-processing includes verification, validation, visualization, and evaluation of retrieved knowledge through data mining, machine learning, and decision-making methods for high accuracy and precision.

All discovered patterns, models, and rules are kept in the knowledge repository, which aids in the entire process of knowledge discovery, especially when assessing the importance of outcome models or concentrating on particular requirements. Preprocessing includes data integration, cleaning, and selection based on correctness and relevance using search and optimization algorithms. Post-processing comprises verification, validation, visualization, and evaluation of retrieved knowledge through data mining, machine learning, and decision-making procedures for high accuracy and precision. Iterative machine learning optimizes feature sets.

Workflow for Customs Data Analytics

A workflow for creating a neural network-based model for customs data analytics is shown in Fig. 2. The procedure is separated into its principal parts: customs data selection and preprocessing feedforward neural network model creation, model training, model evaluation. The fraud detection problem aims to find the patterns behind the features in predicting the target label Fraud. Data is split into three pieces. We assign the first 12 months of data to the training set, the following three months to the validation set, and the last three months to the test set. Categorical variables are label-encoded and numerical variables are min-max scaled.

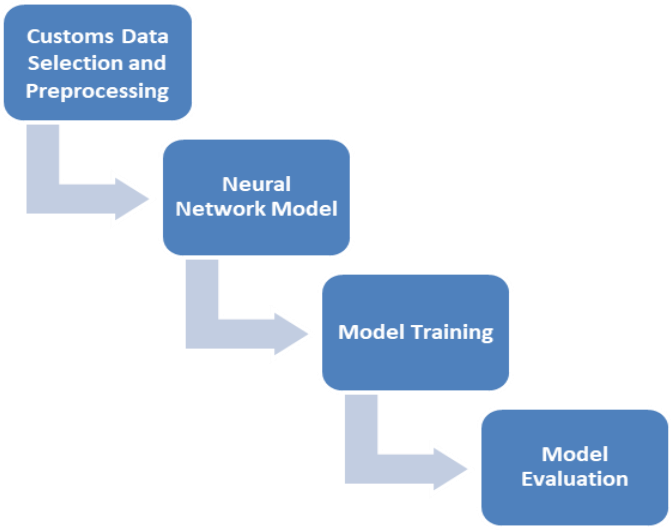


Figure 2. Customs data analytics workflow

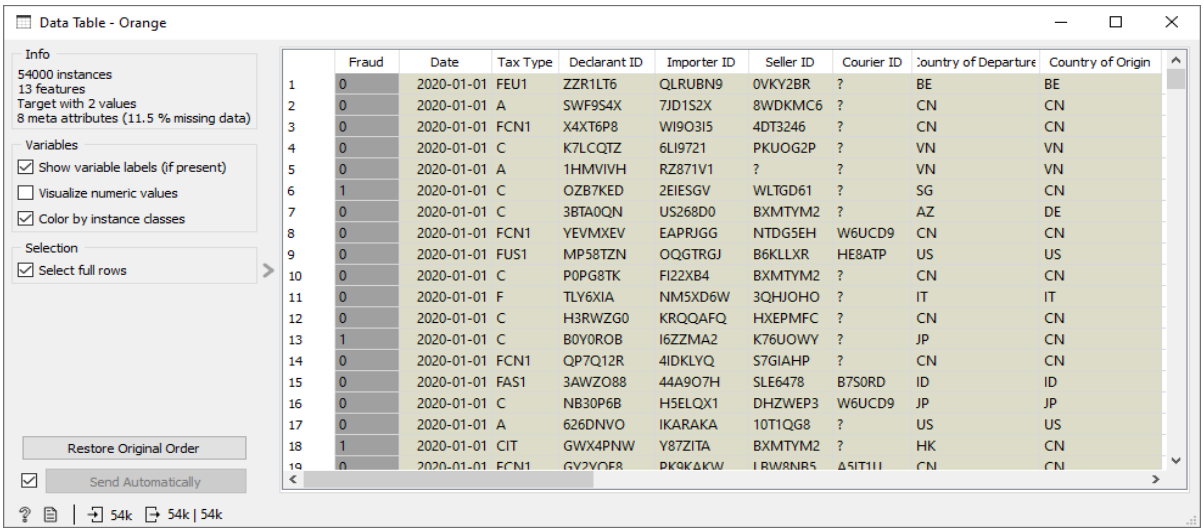
Customs Dataset Selection and Preprocessing

Customs data processing covers the electronic submission of customs declarations and/or the provision of paper declarations. Additional documentation related to the implementation of processing procedures leading to the release of goods and the sharing of information with participating parties is also provided. The initial stage of the proposed workflow involves choosing customs data from databases. Preparation of the data for analysis is the aim of this phase. It could involve normalizing and cleaning the data, such as removing outliers and missing numbers. The original data have to be formatted such that the neural network can use it in the categorizing process.

Table 1. Customs dataset parameters

No	Attribute	Description
1	Declaration ID	Primary key
2	Date	Date of the declaration
1	Declaration ID	Primary key
2	Date	Date of the declaration
3	Office ID	Customs office
4	Process type	Declaration process type
5	Import type	Import type code
6	Import use	Import use code
7	Payment type	Determine the type of tariff payment
8	Mode of transport	Nine modes of transport
9	Declarant ID	Person declaring the item
10	Importer ID	Person who imports the product
11	Seller ID	Foreign business associate
12	Courier ID	Provider of delivery services
13	HS6 code	Six-digit code for the product
14	Country of departure	Nation from which a shipment has left or is expected to leave
15	Country of origin	Manufacturing, production, or design nation
16	Tax rate	Tax rate of item
17	Tax type	Tax category of item
18	Country of origin indicator	A means of designating the nation of origin
19	Net mass	Bulk without any packaging
20	Item price	Evaluated worth of a thing
21	Fraud	Dishonest attempt to lower the customs charge
22	Critical fraud	A crucial case that could endanger public safety

An import declaration dataset including 54,000 artificially created transactions with 22 important features that support correlation is used for the experimental validation of the proposed solution (Chaeyoon et al., 2023). The production process reduces the potential identity risk that could be present in trade statistics, and the released data is distributed similarly to the source data, allowing it to be utilized for a variety of purposes. This dataset was synthesized from 24.7 million customs declarations that were reported in the 18 months between 1 January 2020 and 30 June 2021, with the validated (i.e., labelled) part of the reports being used. Every row contains the report for a single item. The data contains 22 typical qualities incl. fraud and critical fraud, among the 62 attributes listed in the import declaration form. The following are full descriptions of the dataset. Table 1 presents the customs dataset parameters. Sample data records are shown in Fig. 3.



	Fraud	Date	Tax Type	Declarant ID	Importer ID	Seller ID	Courier ID	Country of Departure	Country of Origin
1	0	2020-01-01	FEU1	ZZR1LT6	QLRUBN9	0VKY2BR	?	BE	BE
2	0	2020-01-01	A	SWF9S4X	7JD1S2X	8WDKMC6	?	CN	CN
3	0	2020-01-01	FCN1	X4XT6P8	WI9O3I5	4DT3246	?	CN	CN
4	0	2020-01-01	C	K7LCQTZ	6LI9721	PKUOG2P	?	VN	VN
5	0	2020-01-01	A	1HMVIVH	RZ871V1	?	?	VN	VN
6	1	2020-01-01	C	OZB7KED	2EIESGV	WLTGD61	?	SG	CN
7	0	2020-01-01	C	3BTA0QN	US268D0	BXMTYM2	?	AZ	DE
8	0	2020-01-01	FCN1	YEVMXEV	EAPRJGG	NTDG5EH	W6UCD9	CN	CN
9	0	2020-01-01	FUS1	MP58TZN	OQGTRGJ	B6KLLXR	HE8ATP	US	US
10	0	2020-01-01	C	P0PG8TK	FI22XB4	BXMTYM2	?	CN	CN
11	0	2020-01-01	F	TLY6XIA	NM5XD6W	3QHJOHO	?	IT	IT
12	0	2020-01-01	C	H3RWZG0	KRQQAFQ	HXEPMFC	?	CN	CN
13	1	2020-01-01	C	B0Y0ROB	I6ZZMA2	K76UOWY	?	JP	CN
14	0	2020-01-01	FCN1	QP7Q12R	4IDKLYQ	S7GIAHP	?	CN	CN
15	0	2020-01-01	FAS1	3AWZO88	44A9O7H	SLE6478	B7S0RD	ID	ID
16	0	2020-01-01	C	NB30P6B	H5ELQX1	DHZWEP3	W6UCD9	JP	JP
17	0	2020-01-01	A	626DNVO	IKARAKA	10T1QG8	?	US	US
18	1	2020-01-01	CIT	GWX4PNW	Y87ZITA	BXMTYM2	?	HK	CN
19	0	2020-01-01	FCN1	GY2YOF8	PK9KAKW	LBW8NB5	A5IT1U	CN	CN

Figure 3. Sample customs declaration dataset

Feed Forward Neural Network Model for Customs Data Analytics

Model Creation

Neural Network Model: For this objective, a feed forward neural network model is designed, as seen in Fig. 4. This type of model processes inputs layer by layer as it moves forward. The architecture of the neural network captures the complex correlations that occur between the input features and the intended outcomes. A more nuanced interpretation of the data is made possible by the probabilistic output of the model, where probability acts as a stand-in for forecast certainty. The architecture includes following elements:

- An input layer with eight neurons that match the thirteen input parameters selected from the database.
- A 32-neuron hidden layer that makes it possible to extract more intricate features from the input data.
- An additional hidden layer with 16 neurons to further process the features that the preceding layer had recovered.
- An output layer with two neurons is meant to represent the probability result for a fraud and non fraud, respectively.
- The data is split into training and test data in an 80/20 ratio.

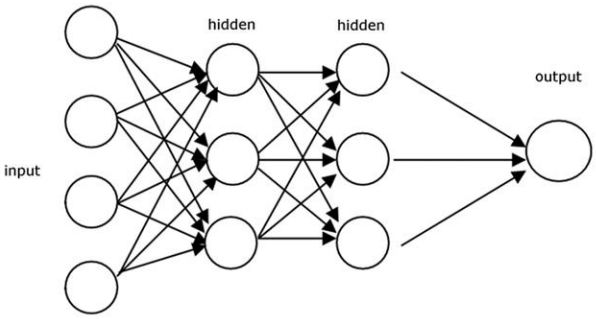


Figure 4. Neural network model

Model Training

Training Process: The neural network is trained in this stage. In order to reduce the discrepancy between the expected and actual results, the network's weights and biases are adjusted throughout training. Usually, backpropagation techniques are combined with an optimization technique to achieve this. The neural network is trained by feeding it training and test vectors; the 200 epochs represent the number of times the network processes the full dataset. The model can progressively learn from the data thanks to this iterative procedure. To avoid underfitting, which occurs when the model learns the data insufficiently, and overfitting, which occurs when the model learns the training data excessively.

Model Validation and Evaluation

Validation: A new, distinct dataset that was not used for training is used to assess performance and usefulness of the model.
Evaluation: Lastly, the performance of the model is evaluated using metrics such as accuracy. Once all training epochs have been read, the accuracy of the model is displayed against the training and validation sets of data (Fig. 5). As expected, accuracy increases with the number of epochs; it is higher for training data and lower for data from validation tests. The trained model can then be saved, loaded onto an alternative system, and checked for weight values.

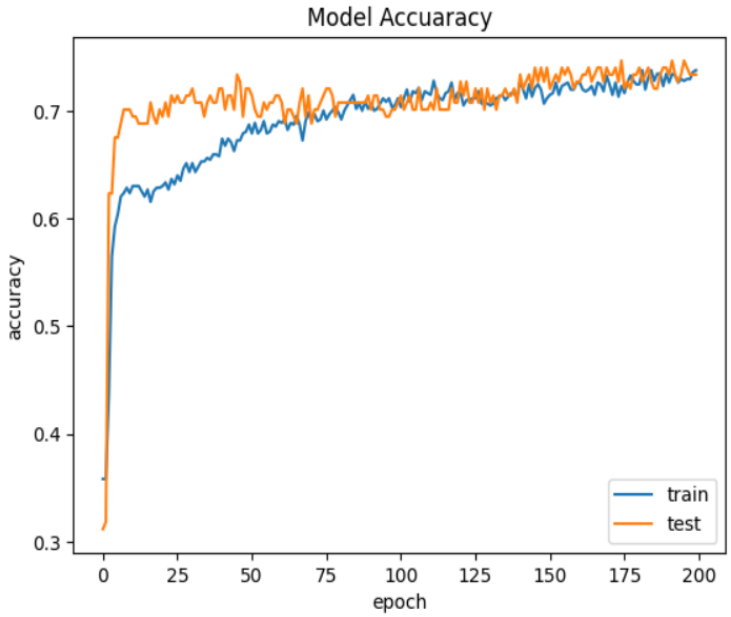


Figure 5. Accuracy of model

This methodology guarantees a rigorous approach to model creation, beginning with data collection and continuing through training and validation to generate a neural network that can assist in customs fraud identification. A solid foundation for creating reliable and effective customs data analytics solutions is laid by the emphasis on careful model design along with adequate training and validation.

As the model quickly learns the underlying patterns in the data, the accuracy increases quickly in the first 50 epochs, as is normal. Following that, test accuracy displays more random variations while training accuracy keeps getting better. This shows that while the model fits the training data well, overfitting or noise in the test data may provide some challenges. Convergence of the training and test accuracy around similar values indicates a strong generalization of the model.

The loss for both training and test data drops very swiftly in the first 50 epochs, suggesting that the model picks up on the underlying patterns in the data quite quickly. Subsequently, the test data loss exhibits a little slower drop and stabilizes early, whereas the training data loss keeps decreasing, albeit more slowly. In the final 50 epochs, both losses even off, with the training loss being marginally greater than the test loss. This suggests that there is little to no overfitting and that the model is well-balanced.

Results and Discussion

The trained model can be deployed in practical situations to assist in detecting customs fraud through several specific applications:

Using the Trained Model

- Handling New Input Parameters: It is crucial that any new data inputs maintain the exact structure and format as those used during the training and validation phases. This requires keeping the input vector's parameters in the same type and order.
- Making Predictions: Once the new input data is properly formatted and entered into the model, the neural network processes it using the learned weights and biases. The model then outputs a probability distribution indicating the likelihood of a customs violation.

Importance of Consistency in Data Format

- Model Accuracy: The accuracy of the model's predictions is heavily influenced by the consistency of the input data format. Neural networks are trained to recognize patterns based on the specific arrangement and type of data. Any deviations in the structure of new input data can lead to inaccurate predictions, as the model may misinterpret the information.
- Streamlined Integration: Keeping a consistent parameter sequence simplifies the integration of the model into automated systems or workflows. This consistency allows experts or automated data collection tools to prepare and input customs data into the model without needing to adjust the data structure for each new case.

Practical Application

- Customs Data Collection: Gathering the necessary customs parameters from new datasets.
- Data Preparation: Ensuring that the collected data adheres to the training format, including the order of the parameters.
- Making Predictions: Inputting the prepared data into the model to obtain a probability distribution indicating the likelihood of customs violations.
- Interpretation of Results: Customs professionals can use the model's output as part of their decision-making process, alongside other assessments and tests, to make informed decisions about potential customs violations.

This use of artificial intelligence and machine learning techniques demonstrates their potential to significantly enhance customs services, particularly in identifying fraud and customs infractions, provided the data is accurately prepared and the model is appropriately used.

Conclusion

The research described in this paper aims to propose a computer-aided system for customs fraud analytics based on artificial intelligence techniques. For the purposes of customs data analysis case study, a feedforward artificial neural network is created. During the training phase, the input data are divided into training and test data. We quantify the training error and show how it affects the neuron weights in the network. A set of statistics on customs data obtained from declaration analysis has been used as experimental data. For the study, 13 attributes from the original database are employed. Additionally, the data is split in a 0.8 to 0.2 ratio. The first 80% of the data are used to train the neural network, and the remaining 20% were used to test the trained network. The calculated accuracy rises with the number of epochs added. As the number of epochs increases, the estimated accuracy decreases for validation test data and increases for training data. The trained model can then be saved, loaded onto an alternative computer, and examined to verify the weight values. To compute new input parameters that are not used during training or validation, the system makes use of the trained model.

Scientific Ethics Declaration

The authors declare that the scientific ethical and legal responsibility of this article published in EPSTEM journal belongs to the authors.

Acknowledgements

* This article was presented as a poster presentation at the International Conference on Research in Engineering, Technology and Science (www.icrets.net) held in Thaskent/Uzbekistan on August 22-25, 2024.

* This research is financially supported by the Bulgarian Ministry of Education and Science, National Science Fund - grant KP-06-N35/12.

References

Dangsawang, B., & Nuchitprasitchai, S. (2024). A machine learning approach for detecting customs fraud through unstructured data analysis in social media. *Decision Analytics Journal*, *10*, 100408.

Digiampietri, L. A., Roman, N. T., Meira, L. A., Jambeiro Filho, J., Ferreira, C. D., Kondo, A. A., & Goldenstein, S. (2008, May). Uses of artificial intelligence in the Brazilian customs fraud detection system. In *DG. O* (pp. 181-187).

Jaccard, N., Rogers, T. W., Morton, E. J., & Griffin, L. D. (2016, November). Automated detection of smuggled high-risk security threats using deep learning. In *7th International Conference on Imaging for Crime Detection and Prevention (ICDP 2016)* (pp. 1-6). IET.

Jeong, C., Kim, S., Park, J., & Choi, Y. (2022). Customs import declaration datasets. *arXiv preprint arXiv:2208.02484*.

Kim, S., Mai, T. D., Han, S., Park, S., Nguyen, D. T., So, J., & Cha, M. (2022). Active learning for human-in-the-loop customs inspection. *IEEE Transactions on Knowledge and Data Engineering*, *35*(12), 12039-12052.

Kim, S., Song, S. K., Cho, M., & Shin, S. H. (2021). Transaction Pattern Discrimination of Malicious Supply Chain using Tariff-Structured Big Data. *The Journal of the Korea Contents Association*, *21*(2), 121-129.

Kim, S., Tsai, Y. C., Singh, K., Choi, Y., Ibok, E., Li, C. T., & Cha, M. (2020, August). DATE: Dual attentive tree-aware embedding for customs fraud detection. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2880-2890).

Author Information

Veska Gancheva
Technical University of Sofia
Kliment Ohridski 8, Sofia, Bulgaria
Contact e-mail: *vgan@tu-sofia.bg*

George Popov
Technical University of Sofia
Kliment Ohridski 8, Sofia, Bulgaria

Kamelia Raynova
Technical University of Sofia
Kliment Ohridski 8, Sofia, Bulgaria

Antoaneta Popova
Technical University of Sofia
Kliment Ohridski 8, Sofia, Bulgaria

Ivaylo Georgiev
Stephan Angeloff Institute of Microbiology
Bulgarian Academy of Science
Acad. Georgi Bonchev str. 2, blok 26, Sofia, Bulgaria

To cite this article:

Gancheva, V., Popov G., Raynova K., Popova A. & Georgiev I. (2024). Computer-aided system for customs fraud analytics based on artificial intelligence techniques. *The Eurasia Proceedings of Science, Technology, Engineering & Mathematics (EPSTEM), 29*, 173-181.